

Describing Human Activities in Video Streams



Nouf Mezel Al Harbi

Department of Computer Science
The University of Sheffield

A dissertation submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

Supervisor: Dr. Yoshihiko Gotoh

May 2017



In the name of Allah, the Most Gracious, the Most Merciful

I would like to dedicate this thesis to my parents, who have been the biggest and continuous source of motivation throughout my life. . .

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Nouf Mezel Al Harbi

May 2017

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful. Thanks to Allah who is the source of all the knowledge in this world, and imparts as much as He wishes to anyone He finds suitable. I would like to express my deepest gratitude to my supervisor, Dr. Yoshihiko Gotoh, who has supported me throughout my research work with his knowledge and guidance. Special thanks to my panel members, Jon Barker and Mike Stannett for their useful suggestions and for being so kind to me. I am thankful to my examiners Amir Hussain and Heidi Christensen for useful discussion during the final viva of my defence. Also, I am indebted to all the members of the Speech and Hearing group at Sheffield University who have been always a continuous source of knowledge and friendship.

I am hugely grateful to my parents for their support, prayers, love and care throughout my life; they have played a vital role in helping me to reach this milestone. My husband, Rami who has constantly encouraged me and extended his wholehearted support especially during my PhD studies, which I could not have completed without him. To my sons Abdulrahman and Hazem, who born during my study, you have played your part very well as you always been the source of my happiness. I owe special thanks to my sisters and brothers for their continuous love and prayers.

Last but not the least I thankfully acknowledge the financial funding for this work from Taibah University in Saudi Arabia, who gave me a scholarship to pursue my PhD studies.

Abstract

This thesis outlines and advances a video description framework that describes human activities and their spatial and temporal relationships that can be used for video indexing, retrieval and summarisation applications. Generating natural language description of video streams demands the extraction of high-level features (HLFs) that sufficiently represent the events. This research centres around the issues combined with this task. One of these issues relates to identifying and segmenting participant human objects and identifying their visual attributions due to a broad range of scene setting variations, occlusion and background clutter.

To that end a five-stage approach is developed to investigate the video description task. Firstly, a proper corpus that can be used for development and evaluation is created which contains relatively long video clips of human activities crafted from the Hollywood2 dataset, depicting a variety of action classes along with human textual annotations for each. Extensive analysis of the hand annotations associated with this corpus results in the conclusion that annotators are most interested in human presences and their visual attributions in the video stream, especially their actions, and interaction with other objects.

Secondly, based on analysis outcome a novel framework that can detect, segment and track human body regions over video frames is proposed in order to efficiently describe video semantic. The proposed framework leverages the advances of low-level image cues and high-level part detectors information. Thirdly, the visual attributions of extracted human objects are extracted as an efficient human action recognition framework is introduced. The video representation is improved by using extracted spatio-temporal human regions combined with the extended spatio-temporal locality-constrained linear coding (LLC) technique in order to identify the action class. Human action classification benchmarks are used to assess the performance of this model. The results reveal that the outcome of this approach outperforms the state-of-the-art, owing to its efficient representation of complex actions in video stream.

Fourthly, as spatial and temporal relations of prominent objects play a vital role in describing video semantic content, a comprehensive representation is developed to efficiently extract spatial and temporal relations between interacted objects present in a video clip using their approximate oriented bounding box. The final stage aims to convert extracted HLFs into sentential descriptions using a template-based approach. By calculating the overlap between

descriptions produced by machine and those annotated by humans, it can be confirmed that context information is captured by automatic descriptions, which means that these descriptions are compatible with human viewing abilities. Finally, a video retrieval task based on textual query is designed to evaluate the generated natural language descriptions. The experimental outcome shows that the approach is able to retrieve relevant video segments and capture the main aspect of video semantic.

Table of Contents

List of Figures	xix
------------------------	------------

List of Tables	xxv
-----------------------	------------

1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Research Focus	3
1.4 Thesis Contributions	6
1.5 Thesis Overview	9
1.6 Published Work	10
2 Corpus Generation and Analysis	13
2.1 Introduction	13
2.1.1 Motivations	14
2.1.2 Corpus Generation and Analysis: Overview	15
2.1.3 Corpus Generation and Analysis: Contributions	16
2.2 Related Work	16
2.3 Corpus Generation	20
2.3.1 Collecting Textual Video Descriptions	21
2.4 Annotation Analysis	22
2.4.1 Human Related Features	25
2.4.2 Objects and Scene Settings	27
2.4.3 Spatial Relations	27
2.4.4 Temporal Relations	29
2.4.5 Similarity between Descriptions	30
2.5 Action Classification Experiments	34
2.5.1 Experimental Setup	34

2.5.2	Results	35
2.6	Findings from the Annotation Analysis	36
2.7	Summary	37
3	Spatio-temporal Human Body Segmentation	39
3.1	Introduction	40
3.1.1	Motivations	41
3.1.2	Spatio-temporal Human Body Segmentation: Overview	42
3.1.3	Spatio-temporal Human Body Segmentation: Contributions	42
3.2	Related Work	43
3.2.1	Object Detection	43
3.2.1.1	Appearance-based methods	44
3.2.1.2	Motion-based methods	44
3.2.1.3	Hybrid methods	48
3.2.2	Object Tracking	49
3.2.2.1	Point Tracking	50
3.2.2.2	Kernel Tracking	51
3.2.2.3	Silhouette Tracking	51
3.3	Human Body Segmentation Framework	52
3.3.1	Estimation of Human Body Region at Frame Level	53
3.3.1.1	Globalisation.	54
3.3.1.2	Graph setup: pixel and part relations.	56
3.3.1.3	Output: decoding eigenvectors.	56
3.3.2	Tracking of Detected Regions over Video Stream	57
3.4	Experiments and results	58
3.4.1	Experimental Setup	58
3.4.2	Human Detection Results	59
3.4.3	Human Segmentation Results	60
3.5	Conclusion	63
4	Human Action Recognition Framework and Visual Attributes Identification	65
4.1	Introduction	66
4.1.1	Motivations	68
4.1.2	Human Action Recognition Framework: Overview	69
4.1.3	Human Action Recognition Framework: Contributions	69
4.2	Related Work	69
4.2.1	Bag-of-Words model	71

4.2.1.1	Hard assignment coding	71
4.2.1.2	Soft assignment coding	71
4.2.2	Fisher Coding	72
4.2.3	Linear Coordinate Coding (LCC)	72
4.2.3.1	Vector Quantisation (VQ)	73
4.2.3.2	Sparse Coding (SC)	73
4.2.3.3	Locality-constrained Linear Coding (LLC)	74
4.3	Human Action Representation	75
4.3.1	Detecting and Tracking Human Body Regions	76
4.3.2	Non-human Object Detection Regions	77
4.3.3	Describing Detected Regions	78
4.3.4	Learning Feature Sets	78
4.4	Experiments	80
4.4.1	Datasets and the experimental procedure	80
4.4.2	Experimental Results	82
4.4.3	Discussion	86
4.5	Conclusion	87
5	Extraction of Qualitative Spatial and Temporal Relations	89
5.1	Introduction	90
5.1.1	Motivations	91
5.1.2	Qualitative Spatial and Temporal Relations: Overview	92
5.1.3	Qualitative Spatial and Temporal Relations: Contributions	92
5.2	Related Work	93
5.3	Individual Aspects of Qualitative Spatial and Temporal Relations	97
5.3.1	Topology	100
5.3.2	Size	102
5.3.3	Direction	103
5.3.4	Distance	103
5.3.5	Temporal Relations	104
5.4	Experiments	105
5.4.1	Video Data	106
5.4.2	Feature Selection and Clustering	106
5.4.3	Temporal Change Identification	107
5.4.4	Human Action Classification	108
5.5	Mapping from QSTR into Natural Language Terms	109
5.6	Conclusion	111

6	Generation of Textual Video Descriptions	113
6.1	Introduction	114
6.1.1	Motivations	115
6.1.2	Generate Textual Descriptions for Video Content: Overview	115
6.1.3	Generating Textual Descriptions for Video Content: Contributions	116
6.2	Related Work	116
6.2.1	Textual Image Descriptions	116
6.2.2	Textual Video Descriptions	117
6.3	Framework for Generating Textual Video Description	120
6.3.1	Visual recognition of Subjects	121
6.3.2	Visual recognition of Objects	122
6.3.3	Visual recognition of Verbs	122
6.3.4	Visual recognition of Prepositions	123
6.3.5	Visual recognition of Scene Settings	123
6.3.6	Zero-shot Language Statistics	124
6.3.7	Sentence Generation	125
6.3.8	Creating Cohesive Descriptions	125
6.4	Experiments and results	127
6.4.1	Frame-based Video Description Baseline	127
6.4.2	Evaluation with ROUGE Metric	128
6.4.3	Human Evaluation	130
6.4.4	Discussion	130
6.5	Conclusion	133
7	Human Action Retrieval via Textual Descriptions	135
7.1	Introduction	136
7.2	Related Work	137
7.2.1	Metadata retrieval	137
7.2.2	Low-level features retrieval	138
7.2.3	Semantic content retrieval	139
7.3	Human Action Text-based Retrieval Framework	140
7.3.1	Off-line Video Segmentation, Annotation and Indexing	140
7.3.2	On-line Searching and Ranking the Results	143
7.4	Experiments and results	144
7.4.1	Evaluation scheme	144
7.4.2	Results	145
7.5	Conclusion	148

8 Conclusion	151
8.1 Original Contributions	152
8.2 Future Work	154
References	157
Appendix A Identification of Additional Visual Attributions	177
A.1 Gender Identification	177
A.2 Age Identification	178
A.3 Facial Emotions Recognition	181
A.4 Scene Setting Identification	182

List of Figures

2.1	A montage of 3 minutes in an DriveCar video and three sets of hand annotations. This video segment of four shots depicting sequence of actions performed by four persons – extracted from Hollywood2 dataset ‘actionclipautoautotrain00094’	24
2.2	A montage of 2 minutes in an AnswerPhone video and three sets of hand annotations. This video segment of one shot depicting sequence of actions performed by one person – extracted from Hollywood2 dataset ‘actionclipautoautotrain00614’	25
2.3	Human related features found in hand annotations. The features are categorised into 8 groups as follows: gender, age, body parts, identity, emotions, grouping, dressing, and actions. For each group a list of high level concepts associated with their occurrences are extracted.	26
2.4	Non-human related features in the hand annotations. The features are divided into 6 groups as follows: man made objects, natural objects, scene settings, locations, size, colours. For each group a list of high level concepts associated with their occurrences are presented.	28
2.5	List of frequent spatial relations manually calculated from hand annotations associated with their frequency counts.	29
2.6	List of frequent temporal relations with their frequency counts.	30
2.7	The similarity degree across hand annotation pairs obtained by ten human subjects via Mturk questionnaire experiment.	33
2.8	Confusion matrix across 12 classes of hand annotations using Naive Bayes classifier with <i>tf-idf</i> features, where each column represents the instances in a predicted class and each row represents the instances in an actual class . .	37
3.1	Categories of object tracking methods. Figure adapted from Yilmaz et al. (2006)	49

3.2	Examples of different object tracking approaches. (a) represents point tracking, (b) kernel tracking, (c) and (d) silhouette tracking. Figure adapted from Yilmaz et al. (2006)	50
3.3	The two-step approach to segment human volume. The human body detected in the initial stage can be replicated in the subsequent video frames in the second stage.	53
3.4	Human body detection at frame-level achieved by combined low-level cues from Fowlkes et al. (2003) with top-down parts detector of Bourdev and Malik (2009)	54
3.5	Graph Setup: Pixel and Part Relations.	54
3.6	Sample of human detection results of ‘actionclipautoautotrain00463’ video clip from the AnswerPhone class.	60
3.7	Low-level cues enhanced the detection results that missed by top-down poselets human detector. Detection represented by bounding box and foreground model. Low-level cue helps in unusual pose (top row) and partial occlusion (bottom row).	62
3.8	Sample segmentations. The first row shows key frames from two video clips. The second and the third rows respectively present the results of key segments and the corresponding segmentation using the approach in this paper. The last two rows show the same attempts using the implementation by Lee et al. (2011) . Best viewed on pdf.	63
4.1	Typical stages via which image classification is performed by the BoW model	70
4.2	Comparison between the three linear coordinate coding techniques; the black circles represent the chosen codewords for the feature x_i	74
4.3	Processing flow of the ‘human body region tracking’ approach with visual object recognition (HBRT/VOC).	76
4.4	A sample clip from the NLDHA dataset: GetOutCar action from a video clip ‘actioncliptest00108’. A region was detected using Felzenszwalb et al. (2010) (red bounding box), while a human body was detected by using the HBRT approach (green contour). The car region was included in the action representation as there was an overlap between a car and a human.	77
4.5	Sample frames from the three action recognition datasets, the KTH (top row), the UCF Sports Action (middle) and the Hollywood2: Human Actions and Scenes (bottom), used for the experiments.	81

4.6	(KTH Dataset) Confusion matrix between six action classes using the HBRT/VOC combination approach, where each column represents the instances in a predicted class and each row represents the instances in an actual class.	83
4.7	(UCF Sports Action Dataset) Confusion matrix between ten action classes using the HBRT/VOC combination approach, where columns represent the predicted classes and rows represent the the actual class.	84
4.8	Samples for action localisation and segmentation. The 1st and 2nd rows respectively present the results of key segments and the corresponding segmentation using the HBRT/VOC on the KTH Dataset ('boxing', 'hand waving' and 'walking' actions). The 3rd and 4th rows show the results on the UCF Sports Actions Dataset ('diving', 'walking' and 'lifting' actions).	87
5.1	The difference between the axis-aligned bounding box (AABB) and the orientated bounding box (OBB). Either can be used with CORE-9 representation.	92
5.2	The 13 Allen's temporal relations that exist between two intervals X and Y .	94
5.3	Graphical representation of the Region Connection Calculus (RCC8), where A and B are two objects, and one of the following eight topology relations might occur between them: disconnected (DC), externally connected (EC), tangential proper part (TPP), tangential proper part inverse (TPPi), partially overlapping (PO), equal (EQ), non-tangential proper part (NTPP) or non-tangential proper part inverse (NTPPi).	95
5.4	Two objects A and B and their projections on the left, while the right shows how their projection identify the associated 9 cores (Cohn et al., 2012).	97
5.5	Sample segmentation of a human body volume for a 'hug' action from the TV Human Interaction dataset and an 'approach' action from the Mind's Eye video dataset. The first row shows original frames from two video clips. The second and the third rows respectively present the results of segmentation and the corresponding key segments. These datasets are to be introduced in Section 5.4.	98
5.6	Abstracting RCC-8 to RCC-3.	101
5.7	Different types of direction relations between two objects A and B	104
5.8	Sample frames from the two datasets used for the experiments: the TV Human Interaction dataset (first row), and the Mind's Eye video dataset (second row).	105

6.1	Summary of proposed framework of generation of video description. The framework starts with content planning stage that identify the HLFs such as people, objects, actions, spatial and temporal relations. In the case of zero-shot action recognition, language statistics from large text-based English corpora will be used to predict the missing verb (action class), given detected objects' classes and recognised scene settings. Finally, a template-based approach is used as surface realizer to convert these HLFs into textual descriptions.	121
6.2	Generating video description stages for the 'actionclipautoautotrain00428' video from the SitDown class, with three shots; the process start with the extraction of the HLFs list, followed by shot-based sentences generation. Finally, the paraphrasing stage is applied.	126
6.3	Example of applying post-processing rules to the system-generated description of 'actionclipautoautotrain00463' video from the AnswerPhone category, with two shots.	127
6.4	Sample of textual video descriptions along with their video shots from different categories from the NLDHA dataset.	132
6.5	Computer vision techniques can result in errors: (a) although the man is detected, his action is misclassified as walking rather than as answering the phone; (b) only a man is identified while the woman next to him not detected; (c) the car is not detected and only two sitting men are identified; (d) two persons who are eating are detected correctly from this scene but the rest are missed.	133
7.1	The content-based video retrieval pipeline (Bhat et al., 2014).	139
7.2	A visualisation of dependency parse tree for the sentence 'A man is sitting on a chair' associated with syntactic relations between tokens using Spacy's dependency parser.	143
7.3	The retrieved video clips samples from a 'a man is driving a car. Next, he is getting out the car' query. The top row is the user textual query. Followed by five top relevant retrievals that depict the main actions in the query (drive a car, get out a car) using SI.	147
7.4	The retrieved video clip samples from different queries that involved spatial relations.	148
A.1	Gender classification approach proposed by Bekios-Calfa et al. (2011)	179

A.2	Confusion matrix for gender identification. Columns show the ground truth, and rows indicate the classification results.	179
A.3	Age classification approach proposed by Choi et al. (2011)	180
A.4	Confusion matrix for age identification. Columns correspond to the ground truth, whereas the rows represent the classification results.	181
A.5	Confusion matrix for human emotion recognition. Columns show the ground truth, and rows indicate the automatic recognition results.	182
A.6	Scene recognition result for video clip named ‘actioncliptrain00366’ from ‘DriveCar’ class using system proposed by Zhou et al. (2014)	183

List of Tables

2.1	Similarity scores within 12 hand annotations using the cosine similarity approach. For each class, scores are calculated under three conditions: (A) raw hand annotations without applying any pre-processing; (B) applying the Porter stemmer and removing stop words, without replacing synonyms; (C) without removing stop words, but applying the Porter stemmer and replacing synonyms.	32
2.2	Detailed accuracy results for supervised classification using Naive Bayes classifier with <i>tf-idf</i> features per class.	36
3.1	Detailed detection results of performance evaluations on the NLDHA dataset for each class using combination of low-level cues and top-down detector information. Columns from left to right, first column lists the classes names in our dataset, second is the number of valid human regions in ground-truth for each class, third is the number of human detections for each class using proposed framework, fourth is the number of missed human detection for each class, fifth is false alarm which corresponds to false positive resulted from our system, last two columns correspond to the percentage of correct detections and false alarms respectively over the number of valid human detections per class.	61
3.2	The average number of incorrectly segmented pixels per frame. The video clip name is in the format of ‘sceneclipautoautotrain·····’ where the ‘·····’ part is shown in the table.	62
4.1	(KTH Dataset) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.	83
4.2	(UCF Sports Action Dataset) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.	84

4.3	(Hollywood2: Human Actions and Scenes Dataset) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.	85
4.4	(Hollywood2: Human Actions and Scenes Dataset) Recognition accuracy for individual action classes. Units are in %. The best score for each class is highlighted by bold fonts. The numbers by Laptev et al. and by Bilen et al. were extracted from (Bilen et al., 2011).	85
5.1	Confusion matrix for the action classification task with the Mind’s Eye video dataset achieved by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.	107
5.2	Confusion matrix for the action classification task with the Mind’s Eye video dataset achieved by AngledCORE-9 (Sokeh et al., 2013).	107
5.3	Precision and recall scores for the human action classification task using the Mind’s Eye video dataset by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.	108
5.4	Precision and recall scores for the human action classification task using the Mind’s Eye video dataset by AngledCORE-9 (Sokeh et al., 2013).	108
5.5	Precision and recall scores for the human action classification task using the TV Human Interactions dataset by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.	109
6.1	The set of vocabulary used to produce textual descriptions of video.	122
6.2	The English corpora used to mine the SVOP tuples.	124
6.3	ROUGE scores calculated for the baseline and our approach, with respect to hand annotations. Four different ROUGE metrics are measured: ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-L (longest common subsequence) and ROUGE-SU4 (skip-4 bi-gram). For each ROUGE metric, the recall (R), precision (P), and F-measure (F) are averaged over all twelve categories from the NLDHA dataset.	129
6.4	Human evaluation for the baseline and our approach, with respect to three aspects: grammatical correctness, cognitive correctness, and relevance. . . .	130
7.1	The text-based video retrieval results using two systems the VSM and SI using set of two queries for each of 12 human action classes from the NLDHA dataset.	146

Chapter 1

Introduction

Generating textual descriptions of human activities in video streams is one of the most prominent topics in both computer vision and natural language generation, and has been increasingly studied in recent years; however, these studies are still in their infancy, especially for realistic settings and complex video scenarios, such as relatively long videos containing a variety of actions. The challenge of this task mainly lies in the difficulties associated with extraction of high level features (HLFs) from video data, such as temporal continuity, background clutter and occlusion. The task of generating textual descriptions for human activities within videos demands many aspects of detecting and tracking human subjects, identifying their visual attributions, formalising their spatial and temporal relations, and finally combining this information together into coherent textual form. This thesis studies the translation of visual data into textual descriptions, specifically focusing on parsing and extracting semantic video content that can capture and summarise the main aspects of the human activities shown and identify the main visual information that can be used to describe the video content. The framework is evaluated against a human action retrieval task. This chapter presents a brief background, along with the motivation for the work and an overview of the research focus. The main contributions of this work are subsequently defined. Finally, the thesis structure is illustrated.

1.1 Background

Public search engines such as YouTube and Google Videos tend to be the most frequently visited websites. The number of videos uploaded to these websites is increasing exponentially. Hence, content-based video retrieval is required to meet the users' needs, and retrieve the desired data from huge video databases. One of the major challenges associated with content-based video retrieval systems is identification of representative features from video data and

translating them to another modality that facilitates the indexing and retrieval tasks. Despite the advance in computer vision techniques, computers are still able to reliably identify and recognise only a limited set of objects and activities.

Hence, inspired by the fact that the human activities videos are human-centric, the development of a robust approach that is able to detect the human presence and identify their associated visual attributions as well as the sophisticated relations between them provide a valuable resource for video analysis applications such as video summarisation and video retrieval. One of the potential applications for such a framework is the human action retrieval approach that has become a key new topic in the big data field as it capable of searching for videos based on the human actions performed (Ramezani and Yaghmaee, 2016).

Despite this potential, most current video retrieval systems still rely on matching query against user metadata text only. Unfortunately, such approaches can fail because of the large number of videos marked with inaccurate or incomplete user metadata, or none at all, as the users usually tag their videos with keywords that increase its spread rather than reflecting the actual content of the videos (Cheung and Zakhor, 2003). Searching for and retrieving related videos quickly turns out to be a challenging task. The automatic generation of semantic textual video descriptions that capture and summarise the main aspects of the video has become an urgent need.

1.2 Motivation

It has become essential to properly define video semantics, as there is a dramatic ongoing rise in video data nowadays. This will aid users in obtaining relevant information according to their personal interests. One way to explain video semantics and contents is that they can be changed into some other modality, for instance written text. A natural way of communicating is the use of human language, and useful objects extracted from videos associated with their inter-relations can be syntactically and semantically formed and presented in natural language.

Describing a video in natural language is easy for humans as they have the capability to understand a visual scene. They intuitively use their domain knowledge to describe a scene based on its visual contents. Despite the fact that natural language techniques are being increasingly integrated into vision systems concepts, computers can only identify and recognise a limited set of objects and activities. The majority of earlier studies conducted have been about semantic indexing of videos with the help of keywords (Chang et al., 2007). However, it is usually difficult to represent the semantic video relations between different objects and activities with the use of keywords alone. Natural language textual description is

more human-friendly and is able to simplify context between entities by understanding their relations.

Video content description framework generally comprises of two main stages named content planning and surface realisation (Guadarrama et al., 2013). The former stage is concerned with the identification of visual information that can be used to describe the video semantic content such as, human, actions and their spatial and temporal relation, whereas the latter stage is responsible for translating the predicted sentence components (subject, verb, object) into syntactical correct sentences. The majority of existing approaches aimed at generating a textual description of a video data share the common part of a visual model of the image-based detector (DPM) (Felzenszwalb et al., 2010), which is applied to each frame to augment a store of detected objects, without preserving any temporal dependency between video frames (Bin et al., 2016). As a result, the descriptions generated using this detector suffer from several weaknesses, such as redundancy and lack of coherence.

To remedy this shortcoming, accurate video content representation is proposed that is considered a key step to robustly describe human activities in video data. Inspired by the fact that the majority of video data relates to human figures (Ramezani and Yaghmaee, 2016), a fine-grained spatio-temporal human segmentation can alleviate the shortcoming of image-based detector and produce an accurate semantic video representation. The segmented regions will be utilised to enhance the performance of an action recognition approach and remedy the limitation of point feature-based methods which cannot be used to label the actions of multiple tracks occurring simultaneously. To this extent, this work particularly contributes to the content planning stage of describing video data and investigates the extraction of spatio-temporal entities' segments, the identification of their visual attributions (action, gender, age, emotion), and the formulation of their spatial and temporal relations. Finally, this information will be expressed using natural language generation techniques.

1.3 Research Focus

Extensive studies are being conducted to advance visual recognition techniques and semantic understanding of natural language. However, surprisingly little work has been done to explore the mutual profit of integrating both fields, given that natural language is the common technique for communicating such information to humans. This thesis demonstrates how the different research fields can benefit from each other, through addressing the problem of producing natural language descriptions of human activities in videos. This task could benefit a variety of applications, including generating automatic textual descriptions of movies for visually impaired people, human-computer/robot interaction, and producing textual

summaries of web videos for indexing and retrieval purposes. Additionally, being able to translate visual content into textual form is a crucial step in understanding the relationship between the richest modalities available to humans, namely visual and linguistic information.

Generating textual descriptions for visual content is considered to be an intriguing task that requires integration of both visual recognition fundamentals and natural language generation techniques. While the image description task is relatively well studied (Farhadi et al., 2010; Kulkarni et al., 2011; Mitchell et al., 2012), the majority of video description approaches still use simple video settings, typically depicting one action summarised by one sentence (Barbu et al., 2012; Gygli et al., 2014a; Thomason et al., 2014). Although these researches have started exploring the visual content description domain, two main research questions arise when conducting such task. The first question is about which visual information in videos is usually verbalised by humans. The second concerns the best approach to converting visual information into linguistic expressions.

To answer the first question, a proper corpus that can be used for development and evaluation is created in this study. The generated corpus contains human textual annotations for relatively long video clips of human activities that depict a variety of action classes as the existing video corpora are inappropriate for this task of describing human activities and their inter-relations, as they either contain short clips, each of which depicts one action, meaning that their associated annotation is only one sentence (for example the NLDV corpus (Khan et al., 2012) and ACL2013 dataset (Yu and Siskind, 2013)) or they are more domain-focused, such as the AMI Meeting corpus (Carletta et al., 2006) and the TACoS Cooking dataset (Regneri et al., 2013). Extensive analysis of human annotations provides insights into human interests while watching videos, and concludes that the annotators are mainly interested in human presences and their visual attributions in the video stream. The human annotation will be used as a reference to evaluate the automatic video description.

To answer the second question, a four-stage approach is developed, involving human body regions segmentation, action recognition identification, formalisation of spatial and temporal relations, and finally the combining of this information to generate a textual description, using a template-based approach. Firstly, a robust approach is urgently needed to detect human presences over video frames. The majority of previous works used the image-based detector for this purpose, which is applied to each frame to create a store of detected objects, without preserving any temporal dependency between video frames. However, video data introduces a temporal dimension that needs to be considered in order to produce accurate and compact semantic representations.

Human body segmentation and tracking plays a vital role in a wide range of video analysis tasks. Despite being relatively little studied so far, these topics still have two main

issues that need to be addressed. The task remains challenging as it must ultimately be tackled automatically, with minimal user intervention and a negligible number of pre-defined constraints on video settings, such as the presumption that a static camera is being used rather than a moving camera, or indoor versus outdoor settings. This thesis first introduces a novel fully-automatic human body extraction algorithm that is able to successfully segment out human body regions, and track the corresponding regions over video frames, regardless of video settings. Specifically, the video segmentation task is formulated as a graph partitioning problem and a number of challenges associated with this task are addressed, such as features extraction and avoidance of error propagation through the tracking process. Qualitative and quantitative evaluation results on real-life videos demonstrate the effectiveness of the proposed framework.

Secondly, a novel action recognition framework is proposed, where the video representation is improved by using the extracted spatio-temporal human regions combined with the extended spatio-temporal locality-constrained linear coding (LLC) technique, in order to identify the action class. The majority of previous researches into action recognition were based on space-time interest points, whereas more spatially extended features, such as regions, have received considerably less attention. The performance of this model is assessed through a human action classification experiments, using the KTH, the UCF sports, and the Hollywood2 datasets. The outcome shows that the local region-based approach with LLC coding technique clearly outperforms the feature-based techniques, particularly with the more challenging Hollywood2 dataset.

Thirdly, the spatial and temporal relations of prominent objects are recognised as playing a vital role in describing video semantic content. The extracted object segments are utilised to formalise their spatial and temporal relations, in order to be able to use them later for the generation of natural language descriptions of human activities in video clips. To this extent, an improvement over the AngledCORE-9 approach introduced by [Sokeh et al. \(2013\)](#) is proposed – a comprehensive representation to efficiently extract spatial information between interacting objects in a video clip by utilising their approximate oriented bounding box (OBB). An approach that incorporates the spatio-temporal volume of objects into AngledCORE-9 is introduced, and extends the extracted relations to accommodate the temporal information. As a result, the proposed approach is able to represent interacting objects in a video stream in an efficient manner, as accurate spatial and temporal information can be obtained by precise representation of the shape region and the OBB.

Finally, a framework that produces textual descriptions of videos, based on the semantic video content extracted in the previous stages, is presented. Detected action classes will be rendered as verbs, participant objects will be converted to noun phrases (mainly subject

for humans and object for other), visual properties of detected objects will be rendered as adjectives, and spatial relations between objects will be rendered as prepositions. These HLFs are converted into textual descriptions using a template-based approach. Paraphrasing of the resulting shot-based descriptions is introduced to create compact and coherent video descriptions. The proposed video descriptions framework is evaluated using ROUGE scores and human judgments. A human action retrieval system is developed based on automatic video description, to evaluate the effectiveness of the proposed framework.

1.4 Thesis Contributions

This section identifies the main thesis contributions, along with the motivation behind the work carried out.

Contribution 1: Corpus Generation and Analysis

Motivation. Although the method of annotating through keywords is relatively well researched and established (Bolle et al., 2010), the quality of this approach can be improved through natural language annotation. Approaches to generate natural language descriptions of human activities within video streams can be explored. The initial phase of conducting this study was to create a proper dataset that could be used for development and evaluation, as there is currently no publicly available resource that considers the temporal relations between a video’s entities. Generating such a corpus help to limit the scope of this study to tight and manageable domains, and guides the identification of the main HLFs to be extracted by computer vision techniques. Finally, this resource will be used as a ground truth for evaluation.

Contribution. The video clips are manually selected from the Human Actions and Scenes Dataset (Hollywood2 dataset) (Marszalek et al., 2009a). These videos were chosen based on two main criteria – the number of camera shots and the variety of human actions performed in the video – in order to explore spatial and temporal relationships between individual shot components, and to produce a story for the complete video clip. We did not intend to create a dataset with a full range of video categories, which is beyond the scope of our research. Instead, the aim was to generate a compact dataset that could be used for developing an approach to translate the visual content of human activities into textual descriptions. The 120 video clips selected have at least one human present in each shot. Annotations were made manually by 12 participants in two ways: a title, which consists of a single phrase or

sentence (where a specific topic or primary concept is outlined), and a full textual description using a number of sentences (providing an extensive description of the visual scene).

Contribution 2: Spatio-temporal Human Body Segmentation

Motivation. Extensive analysis of the hand-annotations associated with this corpus brought the conclusion that annotators are most interested in human presences and their visual attributions in the video stream, especially their actions, gender, emotions and their interaction with other humans and objects. Most previous work on video description task relies on identifying HLFs at frame-level, without exploiting temporal information [Khan et al. \(2015\)](#). As a result, the temporal information associated with each video object is discarded, which can lead to incoherent descriptions. In order to efficiently describe video semantics, a strong framework that can detect and segment areas of the human body from a series of video clips must be established, to be able to extract these HLFs.

Contribution. In order to obtain an accurate and compact description of an input video, this description should be based on semantic visual content, taking into account the temporal dimension. Based on the fact that a human figure is the most important variable video component over a temporal dimension, extracting and describing their body volume over the duration of the video sequence leads to a comprehensive and compact description. In this study, a novel approach is presented where human body regions are extracted from a video sequence. The approach detects and segments human body regions by jointly embedding parts and pixels, which utilise the advances of low-level image cues and high-level part detectors information ([Maire et al., 2011](#)). The appearance and shape models are learned for extracted segments, in order to automatically identify the foreground objects across a sequence of video frames ([Lee et al., 2011](#)).

Contribution 3: Human Action Recognition Framework

Motivation. In order to effectively describe human figures in video streams, their actions need to be identified. Feature learning and coding techniques have been extensively studied in the image processing domain to generate global representations with less coding than the original extracted features. However, these techniques ignore the spatial relationships between features. As a result, they are unable to locate an object or capture shapes ([Wang et al., 2010](#)). Many extensions are proposed to alleviate this shortcoming and add locality constraints, to project each descriptor into their corresponding local-coordinate system. Then, the final video signatures are produced by integrating the projected coordinates using

pooling techniques. However, these techniques are still incapable of capturing the temporal information of video sequences at the coding stage.

Contribution. Once the human body segments are determined, the performed actions must be identified. For this task a new action recognition framework is proposed, where the video representation is improved using extracted spatio-temporal human regions combined with the extended spatio-temporal locality-constrained linear coding (LLC) technique, in order to identify the action class. The using of spatio-temporal human regions improves the feature extraction by focusing on accurate position of actors. Additionally, the LLC coding technique successfully represents video content with less coding than the original extracted set of features, which assists in reducing processing time and storage space.

Contribution 4: Extraction of Qualitative Spatial and Temporal Relations

Motivation. Video sequences are composed of temporal information that establishes the relationship between individual frames. In order to build a video description that is sufficiently robust, the spatial relations between interacting objects must be formalised, as well as temporal relations. An approach to develop ways to estimate the relative positions of humans and the bounding boxes of discrete objects from the individual frame has been proposed by [Khan and Gotoh \(2012\)](#). The use of bounding boxes in this case has the advantage of making spatial relations between HLFs comparatively simple to establish using image processing techniques. However, the previous work implemented a small set of spatial relations, still limited to covering only some types of such relations. Further, in that work no temporal information was considered between frames, which is essential for full description of human activities in video sequences.

Contribution. Spatial and temporal relations of prominent objects play a vital role in describing videos' semantic content. To that extent, an approach is designed for formalising the spatial and temporal relations between interacting objects using their spatio-temporal object bounding boxes and the intervals that result from aligning them to the video frames and applying AngledCORE-9 representation [Sokeh et al. \(2013\)](#). By measuring nine cores, spatial information about two objects can be obtained, such as their topology, direction, relative size and the distance between them. Temporal changes can also be extracted from these cores and their associated intervals by processing a sequence of video frames. The approach was able to detect spatial changes that occurred over time, promoting good semantic

understanding of the video content. Additionally, a set of rules was defined to transfer the qualitative spatial and temporal relations into meaningful natural language terms, which will be used later for generating semantic video descriptions.

Contribution 5: Generation of Textual Video Descriptions

Motivation. Generating textual descriptions of visual content is an intriguing task that requires a combination of two major research aspects: visual recognition approaches and natural language generation (NLG) techniques. To generate descriptions for videos and images, a template-based approach is a powerful tool which needs to be identified (Barbu et al., 2012; Gygli et al., 2014a; Khan and Gotoh, 2012). An alternative approach is to retrieve descriptive sentences from a training corpus based on visual similarity, or to utilise external text-based corpora to help rank the visual detections (Das et al., 2013b; Hanckmann et al., 2012; Mitchell et al., 2012). However, the majority of previous works rely on simple setting videos which are described by one sentences.

Contribution. A framework that produces textual descriptions of video, based on the semantic video content at shot-level, is implemented. Detected action classes will be rendered as verbs, participant objects will be converted to noun phrases (mainly subject for humans and object for other), visual properties of detected objects will be rendered as adjectives and spatial relations between objects will be rendered as prepositions. Further, in cases where no verb is assigned for a given track (called zero-shot action recognition), as the action recognition system is unable to identify the performed action because the action has not previously appeared in the training data, a language model is used to infer a missing verb, aided by the detection of objects and scene settings. These HLFs are converted into textual descriptions using a template-based approach. Paraphrasing of the resulting multi-sentences shot-based descriptions is introduced to create compact and coherent video descriptions.

1.5 Thesis Overview

The remaining content and structure of this thesis is summarised below.

- **Chapter 2: Corpus Generation and Analysis.** This chapter presents the initial stage of the proposed framework, which is to create a proper corpus for development and evaluation purposes. The generated corpus contains relatively long video clips of human activities, depicting a variety of action classes along with 12 human textual

annotations for each. The chapter's content is based on the paper [Al Harbi and Gotoh \(2016\)](#) and justifies Contribution 1.

- **Chapter 3: Spatio-temporal Human Body Segmentation.** This chapter introduces a novel framework that is able to identify human body regions over sequences of video frames. The detection is achieved by combining low-level image cues with high-level part detectors information. Consequently, the detected regions can be tracked over frame sequences by learning their colour and shape models. The chapter's content is based on the papers [Al Harbi and Gotoh \(2013b, 2015b\)](#) and justifies Contribution 2.
- **Chapter 4: Human Action Recognition Framework and Visual Attributes Identification.** This chapter proposes an action-recognition framework that is able to identify the action performed by human subjects during video sequences. Also, experiment results on the human action classification task are presented. The chapter's content is based on the papers [Al Ghamdi et al. \(2012\)](#) and [Al Harbi and Gotoh \(2013a\)](#) and justifies Contribution 3.
- **Chapter 5: Extraction of Qualitative Spatial and Temporal Relations.** This chapter utilises the extracted segments and formalises their spatial and temporal relations. In this context, spatial relations specify how objects are related to each other within a sampled frame, while temporal characteristics are used to describe the changes of spatial relations between two objects over the time domain. The chapter's content is based on the paper [Al Harbi and Gotoh \(2015a\)](#) and justifies Contribution 4.
- **Chapter 6: Generation of Textual Video Descriptions.** This chapter introduces a framework for generating textual descriptions of human activities in videos sequences at a shot-based level, relying mainly on semantic visual detections. The automatic descriptions are compared and evaluated using ROUGE scores and human judgment evaluation. The chapter's content is partly based on the paper [Khan et al. \(2015\)](#) and justifies Contribution 5.
- **Chapter 7: Human Action Retrieval via Textual Descriptions.** This chapter presents a new video retrieval system based on automatic video descriptions. To verify the efficiency of automatic descriptions, the human action retrieval task is selected as an application, using classical information retrieval methods.
- **Chapter 8: Conclusion.** This chapter concludes the work of this thesis by presenting a summary, recommendations, and suggestions for future research directions.

1.6 Published Work

This thesis is partly based on the following publications.

Journal Papers:

1. A unified spatio-temporal human body region tracking approach to action recognition. *Nouf Al Harbi, and Yoshihiko Gotoh*. In *Neurocomputing*, 161: 56–64, 2015.
2. A framework for creating natural language descriptions of video streams. *Muhammad Usman Ghani Khan, Nouf Al Harbi, and Yoshihiko Gotoh*. In *Information Sciences*, 303: 61–82, 2015.
3. Describing Qualitative Spatio-Temporal Relations between Object Volumes in Video Streams. *Nouf Al Harbi, and Yoshihiko Gotoh*. *Submitted for Image and Vision Computing*, 2017.
4. Human action retrieval via complex textual queries. *Nouf Al Harbi, and Yoshihiko Gotoh*. (*In preparation for the Machine Vision and Applications*, 2017).

Conferences and Workshops Papers:

1. Spatio-temporal Video Representation with Locality-Constrained Linear Coding. *Manal Al Ghamdi, Nouf Al Harbi, and Yoshihiko Gotoh*, In *12th European Conference on Computer Vision (ECCV), ARTEMIS workshop*, 2012.
2. Action recognition: spatio-temporal human body region tracking approach. *Nouf Al Harbi, and Yoshihiko Gotoh*. In *Proceedings of the Second Workshop on Recognition and Action for Scene Understanding, CAIP*, 2013.
3. Spatio-temporal human body segmentation from video stream. *Nouf Al Harbi, and Yoshihiko Gotoh*. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2013.
4. Describing spatio-temporal relations between object volumes in video streams. *Nouf Al Harbi, and Yoshihiko Gotoh*. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
5. Natural language descriptions of human activities scenes: Corpus generation and analysis. *Nouf Al Harbi, and Yoshihiko Gotoh*. In *Proceedings of the 5th Workshop on Vision and Language (VL'16)*, 2016.
6. Natural language descriptions of human activities in video streams. *Nouf Al Harbi, and Yoshihiko Gotoh*. (*Submitted for the International Natural Language Generation conference INLG*, 2017).

Chapter 2

Corpus Generation and Analysis

This thesis outlines and advances a video description framework that describes human activities and their spatial and temporal relationships. Further, it will validate this framework, by employing a video retrieval application. It is essential to discern the goals of each specific phase and subsequently the way in which they connect with one another.

This chapter presents the first stage of the proposed framework, which is to create a proper corpus that can be used for development and evaluation. The generated corpus contains relatively long video clips of human activities, depicting a variety of action classes along with human textual annotations for each. The existing video corpora are inappropriate for the task of describing human activities and their inter-relations, as they either contain short clips, each of which depicts one action, meaning that their associated annotation is only one sentence (for example the NLDV corpus (Khan et al., 2012) and the ACL2013 dataset (Yu and Siskind, 2013)) or they are more domain-focused, such as the AMI Meeting (Carletta et al., 2006) corpus and the TACoS Cooking dataset (Regneri et al., 2013).

The chapter is structured as follows: Section 2.1 introduces the corpus generation and analysis process along with the motivations for this work and its contributions. Section 2.2 reviews previous work related to the corpora used for video description task. Details of the corpus generation are presented in Section 2.3, while Section 2.4 explains the corpus analysis. A summary of the results obtained from the experiment is given in Section 2.5. Findings based on corpus analysis are presented in Section 2.6. Finally, Section 2.7 provides a concluding discussion.

2.1 Introduction

There has been continuous growth in the volume and ubiquity of video material. It has become increasingly essential to define video semantics in order to aid the searchability, indexing

and retrieval of this data. One way to describe video semantics and content is to transfer them into a different modality, for example, written form. Human language is a normal mode of interaction. Beneficial elements elicited from videos and their interconnections can be portrayed using human language in a linguistically and semantically accurate way.

Humans have the capability to recount the content of a video without any difficulty, using everyday language and their innate ability to comprehend visual imagery. They are able to explain the material being viewed based on their understanding of the subject matter. Conversely, computers can merely identify and classify a selection of objects and specific actions. The majority of prior research has focused on the semantic indexing of videos, utilising keywords. But it is frequently challenging to solely employ keywords to illustrate the connection between different entities and events captured on video. A noteworthy addition to the keyword system is the use of everyday language for the textual description of videos. They have greater levels of user-friendliness. In addition, they can explain the context of keywords by identifying the relationships between them.

Using everyday language to outline the content of videos is crucial for the typical video retrieval process. Considering the large amount of both public and privately generated multimedia materials available, the challenge is that only a minority contain explanatory annotations. In addition, it is not possible to retrieve most of them using queries based on everyday language. The use of descriptions employing natural language is one technique for portraying the content of a video. This also serves to reduce the volume of file space required, as well as the retrieval time. It may be as abstract as a series of keywords, as discursive as a passage of text, or alternatively a complete story. Extensive stories are more beneficial than the use of key phrases because the correlation between the entities is more transparent. Frequently, the scenes depicted in videos are protracted, and occasionally it can be challenging to define them solely through a number of key phrases.

Video synopses can be created by converting video summaries using natural language. They serve to generate a multimedia repository where video analysis, retrieval and summarisation can be developed. The majority of previous research, in particular for video description tasks, has relied upon short video clips that depict one subject performing one action. Therefore, the textual annotations related to them typically comprise a single sentence. By contrast, reality-based video scenarios incorporate various camera shots depicting a range of actions.

2.1.1 Motivations

Although the method of annotating through keywords is relatively well researched and established (Bolle et al., 2010), the quality of this approach can be improved through natural

language annotation. We are exploring approaches to generate textual descriptions of interrelations between human activities within video streams. The first step of the study is to create a proper dataset to be used for development and evaluation, as there is no such publicly available resource that considers the temporal relations between a video's entities. Creation of such corpora will help in limiting this research to tight and manageable domains. Additionally, it leads to the identification of high level features (HLFs) to be extracted by computer vision techniques. Finally, this resource will be used as ground-truth for evaluation purpose.

To this end, a video corpus of video clips and hand annotations with natural language descriptions has been generated. This corpus is named 'NLDHA – Corpus'¹ and consists of hand annotations of 120 video segments. The video clips are manually selected from the Human Actions and Scenes Dataset (Hollywood2 dataset).² These videos were selected based on a number of criteria, such as number of camera shots and variety of human actions performed during the stream, in order to explore spatial and temporal relationships between individual shot components, and to generate a story for the complete video sequence.

2.1.2 Corpus Generation and Analysis: Overview

This chapter outlines the different phases involved in the compilation and analysis of a comprehensive corpus as required for this thesis. The initial task comprises the selection of appropriate video clips, accompanied by the collation of textual annotations for them all. Lastly, an examination of this collection provides an awareness of human behaviour and engagement when viewing videos. It also establishes the list of primary HLFs that need to be derived from image and video processing techniques to define and describe the semantic content of the video.

The primary contribution outlined in this chapter is extended a subset of video clips manually selected from the Hollywood2 dataset by human annotations. In addition, there is a need to identify the key content to be included in the description of these video streams. While the choice of specific content may be a personal matter, a variety of similar and complementary themes emerge when auditing what appeals to individuals when viewing videos. Typically, visual content identification may be categorised into two areas, namely: (i) object identification; and (ii) establishing an object's actions and the interconnection between them. The most significant objects and actions, as well as their interactions, should be recognised for effective semantic visual understanding. In addition, these contents should be expressed using appropriate words drawn from everyday language.

¹ NLDHA stands for Natural Language Descriptions for Human Activities in videos

² This dataset is available from: <http://www.di.ens.fr/~laptev/actions/hollywood2/>

2.1.3 Corpus Generation and Analysis: Contributions

The main contribution of the proposed work in this chapter is the collection of the manual annotation by 12 participants, for a video corpus of relatively long video clips ranging from 1 to 3 minutes in length, selected from the Hollywood2 dataset. The dataset consists of 120 segments of video – 10 segments for each of the following twelve categories: DriveCar, GetOutCar, Eat, AnswerPhone, Kiss, HandShake, Run, FightPerson, HugPerson, StandUp, SitDown and SitUp. The manual annotation done in two ways: a brief synopsis (title) consisting of a single phrase or sentence (where a specific topic or primary concept is outlined); and a full explanation in everyday language, set out using a number of sentences (providing an extensive description of the visual scene).

The main contributions of the proposed work in this chapter is collection of manual annotation by 12 participants for a video corpus of relatively long video clips ranging from 1 to 3 minutes in length selected from Hollywood2 dataset. It consists of 120 segments of video – 10 segments for each of the following twelve categories: DriveCar, GetOutCar, Eat, AnswerPhone, Kiss, HandShake, Run, FightPerson, HugPerson, StandUp, SitDown and SitUp. This annotation done in two ways: a brief synopsis (title) consisting of a single phrase or sentence (where a specific topic or primary concept is outlined); and a full explanation in everyday language, set out using a number of sentences (providing an extensive description of the visual scene).

2.2 Related Work

Most of the video corpora that were used for evaluating event recognition are not appropriate for evaluating sentential descriptions. For instance, as [Blank et al. \(2005\)](#) highlighted, the Weizmann dataset and KTH dataset used by [Schuldt et al. \(2004\)](#) only facilitate with depicting events with one human participant, and do not allow people to interact with other individuals or objects. The textual description in these datasets consists of only one sentence, containing information related to a verb, such as the person is jumping. However, these datasets, along with others such as YouTube dataset ([Liu et al., 2009](#)) and Sports Actions dataset ([Rodriguez et al., 2008a](#)), frequently help in making action class distinctions that are not relevant to the current project, such as Jump vs. Pjump, Kicking-side vs Kicking-front, Tennis-swing vs. Golf-swing. Other datasets, for instance the UCF50 dataset ([Reddy and Shah, 2013](#)) and the ballet dataset ([Wang and Mori, 2009](#)), indicate a relatively larger scale of activities, which include class names of activity that are not appropriate for sentential description, such as Breaststroke, Basketball, Yo Yo, Hula Hoop, Clean-and-Jerk, Military Parade and Horse Race.

However, more recently a number of video corpora and images have been generated geared towards annotation with natural language, where they are specifically designed with certain prerequisites or constraints to fulfil a specific task or purpose. Some of these corpora are reviewed in detail below.

AMI Meeting Corpus. The AMI Meeting Corpus (Carletta et al., 2006) has 100 hours of meeting recordings content. During the meetings, participants also use unsynchronized pens to record what is written during this meeting. The audio character is variable, as the meetings are split between three environments, and though recorded in English, for most participants this is a second language. The corpus also provides manual orthographic transcription for each speaker. Annotation is also given for a variety of features, including linguistic events and behaviours in other modalities, such as dialogue, topic segmentation, named entities, head and hand gestures, extractive and abstractive summaries, directions of gaze in connection with communications intent, the adoption of new positions within rooms, head locations within video frames, and the emotional states of participants. However, this corpus cannot be used for our task, as it was created for the specific purpose of the meeting scenario, and also the annotation includes casual style of English to reflect the speaker transcript, and includes such words as *wow*, *um* and *yeah*.

TREC Video Datasets. TREC video evaluation is a continuous series of yearly workshops concentrating on selected information retrieval (IR) tasks. It supplies a sizeable test collection, a set of uniform scoring procedures, and a research team forum for presenting results, aiming to support related research activities. Among the IR tasks set is HLFs extraction, which attempts to determine whether high level semantic features are present or absent within a specified video stream (Smeaton et al., 2009). Rushes video has been utilised to investigate various video summarisation approaches (Over et al., 2007).

Further, a range of metadata annotations for video datasets are available through TREC video, such as speech recognition transcripts, which usually consist of informal language, master shot reference lists, and shot IDs of HLFs provided for the HLF retrieval task. Each shot (single camera takes) is annotated for the summarisation task. Shots may include numerous variables, such as multiple human actors and objects, and different backgrounds. Usually, annotations are constructed of a few keywords or phrases of varying length, rather than a complete story, which make them inappropriate for our video description task. Features relating to humans are frequently described, including appearance, physical characteristics and movement, as well as technical details (camera angles and movement). However, less

information is provided about objects and events, or about the emotional state of the humans being viewed.

NLDV Corpus. Khan et al. (2012) explains that the purpose of designing this dataset was to generate natural language descriptions of video content. The work emphasises the phase of natural language generation, which is based on visual features to a great extent, extracted in the phase of HLFs processing. TREC Video datasets craft datasets that comprise actions, objects, scenes setting and subjects, which are easy to identify with use of existing visual processing techniques. Thus, the videos that are crafted are short and comprised of a single shot or scene with minimal activity; as a result they cannot be used to describe human activity with inter-relations between scenes. The NLDV dataset contains seven categories with 20 sections for each, totalling 140 video segments, each segment spanning between 10 and 30 seconds in length.

ACL2013 Dataset. Word meanings can be learnt from video that is coupled with sentences, a methodology proposed by Yu and Siskind (2013). A range of combined situations can be compiled into a dataset of 61 short filmed video clips (each 3–5 seconds at 640×480 resolution and 40 fps). Every clip is made up of a combination of a number of synchronous instances, which can involve a subset of up to four different entities: a chair, a garbage can, a backpack and a person. Three contrasting outdoor surroundings were used in filming these clips to ensure that there was sufficient cross-validation. Manual notes were inserted into each video clip to explain what is happening. The corpus of these 159 training examples couples up videos with more than one sentence and sentences with more than one video, but on average there are 2.6 sentences per video.³ These videos are relatively short and some of them depict non-human objects' activities, without human presence, such as an aeroplane landing, which makes this dataset inadequate for the proposed task of video description.

TACoS Cooking Dataset. The TACoS Cooking dataset (Regneri et al., 2013)⁴ is a subset of the videos in MPII Comosites (Rohrbach et al., 2012). This dataset was proposed for addressing the issue of grounding sentences that describe actions in visual information extracted from videos. Moreover, this corpus helps in aligning video clips with multiple natural language descriptions of the actions portrayed in the videos. 127 videos of 26 basic cooking tasks are included in the dataset, for example cutting a cucumber was recorded between 4 and 8 times. A total of 22 subjects were used for recording a corpus in the kitchen

³ The videos and sentential annotations are available at <http://haonanyu.com/research/ac12013/>.

⁴ This corpus is available from: <http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos>

environment. Each video is used for representing only one task that an individual subject executes. 20 different textual descriptions are collected for each video, which results in 2540 annotation assignments. These assignments are published on Amazon Mechanical Turks (MTurk)⁵, where each annotator was responsible for entering a minimum of five or a maximum of 15 complete sentences in English for describing the events in the video. However, this corpus was designed for the specific purpose of cooking, and as a result all actions are centred on the kitchen environment, which make it unsuitable for a general video description task.

SumMe Dataset. SumMe is a new benchmark proposed by Gygli et al. (2014b) for the task of summarising video.⁶ There are in total 25 videos included in the SumMe dataset, which cover sports, events and holidays. Since there is no human presence in some of them, this dataset is inappropriate for our task. These videos are raw or have been edited minimally by the user, which means that they have a relatively higher compressibility than videos that have been already edited. The length of the videos vary approximately between 1 and 6 minutes. The study included a total of 41 participants (19 males and 22 females) that had different educational backgrounds, for summarising the videos' visual content. Around 15 to 18 different people summarised each video.

Synthesized Multi-view IXMAS Activity Dataset.

Synthesized Multi-view IXMAS Activity Dataset. Liu et al. (2016) developed a latent discriminative structural model to detect complex activity and atomic actions, while learning the temporal structure of atomic actions. A synthesized set of complex activities is constructed by concatenating simple actions from the multi-view IXMAS dataset (Weinland et al., 2007)⁷, which contains 12 simple action classes. Each activity video is constructed by concatenating five different simple actions selected from the 12 classes. For each view, eight complex activity classes are synthesized, where different activity classes have five different atomic actions. However, this dataset cannot be used for our study as only one human subject performs specific action per video segment.

⁵ The Amazon Mechanical Turks are available at <https://www.mturk.com/mturk>.

⁶ Dataset and evaluation code are available on: www.vision.ee.ethz.ch/~gyglim/vsum/

⁷ Dataset and evaluation code are available on: <http://perception.inrialpes.fr>

2.3 Corpus Generation

Marszalek et al. (2009b) conducted a study that utilises scene context for human action recognition in video. The proposed framework was validated on a Human Actions and Scenes Dataset (Hollywood2 dataset).⁸ The corpus is built on top of the Hollywood2 dataset, which is collected from 69 different Hollywood movies. The dataset includes 12 action classes from real-life video scenes: DriveCar, GetOutCar, Eat, AnswerPhone, Kiss, HandShake, Run, FightPerson, HugPerson, StandUp, SitDown and SitUp. This dataset was selected as it has realistic and generic video settings including human subjects that show different activities, interactions with other subjects, and emotions. The total length of action clips is around 600k frames, or 7 hours of videos.

Our corpus, NLDHA, consists of basic video clips selected manually from the Hollywood2 dataset and then extended with 12 textual descriptions obtained via Amazon Mechanical Turk (MTurk) for each video. The chosen clips were selected based on their satisfying one of the following conditions: the video should contain various camera shots of human activities, to help in fulfilling the purpose of this study to understand the temporal and spatial association of human activities; or if only one camera shot is present it should consist of a variety of actions, performed either by one or multiple persons.

The aim was not to derive a dataset that covered all the video categories (that is beyond the scope of the work); rather the intention was to develop a compact dataset which can help to develop approaches for translating the video contents of human interaction and their temporal and spatial relations to natural language descriptions. The selected video dataset consists of 10 video for each of the twelve action classes in the Hollywood2 dataset, resulting in 120 video clips in total. These categories can be classified based on human interactions into two main themes:

Humans vs Humans interactions In these videos, multiple humans interact with each other.

This theme includes the following categories: FightPerson, HandShake, HugPerson, Run and Kiss.

Humans vs Objects interactions: The human present interacts with some objects, such as cars, dining tables or chairs, performing some action such as ‘*sitting, driving and eating*’. This includes the following categories: AnswerPhone, DriveCar, SitDown, StandUp, SitUp, Eat and GetOutCar videos.

The majority of selected segments contained multiple camera shots, with 6 shots on average, varying between indoor and outdoor scene-settings. The total length of the selected clips is 225000 frames, with a sampling rate of 25 frames per second, and an average length

⁸ This dataset is available from: <http://www.di.ens.fr/~laptev/actions/hollywood2/>

of 1875 frames for each video. These videos span between 1 and 3 minutes, with an average of 75 seconds for each clip. All categories relate mainly to humans' activities, expressions and emotions. Sequences of actions and activities are performed by one person, depicted in one shot, whereas multiple shot videos depict relations and interaction between multiple humans. Some of them depict humans' interactions with other objects in a variety of indoor and outdoor settings.

2.3.1 Collecting Textual Video Descriptions

Amazon Mechanical Turk (MTurk) was used to conduct this experiment. A Human Intelligence Task (HIT) was created and published on MTurk, using an adapted version of the annotation tool Vatic (Vondrick et al., 2010). In order to study the influence of the annotators subjectivity factor, the general video annotation practice was followed (Khan et al., 2012; Regneri et al., 2013) and for each video we collected 12 different textual annotations, leading to 1440 annotation assignments. The human annotators prepared two different types of manual descriptions for these video segments: title assignment (a single phrase) and full description (multiple sentences). A title, to some extent, can be described as a summary provided in the most compact form, which includes the essential themes, or contents of the video in a short phrase. In contrast, full description is detailed and comprises various sentences with in-depth details of activities, objects and their interactions. In the rest of this thesis they are referred to as 'manual annotations'. A valuable resource for text-based video retrieval and summarising of tasks can be created through the combination of titles and full descriptions.

For each assignment, one video was shown to the annotator, who was then asked to provide a title for the video in one sentence that highlighted the main theme of the video. The annotator was also asked to provide a description of a minimum of 5 and a maximum of 15 complete English sentences to explain the events of the video. In order to help subject annotators understand the task, they were presented with a sample video segment. Some possible textual annotations were provided, with a title for, and complete description of, the video stream. The selection of the HLFs depicted in the video segment was also shown, to help the annotators understand the annotation generation procedure; for example, the list can include nouns, verbs, adjectives and prepositions shown in the video clip.

Instructions were provided in a simple form that allowed individuals to interpret what they may include in the description. The instructions included using an open vocabulary for making annotations, meaning that annotators had the freedom to use any English word. They were asked to not use any computer codes or symbols, as these cannot be used to describe video content. Further instructions were provided to annotators for not using proper

nouns, such as avoiding stating the person’s name, and data collected through audio, since this information can affect the quality of the description of the semantic video content. Additionally, annotators were asked to explain all the human activities they saw in the video, and were allowed to watch each video as many times as they wanted, by forwarding or skipping backwards, according to their preference.

Annotators were encouraged to take notes when watching the video, and to make a draft of the annotation before they entered it, to ensure they had not missed any part of the video clip. Once the annotators became familiar with the video, the final annotations were done by the annotators by watching the complete video, without any opportunity for non-sequential viewing. They were required to enter each sentence immediately when the action described by the sentence had been completed. At the beginning of the sentence input, the video playback was paused automatically.

HIT approval rate is a measure that reflects the overall work quality, based on the proportion of the cumulative number of tasks done by a given worker approved by the requester. Accordingly, a 75% HIT approval rate was required for the task, to ensure the quality of the English language used by annotations. Increasing the HIT approval rate might increase the quality of the manual annotation, but higher approval rates are usually reserved for tasks that need academic skills, such as surveys. For video annotation tasks, the selected HITs approval rate of 75% has proven to be sufficient, as reported in (Regneri et al., 2013), and it ensures that workers with a wide range of experiences are eligible to complete the HIT. An amount of 1.20 USD was paid for each task; before making the payment, the annotations were randomly inspected, and the quality manually checked. The total cost for the collection of annotations amounted to 1,728 USD. The information was collected within a timeframe of three weeks.

2.4 Annotation Analysis

The total number of documents for this corpus was 1440 (12 annotators created descriptions for 120 videos each). The total number of words for descriptions was 67080; hence the average length of one document was roughly 47 words. We counted 5136 unique words and 2336 keywords (nouns and verbs).

The hand annotations were of two types: ‘title’ and ‘description’. The title often consisted of only a couple of words that do not constitute a complete sentence. Most annotators use verbs in particular to express the main theme of the video, like family eating dinner, men fighting, three people driving. The average length of title in our corpus is three words. An extensive analysis was performed on the title part of the annotation, and it became clear

that most annotators were in agreement in identifying the main theme of the video, though there were differences between them in the words used to express the topic. Figure 2.1 is an example of variation between annotators for the same video, in the words used as a title and agreement on the main video theme: ‘furious man crushing a window on the car’, ‘beating a car and running’, ‘smashing car window’. Here, we can see three annotators expressing the same meaning using different words – ‘crushing’, ‘beat up’ and ‘smash’.

On the other hand, the ‘Descriptions’ on average contain four to six phrases or sentences; basically one sentence describes each camera shot. Most sentences are concise, ranging between six and eight words. Descriptions for human, gender, emotion and actions, with their temporal order, are commonly observed. Some details of objects, dress and location are occasionally stated, unless these objects participate in the event. Typically, annotators were varying in the detail included, from very abstract to very detailed descriptions. However, they often preserved the time order of the activities performed in the video clip. Figure 2.1 shows an example of three different hand annotations for the same video ‘actionclipautoauto-train00094.avi’ from the DriveCar action class in outdoor scene. This video segment consists of four different shots depicting multiple actions performed by four men. It can be observed from these descriptions the difference between annotators in the amount of detail included in the description. However, almost all of the annotators maintain the same temporal order of activities executed during the video; the main two activities in this video are smashing and driving.

Figure 2.2 shows another example of a video segment for a human activity and three sets of hand annotations. This video consists of a single shot depicting variety of actions performed by one person in indoor setting. It is interesting to note that all three annotators in ‘Descriptions’ give high attention to the humans present and their activities. All three annotators mentioned the main activities that occur in the video in the same time order (sit, write, stand, get, sit, call) though annotators had differences of opinions about the HLFs in the feature, theme of the video and description styles. ‘Titles’ for this video segment look quite similar for annotators (2) and (3), as they choose making a call as the main theme, while annotator (1) chose writing a diary as the title. Finally, it is interesting to observe the subjectivity within the task; the variety of words were selected by individual annotators to express the same video contents, for example, annotator (1) used ‘gets up’ while annotator (3) used ‘stands up’ for the same action and annotator (2) used ‘person’ while annotator (3) used ‘man’ for the same actor.



Hand annotation 1

(title) Furious man crushing a window on the car;

(description) Furious man is crushing window on the car with iron stick and screaming. After that, we see him and the other two men driving in the car. All except the driver are eating sandwiches. Then we see driver sticking nails in wooden lath.

Hand annotation 2

(title) Beating a car and running;

(description) Angry man begins beating up a car. He breaks the windshield and windows. Then there are three men riding in another car. They are eating and riding somewhere. Then a man is beating nails into a board.

Hand annotation 3

(title) Smashing a car window;

(description) A man is smashing the window of a parked car with a sledge hammer at night. Next, the man who was speaking on the phone is driving a car with two other men as passengers at night. Later, another man is speaking on the phone while hammering nails in a board.

Fig. 2.1 A montage of 3 minutes in an DriveCar video and three sets of hand annotations. This video segment of four shots depicting sequence of actions performed by four persons – extracted from Hollywood2 dataset ‘actionclipautoautotrain00094’.



Hand annotation 1

(title) A man writing a diary;

(description) A man sits on a chair and he writes on his diary. Then, he gets up and gets the phone. Later, he dials to make a call. Finally, he sits back again.

Hand annotation 2

(title) Making a Phone Call;

(description) A person is writing some notes on his diary during night time. Then, he is making a telephone call to someone. Next, he is sitting down on his chair.

Hand annotation 3

(title) Phone call;

(description) A man is sitting at his desk in a dark apartment. He is writing a note. He is standing up and picking up the phone to bring it to his desk. He is sitting down and beginning to call someone.

Fig. 2.2 A montage of 2 minutes in an AnswerPhone video and three sets of hand annotations. This video segment of one shot depicting sequence of actions performed by one person – extracted from Hollywood2 dataset ‘actionclipautoautotrain00614’.

2.4.1 Human Related Features

Figure 2.3 illustrates the human-related information that was highlighted in hand annotations. Annotators emphasised the presence of humans in the video, as the two words (nouns) that were used most frequently include ‘woman’ with 375 occurrences and ‘man’ with 635 occurrences. Full attention was paid to the human presence in the video by the annotators using different words. To identify the main HLFs list related to human, all human presence related words are divided into two main categories based on their gender ‘male’ and ‘female’. In the category of ‘female’ related words that indicate this gender, such as ‘lady’, ‘girl’ and ‘woman’, were included. Similarly, for male related words – for example ‘boy’ and ‘man’ - were combined. This supports that humans and their attributions were the most important and interesting HLFs, as they were presented in the video. By contrast, factors such as identity (for example officer, police, father), age information (such as young, old, child) were not

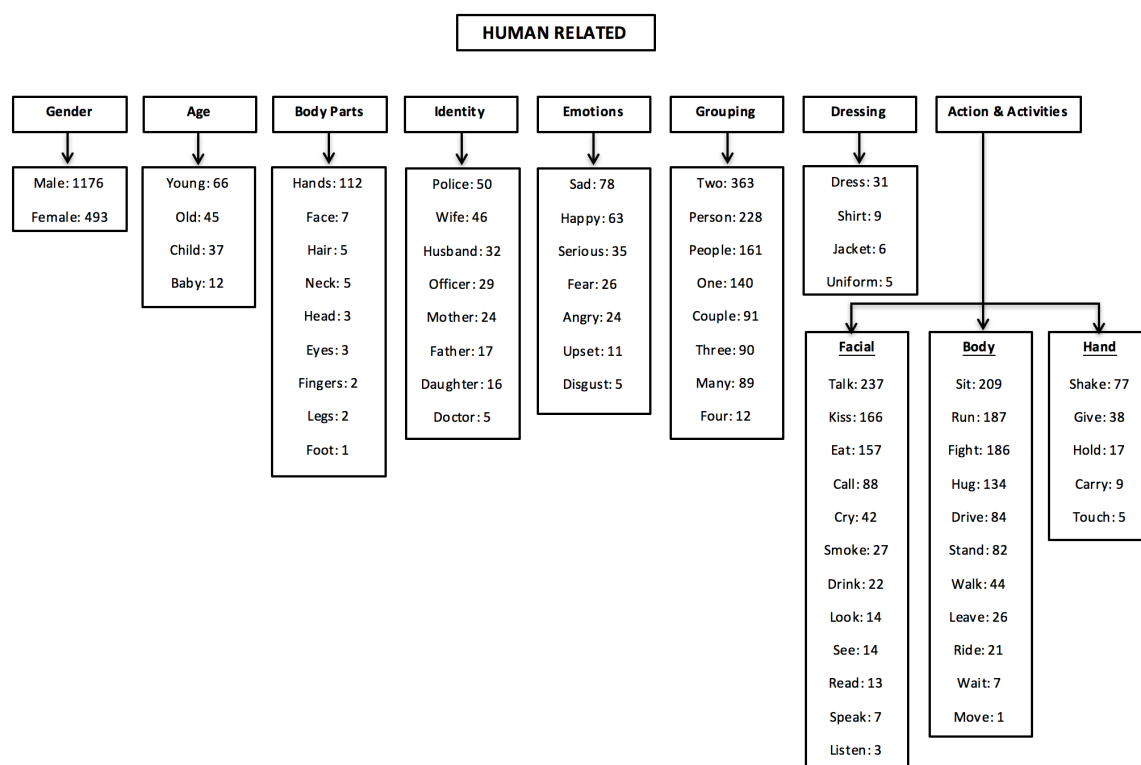


Fig. 2.3 Human related features found in hand annotations. The features are categorised into 8 groups as follows: gender, age, body parts, identity, emotions, grouping, dressing, and actions. For each group a list of high level concepts associated with their occurrences are extracted.

identified very often. Names for human body parts had mixed occurrences, ranging from high ‘hand’ to low ‘foot’.

Ekman (1992) has mentioned that the six basic human emotions are: fear, anger, sadness, surprise, disgust and happiness, which are also the most frequent human facial expressions in hand annotations. Another interesting feature was dress; when an individual was dressed in a unique manner, for instance wearing a formal suit, an army or police uniform or a coloured jacket, it was noted – otherwise the feature was not considered frequently. Video including multiple humans was also very common, and therefore information about human grouping was frequently recognised. Human activities were identified through the involvement of human body parts, including actions such as walking, running, sitting, fighting and standing. Other actions relevant to the human body and posture were identified frequently. Unique identities such as doctor, police and officer were rarely described.

2.4.2 Objects and Scene Settings

Figure 2.4 illustrates the hierarchy that has been developed for HLFs, which was not found in Figure 2.3. Many of the words denoted the artificial objects, and an interaction was found between humans and those objects for completing activities, such as ‘man is sitting on chair’, ‘he is driving a car’ and ‘she is talking on the phone’. Other important factors were information about the place and location (such as office, restaurant, shop, school), which helped in knowing about the human or object’s position in the scene (*e.g.* ‘people are eating in the restaurant’, ‘there is a car on the road’).

In order to identify separate HLFs, information about colour played an essential part, such as ‘she is wearing a white uniform’, ‘a man in a black shirt is walking with a woman with a green jacket’. Considering the great number of colour occurrences, it is evident that humans also have an interest in observing the colour scheme in visual scenes, along with the objects. The interest of annotators in the scene-setting in the background or foreground was reflected through a few hand descriptions. Some annotators were also interested in outdoor/indoor scene-settings. The interest of viewers in high level details of the video, and the association between prominent objects in a visual scene, was demonstrated by these observations, for example in terms of size (‘two boys are seated on a small boat’, ‘a lady with long hair is walking on the road’). Natural objects were rarely described in the hand annotations.

2.4.3 Spatial Relations

Visual scenes in video are best described in terms of spatial relations, which can be defined as how objects are located in space in relation to some reference object. This is very important in describing visual scenes. In a video stream this reference object is usually in the foreground. The competent use of prepositions, such as on, at, inside or above, can facilitate the creation of flowing and concise descriptions in presenting the spatial relations between objects. In order to identify these prepositions, different aspects of spatial relations should be explored. According to [Cohn et al. \(2008\)](#) there is a variety of spatial expressions which can be used to gain accurate spatial representations. These relations can be classified into three main categories: direction, distance and topology. Direction relations are used to describe the direction of one object relative to another, for example (‘left’, ‘under’), while distance relations specify how far the object is away from the reference object, such as (‘far’, ‘near’). Finally, topology relations specify how an object is located in relation to another object in space, such as (‘touch’, ‘inside’).

A record of the most frequently recurring words in the corpus which concern spatial relations is presented in Figure 2.5. It is obvious from the high incidence of these words

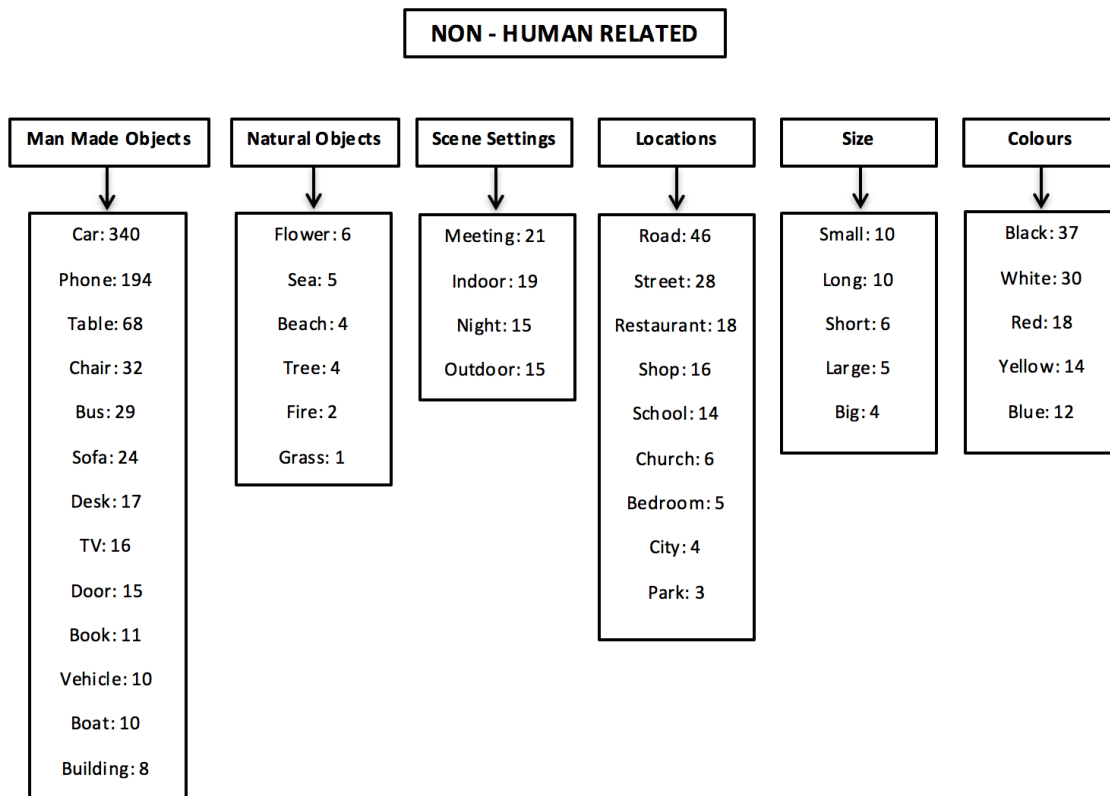


Fig. 2.4 Non-human related features in the hand annotations. The features are divided into 6 groups as follows: man made objects, natural objects, scene settings, locations, size, colours. For each group a list of high level concepts associated with their occurrences are presented.

that people use them regularly to describe visual scenes. The semantics of the visual scenes are better understood because the use of these words contributes to the identification of connections between different HLFs. A record of spatial relations occurrence words has been manually calculated for many reasons, as follows. Firstly, the words denoting spatial relations may have a multitude of alternative uses in addition to spatial relations. The following three phrases demonstrate how the word ‘in’ can be used to represent different purposes: ‘three people are sitting in a car’ represents a spatial relation, whilst ‘the dog in the last shot’ depicts a relationship between various scenes; ‘two people in a dialogue’ augments the ease with which the description can be read. Secondly, the spatial word can be a preposition by itself, for example ‘in’, or it can be syntactically overlapped with another preposition, for example ‘in front of’, as in ‘three persons are talking in front of shops at night’. Finally, there are some preposition words that can be used for both spatial and temporal relations, for example ‘at’ in the following examples: ‘A man is smashing the window of a parked car with a sledge

<p>in: 653; on: 335; with: 235; at: 121; between: 36; around: 26; behind: 25; touch: 23; middle: 21; together: 20; inside: 17; far: 16; in front of: 13; beside: 11; to the right: 10; to the left: 8, near: 6; under: 5; in the middle: 3</p>
--

Fig. 2.5 List of frequent spatial relations manually calculated from hand annotations associated with their frequency counts.

hammer at night’ uses as temporal relation, whereas in the following sentence: ‘There are three people eating dinner at home’, it is used for spatial relation purposes.

2.4.4 Temporal Relations

Temporal expressions such as ‘before’, ‘long’, ‘a while’ or ‘during’, describe when something happened, the duration and how often it occurred (Pustejovsky et al., 2003). Temporal and spatial relations are combined in videos as a sequence of time series data using highly sophisticated multi-dimensional content. A complete and systematic video sequence is created when individual frames are linked with each other. Annotators then use temporal data to combine the narratives for sequential frames and produce a complete account of the video content. Three separate frames can be connected using the two temporal relations, ‘then’ and ‘later’. For example;

*A man and woman are talking and the woman walks out of the house; **then** she sees him through the door as he is passing in the street; **later**, another man enters the house.*

A total of thirteen relations – ‘overlaps’, ‘overlapped-by’, ‘starts’, ‘started-by’, ‘meets’, ‘meet-by’, ‘finishes’, ‘finished-by’, ‘equals’, ‘after’, ‘before’, ‘contains’ and ‘during’ – make up a temporal logic, as identified by Allen and Ferguson (1994), who also proposed that scenarios can be more often described using time intervals than time points. Temporal relations play an important role in picking out activities in videos. According to Allen’s temporal logic, the most prevalent relations in video sequences for an individual person include ‘start’, ‘finish’, ‘before’ and ‘after’.

Based on corpus analysis, there are two themes of videos which can be used for depicting temporal relations. i) obtaining temporal relations from activities of an individual human and ii) multiple humans interacting with each other. Figure 2.6 contains a list of the most commonly used words contained in the annotation corpus regarding temporal relations. It is obvious that the annotators place more importance on the keywords which are connected to numerous human activities. The most typical keyword for distinct activities performed simultaneously by multiple humans is ‘while’, for example, ‘A man is eating while his friend is drinking’.

<p>Single human: then: 125; after: 60; afterwards: 44; before: 42; later: 32; throughout: 32; start: 27; end: 25; next: 25; finish: 25;</p> <p>Multiple humans: while: 87; during: 47; overlap: 12; meanwhile: 12; throughout: 12; in: 11; at: 10; equals: 4,</p>

Fig. 2.6 List of frequent temporal relations with their frequency counts.

Moreover, for individual human activities, it is usually temporal relations words that are used to determine the chronological order of actions occurring, for example, ‘A man comes into the room a little awkwardly. Then he sits on the chair’. It can be observed from Figure 2.6 that the annotators are interested in describing the temporal order of events in a video clip, especially if only one actor is featured. On the other hand, for multiple humans activities videos, corpus analysis shows that the annotators are much more likely pay high attention to the actions carried out simultaneously by different people, rather than describing single human activities, if present. For some temporal relations which relate to both ‘single human’ and ‘multiple human’ videos, such as ‘throughout’, the calculation of occurrences was done manually to prevent overlapping and ensure that none of them were counted twice.

2.4.5 Similarity between Descriptions

Cohen (1960) proposed Cohen’s kappa coefficient (κ), which is a well-established approach for calculating human inter-annotator agreement for qualitative items. Cohen’s kappa κ is used for content analysis, such as in psychiatry, to measure the agreement level between students’ diagnoses on a group of test cases with expert answers (Grove et al., 1981). Carletta et al. (1997) deserve credit for attracting the attention of computational linguistics to the use Cohen’s kappa κ approach in their field.

Cohen’s kappa calculates the agreement between two raters, where each of them classifies N items into C mutually exclusive classes. However, for the task of measuring similarity between hand annotation descriptions, it was not possible to calculate the inter-annotator agreement with the use of the κ approach, as in hand annotations significant subjectivity was involved, and as a result there was clear variation in the description length for each video among annotators.

A more effective and well-known approach to measuring text similarity is based on the Cosine Similarity, which works independent of document length, one of its important properties. The similarity between two documents corresponds to the correlation between the vectors when the documents are represented as term vectors. This can be quantified as the cosine of the angle between vectors. Huang (2008) suggests that one of the most popular

similarity measures that can be applied to text documents is the Cosine Similarity, which can be represented using a dot product and magnitude of term vectors.

Let $D = d_1, \dots, d_n$ be a set of documents and $T = t_1, \dots, t_m$ the set of distinct terms occurring in D . A document is then represented as a m -dimensional vector \vec{t}_d . Let $tf(d, t)$ stand for the frequency of term $t \in T$ in document $d \in D$. Then the vector representation of a document d is:

$$\vec{t}_d = (tf(d, t_1), \dots, tf(d, t_m)) \quad (2.1)$$

Given two vectors \vec{t}_a and \vec{t}_b , their cosine similarity is

$$SIM_C(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{|\vec{t}_a| \times |\vec{t}_b|}, \quad (2.2)$$

where \vec{t}_a and \vec{t}_b are m -dimensional vectors over the term set $T = t_1, \dots, t_m$, the numerator represents the dot product of \vec{t}_a and \vec{t}_b , and the denominator is the product of their Euclidean lengths. Each dimension is used for representing a term with its weight in the document which is non-negative, due to which the cosine similarity is non-negative and bounded within $[0, 1]$. Where the documents are completely different the result is 0, and in the situation where two documents are identical the result is 1.

For the purposes of completing this experiment, a number of standard text processing filtering techniques were used. The first measure was the removal of stop words (Flood, 1999) as there are some words such as *a*, *and*, *are* and *do*, which are non-descriptive for the purpose of this document. The second measure involved stemming, which is mapping words with different endings into a single word using the Porter stemmer (Porter, 1980). As an example, *production*, *produce*, *produces* and *product* will be mapped to the stem *produc*. The prevalent current thinking is that morphological variations of words with the same root/stem can be regarded as a single word because thematically they are the same. Finally, it is usually helpful to minimise the vocabulary of a text by substituting words with common synonyms without affecting meaning and this can be achieved by using NLTK WordNet interface.⁹ Synonyms will reduce the annotators' variation and subjectivity caused by their use of different words for the same concept, and will also increase the occurrence of significant collocations.

The average similarity scores within 12 hand annotations for each of the 120 video across 12 categories are shown in Table 2.1. Individual description scores were used for calculating the average, which was compared with the remaining descriptions in the same category. The results highlighted that while watching videos humans take different interests

⁹www.nltk.org/

	(A)	(B)	(C)	Average
AnswerPhone	0.5294	0.5236	0.5446	0.5325
DriveCar	0.5564	0.5587	0.5632	0.5594
Eat	0.5272	0.5386	0.5386	0.5348
FightPerson	0.4010	0.4104	0.4245	0.4119
GetOutCar	0.4679	0.4607	0.4707	0.4664
HandShake	0.3955	0.4034	0.4187	0.4058
HugPerson	0.4036	0.4216	0.4236	0.4162
Kiss	0.3868	0.4065	0.4187	0.404
Run	0.3996	0.4056	0.4076	0.4042
SitDown	0.3925	0.4065	0.4158	0.4049
SitUp	0.3898	0.3952	0.4023	0.3958
StandUp	0.4043	0.4074	0.4274	0.4130

Table 2.1 Similarity scores within 12 hand annotations using the cosine similarity approach. For each class, scores are calculated under three conditions: (A) raw hand annotations without applying any pre-processing; (B) applying the Porter stemmer and removing stop words, without replacing synonyms; (C) without removing stop words, but applying the Porter stemmer and replacing synonyms.

and observations. All different combinations of setting are tested and the three best results are reported in Table 2.1. In two conditions the calculation was repeated: one was carried out by removing the stop words and applying the Porter stemmer, but without replacing synonyms, and in the other condition the stop words were not removed, but synonyms were replaced and the Porter stemmer was applied. The latter combination of pre-processing techniques resulted in better scores. Thus, with replacement of synonyms the performance was improved, which indicates that humans are likely to express the same concept with the use of different terms. There were some classes such as DriveCar, AnswerPhone and Eat for which the similarity score was relatively higher than the rest of the categories. There was a certain uniformity in the hand annotations for these videos as a result of the presence of major objects associated with humans and their actions, such as car, phone, and dining table. The majority of annotators were in agreement in paying high attention to such objects, and they used the same concept names for their description, leading to high similarities between video descriptions across different annotators. Conversely, for all the other classes the annotators varied in the type of information and details included in their description, even for the same video, although they remained similar in that they included the main verbs for actions performed by the actors.

To determine whether the value of the acquired similarity scores for the hand annotations corpus is sufficient, a human-based judgement experiment was carried out. For this, the similarity outcomes are split into four groups, based on their values. The groups are as

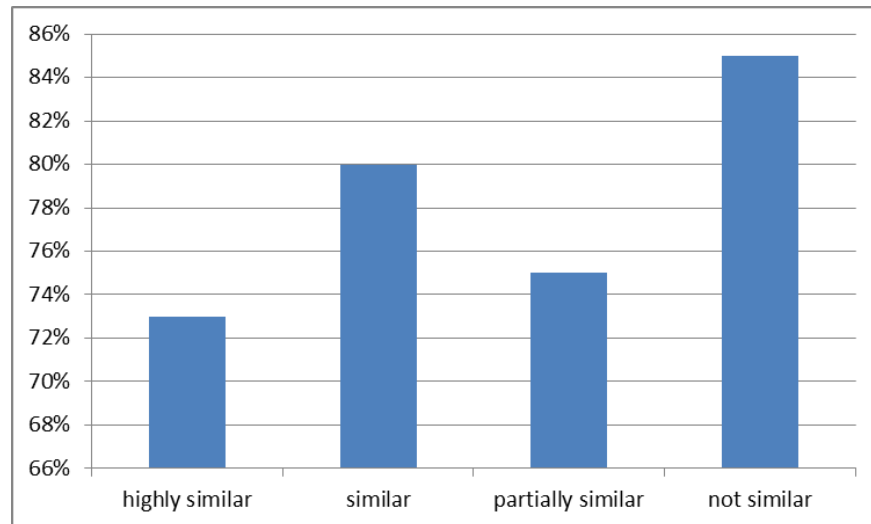


Fig. 2.7 The similarity degree across hand annotation pairs obtained by ten human subjects via Mturk questionnaire experiment.

follows: values greater than 0.75 are classified as highly similar, values between 0.75 and 0.5 classified as similar, values between 0.5 and 0.25 as partially similar, and values below 0.25 classified as not similar. Next, ten pairs were randomly selected from every group and used to answer the following question by human subjects:

Question: How much is the given pair of annotations similar, spanning from ‘highly similar’ to ‘similar’, ‘partially similar’ and ‘not similar’.

A total of ten human subjects carried out this experiment and ranked the extent of similarity shared by 40 pairs of selected hand annotations. The resource Amazon Mechanical Turk (Mturk) was utilised for hiring the required human subjects. To conduct the questionnaire, the human subjects needed to have an HIT approval grade of 75% or more. This was necessary to guarantee the accuracy of gathered outcomes.

The findings from the experiment were analysed by contrasting the outcomes of the questionnaire with the outcomes of the cosine similarity with the same setting of Table 2.1(C) (applying the Porter stemmer and replacing synonyms but without removing stop words) for the same hand annotation pairs. The typical similarity value amongst the ten human subjects was calculated. Figure 2.7 demonstrates that the proportion of instances on which they agreed was larger than expected, ranging between 73% and 85% in terms of the four similarity groupings. In accordance with the outcomes from this analytical strategy, the significant degree of matching among questionnaire results and baseline findings demonstrates that the acquired similarity extent using cosine similarity values is sufficiently accurate, and as a result reflects the sufficiency and good quality of the hand annotation.

2.5 Action Classification Experiments

This section uses an action classification task based on the manual annotations for demonstrating the application of the NLDHA corpus with natural language descriptions. [Dumais et al. \(1998\)](#) explains that textual document features can be expressed through a *tf-idf* score. The relation between document d and term t can be represented through traditional *tf-idf*. The importance of a term within a particular document can be measured, calculated as:

$$tfidf(t, d) = tf(t, d) \cdot idf(d) \quad (2.3)$$

where the term frequency $tf(t, d)$ is given by

$$tf(t, d) = \frac{N_{t,d}}{\sum_k N_{k,d}} \quad (2.4)$$

The number of occurrences of term t in document d is presented through $N_{t,d}$ in the above equation; the sum of the number of occurrences for all terms in document d is given in the denominator, which is the size of the document $|d|$. Further, the inverse document frequency $idf(d)$ is

$$idf(d) = \log \frac{N}{W(t)} \quad (2.5)$$

where N is the total number of documents in the corpus and $W(t)$ is the total number of documents containing term t .

2.5.1 Experimental Setup

For the purposes of completing the experiment we needed to use a number of standard text processing techniques. Firstly, stop words are removed by using the technique available in the Weka machine learning workbench ([Hall et al., 2009](#)), which has 527 stop words. The second measure involved stemming, which is mapping words with different endings into their single root word, and this has been achieved by using the NLTK Porter stemmer.¹⁰

The third measure involved removing words that arise with less than a given threshold frequency, after looking at the effect of allowing infrequent terms in the document. The reason for removing infrequent terms is that often they make little or no contribution to the overall description of the document's subject. In addition, including uncommon terms increases the cost of determining any similarity because of additional noise in the clustering

¹⁰www.nltk.org/

process. For our experiments we therefore chose the top 1000 words, ranked according to frequency of use.

In the experiment, the Naive Bayes probabilistic supervised learning algorithm was applied for classification using the Weka machine learning library. The Naive Bayes classifier is a probabilistic model based on Bayes rule, with the assumption of independence among predictors. Naive Bayes relies on a very simple document representation model, a bag of words, as the order of words is ignored and only word occurrence numbers are considered. Despite the simplicity of the Naive Bayes model, it works very well and outperforms many sophisticated classification techniques. Finally, a ten-fold cross validation was applied. Performance was measured using precision (the fraction of selected instances that are relevant), recall (fraction of correct instances that are retrieved) and F1-measure, which is the harmonic mean of both precision and recall:

$$F1 = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} \quad (2.6)$$

2.5.2 Results

Table 2.2 and Figure 2.8 indicate the outcomes of the monitored classification assessment using *tf-idf* characteristics. The F1 value was greater than expected for a number of groups, for example ‘AnswerPhone’, ‘Eat’, ‘DriveCar’ and ‘Run’ videos, and in the context of human and primary items (dining table, car, and phone). The annotators were to some extent agreed on the main information included in these videos’ description. Also, these classes carried high classification ratings, as they drew a lot of interest from annotators while the video events were being outlined. In comparison, the classification outcomes acquired for ‘SitDown’, ‘SitUp’ and ‘StandUp’ were the lowest.

This was likely due to the fact that the annotators didn’t give these actions great attention, for two main reasons. Firstly, in some videos these human actions are performed very quickly or briefly in the videos. Secondly, another reason noted for high variation between annotators for these classes is because these actions are usually associated or overlapped with another action by different subjects that in the annotator’s view were more important; for example, a person is sitting down or standing up during an eating scene video; in this example the annotator may focus on the eating action and describe it in detail. Generally, the classification outcomes demonstrated the best performance for videos featuring humans engaging with objects. The classification outcomes are positive, which indicates that the corpus gathered is a reliable tool for assessing natural language description frameworks of video clips.

	Precision	Recall	F1-measure
AnswerPhone	0.836	0.85	0.843
DriveCar	0.803	0.85	0.826
Eat	0.855	0.883	0.869
FightPerson	0.786	0.858	0.821
GetOutCar	0.791	0.725	0.757
HandShake	0.817	0.783	0.8
HugPerson	0.921	0.775	0.842
Kiss	0.783	0.783	0.783
Run	0.939	0.9	0.919
SitDown	0.623	0.675	0.648
SitUp	0.686	0.583	0.631
StandUp	0.483	0.575	0.525
Average	0.777	0.77	0.772

Table 2.2 Detailed accuracy results for supervised classification using Naive Bayes classifier with *tf-idf* features per class.

2.6 Findings from the Annotation Analysis

This chapter has presented the analysis of the manual annotation corpus. The corpus is important for the following reasons:

1. It limits this study to a manageable and defined domain.
2. It helps to identify the main HLFs that should be extracted by image processing to describe video semantic content.
3. It helps to prepare for development/test data and ground-truth for evaluation.

With regard to 2 above (identifying HLFs), several conclusions can be drawn based on the analysis of manual annotations. Annotators are most interested in human presences and their visual attributions in the video stream, especially their actions, gender, emotions and their interaction with other humans and objects. More precisely, annotators are very keen to identify the human actions and their relationships in terms of spatial and temporal order. Artificial objects are mostly attached to humans and their activities (*e.g.* ‘man is sitting on the chair’).

Humans are usually considered as part of the foreground. Based on these observations, a list for HLFs for automatic extraction is derived which include human, age, gender, human emotion, human action, objects, scene setting and spatial and temporal relations between interacted objects. Based on the fact that the human is the main element of this list, the next chapter will present a novel framework that can detect, track and segment the human body over video frames. Manual annotation for the same visual scene vary mainly due to the individual influence of the people involved in description generation. In spite of these

	AnswerPhone	DriveCar	Eat	FightPerson	GetOutCar	HandShake	HugPerson	Kiss	Run	SitDown	SitUp	StandUp
AnswerPhone	102	0	0	2	1	0	0	3	0	6	3	3
DriveCar	1	102	0	0	11	0	0	0	2	1	0	3
Eat	0	2	106	0	3	0	0	1	0	0	0	8
FightPerson	0	0	0	103	1	1	1	1	1	3	6	3
GetOutCar	4	16	0	1	87	2	2	0	1	3	1	3
HandShake	1	2	0	1	0	94	0	1	1	11	3	6
HugPerson	0	0	0	2	1	5	93	11	0	6	0	2
Kiss	1	0	1	1	0	2	13	94	0	3	2	3
Run	0	0	0	2	1	1	0	0	108	0	7	1
SitDown	0	0	3	2	0	5	1	4	0	81	3	21
SitUp	6	0	2	13	3	4	1	3	2	5	70	11
StandUp	7	5	12	4	2	1	0	2	0	11	7	69

Fig. 2.8 Confusion matrix across 12 classes of hand annotations using Naive Bayes classifier with *tf-idf* features, where each column represents the instances in a predicted class and each row represents the instances in an actual class .

variations, there exist some similarities among descriptions derived from the same video. It can be argued that the dissimilarity lies in the words used and that similarity can be found in the facts that are included in the description. Manual annotation descriptions can be used as a reference to evaluate the information content of machine generated descriptions. Natural language description of video streams starts with identification of these HLFs, especially humans and their visual attributes.

2.7 Summary

This chapter has discussed the generation of metadata in the form of natural language descriptions for video corpora. 12 annotators produced titles and descriptions for 120 video segments.¹¹ A novel characteristic of this dataset is that rather than consisting of short clips each of which depicts a single action class, this dataset contains much longer streaming video segments that each contain numerous instances of a variety of action classes that often overlap in time and may occur in different portions of the field of view. The annotation that accompanies this dataset delineates not only which actions occur but also their temporal extent. Analysis of this corpus presents insights into human interests and thoughts while

¹¹ We plan to make this corpus publicly available with the following structure, video ID, video class, shot boundary, title, hand annotations

watching videos. This analysis further provides a list of important visual contents, which are referred to as high level features (HLFs).

Chapter 3

Spatio-temporal Human Body Segmentation

The previous chapter presented the first stage of video description framework that is being developed throughout this thesis. In the first stage the NLDHA corpus is generated which contains relatively long video clips of human activities, depicting a variety of action classes along with human textual annotations for each. Extensive analysis of the manual annotations associated with this corpus results in the conclusion that annotators are most interested in human presences and their visual attributions in the video stream, especially their actions, gender, emotions and their interaction with other humans and objects. The majority of previous work on video description task relies on identify HLFs at frame-level without exploiting the temporal information. In order to efficiently describe video semantic content a strong framework that can detect and segment regions of the human body from a series of video frames must be established to be able to efficiently extract these HLFs. A framework will be outlined in this chapter where human body volume will be established from a video stream. By using a range of object detection and tracking methods, the approach starts by detecting and segmenting areas of the human body at frame-level using joint embedding of parts and pixels that leverage the advances of low-level image cues and high-level part detectors information. Next, to be able to automatically extract the primary foreground objects throughout a series of video frames, the appearance and shape models are learned for all extracted segments to allow tracking these detections over sequence of frames.

The chapter is structured as follows: Section 3.1 introduces the human body segmentation framework combined with the motivations for this work and its contributions. Section 3.2 reviews previous work related to the human detection and tracking tasks in video stream. The implementation of the proposed human body segmentation is presented in Section 3.3,

while Section 3.4 presents a summary of the results obtained from the evaluation experiment. Finally, Section 3.5 provides a concluding discussion.

3.1 Introduction

There are several problems associated with computer vision, the most significant of which regards the automatic interpretation of video clips. This is a complex task that requires video contents to be detected and analysed, as well as identifying persons and objects, and their individual actions. Such tasks are typically conducted by working on individual video frames at a time, but videos, in reality, should be interpreted beyond a specified time frame, since the information they produce is temporal in nature. Such clues include the movement of objects, the interaction between people and objects over time, the movement of time through the video clip, and event relationships between people and objects.

Recently, several studies have been conducted to organise video pixels into coherent regions based on their similarity. However, this task proved to be a difficult and not free of errors. Video segmentation techniques aim to segment pixels content into spatio-temporal unity based on their similarity in terms of contents and individual movements (Yadav et al., 2011). Object segmentation plays a vital role in a wide range of video analysis tasks such as activity recognition, content-based retrieval. However, the complexity of the video segmentation process task stems from the nature of video data that correspond to temporal coherence. The frame-based segmentation approaches generally results in a choppy and inconsistent regions, since individually segmented frames are difficult to patch back into a video stream due to a lack of coherence in movement.

An approach consisting of two spatial and one temporal dimension to extract (3D) human body volume will be investigated in this chapter. We will use a video segmentation method that is in accordance with tracking-based methods, which can thus detect and segment areas of the human body in a video stream by using the joint embedding of parts and pixels (Maire et al., 2011). Next, the appearance and shape models of each extracted segment will be learned, which is crucial to be able to automatically identify key foreground objects in the video frames. It will concentrate on human contours, especially those adapted from Lee et al. (2011) category-independent segmentation technique. To the best of our knowledge there is no dataset or benchmark especially designed for human object segmentation task; hence, the NLDHA dataset will be used to evaluate this approach. Furthermore, it is demonstrated by the experimental findings that the approach is more effective by creating consistently better segmentation than the state-of-the-art approach of Lee et al. (2011) .

3.1.1 Motivations

Over the past decade, various researchers have proposed different techniques to automatically segment objects from background in video streams. One common approach is for such segmentation to be performed in interactive manner, where a user requires object boundaries to be annotated in some of the key frames and then propagated to subsequent frames while another user stands by to correct errors (Bai et al., 2009; Price et al., 2009). Tracking-based approaches try to alleviate the supervision to a manual segmentation needed to just one frame (Ren and Malik, 2007; Tsai et al., 2012). However, such methods rely on user input and may be affected by the user's lack of annotation expertise.

As an improvement, bottom-up approaches are proposed to segment video objects in a fully automatic manner based on low-level cues such as motion and appearance features. Motion segmentation techniques generally use three models: background subtraction (Stauffer and Grimson, 1999; Zhao and Nevatia, 2003), temporal difference (Paragios and Deriche, 2000), and optical flow (Fejes and Davis, 1998; Wixson, 2000). The background subtraction method is a common approach used for detecting video objects in motion from stationary cameras. However, for realistic video settings such as camera motion, lighting alteration or interesting but static objects this technique is inappropriate. Recently implemented approaches either accomplish a spatio-temporal pixel-level segmentation for video data from scratch (Grundmann et al., 2010), starting by applying image segmentation at frame-level and then matching resulted regions across successive frames (Vazquez-Reina et al., 2010), or utilise dense flow in order to create motion trajectories (Brox and Malik, 2010). However, these approaches tend to fail without prior top-down knowledge as over-segmentation issues arise, producing regions that lack of semantic meaning.

To overcome the limitations of the previous segmentation methods, Maire et al. (2011) proposed an image segmentation approach that is achieved by developing a perceptual grouping framework. In this framework two types of plug-in state-of-the-art components are combined. The first one is a high quality bottom-up image segmentation method which concentrates on low-level cues such as colour, texture and edges and looks at the statistics of local image patches. Then these low-level cues are combined with the resulted information from poselets (Bourdev and Malik, 2009), a top-down vocabulary detector of part detection for finding people in still images. However, this work has been defined for image processing applications that consider only spatial information and disregard any temporal information. In this chapter, this approach is extended to automatically segment human objects discovered in video stream. The detection of human body regions is achieved by combining low-level cues with a top-down detector. Once the human body regions have been identified at frame level, the next step is tracking them over the video stream and this will be achieved by

building a foreground model which combines colour and shape estimated models. Lastly, pixel-wise segmentation is achieved by utilising the space-time Markov Random Field (MRF) to extricate the human body regions from the surround from video clip on a frame-by-frame basis.

3.1.2 Spatio-temporal Human Body Segmentation: Overview

This chapter presents a spatio-temporal human body segmentation approach that detects, segments and tracks human body regions across video frames. The approach consists of two principle stages. The first part of the proposed segmentation approach involves identification of human body objects at frame level. This is achieved by using the joint embedding of parts and pixels approach proposed by [Maire et al. \(2011\)](#). The joint embedding of parts and pixels approach detects and segments human objects by solving a single grouping problem. This framework is considered as a perceptual organisation stage that combines information from low-level image cues, such as colour and textures, with high-level part detectors information. The pixels and parts represent nodes in a graph, whereas affinity and ordering relationships are encoded by edges.

The second stage, once the human body regions have been identified at frame level, is tracking them over the video stream and this will be achieved by building a foreground model which combines colour and shape estimated models. The consistency of recurring foreground objects also been exploited over the time domain. In particular, we have a robust localisation prior for subsequent frames. To achieve the consistency, we project the detected segments onto each frame in the video by matching detected regions boundaries using Boundary Preserving Local Regions (BPLR) ([Kim and Grauman, 2011](#)), a densely local feature extract which maintains object boundaries and partial shape.

The approach is evaluated on the NLDHA dataset using the human detection and segmentation tasks. The segmentation accuracy of the proposed approach is compared against a recent state-of-the-art segmentation method proposed by [Lee et al. \(2011\)](#). The outcome from the experiment indicates that the proposed approach is able to create better segmentation than the state-of-the-art implemented work.

3.1.3 Spatio-temporal Human Body Segmentation: Contributions

The contributions of the proposed work in this chapter can be summarised as follows:

- combination of the joint embedding of parts and pixels approach, the effective segmentation method in the image processing field with state-of-the-art video tracking approach to be used for the video segmentation task;

- provision of a sturdy automatic approach for segmenting human objects discovered in video clips, which can be utilised for the future visual actor retrieval research;
- application of the spatio-temporal human body segmentation approach for the human detection and segmentation task produced consistently better segmentation results than the the state-of-the-art approach implemented by [Lee et al. \(2011\)](#).

3.2 Related Work

The central research areas in computer vision that have supported a range of critical applications, such as human behavioural analysis, object recognition, and visual surveillance, are object detection, tracking, and segmentation. A task constituting a significant challenge is the segmentation and tracking of several human objects accurately and in a consistent manner when they overlap with each other under occlusion in a complex setting; this is an even greater challenge than when the targets are separated. This can be attributed to the non-rigid motion of deformable objects, including colour distribution, shape, and visibility. The purpose of the current section is to provide a review of the previous research by examining two central themes: interesting object detection and object tracking.

3.2.1 Object Detection

Object detection is a fundamental step that facilitates the further examination of video. All approaches to tracking mandate the application of an object detection process for each frame or, alternatively, at the point where the object initially occurs in the video. This relates to the segmentation of objects that are in motion from stationary background objects ([Paragios and Deriche, 2000](#)). The central area of focus in this task is high-level processing, and it also reduces the computation complexity. Owing to a range of environmental aspects, such as illumination modifications, something that emerges as a considerable challenge is shadow object segmentation.

A frequently employed method by which object detection takes place is to employ information in a single frame, but it should be noted that certain object detection approaches utilise the temporal data computed from successive frames in order to reduce the frequency of false positives ([Elgammal et al., 2002](#)). Such temporal data is typically used in a frame differencing format, highlighting the regions that are subject to dynamic modification in successive frames. Following the discernment of the object regions within the frame, the tracking process role is to undertake object correspondence for successive frames, thereby generating the tracks. The central focus of this section is to provide a review of the most

prominent object detection techniques, which can be classified into three major groups: appearance-based methods, motion-based methods and hybrid methods.

3.2.1.1 Appearance-based methods

Edge information is used quite often in the scenario where the detection of objects in images is required. This method is known as an appearance-based approach. Such approaches can be used to train specialised detectors on huge pedestrian datasets (Geronimo et al., 2010). These kinds of algorithms are designed to look for similarities and configurations in each individual frame that fit patterns that have already been identified, and this technique can also be utilised in non-static cameras, such as those used for helping automatic driving (Gavrila and Munder, 2007). Examples of where this has been implemented successfully include Haar features (Paisitkriangkrai et al., 2008), the Viola and Jones algorithm (Viola and Jones, 2001) and the histogram of oriented gradients (HOG) (Dalal and Triggs, 2005).

The individual shape of any person can be identified utilising HOG and the variations in edge data, as highlighted by the changes in intensity. A support vector machine (SVM) classifier is often used with HOGs to assist in this process. The main issue with HOGs is that they vary as the object shifts in time and space, even though they are resistant to any variations in size and brightness. Several of the proposed extensions of this technique have been implemented, which for example involve the use of the HOG representations for individual parts of the body and then joining them together (Felzenszwalb et al., 2010); an alternative method which is often used is a top-down approach for segmenting based on local and global combination cues (Leibe et al., 2005). The technology supporting face recognition based on the Viola and Jones algorithm can also be used for identifying the outlines of pedestrians (Viola et al., 2005).

The performance levels of HOGs in the INRIA dataset can be emulated through the development of computationally effective algorithms that are a combination of Haar and covariance characteristics (Paisitkriangkrai et al., 2008). Haar features were employed in (Elkerdawi et al., 2014) in order to find vehicles and train a cascaded Adaboost classifier. The benefit of this feature is its sensitivity to vertical, horizontal and symmetric structures, meaning it is highly suitable for real time application. The downside, however, is the limited computational efficiency.

3.2.1.2 Motion-based methods

Detection techniques which use the movement of body parts and time bound data for identifying human movement are deemed to be motion-based; for example, the repeated

movement of arms and legs in a static camera scene use motion-based features to recognise the moving object (Bouwman, 2011). There are three classes for these motion-based techniques, namely background subtraction, temporal differencing and optical flow.

A. Background Subtraction

Considerable modifications in an image region from the background model signals of an object in motion and the pixels comprising this region are subject to modification and logged for continued processing. It is often the case that a detection algorithm is implemented for the purpose of obtaining connected regions that match along with the objects, and this is the procedure known as background subtraction (Elgammal et al., 2002).

Heikkilä and Silvén (2004) proposed a framework for the monitoring of cyclists and pedestrians. A motion detection is achieved by background subtraction following this procedure. At the beginning, the system reference background is subject to initialisation with the introductory frames of the video, and these are brought up to date for the purpose of adapting to the dynamic modifications in the scene. For every new frame, foreground pixels are identified through the subtraction of intensity values from the background and, following this, filtering absolute value of differences with a dynamic threshold for each pixel. The threshold and reference background are brought up to date by employing foreground pixel information. Notably, this engages in the detection of regions in motion through the subtraction of the current image on a pixel-by-pixel basis from a reference background image that is generated by taking the averages of images over a certain timeframe in an initialised period. A foreground is deemed to be the pixels where the variance is above a threshold and, after generating the foreground pixel map, a number of morphological post-processing operations, including erosion, dilation, and closing, are carried out for the purpose of reducing the impacts of noise and enhancing the detected regions. In turn, the reference background is brought up to date with new images over a certain timeframe, thereby adapting to the dynamic scene modifications.

A pixel is logged as foreground in I_t if the following inequality is satisfied:

$$|I_t(x,y) - B_t(x,y)| > T \quad (3.1)$$

where T represents a pre-defined threshold. Here, the background image B_t is brought up to date through the employment of a first-order recursive filter as illustrated in the following equation:

$$B_{t+1} = \alpha I_t + (1 - \alpha)B_t \quad (3.2)$$

where α represents an adaptation coefficient. Here, the fundamental idea is to feed the new incoming data into the current background image and, following this, the rapid, novel modifications in the scene are brought up to date in the background frame.

The Eigen background subtraction method was formulated by [Oliver et al. \(2000\)](#) to present an Eigen space model for moving object segmentation. The method is characterised by the construction of the space dimensionality from sample images, which is decreased by employing Principal Component Analysis (PCA). The space that has been subject to reduction following the PCA is intended only to represent the stationary components of the scene, keeping the objects that are in motion in situations where an image is projected onto the space. An accurate and real-time approach for moving object detection and tracking, using the reference background subtraction and dynamic use of threshold value, is put forward by [Tah et al. \(2017\)](#). This method allows for the effects of luminescence alterations to be eliminated. Because of deployment of the Kalman filter, this rapid algorithm is simple to implement for the detection of moving objects to a more effective degree, with wide application potential.

B. Temporal Differencing

The temporal differencing method employs the pixel-wise variance between two or among three consecutive frames in video imagery in such a way as to extract the regions that are in motion. Despite the fact that it is a highly versatile method for dynamic scene modifications, it does not succeed in extracting all of the necessary pixels of a foreground object, and this is particularly the case in instances where the object has a uniform texture or is in motion in a slow manner ([Paragios and Deriche, 2000](#)). In situations when a foreground object becomes stationary, the temporal differencing method is ineffective in logging the change between successive frames and, as a result of this, fails to track the object. Let $I_n(x)$ denote the gray-level intensity value at pixel position x and at time instance n of video image sequence I , which is in the range $[0, 255]$. Furthermore, let T represent the threshold initially set to a pre-determined value. In light of this, [Lipton et al. \(2005\)](#) formulated a two-frame temporal differencing algorithm suggesting that a pixel is in motion in situations where it satisfies the equation below:

$$|I_n(x) - I_{n-1}(x)| > T \quad (3.3)$$

It should be noted that, in terms of computation, this method is complex and flexible to a lesser degree with regard to dynamic modifications in the video frames. For the temporal difference technique, the removal of moving pixels is straightforward and can be achieved rapidly, but it should be noted that it can leave holes in foreground objects. Furthermore, it is sensitive to a greater degree to the threshold value in situations when it attempts to determine

the modifications within difference of successive video frames. It should also be noted that temporal difference needs a special supportive algorithm in order to identify objects that have progressed from being in motion to being static. [Barbu \(2014\)](#) put forward an approach which allowed for multiple humans to be detected and tracked. A moving person identification technique was the initial offering, where the video objects are detected through a cutting-edge temporal differencing approach, along with a number of mathematical morphology-based processes. Following this, the moving image objects portray pedestrians established by examinations of numerous conditions linked with human bodies and the detection of skin regions within the video frames.

[Paul et al. \(2017\)](#) proposed a reliable and computationally affordable option that can detect moving objects in video. This method involved three stages, starting with the calculation of difference images with temporal information. Difference images are computed by subtracting two input frames at every pixel position. Rather than producing difference images with the commonly used continuous frame difference approach, a set number of different frames based on the current frame are employed. This method allows for the reduction of computational difficulty, with no drawback when it comes to difference image quality. Following the computation of difference images, an innovative post-processing scheme is used through the gamma correction factor and Mahalanobis distance metric, in order to limit false positives and false negatives. Finally, object segmentation is undertaken for the refined difference image with a local fuzzy thresholding scheme. This overcomes the issues related to hard thresholding, and particularly pixel misclassification, which is considered the most critical.

C. Optical Flow

The capability and performance of the detection process can be enhanced significantly compared to the corresponding static detector through the addition of motion characteristics. [Haritaoglu et al. \(2000\)](#) described various observation-centred detectors incorporating the flow-based movement identification technique. In a study by [Efros et al. \(2003\)](#), while numerical data was not presented they did investigate form and flow characteristics in a model-based detector that captured distance images of sports participants. A detector system which incorporates mobile cameras and backgrounds for detecting mobile and static people was developed by [Dalal et al. \(2006\)](#) and using this system and a number of various movement coding techniques, they were able to deduce that the best results were achieved by employing orientated histograms of separated optical flows. Joining these descriptors with the HOGbased one results in a detector which is much superior.

A technique for the detection and segmenting of mobile entities employing long-term point tracks and optical flow was introduced by [Brox and Malik \(2010\)](#). The mobile entities

can be grouped together by identifying the pairwise space distance between the points' trajectories, which in a video shoot causes uniform, sequential segmenting of these mobile entities. [Hossen and Tuli \(2016\)](#) suggested a straightforward and effective surveillance system that used motion detection with motion vector estimation from surveillance video frames. The motion is found through a new approach-edge region determination, allowing for more rapid detections. Following this, the surveillance video is processed for motion estimation through optical flow with the Horn-Schunck algorithm, to calculate the motion vector through its steady performance and simplicity. This approach is computationally faster and does not need any unique hardware for image processing, meaning it is a suitable choice for embedded systems.

3.2.1.3 Hybrid methods

The best and most accurate human detection approaches incorporate a combination of motion and appearance ([Dalal, 2006](#); [Viola et al., 2005](#)) which together allow a unique investigation of the information using a decision function to yield the ultimate output, or alternatively apply a still-image identifier to areas where there is possibly human movement such as in a blob tracker. The disadvantage is that the mathematical difficulty level is raised significantly because appearance-based detectors carry out a dense examination of the complete frame. The solution for this issue lies in reducing the sample rate of frames per second, which results in a decreased level of mathematical difficulty and reduces the incidence of false positives. In relation to false positives and false negatives the most advantageous detection rates are achieved through the techniques that use a background model to separate out individual pedestrians.

A Viola and Jones detector ([Viola and Jones, 2001](#)), which works on variations linking image pairs, and an AdaBoost trainer are joined to create a product that exhibits both characteristics. In the AdaBoost process a selection is made from the various movement and appearance characteristics with the smallest inaccuracy within the training samples. Pixel density and movement data are collected for the last identification so that the detection rates are at the highest level and the precision may be enhanced. The downside of this method is that a higher frame rate is necessary ([Jones and Snow, 2008](#)). [Dalal \(2006\)](#) and [Bouchrika et al. \(2010\)](#) used a similar technique when they joined HOGs with movement and walking examples. This idea can also be applied to situations where individual behaviour can be used as a detection cue ([Ge et al., 2012](#)), where blobs with related movement are more likely to indicate the presence of pedestrians.

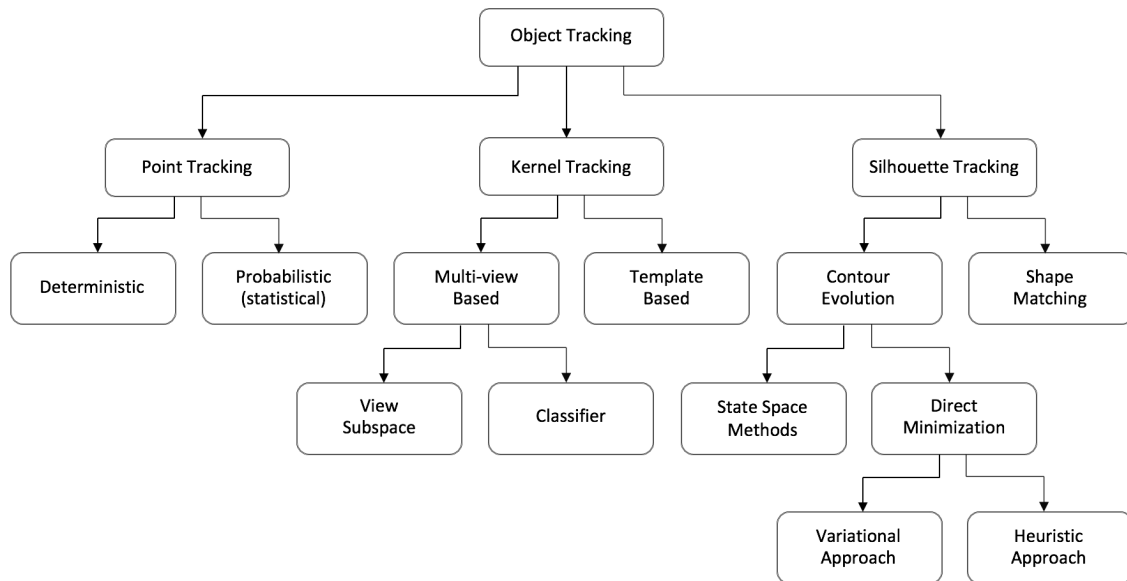


Fig. 3.1 Categories of object tracking methods. Figure adapted from [Yilmaz et al. \(2006\)](#)

3.2.2 Object Tracking

The central factor for consideration in human motion analysis is object tracking, characterised as a higher-level computer vision issue. Tracking constitutes the alignment of detected foreground objectives between successive frames through the employment of object features including motion, speed of motion, texture, and colour. Object tracking is the procedure utilised to trace the object's movement over time, and this takes place by identifying its position in each frame within the video sequence; see Figure 3.1. In addition, object tracking can calculate the entire region that is occupied by the tracked object at every point in time ([Yilmaz et al., 2006](#)). For the tracking procedure, the tracked objects are denoted by employing shape or appearance models. In ([Mandellos et al., 2011](#)) shape based tracking with Kalman filtering was used to match simple regions.

Notably, the model chosen to denote object shapes places limitations on the kinds of motion that can be tracked. For instance, if the model selected means that the object is denoted as a point, a translational model is the only effective approach that can be employed. If a geometrical shape such as an oval is used to denote the object, more suitable models including parametric motion models can be used, which include affine or projective transformations ([Isard and MacCormick, 2001](#)). It should be taken into account that such representations can generate approximations of the motion of rigid objects in the scene but, for non-rigid objects, silhouettes or contours are the most descriptive representation ([Sun et al., 2015](#)). Thus, for

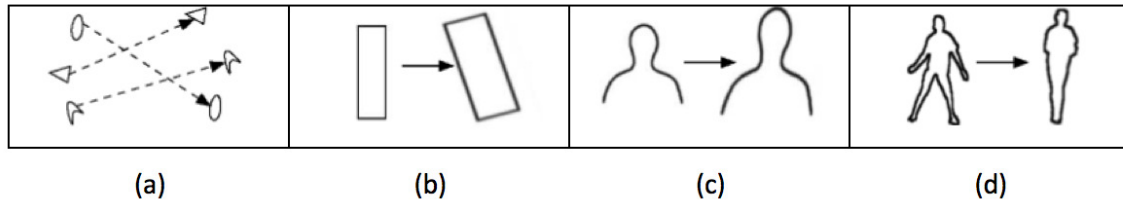


Fig. 3.2 Examples of different object tracking approaches. (a) represents point tracking, (b) kernel tracking, (c) and (d) silhouette tracking. Figure adapted from [Yilmaz et al. \(2006\)](#)

non-rigid objects, parametric and non-parametric models are suitable for the specification of their motion. A range of popular object tracking approaches shown in Figure 3.2 are discussed below.

3.2.2.1 Point Tracking

This approach is characterised by robustness, reliability, and accuracy, and it was proposed by [Veenman et al. \(2001\)](#). It is most commonly employed for vehicle tracking, and an effective level of fitness for the detected object is required for the approach to be successful. It is necessary to employ deterministic or probabilistic approaches for this to proceed effectively. The object subject to tracking is based on a point and this is denoted in the detected object in successive frames; the relationship between the points is founded on the previous object state, and this can involve the object's position and motion. It is worth mentioning that point tracking necessitates the inclusion of an external mechanism that can be used to detect the objects in each frame.

Commonly used point tracking algorithms, such as the KLT, employ local 2D information aggregation to detect and track features, and as a result their performance falls at the object boundaries, which split up various objects. In a new study, CoMaL Features have been suggested ([Ravindran and Mittal, 2016](#)) to deal with these situations. Alternatively, a simple tracking framework was put forward, under which the points are re-detected in every frame and matched. Clearly, this is not efficient and can lead to losses which cannot be re-detected and made up for in the following frame. [Ramakrishnan et al. \(2017\)](#) suggested a new tracking algorithm which can reliably track CoMaL points with high efficiency. In this situation, the level line segment linked with the CoMaL points is suited to MSER segments in the following frame through shape-based matching, and these are then filtered once again through texture-based matching.

3.2.2.2 Kernel Tracking

For this method, the kernel requires the shape and appearance of the object (Elhabian et al., 2008), and any characteristic of the object is employed for the purpose of tracking it as a kernel-like rectangular template or an oval-like template with a related histogram. Following the computation of the motion of the kernel with regard to successive frames, object tracking is achieved. For mean-shift tracking, this is founded on the kernel tracking technique that is employed (Comaniciu and Meer, 2002). In this approach, an E-kernel is utilised, and this denotes the histogram characteristic based by spatial masking on the basis of an isotropic kernel.

Hare et al. (2016) presented a framework for adaptive visual object tracking that uses structured output prediction. Through specifically allowing the output space to show the requirements of the tracker, an intermediate classification stage is not necessary. This approach employed a kernelised structured output support vector machine (SVM) that learned online to offer adaptive tracking.

3.2.2.3 Silhouette Tracking

For this method, a silhouette is taken from the detected object. Through shape matching or contour evolution, the silhouette is then tracked through the calculation of the object region in consecutive frames until the tracking is completed. Such approaches utilise the information stored within the object region in order to facilitate tracking, and this information can relate to appearance, density, and shape models (Elgammal et al., 2002).

Mondal et al. (2016) proposed an algorithm which is able to provide efficient tracking for the contour taken from the outline of the moving object in a specific video sequence, with local neighbourhood information and a fuzzy k-nearest-neighbour classifier. In order to categorise every unlabelled sample in the target frame, a subset of the training set is used, based on the level of motion of the object across the previous two consecutive frames. This approach allows for the classification process to move with greater rapidity, and can boost classification accuracy. Transition pixels from the non-object region to the object silhouette (and the opposite) are considered and handled as boundary or contour pixels of the object, which are taken through the linking of boundary pixels.

In object models, tracking revolves around a consideration of features and, therefore, necessitates the selection of the correct features. Generally, the object characteristics considered must be unique since this distinctiveness facilitates the discernment of the object in the feature space. The consideration of these features is thus crucial for the process to be successful, and the range of features discussed below are employed:

- **Colour:** An object's apparent colour is impacted mainly by physical factors such as the spectral power distribution of the illuminate and the object's surface reflectance characteristics (Stauffer and Grimson, 2000). For image processing domain, the RGB colour space is most commonly employed for the representation of colour.
- **Edges:** It is usually the case that the perimeter of an object produces marked alterations in image intensities (Wren et al., 1997). Edge detection is the method employed in silhouette tracking that can discern such alterations, and a key feature of edges is that, when considered in relation to colour features, they are sensitive to a lesser degree with regard to illumination changes.
- **Centroid:** The centroid refers to the centre of mass, and this is a vector of $1 \times n$ dimensions in length that delineates the centre point of a particular region. It should be noted that, for every point, the initial element of the centroid is the x -coordinate of the centre of mass while the following element is the y -coordinate (Elgammal et al., 2002).
- **Texture:** This feature is used both for the purposes of classifying and tracking, and it is useful in identifying regions or objects in which the surveillance personnel are interested. Texture serves to measure the intensity variation of a particular surface, and this provides a quantification of characteristics including the extent to which an object is smooth and regular (Yilmaz et al., 2006). When considered in relation to colour, texture necessitates a processing stage for the generation of the descriptors. With respect to all of the above features, the most commonly employed for object tracking are colour and texture, and it should be noted that colour bands are characterised by sensitivity to illumination variation.

3.3 Human Body Segmentation Framework

The aim of this chapter is to segment human body region from an unlabelled video. There will be two key steps in our approach, which can be seen in Figure 3.3. The first step will involve human body objects being segmented frame-by-frame by using a collaboration of low-level cues (Fowlkes et al., 2003) and top-down part-based person detector (Bourdev and Malik, 2009), which will thus create grouped patches of proposed human body regions. Secondly, the temporal consistency of detected foreground objects using colour models and local shape matching (outlined by Lee et al. (2011)) will be used to reproduce detected segments in subsequent video frames. This will result in a final output of a spatio-temporal segmentation of the human body in a video stream. Each step will now be discussed in detail.

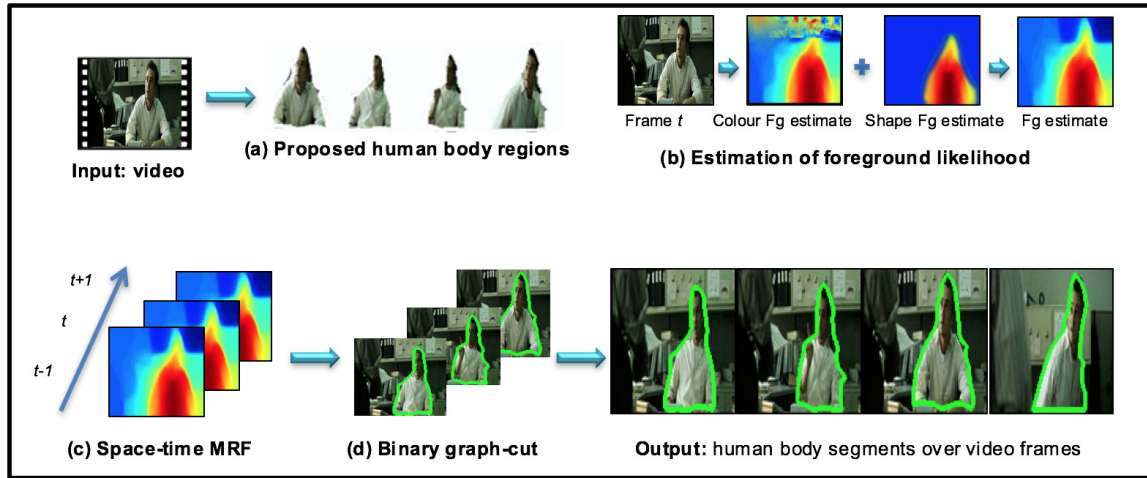


Fig. 3.3 The two-step approach to segment human volume. The human body detected in the initial stage can be replicated in the subsequent video frames in the second stage.

3.3.1 Estimation of Human Body Region at Frame Level

This stage builds on the graph-based image segmentation technique of [Maire et al. \(2011\)](#). In this stage the segmentation will be achieved by developing a perceptual grouping framework. The perceptual grouping framework can be described in terms of relationships between and among parts and pixels (see Figure 3.4). Each pixel is going to be connected to its neighbours based on low-level appearance cues using an intervening contour cue ([Fowlkes et al., 2003](#)). In particular, each pixel will be connected with high affinity to neighbours within certain radius if they share similar low-level cues (colour, texture and brightness). Similarly, poselets parts detectors ([Bourdev and Malik, 2009](#)), where human body parts are tightly connected based on appearance cues and configuration space, will be run in order to find human body part such as head and feet across the image. Then, the resulting parts will be connected according geometric compatibility. As a result, a grouping of parts and pixels is established during this stage based on the following concepts:

- Pixel connectivity depends on low-level cues to achieve region consistency.
- Detected segments that are part of the same object are grouped together.
- The regions belonging to a part are included in the foreground, whereas the remaining regions are pushed to the background.

Next, these two types of information are going to be combined by creating a graph in which pixels and parts appear as nodes and the links encode affinity relationships, as shown in Figure 3.5. In addition to affinity relationships a new type of relationship is added to ensure that the detected parts belong to image foreground: this layer is called the surround layer. So, one surround per part that represents the area of the image that immediately surrounds

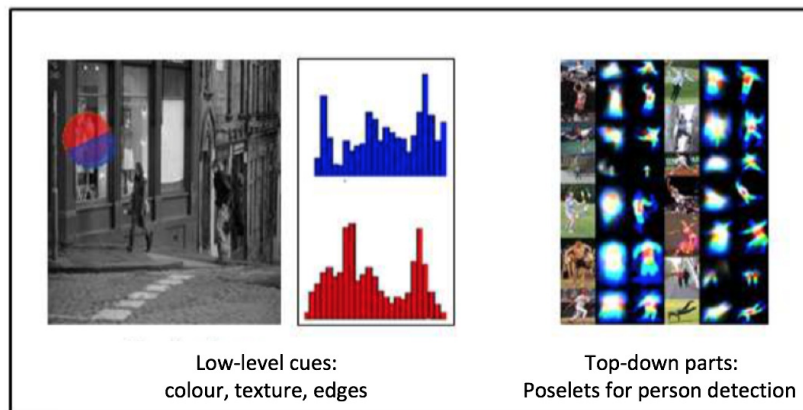


Fig. 3.4 Human body detection at frame-level achieved by combined low-level cues from Fowlkes et al. (2003) with top-down parts detector of Bourdev and Malik (2009).

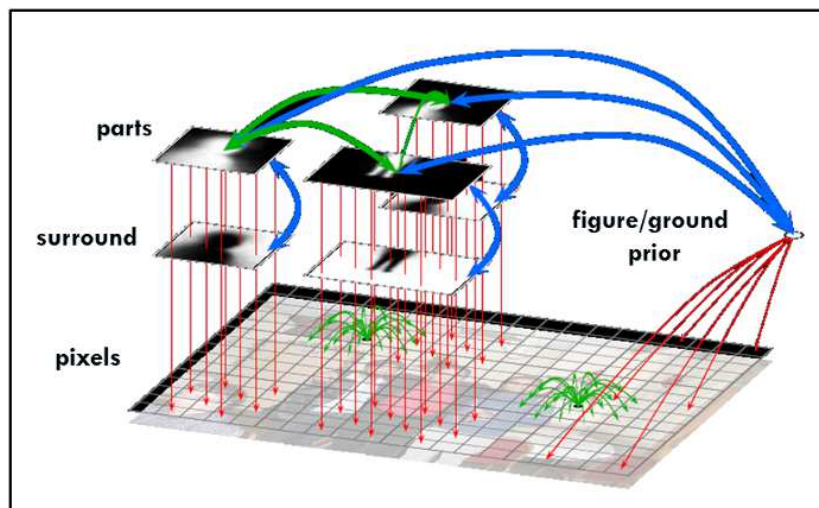


Fig. 3.5 Graph Setup: Pixel and Part Relations.

the part. Also, parts are going to be connected to set of pixels under a set of constraints and this is going to allow grouping of the parts layer to influence a grouping of the pixel layer. And finally all irrelevant regions of images where none of the part detectors is fired will be thrown in to the background. A brief description of this stage is given below.

3.3.1.1 Globalisation.

As a result of this representation the scene can be interpreted in terms of human object segmentation. A way to transform the graphs information from these local relationships between nodes into a globally coherent analysis of the scene is required, and to achieve this, the issue with embedding needs to be addressed using an angular embedding (AE) (Stella,

2009), which is good at combining affinity and ordering relationships. These graph nodes must be plotted into the complex number space. This embedding will allow interpretations of the scene to be made according to human object segmentation.

The angular embedding (AE) algorithm is used as a globalisation framework, where the delineation of an AE problem's input is facilitated by a pair (C, Θ) of real-valued matrices, and these matrices also facilitate the acquisition of pixel-pixel and part-part relationships. While the latter matrix Θ , referred to as the skew-symmetric matrix, defines relative ordering relationships, the function of the former matrix C , referred to as a symmetric matrix, is to delineate the confidence linked to the relationships. Confidence is established in line with the degree of attraction, whereas relative ordering is established at zero, and this facilitates the encoding of pairwise affinity.

The output of an AE problem draws on a complex number space to represent pixels and parts, and distances in this space are reflective of the concept of grouping, while the encoding of a global ordering is facilitated by phases of complex numbers. Following the imposition of specifications, which are formulated in the form of a sparse matrix U , on the embedding solution space, it is possible to acquire pixel-part associations (Yu and Shi, 2004). It is also worth noting that the columns of the matrix U delineate linear constraints associated with pixels and parts.

Given a set of matrices C , Θ and U , the complex eigenvectors, z_0, \dots, z_m is solved corresponding to m largest eigenvalues $\lambda_0, \dots, \lambda_m$ of constrained AE problem:

$$QPQz = \lambda z \quad (3.4)$$

where P is a normalised weight matrix and Q is a projector onto the feasible solution space defined by:

$$P = D^{-1}W, \quad Q = I - D^{-1}U(U^T D^{-1}U)^{-1}U^T \quad (3.5)$$

Here D and W are defined based on C and Θ :

$$D = \text{Diag}(C1_n), \quad W = C \bullet e^{i\Theta} \quad (3.6)$$

where n indicates the number of nodes, 1_n represents a column vector of ones, I is the identity matrix, $\text{Diag}(\cdot)$ is a matrix with an argument on the main diagonal. The matrix Hadamard product is represented by \bullet , $i = \sqrt{-1}$ and exponentiation performed element-wise. The largest eigenvalues are represented by the Eigenvectors z_0, \dots, z_m which thus transfers the pixel outputs into \mathbb{C}^m .

3.3.1.2 Graph setup: pixel and part relations.

Four node types will be applied in the creation of the image segmentation graph, namely pixels (p), parts (q), surround (s) and figure/ground prior (f). These will be used in a block structure established by using $n \times n$ matrices, C and Θ . Analysis of the contour between pixels will determine the colour and texture pixel-pixel affinity C_p and pairwise part-part compatibility and poselets detection scores will be used to establish geometric compatibility C_q (the part-part affinity). Augmentations in repulsion between part and the surround (C_s, Θ_s) is determined using these part-part detection scores; the latter of which focuses on the global surround node (C_f, Θ_f). Part-embedding that is equal to the mean embedding of total pixels that make up the part restricts U . Furthermore, the part/surround nodes must agree with pixels allocated to each node.

$$C = \begin{bmatrix} C_p & 0 & 0 & 0 \\ 0 & \alpha \cdot C_q & \beta \cdot C_s & \gamma \cdot C_f \\ 0 & \beta \cdot C_s^T & 0 & 0 \\ 0 & \gamma \cdot C_f^T & 0 & 0 \end{bmatrix}, \quad \Theta = \Sigma^{-1} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -\Theta_s & -\Theta_f \\ 0 & -\Theta_s^T & 0 & 0 \\ 0 & -\Theta_f^T & 0 & 0 \end{bmatrix} \quad (3.7)$$

where C_p facilitates the storage of pixels affinity, C_q stores affinity regarding parts, (C_s, Θ_s) facilitates the encoding of the separation regarding part and surround, and (C_f, Θ_f) facilitates the encoding of the figure/ground prior. In addition to this, weights α, β , and γ exchange the relative significance of the relationships types over the course of globalisation, and Σ denotes a normalisation factor which incorporates the total of the absolute value of the Θ_s and Θ_f entries.

3.3.1.3 Output: decoding eigenvectors.

The nodes in \mathbb{C}^m are of significant importance, since they are based on pixels and parts plugged into the graph in accordance with the eigenvectors. The areas containing each single human body object in a frame can be identified by applying eigenvectors. The following equation determines the assignment of every pixel p_k to a relevant part:

$$p_k \longrightarrow \operatorname{argmin}_{Q_i} \left\{ \min_{\substack{q_j \in Q_i \\ p_k \in M_j}} \{D(p_k, q_j)\} \right\} \quad (3.8)$$

where M_j indicates the area of the image overlapped by a part q_j . Every part is subsequently assigned to an accepted human object Q_i , which is confirmed by scoring each

hypothesis using a linear discriminant classifier (Bourdev et al., 2010). This is the symbol that signifies the number of confirmed objects detected. A score is subsequently given to the segments in every frame. A set of $N \times F$ is then used to repeat this step, where each N represents how many human body objects are present in each frame and the number of frames is represented by F . A group of hypotheses h will then be identified using the results of these steps, and then the spatio-temporal segmentation of human body parts throughout the whole video can be established using these hypotheses.

3.3.2 Tracking of Detected Regions over Video Stream

Once the human body regions have been identified at frame level, the next step is tracking them over the video stream and this will be achieved by building a foreground model which combines colour and shape estimated models (see Figure 3.3(B)). Preserving the consistency of recurring foreground objects viewed over time is accomplished by matching using Boundary Preserving Local Regions (BPLR) (Kim and Grauman, 2011), a densely-extracted local feature that preserves object boundaries and partial shape.

Each of the hypotheses (h) identified in the previous stage defines a foreground (human body) and a background (surround) model. Each object-like region in each frame is replaced by a human body, following the method of Lee et al. (2011). Pixel-wise segmentation is used to extract the human body segments from the surround in the video stream on a frame-by-frame basis, using the space-time Markov random field (MRF) described below. For each frame, the space-time graph of a pixel is defined, where the pixel is represented by a node and the edge between two nodes equates to the cut between two pixels. Each hypothesis h has an energy function, which can be determined by:

$$E(f, h) = \sum_{i \in S} D_i^h(f_i) + \delta \sum_{i, j \in \mathcal{N}} V_{i, j}(f_i, f_j) \quad (3.9)$$

where f represents the pixel nodes, $S = \{p_1, \dots, p_n\}$ is a set of n pixels in the video, and i and j index the pixels in space and time. Each pixel is then assigned to the foreground or background by setting p_i of each pixel to $f_i \in \{0, 1\}$, where 0 = background and 1 = foreground. The neighbourhood term $V_{i, j}$ is used to enhance smoothness in space and time between the pixels in adjacent frames. For neighbourhood pixels \mathcal{N} four spatial neighbours are assigned to each pixel per frame. Two temporal neighbours are assigned in the preceding and subsequent frames; each of these is then given an optical flow vector displacement. Neighbouring pixels of the same colour are labelled using standard contrast dependent

functions (Rother et al., 2004), with the cost of labelling defined by:

$$D_i^h(f_i) = -\log(\rho \cdot U_i^c(f_i, h) + (1 - \rho) \cdot U_i^l(f_i, h)) \quad (3.10)$$

where $U_i^c(\cdot)$ is the colour-induced cost, and $U_i^l(\cdot)$ is the local shape match-induced cost defined in (Rother et al., 2004). The segments detected in each frame on the basis of their parts and pixels are projected onto other frames by local shape matching, with a spatial extent which defines the location and scale prior to the segment, whose pixels can subsequently be labelled as foreground or background. Optical flow connections are used to maintain frame-to-frame consistency of the background and foreground labelling of propagated segments. For each hypothesis h , the foreground object segmentation of the video can be labelled by using binary graph cuts to minimise the function in Equation (3.9). Each frame is labelled in this way, using a space-time graph of three frames to connect each frame to its preceding and subsequent frames. This is more efficient than segmenting the video as a whole.

3.4 Experiments and results

The ability to detect, segment and track the movement of people in video clips is a very important topic for a range of computer vision tasks which involve human interaction, such as event detection, video surveillance and semantic video retrieval. In this section the experiments on human body detection and segmentation are carried out and the proposed framework is evaluated. The main objective is to identify the level of accuracy in our methodology in segmenting the foreground (human body object). To date, as reported in Perazzi et al. (2016) the only available benchmark dataset proposed for video segmentation and tracking tasks is SegTrack database (Tsai et al., 2012), that consists of six low-resolution video clips including the following categories: parachute, girl, monkeydog, penguin, birdfal, cheetah. However, this dataset unsuitable for this task as the approach is human-centric. Thus, the approach is evaluated on the NLDHA dataset using the human detection and segmentation tasks.

3.4.1 Experimental Setup

Before carrying out this experiment three pre-processing techniques should be applied. The first is frame extraction with the FFmpeg,¹ which is used to stream, convert and record a range of multimedia formats. A range of commands, free programs and open source material

¹<http://www.ffmpeg.org/>

are available through libavcodec, which is the audio and video codec library, and libavformat which is the input reading library. The sample rate of 1 fps is used in this experiment.

Secondly, for computation efficiency, the segmentation at frame-level is applied only for keyframes within each shot and then the segmented regions tracked over sequence of frames until next keyframe. In order to identify the keyframes of each shot the approach that based on histogram differences is applied.²

Finally, the ground-truth set is made ready using the shot boundary information contained within the original Hollywood2 dataset, represented by the starting and ending frame number of each shot. The first frame in each shot is annotated manually with bounding box to highlight the presence of a human objects. The parameters control the detection and segmentation processes set as follows: the parameters govern the subproblems importance $\alpha = 0.002$, $\beta = 0.5$ and $\gamma = 0.1$ in Equation (3.7), whereas $\delta = 4$ in Equation (3.9) is represents the smoothness term and $\rho = 0.5$ in Equation (3.10) is used for graph-cuts minimisation.

3.4.2 Human Detection Results

To determine the accuracy of the proposed framework it is necessary to look at the differences between the detection of human objects for the first frame in each shot and the manual developed ground-truth. The detection accuracy of human detection is evaluated at image level on the PASCAL VOC 2010 in Maire et al. (2011) and achieves an 11% relative boost over the human detector of Bourdev and Malik (2009). In this experiment, a match is found and recorded when there is more than a 50% overlap between a human region in the solution and a human region in the ground-truth with mandatory one-to-one mapping. False alarms are the occurrence of unmatched entities and miss-detections are the unmatched objects in the ground-truth and those human shapes which partially appear in the scene are not recorded at all. The results are shown in Table 3.1.

The missed detections occur mainly due to difficulties with the low-level features computation where clothing merges with the background as they have similar colours or when several people are standing close to each other and are seen as ambiguous blobs. Another main cause for a false alarm is including significant non-human areas which are relatively large compared to a human's size in the foreground, sample of detection results is shown in Figure 3.6.

The process of human detection using a poselets detector is aided by incorporating low-level cues and this is illustrated in Figures 3.7. The system deals with difficult and extraordinary poses which are not capable of being detected using top-down poselets models

²https://uk.mathworks.com/matlabcentral/newsreader/view_thread/305276

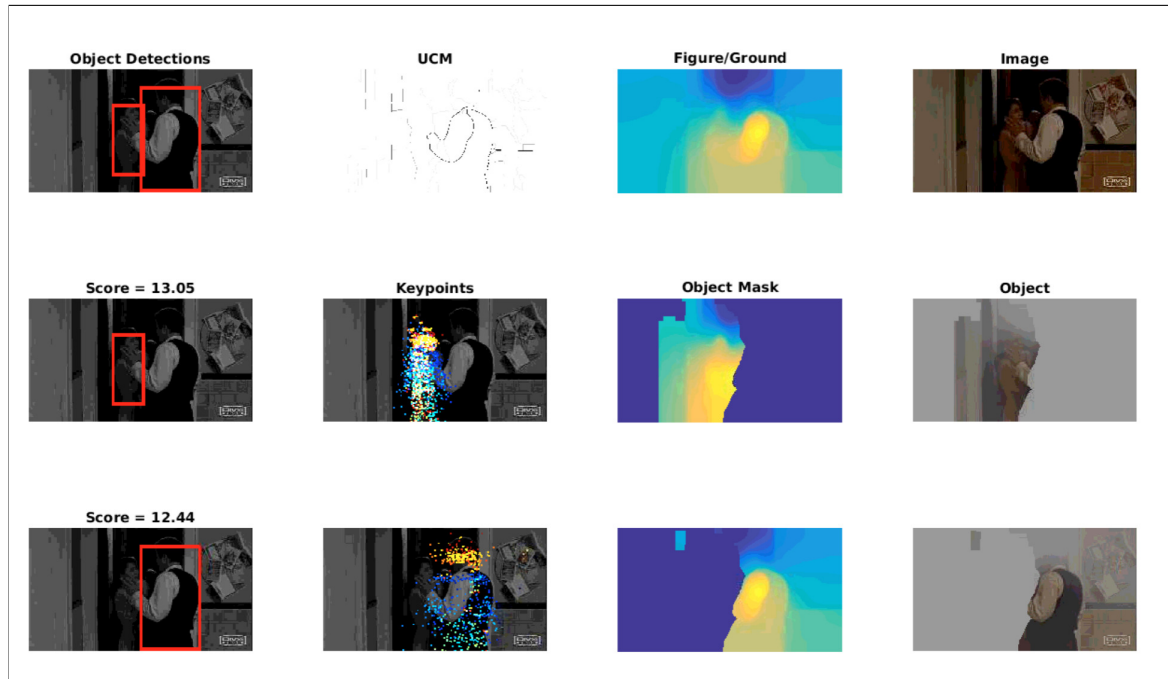


Fig. 3.6 Sample of human detection results of ‘actionclipautoautotrain00463’ video clip from the AnswerPhone class.

by themselves. It is important to note that the low-level cues helps when there is partial occlusion by filtering objects (such as, car in bottom row of Figures 3.7. Further, low-level cues helps the system in part grouping to locate people not identified in the top-down scanning of poselets detector.

3.4.3 Human Segmentation Results

In this section, the performance of the human body segmentation approach is compared with a recent object segmentation approach proposed by [Lee et al. \(2011\)](#). Following the popular video segmentation benchmark setting ([Tsai et al., 2012](#)) ten video clips are chosen from the Hollywood2 dataset to conduct this experiment. Each video frame is manually annotated, with human presence to be used as the ground-truth for evaluation purposes. All the selected scenes are set in an office environment and feature a broad range of motions as well as a variety of temporal changes, thus creating a challenging video segmentation task. The selected clips vary in duration up to 2 minutes, with at least one human present in each shot; there are many shots showing multiple human figures. For each clip, video frames are extracted using a `ffmpeg`³ decoder, with a sample rate of one frame per second. The purpose

³www.ffmpeg.org/

Class	Valid humans	Correct detections	Missed detections	False alarms	Detection rate	False alarm rate
AnswerPhone	59	39	20	5	66.10%	8.47%
DriveCar	60	44	16	2	73.33%	3.33%
Eat	62	50	12	3	80.65%	4.84%
FightPerson	47	36	11	1	76.60%	2.13%
GetOutCar	42	31	11	2	73.81%	4.76%
HandShake	48	35	13	3	72.92%	6.25%
HugPerson	33	24	9	2	72.73%	6.06%
Kiss	46	32	14	4	69.57%	8.70%
Run	33	20	13	1	60.61%	3.03%
SitDown	42	32	10	2	76.19%	4.76%
SitUp	34	27	7	1	79.41%	2.94%
StandUp	44	32	12	4	72.72%	9.09%

Table 3.1 Detailed detection results of performance evaluations on the NLDHA dataset for each class using combination of low-level cues and top-down detector information. Columns from left to right, first column lists the classes names in our dataset, second is the number of valid human regions in ground-truth for each class, third is the number of human detections for each class using proposed framework, fourth is the number of missed human detection for each class, fifth is false alarm which corresponds to false positive resulted from our system, last two columns correspond to the percentage of correct detections and false alarms respectively over the number of valid human detections per class.

of this comparison is to show the rough position of our approach among the state-of-the-art techniques in the context of the object segmentation task.

Accuracy is the commonly used measure for evaluating video segmentation tasks. In this work we adopt the average per-frame pixel error rate (Tsai et al., 2010) for evaluation of the approach. Let F denote the number of frames in the video, and S and GT represent the number of pixels in the segmented region and in the ground-truth across the frame sequence respectively. The error rate is calculated using the exclusive OR operation:

$$E(S) = \frac{|XOR(S, GT)|}{F} \quad (3.11)$$

The equation is used under the general hypothesis that object and ground-truth annotation should match.

Figure 3.8 presents sample outcomes of segmentation using (a) the approach in this chapter and (b) the implementation by Lee et al. (2011).⁴ We tested their implementation with our office scene dataset. This was perhaps not totally a fair comparison because the purpose of their work was an unsupervised approach to key object segmentation from unlabelled video, where the number of object was restricted to one, while we focused on extraction of human volume. It shows that accurate segmentation of humans was made by our approach. Implementation by Lee et al. (2011) could not extract a complete human body

⁴ Program code available from www.cs.utexas.edu/~grauman/research/software.html

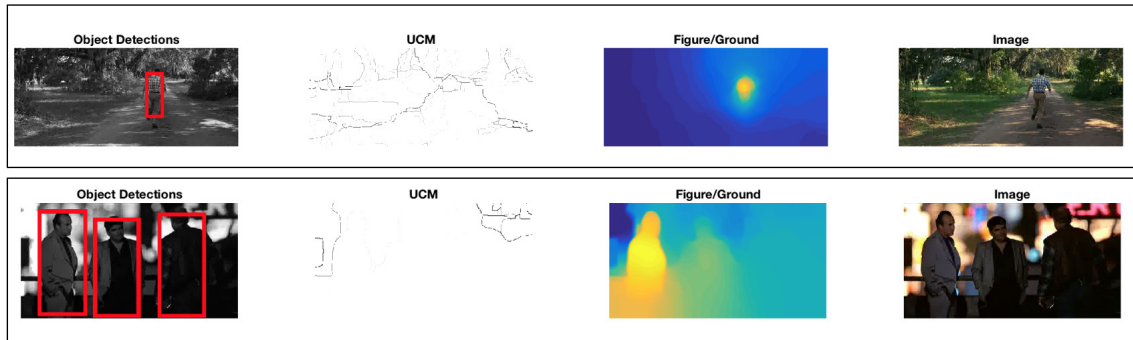


Fig. 3.7 Low-level cues enhanced the detection results that missed by top-down poselets human detector. Detection represented by bounding box and foreground model. Low-level cue helps in unusual pose (top row) and partial occlusion (bottom row).

clip name	this approach	Lee <i>et al.</i>
00007	1172	1875
00062	9829	42532
00099	6996	20858
00105	10870	13949
00107	2096	6919
00181	1265	9624
00187	8900	19112
00319	520	4659
00405	3400	30513
00431	11585	45361

Table 3.2 The average number of incorrectly segmented pixels per frame. The video clip name is in the format of ‘sceneclipautoautotrain·...’ where the ‘...’ part is shown in the table.

although it discovered some parts. That was mainly due to the fact that the system relies on low-level cues (appearance and motion) only for the purpose of object detection. Table 3.2 shows that the proposed approach in this chapter produced consistently better segmentation than the one implemented by Lee *et al.* (2011). We observed that the typical cause of failed segmentation by our approach was the presence of high occlusion, where the human body is mostly occluded by other objects.

On the other hand, segmentation by Lee *et al.* (2011) was unsuccessful especially when there was more than one person present in the scene. Moreover, they use low-level cues for the detection stage results in over segmentation where the targeted object is fractured into multiple components. Analysis of error patterns shows that low-level cues alone failed to

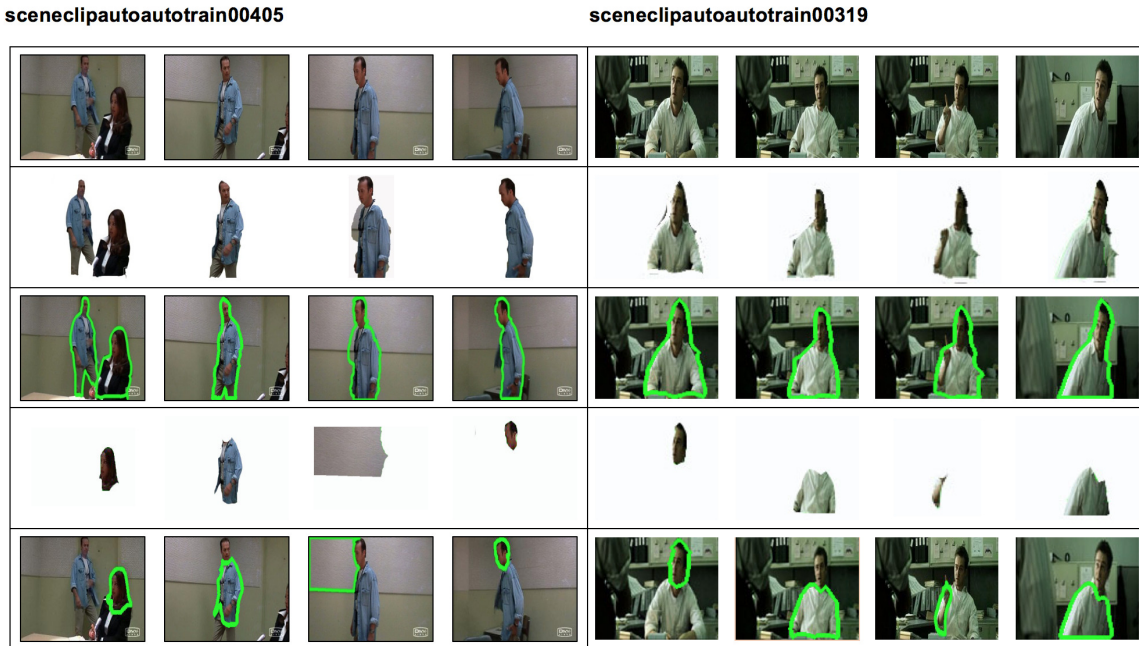


Fig. 3.8 Sample segmentations. The first row shows key frames from two video clips. The second and the third rows respectively present the results of key segments and the corresponding segmentation using the approach in this paper. The last two rows show the same attempts using the implementation by [Lee et al. \(2011\)](#). Best viewed on pdf.

detect partially occluded human objects. However, the top-down poselets detector which is used in our approach performs very well on detection in occlusion states, especially when there are some prominent poselets still be recognised from that occluded objects in the image.

3.5 Conclusion

In this chapter we have presented the two-stage approach to spatio-temporal human body segmentation by extracting a human body at a frame level, followed by tracking the segmented regions using colour appearance and local shape-matching across the video frames. By detecting and segmenting human body parts, we overcame the limitations of the bottom-up unsupervised methods that often over-segment an object. Using challenging video clips derived from the Hollywood2 dataset, we were able to obtain consistently better segmentation results than state-of-the-art implementations in the field.

The next chapters will be dedicated to the extraction of high-level visual attributes features for the human body regions. These features include actions, age, gender, emotions, spatial and temporal relationships between these regions in terms of space and time domains. Finally this information will be utilised to generate a natural language description for the video clip.

Chapter 4

Human Action Recognition Framework and Visual Attributes Identification

The ultimate goal of this thesis is generate a natural language description of human activities in video stream. In the previous chapter human objects are extracted from a video stream as the human usually is the main subject in daily activities. In this chapter, the visual attributions of the extracted human object, which constitute a list of high level features (HLFs), are extracted. These HLFs include action, emotion, gender, age, scene setting and other non-human objects involved in these activities. The main contribution of this chapter is the development of a new action recognition framework where the video representation is improved by using extracted spatio-temporal human regions combined with the extended spatio-temporal locality-constrained linear coding (LLC) technique in order to identify the action class. Most of the previous studies in this field were based on space-time interest points, whereas more spatially extended features, such as regions, have received considerably less attention. This is due to the fact that, in a local region-based approach, the motion flow information pertaining to a particular region must be subject to temporal collation. This study addresses the matter by applying a sturdy region tracking method, demonstrating that region descriptors can be attained for the action classification task. Based on the assumption that a frame consists of an individual as the principal actor and, as such, their body regions constitute the regions of interest, a cutting-edge human detection method is applied to generate a model incorporating generic object foreground segments. These segments are extended to include non-human objects which interact with human in the video scene to capture the action semantically. Extracted segments are subsequently described using HOG/HOF descriptors in order to delineate their appearance and movement. Next, the spatio-temporal extension of the LLC technique is implemented to optimise the codebook, the coding scheme projecting every one of the spatio-temporal descriptors into a local coordinate representation developed

via max pooling. To assess the performance of this model human action classification experiments are conducted using the KTH, the UCF sports and the Hollywood2 datasets. The outcome shows that the local region-based approach with LLC coding technique clearly outperforms the state-of-the-art, point feature-based techniques, particularly with the most challenging Hollywood2 dataset. Additionally, the identification of the rest of the high-level visual features list including age, gender, emotion and scene setting will be presented in detail in Appendix A as they will be used later as an input for language generation model. Identification of these features is achieved by utilising off-the-shelf software packages.

The chapter is structured as follows: Section 4.1 introduces the human action recognition framework combined with the motivations for this work and its contributions. Section 4.2 reviews previous work related to features coding in video representation. The implementation of the proposed action recognition framework is presented in Section 4.3, while Section 4.4 presents a summary of the results obtained from the evaluation experiment and a comparison with the state-of-the-art methods. Finally, Section 4.5 provides a concluding discussion.

4.1 Introduction

It is not a difficult task for a human to comprehend the actions occurring in a video clip, regardless of the scene context, individuals in the scene or the camera angles with which the scene is presented. Furthermore, viewers can follow an extensive series of actions no matter how complex they are. From the computational point of view, however, action representation poses considerable challenges. To provide a solution to this problem, most existing approaches are geared towards the description of motion information within a scene. Descriptors for motion information are highly important; in recent years methods used to garner space-time interest point features have been greatly improved (Marszalek et al., 2009a). This chapter presents a departure from the point feature-based approaches; instead a spatio-temporal region-based approach to motion description from a video stream, which was extracted from previous chapter, is explored. Furthermore, a spatio-temporal extension of the LLC coding technique is presented to reduce the dimensionality of extracted low-level features.

It has been demonstrated that high-level models, which function on representations based on tracked objects, their features and/or interaction, are capable of identifying complex actions (Brendel et al., 2011; Sridhar et al., 2010c). One such model, relying on interaction primitives, was proven to be highly effective when extracting appearance and space-time interest point (STIP) primitives. STIP primitives were outlined on person and object trajectories, and attained via a flexible part-model detector (Packer et al., 2012). Despite their efficiency,

such models remain incapable of reliably tracking interacted objects or of functioning in a variety of observation conditions.

Extraction of a multitude of distinctive features is done in the case of video processing. However, retrieval outcomes related to these features are usually negative, particularly when the representation of high-level concepts established according to users' interests based on low-level features is challenging. To bring features and concept closer together and avoid having to use the feature space, [Jiang et al. \(2005\)](#) sought to discover a hidden semantic 'concept' space. It was anticipated that such a hidden space would have greater representativeness and its processing and management would be easier because its dimensions would not be as extensive as those of the initial space. In order to obtain this hidden space, extraction of low-level features and their conversion into mid-level representations was the approach adopted by the majority of applications ([Liu et al., 2008](#)). Bag-of-words (BoW) ([Csurka et al., 2004](#)), spatial pyramids ([Lazebnik et al., 2006](#)), and sparse coding (SC) ([Yang et al., 2009](#)) are all common types of mid-level features. However, considerable limitations are presented by this approach; one major limitation is the fact that it disregards the order of the features in space, and therefore it is unable to determine where the represented object is situated. To address such limitations, several solutions have been proposed, such as LLC ([Wang et al., 2010](#)), which undertakes projection of features into a local-coordinate system based on locality restrictions followed by pooling-based incorporation of features to produce a representation of greater compactness. This solution was successful when it was applied in the context of image processing. Nonetheless, in terms of temporal data, significant progress still needs to be made.

The spatio-temporal information supplying content of semantic coherence is the defining trait of a video sequence ([Singh et al., 2009](#)). Preceding frames pass on objects with specific spatial relations and movement data to the frames that are arranged in a chronological manner. The definition of the video content depends significantly on the temporal trajectories of inter-object spatial relationships as well as those of individual objects for activity recognition. Hence, in the case of video sequence applications comprising such complex relations, it is not enough to consider just the spatial data. However, even though they are clearly important, temporal and spatial traits have been given only limited attention in the existing coding methods.

Given the above considerations, the present chapter will approach movement description from a video stream via a technique based on a spatio-temporal 'region', rather than the common technique based on point features. Moreover, to decrease the dimensionality of extracted low-level features and move them to mid-level representation, the chapter considers a spatio-temporal extension of the LLC coding method.

4.1.1 Motivations

Action detection in real-life videos is the focus of the suggested model. Unlike space-time interest points, which have formed the basis of the majority of earlier studies, spatially extended features (*e.g.* regions) have not attracted the same level of interest, because the motion flow data associated with a certain area must undergo temporal collation within an approach based on local region. To prove that it is possible to obtain region descriptors for the task of action classification, a robust region tracking technique is employed in this chapter to deal with this issue.

Among the methods that have been demonstrated to be useful for image and video processing are bag of words (BoW) and sparse coding (SC). Representing an image with a histogram of its local features, BoW has a good performance when it comes to both spatial translation of features and image classification tasks. However, the method is deficient with regard to capture of shapes or detecting where objects are because it does not consider the spatial relation between local features. To solve this issue, the suggested coding method called spatial pyramid matching (SPM) is believed to have good potential as far as image processing tasks are concerned. Nonetheless, to achieve a high performance, SPM must be used alongside classifiers and non-linear kernels. This requires further complex computation that is expensive and consequently this method does not have good scalability for video applications.

LLC was created as a representation for image features of greater speed and efficiency in order to overcome the issues faced by other coding techniques, such as loss of relationship data and computational cost. In spite of this, the problem remains that only spatial information is taken into account by existing research, while temporal information is overlooked. Furthermore, a mid-level representation encoding the extracted features with not as many codes but no less informative must be established for the video sequences application. The numerous interest points typically extracted by detectors and descriptors of features are compacted by this mid-level to a series of codes produced based on a codebook outlined from a selection of those features. Consequently, a descriptor coding is considered as a mediating step in the representation of the entire video, instead of every spatio-temporal descriptor being labelled as a distinct feature. From a computational perspective, using the codes as video representation is advantageous because it makes handling less time-consuming and enables more effective storage of visual descriptors.

4.1.2 Human Action Recognition Framework: Overview

In order to process complex actions which are challenging to track efficiently using conventional descriptors, a new model for action representation that relies on generic object segments is investigated. To this end, this chapter utilises the extracted spatio-temporal human body regions from previous chapter to limit the search space of features and incorporate the spatio-temporal extension of the LLC scheme to develop an action recognition framework. The LLC is an image processing coding scheme proposed by Wang et al. (2010) to encode a set of features extracted using 2D SIFT with a smaller (than the original set of features) set of codes based on the spatial relationship between the features. To detect interest feature points, the dense 2D SIFT is replaced, from the original work, with HOG/HOF descriptors. The LLC is able to represent these points with fewer codes using the spatio-temporal relationship between the descriptors and the basis codebook. The approach consists of two principle stages. The first involves extraction of the motion and appearance features from detected spatio-temporal regions using HOG/HOF descriptors (Laptev et al., 2008) that formulates a descriptor encompassing the static and dynamic features of detected segments. In the second stage, the LLC coding scheme is applied to the extracted descriptors in order to encode the local descriptors with similar basis from a codebook. The approach is evaluated using a human action classification task as a benchmark.

4.1.3 Human Action Recognition Framework: Contributions

The main contributions of the proposed work in this chapter can be summarised as follows:

- development of an efficient and robust schema to represent a human action signal by combining spatio-temporal regions with the LLC coding scheme;
- application of the spatio-temporal region-based approach to the action classification task with the Hollywood2, one of the most challenging real-world datasets, demonstrating that the approach outperforms the state-of-the-art interest point-based techniques by a clear margin.

4.2 Related Work

An image or video is assigned one or multiple category labels in a process called image or video classification. Among the various applications of image and video classification, which is a major issue in computer vision and pattern recognition, are video surveillance (Bouwman and Zahzah, 2014), image and video retrieval (Veltkamp et al., 2013) and human-computer interaction (Rautaray and Agrawal, 2015). However, recent studies have focused

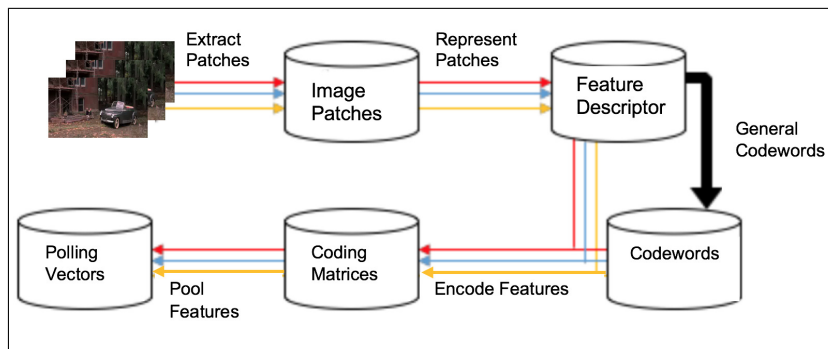


Fig. 4.1 Typical stages via which image classification is performed by the BoW model

particularly on the bag-of-words (BoW) (Ensaft et al., 2014), an image classification model of great efficiency and popularity. Figure 4.1 illustrates the five typical stages via which image classification is performed by the BoW model. Huang et al. (2014) summarise these stages as follows:

1. **Patch extraction:** In this stage, the images and image patches respectively represent the input and output. In this procedure, local image areas are sampled densely or sparsely, by employing fixed grids or feature extractors, respectively.
2. **Patch representation:** The feature descriptors (vectors) of the image patches constitute their outputs and the procedure involves performance of statistical analysis over image patch pixels, such as the commonly used scale-invariant feature transform (SIFT) descriptor (Susan et al., 2015), local binary pattern (Devi et al., 2015), and oriented gradient histogram (Dhamecha et al., 2014).
3. **Codeword generation:** The feature descriptors derived from every training image and codeword respectively represent the inputs and outputs of this stage. Random sampling of a subseries of descriptors from all the descriptors is undertaken in real application to make computation more efficient. Clustering (*e.g.* K-means (Cai et al., 2013)) over feature descriptors or supervised (Yang et al., 2010) or unsupervised (Jiang et al., 2012) codeword learning are the main methods for generating codewords.
4. **Feature encoding:** In this stage, the input is represented by the feature descriptors and codewords, while the coding matrix is the output. The process involves production of a coding vector through activation of several codewords by the feature descriptors. This vector is as long as the number of codewords. The manner of codeword activation (*e.g.* the specific codewords that are activated and the size of their response amplitude) is what differentiates the diverse coding algorithms.
5. **Feature pooling:** The input and output for this stage are the coding matrix and a pooling vector for every image (*e.g.* final image representation), respectively. The

procedure involves integration of all responses on every codeword into a single value via methods such as average pooling, which maintains the average response (He et al., 2015), and MAX pooling, which maintains the maximum response (Cho et al., 2014).

Feature coding is the most important of the above stages, largely determining how precise and fast the image classification process is; due to the fact that it connects feature extraction and pooling. There are a range of coding techniques for several benchmarks, including the BoW framework, with both hard and soft assignment, Fisher coding, and the linear coordinate coding family consisting of vector quantisation (VQ), SC and LLC. Each of these techniques is discussed below.

4.2.1 Bag-of-Words model

Among the initial implementations in object recognition, scene matching and image categorisation was the BoW framework (Csurka et al., 2004). A codebook is generated as a result of the clustering of the local features in the training stage. The pre-learned codebook supplies the visual words used to code the extracted local features, thus producing the representation. Hard and soft assignments are just two of the different methods that have been suggested for BoW coding.

4.2.1.1 Hard assignment coding

Hard assignment coding entails the use of a particular distance metric to allocate feature x_i to its closest word in codebook V as the base of every descriptor (Lazebnik et al., 2006). For instance, the coding takes the form below if the Euclidean distance is employed:

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } j = \arg \min_{j=1,\dots,k} \|x_i - v_j\|_2^2 \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

In the above relationship, the code correlated with descriptor x_i is denoted by $\alpha_{i,j}$ while the ℓ_2 - norm is represented by $\|\cdot\|_2$ and K represents the closest neighbours.

4.2.1.2 Soft assignment coding

There are two drawbacks to the traditional codebook model: code-word uncertainty and codeword plausibility. Both of these drawbacks stem from the hard assignment of visual features to a single codeword. Allowing a degree of ambiguity in assigning codewords shows that the categorisation performance is improved.

The extent to which a descriptor x_i is a member of the j th codeword v_j constitutes the base of $a_{i,j}$ in the context of soft assignment coding (Van Gemert et al., 2008):

$$\alpha_{i,j} = \frac{\exp\left(-\beta \|x_i - v_j\|_2^2\right)}{\sum_{l=1}^k \exp\left(-\beta \|x_i - v_l\|_2^2\right)} \quad (4.2)$$

In the above, the parameter for regulation of assignment softness and the ℓ_2 - norm are respectively denoted by β and $\|\cdot\|_2$.

4.2.2 Fisher Coding

A signal with a gradient vector obtained from its probability density function is defined by the Fisher kernel method (Jaakkola et al., 1999), which constitutes the basis for Fisher coding (Perronnin and Dance, 2007). To ensure optimal compatibility with the data, the gradient vector informs the direction of adjustment for the parameters. For the purpose of image classification, feature coding is undertaken based on the gradient vector, while the signal represents an image. Several extended versions of the initial Fisher coding have been developed, such as the improved Fisher kernel (IFK) (Perronnin et al., 2010) which is the most efficient, as far as the current researcher is aware.

The Gaussian mixture models (GMM) define the probability density distribution of features in the context of IFK. The expectation maximisation (EM) algorithm is usually applied to determine the GMM parameters, namely, $\theta_m = \omega_m, \mu_m$, and Σ_m , which respectively represent the weight, mean vector and covariance matrix of the m th Gaussian distribution (Liu et al., 2017). The expression of an image may take the form of the log-likelihood of the entirety of extracted features, if no feature is dependent on another:

$$L(\mathcal{X}|\theta) = \sum_{n=1}^N \log p(x_n|\theta) \quad (4.3)$$

In the above equation, the probability density function underpinned by GMM is denoted by $p(x_n|\theta)$.

4.2.3 Linear Coordinate Coding (LCC)

Using codeword constraints to solve an optimisation problem based on least-squares and thus to achieve a codeword-based feature reconstruction is the principle underlying LCC techniques. Let $X = x_1, x_2, \dots, x_N \in \mathbb{R}^{D \times N}$ be a set of D -dimensional descriptors representing

the image and $B = b_1, b_2, \dots, b_M \in \mathbb{R}^{D \times M}$ be a codebook of size M . Then $S = s_1, s_2, \dots, s_N \in \mathbb{R}^{M \times N}$ will be a set of codes where each descriptor in X will be converted to an M -dimensional using coding scheme to formulate the ultimate image representation. In this section three different coding schemes of this type are discussed,

4.2.3.1 Vector Quantisation (VQ)

VQ coding is employed by conventional SPM, providing a solution to the problem of constrained least squares fitting below (Wang et al., 2010):

$$\operatorname{argmin}_C \sum_{i=1}^N \|x_i - B_{c_i}\|^2, s.t. \|c_i\|_{\ell^0} = 1, \|c_i\|_{\ell^1} = 1, c_i \succeq 0, \forall i \quad (4.4)$$

In the above, the code set for X is denoted by $C = c_1, c_2, \dots, c_N$. The fact that every code c_i will contain just one non-zero element equivalent to the quantisation id of x_i is indicated by the cardinality constraint $\|c_i\|_{\ell^0} = 1$. For x , the coding weight is 1, as indicated by the non-negative, ℓ^1 constraint $\|c_i\|_{\ell^1} = 1, c_i \succeq 0$. Practically, identification of the closest neighbour is needed to determine the one non-zero element.

4.2.3.2 Sparse Coding (SC)

A sparsity regularisation term enables the relaxation of the restrictive cardinality constraint $\|c_i\|_{\ell^0} = 1$ in Equation 4.4 in order to alleviate the quantisation loss of VQ. Such a term is chosen as the ℓ^1 - norm of c_i in ScSPM (Yang et al., 2009), and therefore the coding of every local descriptor x_i takes the form of conventional sparse coding (SC) (Lee et al., 2006) problem:

$$\operatorname{argmin}_C \sum_{i=1}^N \|x_i - B_{c_i}\|^2 + \lambda \|c_i\|_{\ell^1} \quad (4.5)$$

Among the key functions of the sparsity regularisation term are ℓ^1 regularisation to provide a singular solution to the under-determined system as the codebook B is typically over-complete (*e.g.* $M > D$); sparsity prior let capturing of pertinent local descriptor patterns by learned representation; and reduction of quantisation error compared to VQ. Hence, the performance of ScSPM significantly exceeds that of nonlinear SPM approach on benchmarks such as Caltech-101, even when the linear SVM classifier is used (Yang et al., 2009).

Owing to its advantages, SC has been commonly employed in image representation, machine learning and signal processing (Yang et al., 2009). These advantages include reduced reconstruction error compared to VQ due to the use of fewer constraints, effective

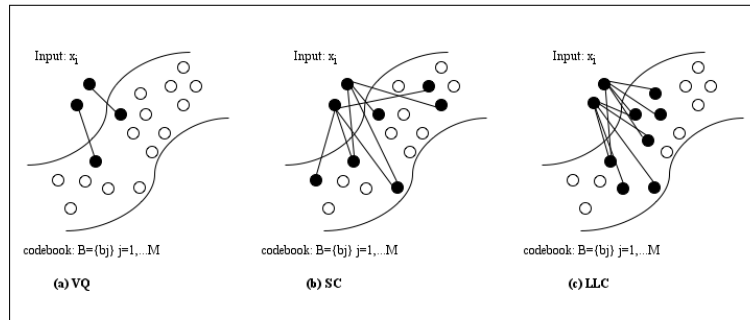


Fig. 4.2 Comparison between the three linear coordinate coding techniques; the black circles represent the chosen codewords for the feature x_i .

modelling of the most important image features, and the fact that the sparsity of images has been demonstrated by statistical investigation of images.

However, a number of issues are yet to be resolved, even though numerous algorithms for classification and recognition have been created (Shi et al., 2017; Zhang et al., 2014). One major issue is that model selection and learning are made highly complex because the feature extraction technique usually has an inclination towards sparsity. This has prompted the formulation of various strategies that integrate methods of features learning, dimensionality reduction methods and clustering techniques. For example, to address the issue of SC as sparsity-constrained robust regression, Yang et al. (2011a) proposed the robust sparse coding (RSC) model, which helped the original SC perform better and was effective in the management of facial occlusions.

4.2.3.3 Locality-constrained Linear Coding (LLC)

To facilitate projection of distinct descriptors on their corresponding local-coordinate systems, Wang et al. (2010) put forth the LLC coding scheme. This scheme attributes greater importance to locality compared to sparsity because sparsity is implied by locality but locality is not implied by sparsity. Several beneficial qualities may be derived for this preference for the locality constraint instead of the sparsity constraint, as follows:

- **Improved reconstruction:** As shown in Figure 4.2a, representation of every descriptor in VQ has just one basis in the codebook. The VQ code may not be the same for similar descriptors owing to extensive quantisation errors. Additionally, the correlations among various basis are not considered by the VQ process and therefore this loss of data must be rectified by non-linear kernel projection. By contrast, in LLC, more than one base

is used for precise representation of every descriptor and the basis are shared to enable the relationships between similar descriptors to be captured (Figure 4.2c).

- **Local smooth sparsity:** Just like LLC, multiple basis are also employed by SC to reduce reconstruction errors. However, in SC, the ℓ^1 – norm regularisation term is not smooth. Relationships between codes may be lost because, in order to promote sparsity, different basis may be chosen by the SC for similar patches, as the codebook is over-complete (Figure 4.2b). However, similar codes for similar patches are guaranteed in LLC thanks to the explicit locality adaptor.
- **Analytical solution:** The practical implementation of LLC is much quicker, unlike sparse coding, which involves procedures that are challenging from a computational perspective.

An online method of learning facilitates the incorporation of a codebook learning step into LLC (Wang et al., 2010). The updating of B , the original codebook trained on the basis of k – means clustering, is performed gradually, with the iteration of the training descriptors. Single or small-batch examples x_i are taken up for every increment and employed to obtain the necessary solution, generating the LLC codes related to codebook B . The procedure keeps just one set of basis B_i with weights greater than a pre-defined constant and therefore it functions as a feature selector. Refitting of the x_i values is subsequently undertaken in the absence of locality constraints. The basis is afterwards updated by the code via a gradient descent. Last but not least, the code is submitted to multi-scale pyramid max pooling, producing the representation outcome (*e.g.* feature representation). This strategy is advantageous because it is fast, simple, and scalable, without lowering SPM performance (Yang et al., 2009).

4.3 Human Action Representation

There are numerous instances in which, in addition to the direct observation of a human body in motion, the characteristics of related objects can also contribute to the identification of human actions. The aim of this chapter is to address this issue and suggest a multi-feature method of determining human actions. This study addresses the matter by applying a sturdy region tracking method, instead of the conventional space-time interest point feature-based techniques, demonstrating that region descriptors can be attained for the action classification task. A cutting-edge human detection method is applied to generate a model incorporating generic object foreground segments. These segments can be extended to include non-human objects (*e.g.* car, chair) which interact with a human in a video scene to capture the action semantically. Extracted segments are subsequently expressed using HOG/HOF descriptors

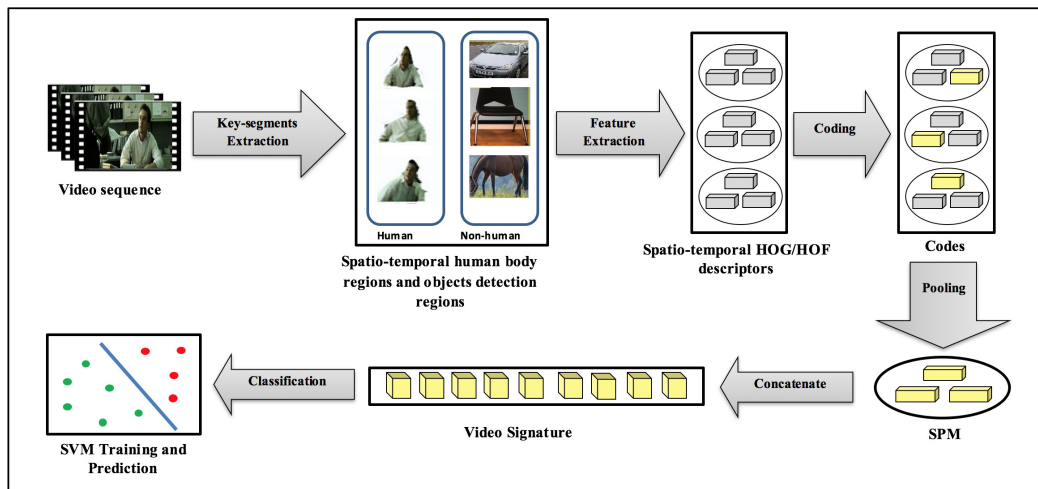


Fig. 4.3 Processing flow of the 'human body region tracking' approach with visual object recognition (HBRT/VOC).

to delineate their appearance and movement. LLC coding is employed to optimise the codebook, the coding scheme projecting every one of the spatio-temporal descriptors into a local coordinate representation developed via max pooling. Figure 4.3 illustrates the processing flow of the technique presented in this chapter, which is later on referred to as the 'human body region tracking' approach with visual object recognition (or HBRT/VOC). Each of these stages is described in turn below.

4.3.1 Detecting and Tracking Human Body Regions

A robust approach introduced in [Al Harbi and Gotoh \(2013b\)](#) is utilised for the purpose of extracting a key segment of the human body regions from video sequence. This approach involves using an unannotated video as input, generating a list of inferred space-time segmentation of the key regions of the human body which are brought into focus by the activity. The key-segment extraction approach consists of several stages. A preliminary list of targeted regions is identified at frame level using low-level cues and a top-down part-based person detector. Next, the temporal consistency of detected foreground objects using colour models and local shape matching are used to reproduce detected segments in subsequent video frames. This will result in a final output of a spatio-temporal segmentation of the human body in a video stream. More details are presented in Chapter 3.



Fig. 4.4 A sample clip from the NLDHA dataset: GetOutCar action from a video clip ‘actioncliptest00108’. A region was detected using Felzenszwalb et al. (2010) (red bounding box), while a human body was detected by using the HBRT approach (green contour). The car region was included in the action representation as there was an overlap between a car and a human.

4.3.2 Non-human Object Detection Regions

In many instances, human activities can be effectively presented by collaboration and interaction between human and non-human objects. Eating action, for example, can be illustrated by describing a person who sits around a dining table and grasps the food. Consequently, action classification will operate more effectively if non-human objects are incorporated into the zone of interest. A number of studies have been conducted for visual object recognition tasks¹ (VOC) that can be employed as a front-end processor for the human body region tracking (HBRT) approach. In this study a detector developed by Felzenszwalb et al. (2010) is adopted, creating a store of the following object classes: bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, person, bottle, chair, dining table, potted plant, sofa, phone and tv/monitor. A window is tightly fitted to the identified object segment and a bounding box is drawn on the window. In order to guarantee the consistency of the object segments perceived with the human body regions in the video, certain spatial restrictions are imposed; if a confluence of human body and non-human regions exists, the segments are included as key-segments regions (see Figure 4.4).

¹The PASCAL visual object classes available at: pascallin.ecs.soton.ac.uk/challenges/VOC/

4.3.3 Describing Detected Regions

Once key-segments are determined, they must be described as the identified hypotheses. To encompass the appearance and motion patterns of the regions of interest throughout a video clip, the HOG/HOF descriptor (Laptev et al., 2008) is employed in this study. The 162-bin descriptor is composed of a histogram of oriented gradients (HOG) and a histogram of oriented flow (HOF). To outline the movement and appearance of selected features, the histogram descriptors of space-time volumes are positioned in the proximity of the identified points. Each volume is subsequently separated into a $n_x \times n_y \times n_t$ grid of cells²; for each cuboid a coarse HOG with 4-bin histogram and a HOF with 5-bin histogram are generated. Normalised histograms are integrated into HOG/HOF descriptor vectors, exhibiting certain similarities with the SIFT (scale invariant feature transform) descriptor by Lowe (2004). As an additional note, in the event that key-segment hypotheses are not produced for some video clip due to a failure of human detection, the HOG/HOF descriptor is augmented with space-time interest points.

4.3.4 Learning Feature Sets

A training set consists of N videos, and we define $X = \{x_1, x_2, \dots, x_N\}$ where x_i represents a D -dimensional spatio-temporal descriptor for each video. A sufficient number of features are randomly selected and grouped together using k -means in order to attain a preliminary codebook B of the fixed vocabulary size for spatio-temporal features. LLC is subsequently applied to enhance the codebook, which consists of the following three steps; representing video signals with spatio-temporal local descriptors X , generating the locality-constrained sparse code S , and finally optimising the codebook B . The codebook basis and LLC coefficients should efficiently approximate spatio-temporal descriptors. The following objective function (Wang et al., 2010) is employed:

$$\operatorname{argmin}_{S, B} \sum_{i=1}^N \{ \|x_i - B s_i\|^2 + \lambda \|d_i \odot s_i\|^2 \} \quad \text{st.} \quad 1^\top s_i = 1, \forall i \quad (4.6)$$

where \odot denotes element-wise multiplication and λ is a weight parameter to control the locality constraint. The constraint, $1^\top s_i = 1$, meets the requirement of shift-invariance for the LLC coding scheme. The locality constrained parameter d_i represents every basis vector in the codebook with different freedom based on its similarity to the spatio-temporal descriptor

² The parameter values employed in this study are $n_x = n_y = 3$ and $n_t = 2$, following the setup described in Laptev et al. (2008).

x_i :

$$d_i = \exp\left(\frac{\text{dist}(x_i, B)}{\sigma}\right) \quad (4.7)$$

with $\text{dist}(x_i, B) = \{\text{dist}(x_i, b_1), \dots, \text{dist}(x_i, b_N)\}^T$

where $\text{dist}(x_i, b_j)$ represents the Euclidean distance between the spatio-temporal descriptor and the basis codebook B , and σ is a weight parameter to control the locality constraint.

The initial codebook B is optimised by approximating each spatio-temporal descriptor by the product of LLC coefficients and codebook. The optimal codebook can be generated by solving Equation 4.6. This is a convex problem in B only or in S but not in both together, and can be iteratively solved by the coordinate descent method, as follows:

1. Initialise the dictionary B with the codebook generated by K-means clustering:

$$B \leftarrow B_{init} \quad (4.8)$$

2. For each spatio-temporal descriptor x_i , compute the new LLC coefficient s_i using the current B :

$$s_i \leftarrow \underset{s}{\operatorname{argmax}} \{ \|x_i - Bs\|^2 + \lambda \|d \odot s\|^2 \} \quad \text{st.} \quad \mathbf{1}^\top s = 1 \quad (4.9)$$

3. For each basis in B , remove the column B_i with weights that exceed a predefined threshold:

$$id \leftarrow \{j | \text{abs}(s_i(j)) > 0.01\}, \quad B_i \leftarrow B(:, id) \quad (4.10)$$

4. Refit x_i without the locality constraint with an approximated code \tilde{s}_i to speed up the coding process:

$$\tilde{s}_i \leftarrow \underset{s}{\operatorname{argmax}} \|x_i - Bs\|^2 \quad \text{st.} \quad \sum_j s(j) = 1 \quad (4.11)$$

5. Update the current dictionary, only if the computed LLC coefficient value is greater than a predefined threshold:

$$\Delta B_i \leftarrow -2\tilde{s}_i(x_i - B_i\tilde{s}_i) \quad (4.12)$$

$$\mu \leftarrow \sqrt{\frac{1}{i}} \quad (4.13)$$

$$B_i \leftarrow B_i - \frac{\mu \Delta B_i}{|\tilde{s}_i|_2} \quad (4.14)$$

6. Project the computed dictionary onto the output matrix:

$$B(:, id) \leftarrow \text{proj}(B_i) \quad (4.15)$$

The features are quantised on the basis of the vocabulary, with the purpose of creating a feature histogram which represents the vector for feature categorisation. A non-linear support vector machine (SVM) classifier with a χ^2 -kernel is used to learn a model from feature vectors for each action (Fan et al., 2008).

4.4 Experiments

This section assesses the effectiveness of the spatio-temporal HBRT approach and its extension HBRT/VOC, using three action recognition datasets. For all datasets we apply HOG/HOF descriptors with 162 dimensions. 100,000 features are randomly selected for initialisation of the codebook with the vocabulary size of 4,000 words (the key parameter for dictionary training), and the number of neighbours is $K = 5$. In Equation (4.6), $\lambda = 500$ is selected, and $\sigma = 100$ is set for Equation (4.7). To evaluate the outcome of the action classification task, accuracy per class is calculated using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (4.16)$$

where TP , TN , FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively.

4.4.1 Datasets and the experimental procedure

Recognizing human activities has become an important topic in the past few years. A variety of techniques for representing and modelling different human activities have been proposed, achieving reasonable performances in many scenarios. The recent benchmark introduced by Caba Heilbron et al. (2015) is named ActivityNet, a new large-scale video benchmark for human activity understanding. This benchmark aims at covering a wide range of complex human activities that are of interest to people in their daily living. In its current version, ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video. Further, Shahroudy et al. (2016) introduced a large-scale dataset for RGB+D human action recognition with more than 56 thousand video samples and 4 million frames, collected from 40 distinct subjects. This dataset contains 60 different action classes, including daily, mutual, and health-related actions.



Fig. 4.5 Sample frames from the three action recognition datasets, the KTH (top row), the UCF Sports Action (middle) and the Hollywood2: Human Actions and Scenes (bottom), used for the experiments.

However, due to a lack of published results on these dataset, the body region tracking approaches (HBRT and HBRT/VOC) were comprehensively evaluated in this study using three popular action datasets selected from the KTH, the UCF sports and the Hollywood2 video data. The datasets encompassed a variety of locations and scene settings shown in video clips, including controlled experimental settings, popular films and televised sporting events. The assessment incorporated a range of variations resulting from different resolutions, perspectives, lighting shifts, occlusion, background disorder, and irregular motion. Overall, more than 4,000 video segments were assessed with 28 action classes. The sample frames are presented in Figure 4.5.

KTH Dataset³ (Schuldt et al., 2004)

This dataset comprises 2391 video segments, with six types of human action: ‘walking’, ‘jogging’, ‘running’, ‘boxing’, ‘waving’, and ‘clapping’. Each action was carried out for a number of times by 25 people and filmed in a variety of settings: outside, outside with scale variation, outside in changed clothing and inside. In most of the segments the background was monotone and still. Segments were resized to a spatial resolution of 160×120 pixels and the mean duration of video clips was four seconds. We adhered to the experimental format of the existing studies by splitting the samples into a test set (nine subjects: 2, 3,

³<http://www.nada.kth.se/cvap/actions/>

5, 6, 7, 8, 9, 10, and 22) and a training set (the other 16 subjects). Emulating the original paper, we trained and assessed a multi-class classifier (Schuldt et al., 2004), and calculated the accuracy for each class and finally reported the average accuracy over all classes.

UCF Sports Action⁴ (Rodriguez et al., 2008b)

This dataset contains ten human actions: ‘swinging’ (both on a pommel horse and on the ground), ‘diving’, ‘kicking a ball from the front and the side’, ‘lifting weights’, ‘horse-riding’, ‘running’, ‘skateboarding’, ‘swinging from the high bar’, ‘gold swinging from the back, front and side’, and ‘walking’. Nearly 200 video segments were used with a resolution of 720×480 , indicating significant intra-class variability. As with the KTH set, we employed a multi-class classifier and reported the average accuracy in all classes.

Hollywood2: Human Actions and Scenes Dataset⁵ (Marszalek et al., 2009a)

This data has been collected from 69 different Hollywood movies. It consisted of the following 12 action classes to be identified from real-life film scenes: ‘answering a phone’, ‘driving a car’, ‘eating’, ‘fighting’, ‘getting out the car’, ‘hand shaking’, ‘hugging’, ‘kissing’, ‘running’, ‘sitting down’, ‘sitting up’ and ‘standing up’. In total there were 1707 video sequences divided into a training set (823 sequences) and a test set (884 sequences), with the average length of ten seconds. Training and test sequences were mutually exclusive. The experiment was performed on this dataset with a spatial resolution of 360×288 pixels and a sample rate of 4.6 fps (frames per second) as suggested by Wang et al. (2009). A one-against-all SVM categorisation was applied where a binary classifier recorded every action (Fan et al., 2008).

4.4.2 Experimental Results

Table 4.1 presents the comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based techniques using the KTH dataset (Schuldt et al., 2004). To date, it is probably the most frequently used dataset in assessment of action recognition. The region tracking approaches performed well; the HBRT method achieved 97.2%, and its integration with VOC reached 98.5%, outperforming the reported outcome of Sadanand and Corso (2012) by a small margin. Figure 4.6 illustrates the confusion matrix associated with the HBRT/VOC approach. It still made occasional, although rare, confusion between ‘jogging’, ‘running’ and ‘walking’ actions. The effectiveness of the HBRT/VOC resulted from the shrewd targeting of interest points represented by human body

⁴<http://server.cs.ucf.edu/vision/data.html>

⁵<http://lear.inrialpes.fr/data>

	Box	Clap	Wave	Jog	Run	Walk
Box	100	0	0	0	0	0
Clap	0	100	0	0	0	0
Wave	0	2	98	0	0	0
Jog	0	0	0	97	3	0
Run	0	0	0	2	97	1
Walk	0	0	0	0	1	99

Fig. 4.6 (**KTH Dataset**) Confusion matrix between six action classes using the HBRT/VOC combination approach, where each column represents the instances in a predicted class and each row represents the instances in an actual class.

method	accuracy (%)
point feature based:	
Laptev et al. (2008)	91.8
Le et al. (2011)	93.9
Gilbert et al. (2009)	94.5
Sadanand and Corso (2012)	98.2
local region tracking:	
HBRT	97.2
HBRT/VOC	98.5

Table 4.1 (**KTH Dataset**) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.

regions, which allowed the action to be pre-determined, and eradicated superfluous and noisy background.

Table 4.2 makes the same comparison, but this time using the UCF Sports Action Dataset (Rodriguez et al., 2008b). For the region tracking approaches, the overall accuracy was 90.8% with the HBRT, which was further improved to 96.2% with the HBRT/VOC; the latter clearly outperformed the recent state-of-the-art (95.0%) by Sadanand and Corso (2012). For the HBRT/VOC, the confusion matrix between ten action classes is presented in Figure 4.7. The figure shows some confusion pairs such as ‘lifting’/‘skating’ and ‘running’/‘walking’ as their action representation was quite similar. The outcome indicates that the local region tracking is an effective new approach to capturing human activity on video, and possesses the great potential to achieve consistent performance in realistic conditions.

The KTH and the UCF Sports Action are both relatively small datasets. Hollywood2: Human Actions and Scenes (Marszalek et al., 2009a), on the other hand, was substantially more difficult dataset to process because of several reasons, for example, more classes, the larger number of videos, actions with more realistic background involving multiple

	Dive	Golf	Kick	Lift	Rid	Run	Skate	SwingBench	SwingSide	Walk
Dive	100	0	0	0	0	0	0	0	0	0
Golf	0	100	0	0	0	0	0	0	0	0
Kick	0	0	100	0	0	0	0	0	0	0
Lift	0	0	0	92	0	0	8	0	0	0
Rid	0	0	0	0	100	0	0	0	0	0
Run	0	0	0	0	0	95	0	0	0	5
Skate	0	0	0	6	0	0	94	0	0	0
SwingBench	0	0	0	0	0	0	0	100	0	0
SwingSide	0	0	0	0	0	0	0	10	90	0
Walk	2	0	0	0	0	7	0	0	0	91

Fig. 4.7 (UCF Sports Action Dataset) Confusion matrix between ten action classes using the HBRT/VOC combination approach, where columns represent the predicted classes and rows represent the the actual class. .

method	accuracy (%)
point feature based:	
Le et al. (2011)	86.5
Kovashka and Grauman (2010)	87.3
Wu et al. (2011)	91.3
Sadanand and Corso (2012)	95.0
local region tracking:	
HBRT	90.8
HBRT/VOC	96.2

Table 4.2 (UCF Sports Action Dataset) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.

objects, camera motions. Table 4.3 compares various approaches using the challenging Hollywood2 data. Laptev et al. (2008) presented a technique where space-time interest points were identified by the Harris-Laplace detector and described with HOF. Another technique, based on the motion region, was proposed by Bilen et al. (2011). The table shows the significant improvement made by the HBRT and the HBRT/VOC, that are the local region-based approaches. In particular the latter achieved improvement of more than 4% absolute over the recent state-of-the-art by Laptev et al. (2008).

Additionally Table 4.4 presents the comparative analysis of the performance for individual classes by the region and point feature-based approaches. The complexity of the Hollywood2 data resulted in the low performance with several action classes, in particular with ‘AnswerPhone’, ‘GetOutCar’ and ‘SitUp’. Some videos contained a variety of camera motions, peripheral actions as well as a multitude of viewing angles and action sequences.

method	accuracy (%)
point feature based:	
Laptev et al. (2008)	44.4
Bilen et al. (2011)	41.3
local region tracking:	
HBRT	44.4
HBRT/VOC	48.6

Table 4.3 (**Hollywood2: Human Actions and Scenes Dataset**) Comparison of the local region tracking approaches (HBRT and the HBRT/VOC) with the state-of-the-art, point feature-based methods.

action class	Laptev <i>et al.</i>	Bilen <i>et al.</i>	HBRT	HBRT/VOC
AnswerPhone	19.1	21.9	15.2	18.4
DriveCar	80.2	84.5	70.6	86.6
Eat	60.2	49.6	61.2	72.4
FightPerson	72.4	59.2	70.9	71.1
GetOutCar	25.6	24.0	18.2	28.7
HandShake	18.9	12.3	29.3	31.3
HugPerson	32.1	21.4	33.1	33.6
Kiss	47.8	49.3	50.3	52.3
Run	68.8	61.8	61.0	62.2
SitDown	49.2	40.9	50.9	53.3
SitUp	9.9	20.8	21.3	23.5
StandUp	49.0	50.4	50.2	50.3

Table 4.4 (**Hollywood2: Human Actions and Scenes Dataset**) Recognition accuracy for individual action classes. Units are in %. The best score for each class is highlighted by bold fonts. The numbers by Laptev et al. and by Bilen et al. were extracted from (Bilen et al., 2011).

Even a human could fail the classification task when a subject was far from a camera position. The HBRT has proved to be highly effective, particularly when processing subtle actions such as ‘SitUp’ and ‘SitDown’. It is interesting to note that, according to Table 4.4, the HBRT result was improved by the HBRT/VOC; the latter extended the region of interest to accommodate non-human objects such as, car, dining table and chair. The significant improvement was observed with classes such as, ‘Eat’, ‘GetOutCar’ and ‘DriveCar’, indicating that the HBRT and VOC were complementing each other, especially when a human interacted with other objects in the video scene (*e.g.*, a human and a car, a human and a dining table).

4.4.3 Discussion

The local region tracking schemes performed better than the point feature-based techniques for relatively small and well studied datasets such as the KTH and the UCF Sports Actions. Figure 4.8 presents several examples for action localisation and segmentation with these two datasets. Interestingly, the accuracy by the HBRT/VOC improved over the HBRT even for datasets such as the KTH that did not contain any non-human objects. This was probably due to the person detector module of VOC, which contributed to successful human detection.

The local region tracking scheme showed its clear advantage when processing the complex and large dataset of Hollywood2, although the contribution of region tracking varied among action classes. It can be observed in Table 4.4 that, for ‘FightPerson’ and ‘Run’ actions, the space-time interest point features (Laptev et al., 2008) performed better than the region-based approaches. This was because the point features were able to provide more compact and abstract representation of video signals than the HBRT or the HBRT/VOC that relied on motion segmentation. The interest point features were useful when it was difficult to spatially localise the action using the region-based approaches.

The region tracking schemes clearly showed state-of-the-art performance with the Hollywood2 dataset, in particular when the action could be fully identified using mainly human body regions – this included actions such as ‘Kiss’, ‘HandShake’ and ‘HugPerson’. It was demonstrated that further improvement could be made by incorporated a trained model of visual object classes recognition (VOC) regions. The ‘DriveCar’ class from the Hollywood2 Dataset presented one such example, in which the regions of interest could be presented by multiple objects (*e.g.* , a human and a car).

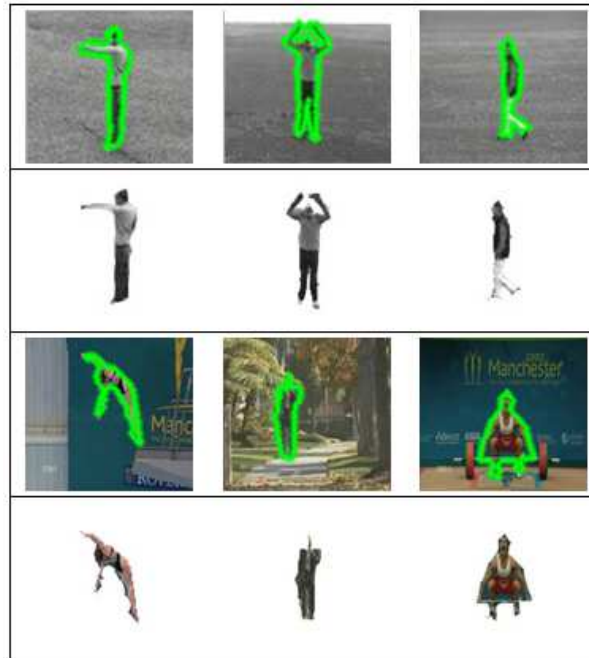


Fig. 4.8 Samples for action localisation and segmentation. The 1st and 2nd rows respectively present the results of key segments and the corresponding segmentation using the HBRT/VOC on the KTH Dataset ('boxing', 'hand waving' and 'walking' actions). The 3rd and 4th rows show the results on the UCF Sports Actions Dataset ('diving', 'walking' and 'lifting' actions).

4.5 Conclusion

The present chapter has put forward the human body region-based approach associated with extended spatio-temporal LLC to action classification task. The approach was further extended to accommodate non-human objects, resulting in the HBRT/VOC scheme. We showed that description of a human body volume using a spatio-temporal descriptor (HOG/HOF) and extended LLC coding scheme generated stable representation for the appearance and motion patterns underpinning comprehension of the actions carried out. Three widely used datasets were processed for evaluation and the region-based approach was able to outperform the recent state-of-the-art, point feature-based techniques with all three datasets. The resulting action class will be rendered as verbs later for description generation framework. Furthermore, the list of extracted semantic HLFs from the video clip has been extended to accommodate gender, age, emotion and scene setting in order to use them in later stage to generate natural language description of human activities.

Chapter 5

Extraction of Qualitative Spatial and Temporal Relations

Spatial and temporal relations of prominent objects play a vital role in describing video semantic content. This chapter utilises the extracted object segments from the previous chapters and formalises their spatial and temporal relations in order to be able to use them later for the generation of natural language description for a human activities in video clips. In this context spatial relations specify how objects are related to each other within a sampled frame, while temporal characteristics are used to describe the changes of spatial relations between two objects over the time domain.

To this extent, an improvement over the AngledCORE-9 approach introduced by [Sokeh et al. \(2013\)](#) is proposed – a comprehensive representation to efficiently extract spatial information between interacting objects present in a video clip using their approximate oriented bounding box (OBB). Spatial information is important for the identification of relations between multiple objects; hence the work is a step forward for tasks such as semantic content analysis and visual information retrieval. To that end we propose an approach that incorporates the spatio-temporal volume of objects into AngledCORE-9, and extends the extracted relations to accommodate the temporal information. As a result, the proposed approach is able to represent interacting objects in a video stream in an efficient manner, as accurate spatial and temporal information can be obtained by precise representation of the shape region and the OBB. A human action classification task is adopted to assess the efficiency of the proposed approach. The experiment carried out on two challenging datasets indicates that this approach outperforms the existing methods. Finally, we propose a set of rules to facilitate the mapping between qualitative relations and natural language terms, as the ultimate goal of this thesis is to improve visual identification of spatial and temporal relations that leads to the generation of a natural language description of video semantics.

The chapter is structured as follows. Section 5.1 introduces the qualitative spatio-temporal relations framework, along with the motivation for this work and its contributions. Section 5.2 reviews previous work related to qualitative spatio-temporal relations extraction and representation. The implementation of the proposed spatial and temporal relationships extraction framework between interacted objects in video streams is presented in Section 5.3, while Section 5.4 presents a summary of the results obtained from the evaluation experiment and a comparison with the recent implemented methods. A set of rules to facilitate the mapping process between extracted spatial and temporal relations and natural language terms is introduced in Section 5.5. Finally, Section 5.6 provides a concluding discussion.

5.1 Introduction

Overwhelming quantities of video data are available on the web and are being recorded on a regular basis using camera phones, surveillance cameras and other forms of video recording. Semantic analysis of these videos is essential for a diverse field of applications. In order for the information in these videos to be utilised in an effective fashion, it needs to be detected, organised and stored systematically for future use and easy access. The area in question in the current case, namely, qualitative spatial and temporal reasoning (QSTR) is no longer an emerging one; as noted in [Cohn et al. \(2008\)](#), a number of calculi have been defined and there has been significant research undertaken to study their computational properties. Today these calculi are utilised in a wide range of areas, such as Geographic Information System (GIS), language research, robotics, as well as in-depth analysis and assessment of semantic information located on video, especially for the video classification task ([Sridhar et al., 2011, 2010c](#)).

This utilisation has found many admirers and advocates who espouse the view that there are advantages to be gained in factors such as noise and unimportant variations in occurrences of video events. Such information can be abstracted away using QSTR techniques. As a consequence the same kinds of events can be rendered approximately in the same manner. It is particularly challenging to analyse the information stored on video in an automatic fashion and it is absolutely critical that essential information can be extracted, compactly represented, and then classified. This facilitates the indexing, searching and retrieving tasks in the video processing field. Of particular interest is the characterisation of objects contained in video footage and, in particular, their spatial features and qualities, as well as the evolution of spatial relations between objects featured over time.

As is often observed in these cases, the temporal evolution of characteristics for objects and their interconnection are commonly tied to certain events and behaviours. As [Sridhar](#)

et al. (2010c) have observed, from these patterns certain norms and rules may be derived, which can enable us to better understand moves by potentially multiple objects and their inter-relations over time. For video clips, certain variables such as the positioning of a camera and angles of refraction, as well as the inherent characteristics of individual devices, may affect the perception and placing of objects observable to the viewer. The consequence is that it is not particularly useful to attempt a precise numerical representation of these data, such as their exact coordinates (Sokeh *et al.*, 2013).

Rather, a qualitative analysis is of more use, examining the interconnections of relevant objects. This is done via examination of the topology; their direction, their distance, changes and alterations which occur over time. A qualitative calculus is employed to represent each of these aspects. A useful and all-inclusive representation of spatial data was created by Cohn *et al.* (2012) that includes and collates all of the relevant information, rather than locating each individually; this system was called CORE-9. This comprehensive representation depends solely on the position of the minimal bounding boxes for the objects in video frames and so it is especially suited to the processing of video data and their representation. If objects in video frames are detected and tracked in a clear and efficient fashion, there is greater impact in various applications in the computer vision field. CORE-9 offers a useful tool for these tasks; the relevant and particular spatial facets and calculi are thus determined.

5.1.1 Motivations

The majority of studies in the QSTR field assume that the object tracks from image or video data already exist. However, the work in this chapter focuses on bridging the semantic gap between computer vision and knowledge reasoning fields. For video, the gap issue arises from the inherited complexity of image processing and adding time as an extra data dimension for the purpose of interpretation of movement using temporal information.

Furthermore, there is a major shortfall associated with CORE-9 (Cohn *et al.*, 2012) introduced early in this chapter; the use of the Axis-Aligned Bounding Box (AABB) is far from being sufficiently precise and this flies in the face of the otherwise all-inclusive and efficient CORE-9 process. This problem can be seen from the fact that rough approximation of objects' physical positions may cause an inadvertent overlap of their bounding boxes even for some cases where these objects are visually separated.

AngledCORE-9 (Sokeh *et al.*, 2013) was put forward as a mean of alleviating this issue; this model relies on the more compact approximation of the region shapes by drawing OBBs of various angles, using Principal Component Analysis (PCA). Refer to Figure 5.1 for an illustration of the difference between using CORE-9 representation with both the AABB and the OBB. In the work presented in this chapter we aim to enrich the representation

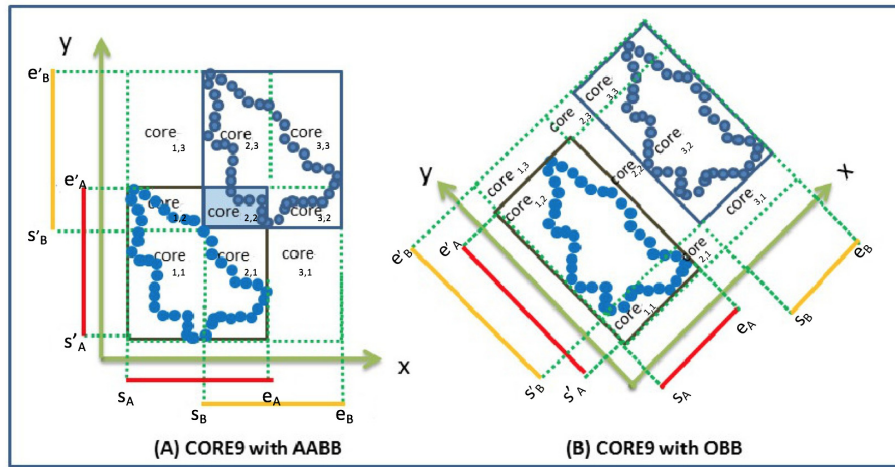


Fig. 5.1 The difference between the axis-aligned bounding box (AABB) and the orientated bounding box (OBB). Either can be used with CORE-9 representation.

of AngledCORE-9. Moreover, in order to obtain a meaningful set of spatial and temporal relations, a set of rules is proposed to map between these qualitative relations and natural language terms in order to utilise them later for the description generation process.

5.1.2 Qualitative Spatial and Temporal Relations: Overview

To efficiently and accurately represent the relationships between interacting objects present in a video stream, a series of modifications is applied to AngledCORE-9. Firstly an approximated region of OBB (Sokeh et al., 2013) is replaced with a space-time volume for objects by tracking and segmenting them from a video stream in an efficient manner. Then, for each extracted volume, a tight OBB is drawn using the adapted hybrid approach (Chang et al., 2011). Finally the compact CORE-9 representation is used to extract the spatial and temporal aspects from multiple, inter-related object bodies.

Once certain areas have been identified and isolated, the conventional core scheme for CORE-9 is employed to calculate the spatial relations. Compared to the commonly used representation CORE-9, the proposed object volume-based method has a higher chance of generating more reliable results regarding the direction of objects, topologies, size, distances and temporal changes.

5.1.3 Qualitative Spatial and Temporal Relations: Contributions

The contributions of the proposed work in this chapter can be summarised as follows:

- enriches the AngledCORE-9 representation by utilising spatio-temporal object tracks extracted from video clips;
- extends the set of extracted relations to accommodate temporal characteristics between spatial relations over a sequence of frames;
- proposes a set of rules to facilitate mapping resultant qualitative relations into corresponding natural language terms;
- applies the spatial and temporal relations extraction approach to the action classification task with the Mind’s Eye video dataset and TV Human Interactions dataset, demonstrating that the approach is able to efficiently discriminate the actions based on similarity of change patterns.

5.2 Related Work

The QSTR provide a prospective avenue to bridge the gap between activities at low-level features and high-level descriptions. Qualitative primitives configure quantitative measures into more clear-cut classes, thus delineating and defining qualitatively stimulating ideas from quantitative measures (Cohn et al., 2002). Qualitative spatial and temporal depictions are developed to characterise and justify the two main essential facets of knowledge, more specifically that of time and space. Qualitative temporal reasoning is a fundamental constituent of knowledge and is mirrored in natural language, for example in instances where particular events occur during or before one another. Allen’s Interval Algebra (Allen and Ferguson, 1994) was established as a basis for temporal reasoning; this calculus outlines 13 probable base relationships among intervals on a fixed line. Allen’s temporal algebra delivers a table containing configurations that can be utilised for describing temporal relations of activities, see Figure 5.2.

Numerous qualitative spatial calculi were generated from the interval algebra for the purposes of reasoning about objects which are associated to one another in space. The three key features of spatial relationships are topology (*e.g.* touch, inside), direction (*e.g.* left of, above) and distance (*e.g.* near, far). These associations are applied in natural language to define facets around us, in space and in a qualitative manner.

It is beneficial to use a group of qualitative binary base relations, which have the inherent property of being comprehensive and disjointed in pairs, if we are to apply logical reasoning to relationships in space. For example, between two spatial objects, exactly one of the base relationships hold. The Region Connection Calculus or more precisely the RCC-8 (Randell et al., 1992) are the best-known calculi for topological relationships. The RCC-8 is based on the topology spatial theory, which provides an abstract spatial conformation of two








Relation	Illustration	Interpretation
$X < Y$ $Y > X$		X takes place before Y
$X m Y$ $Y mi X$		X meets Y (i stands for inverse)
$X o Y$ $Y oi X$		X overlaps with Y
$X s Y$ $Y si X$		X starts Y
$X d Y$ $Y di X$		X takes place during Y
$X f Y$ $Y fi X$		X finishes Y
$X = Y$		X is equal to Y

Fig. 5.2 The 13 Allen's temporal relations that exist between two intervals X and Y .

physical areas in relation to a group of mutually extensive pairwise disjoint (jepd) qualitative relationships. These have the ability to embrace the area in between these regions. Eight distinct topological relationships between two regions A and B are well demarcated based on the constituents they share, as shown in Figure 5.3.

Qualitative spatial relationships can be delineated manually or learned from the gathered data. [Fernyhough et al. \(2000\)](#) suggest that primitive spatial relations such as right, ahead, behind, which are typically seen in road traffic scenarios, are manually demarcated. Composite events such as following and pulling out, which are analogous to sequences of primitive actions, are considered to be learned from data. The approach proposed by [Galata et al. \(2002\)](#) is comparable to that of [Fernyhough et al. \(2000\)](#). However, this method automatically learns the qualitative spatial relationships from data and overtly computes the possibilities related with these events.

[Southey and Little \(2007\)](#) suggest that learned qualitative spatial relations between objects can be achieved through the use of an entropy model as well as proximity features (touching,

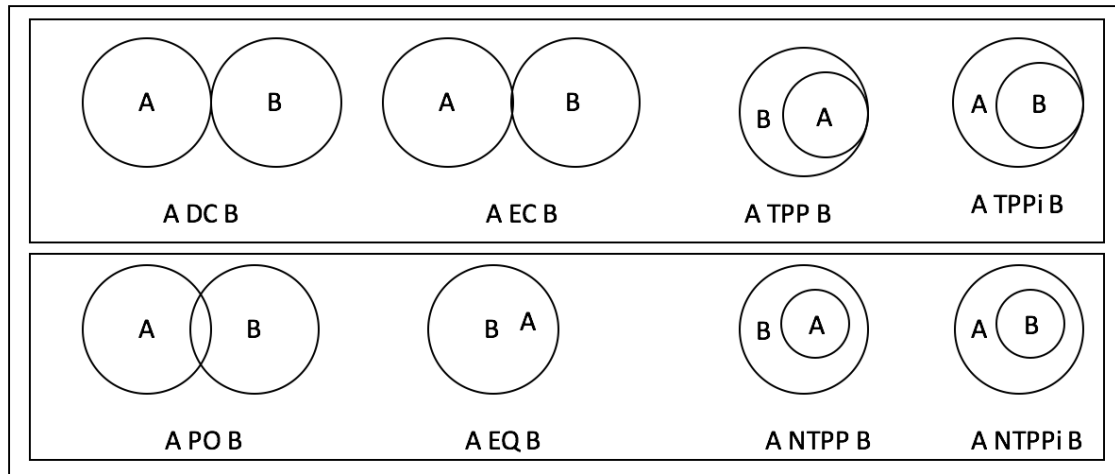


Fig. 5.3 Graphical representation of the Region Connection Calculus (RCC8), where A and B are two objects, and one of the following eight topology relations might occur between them: disconnected (DC), externally connected (EC), tangential proper part (TPP), tangential proper part inverse (TPPi), partially overlapping (PO), equal (EQ), non-tangential proper part (NTPP) or non-tangential proper part inverse (NTPPi).

near, mid, far), so that interactions between such objects can be modelled. One of their examples is that of an unidentified object in the middle, in addition to a fork to the left and a knife to the right of it. It is hypothesised that a model of qualitative spatial relationships between objects in a given situation can assist in the provision of more appropriate cues, to distinguish objects such as a plate, which is surrounded by other objects like the fork and knife. They determine its application for recognising objects based on learned contextual cues.

Recently, there is growing significance of methods that involve knowledge representation techniques in the field of video analysis and understanding. Earlier approaches, such as [Sridhar et al. \(2008, 2010a\)](#), created a qualitative spatio-temporal graph to portray activities observed in visual data. They incorporated spatial relations centered around spatial calculus ([Cohn et al., 2008](#); [Randell et al., 1992](#)) by region connection calculus (RCC), in combination with temporal relations built on Allen's interval algebra ([Allen, 1983](#)). Spatio-temporal relationships were also described and used by [Sridhar et al. \(2010b\)](#) as a relational graph, which can define video activities and cluster similar activities together. [Morariu and Davis \(2011\)](#) utilised logic to decipher knowledge pertaining to spatio-temporal structures and automatically analyse video and detect activities.

The RCC spatial calculus, in conjunction with Allen's Interval Algebra, has been employed in [Dubba et al. \(2010\)](#). Their work was based on pre-defined knowledge of the object

categories along with the spatio-temporal features. Using inductive logic programming (ILP) it was able to successfully learn and identify human activities, with the major advantage of preventing an over-fitting problem against the training dataset. However, its dependence on pre-defined knowledge of the object categories, together with its strict classification methodology because of ILP, meant that performance problems could arise in their absence. A hierarchical approach was built with variable length Markov models using the contours of the human body as observations that acted as control points. It was tested with the task for determining exercise activities which did not involve object interactions (Galata et al., 1999, 2001). It is not clear whether the approach can handle activities involving object interactions, as it relies on the contour of single object present in video frame.

A number of methodologies make use of interest point detectors (Dollár et al., 2005; Xia and Aggarwal, 2013). A 3D cuboid was extracted and a descriptor was calculated at every interest point; descriptors sharing similarities were moved together. This created a feature descriptor codebook, which was not unlike the commonly used bag-of-words approach. A probabilistic approach (Zhang and Parker, 2011) was built on this method; it used domain knowledge to model every action as a distribution over the codewords, and every video as a distribution over the activities. The benefit of these methodologies was that image descriptors did not require skeleton or object tracks to define the activity. However, they could not consider spatio-temporal connections among the various relevant entities in the scene, which were crucial factors when learning and identifying human activities (Cohn et al., 2008; Forbus, 2008). To counter this problem Zhang et al. (2012) proposed ‘spatio-temporal phrases’, mixtures of local words in a specific spatial and temporal structure, involving their order and relative positions. This method was in line with the idea of the graphical representation defined earlier in Sridhar et al. (2008, 2010a). However, the spatio-temporal phrase was unable to involve qualitative spatial relation; additionally, the temporal relations are less versatile than the Allen’s Interval Algebra employed in the graphical method.

More recently, Tavanai et al. (2013) suggested a domain-based method that identifies objects by modelling the person-carried object relationship that is distinctive of the carry event. To distinguish a standard class of carried objects, they suggested the use of geometric shape models, rather than using pre-trained object class models or solely relying on protrusions. To help track the carried objects, an optimisation process is used, which unites the spatio-temporal consistency, distinctive of the carry event. Conventional features, such as appearance and motion smoothness, are also considered. The hypothesised method considerably outperforms a recent state-of-the-art method on two challenging datasets – the PETS 2006 and Mind’s Eye video dataset. However, this approach was domain-specific and focus on particular set of events.

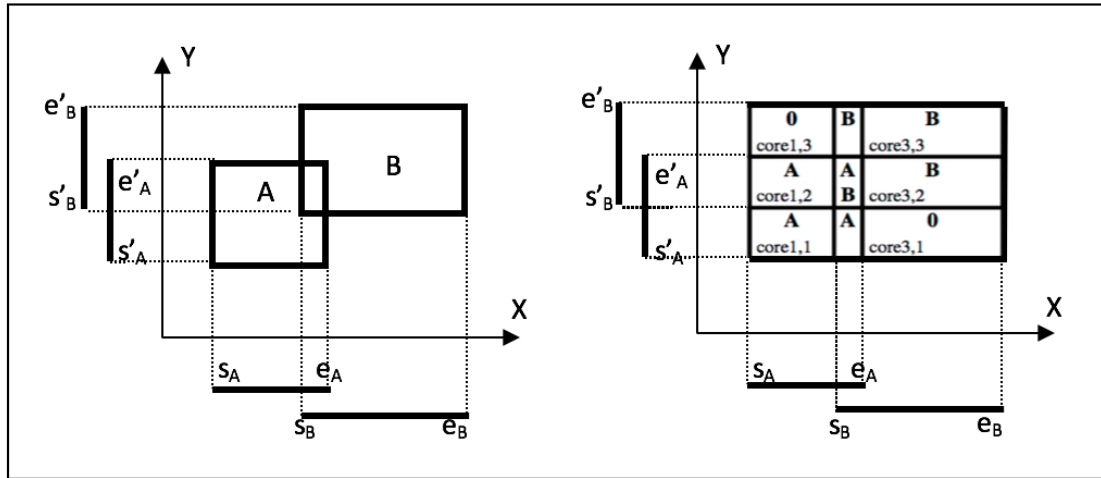


Fig. 5.4 Two objects *A* and *B* and their projections on the left, while the right shows how their projection identify the associated 9 cores (Cohn et al., 2012).

Cohn et al. (2012) propose a comprehensive representation technique named CORE-9 which is able to represent the spatial relations between interacted objects over a video stream; see Figure 5.4. However, Sokeh et al. (2013) have proposed a more accurate version of CORE-9, which they referred to as AngledCORE-9. In this approach the OBB is employed instead of AABBs, as they are more accurate for the purpose of drawing a tight bounding box around a targeted object. The direction of OBB is estimated by applying PCA. This technique is able to successfully extract and store spatial relations to use them for video classification task.

5.3 Individual Aspects of Qualitative Spatial and Temporal Relations

Accuracy of vision processing at its lower level, such as the detection of objects and tracking of their moves, impacts significantly upon the higher level of reasoning; one such example is a study in semantic analysis of visual scenes, incorporating steps such as localisation, identification, understanding and reasoning. The first step is to represent video as a collection of spatio-temporal volumes, from which inter-relation of the relevant objects may be derived along the space and time axes. A set of methods are employed (refer to Chapters 3 and 4 for more detail), one to identify human bodies and the other to extract non-human regions and as a result a set of tracks T is generated. The following stages encode the qualitative spatio-temporal relations between the resulting tracks present in video clip.

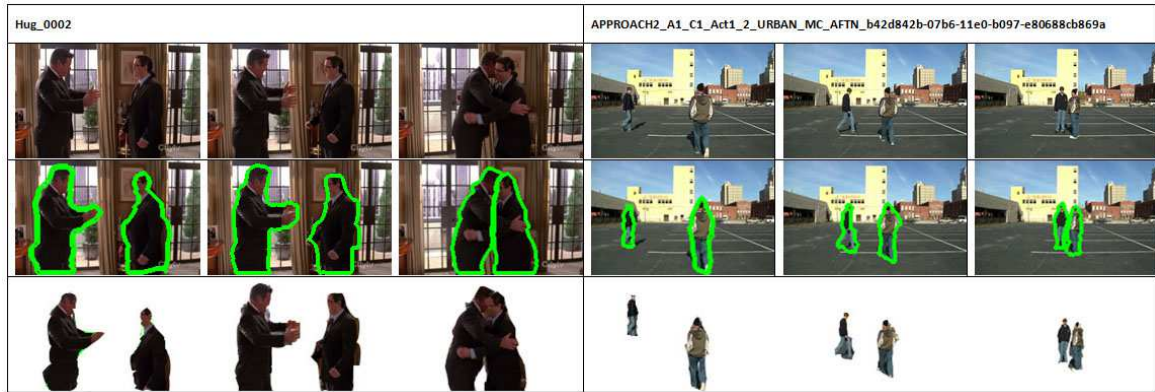


Fig. 5.5 Sample segmentation of a human body volume for a ‘hug’ action from the TV Human Interaction dataset and an ‘approach’ action from the Mind’s Eye video dataset. The first row shows original frames from two video clips. The second and the third rows respectively present the results of segmentation and the corresponding key segments. These datasets are to be introduced in Section 5.4.

In CORE-9 (Cohn et al., 2012), AABBs are employed when extracting spatial information; this is achieved by drawing tight bounding boxes around regions visually occupied by an object. One major benefit of using AABBs is their simplicity, as they are aligned with the axes of the coordinate system for video frames. For each individual pair of objects, say A and B , the CORE-9 scheme creates the relevant AABBs, a and b , and the four corners of each bounding box are given by their coordinates. Using corner positions we draw a 3×3 grid around two objects, where each box area segmented by grid lines is referred to as a core. Moreover, the six intervals are set by the distance between these corners coordinates over the two axes.

There are eight corners for two objects (*e.g.*, four corners of a bounding box for each object), which are sufficient for extracting the qualitative spatial relations as well as their temporal changes. Two objects can be either visually disconnected or overlapped; depending on their relative positions, each of nine cores may belong to either zero, one or two objects. Nine cores and six intervals are identified by corner positions, defined by coordinates, four coordinates on the x -axis and four on the y -axis. By measuring nine cores, spatial information on two objects can be obtained such as the topology, the direction, the relative size and the distance between them. Temporal changes can also be extracted from these cores and their associated interval by processing a sequence of video frames.

Sokeh et al. (2013) have developed a more accurate version of CORE-9, which they referred to as AngledCORE-9. It employs OBBs instead of AABBs, so as to draw a more accurate box around the object. It may be seen that an OBB is the general case with arbitrary rotation of an AABB. The direction can be estimated using PCA. Although providing an

accurate representation, it requires more complex calculations when testing for overlap or disconnection of objects. Another shortfall is that PCA cannot locate possible hyperplanes in such cases as when the eigenvalues of covariance matrices are not distinguishable from one another, and there is little symmetry in the shape. There are many such cases in real-life videos, where the AngledCORE-9 scheme does not improve very much over CORE-9. That said, CORE-9 is a fast and relatively simple algorithm, which is advantageous in various applications.

To better estimate the OBB for a group of coordinates, a hybrid approach¹ (Chang et al., 2011) has been employed. With the hybrid approach, a beneficial amount of convergence is found while maintaining the satisfactory scope for the search space in order to evade local minima. Object segments are identified and then inscribed in a close-fitting OBB via utilisation of the hybrid process. Once certain areas have been identified and segmented, the conventional core scheme for CORE-9 is employed to calculate the spatial relations. A combination of the commonly used CORE-9 and the extracted object tracks has a higher chance of generating more reliable results regarding the direction of objects, topologies, size, distances and temporal changes involved. There are a number of features for consideration with QSTR:

Spatial relations show the characteristics and relations between multiple objects which are present in space. Spatial features include many aspects, such as poses of objects, relative distance of objects compared to other objects, and absolute or relative direction of motion.

Temporal relations are characteristics and relations of the objects or activities themselves along the time axis. Examples of temporal relations between two events can be defined by Allen's interval algebra, the time an activity begins and how long it runs for.

Qualitative features are described as properties 'relating to, measuring, or measured by the quality of an item instead of its quantity'. A qualitative spatial feature can be described as where two objects overlap to a small extent without defining how much overlap there is. A qualitative temporal feature is where an activity begins and ends prior to another activity. RCC and Allen's Interval Algebra are both qualitative relational techniques.

For each type of relation, the calculations take place at each frame, $F = \{f_1, f_2, \dots, f_n\}$, where $T = \{t_1, t_2, \dots, t_m\}$ object tracks are present and the computations take place between pair of tracks $t_i, t_j \in T$, ($i \neq j$). This produces an $r \times n$ matrix for each type of relation, where r represent the number of all possible pairwise combinations of the tracks present at n number of frames. Finally, the resultant matrix is optimised by deleting redundant columns, where

¹Source code is available from: perso.uclouvain.be/chia-tche.chang/code.php

the spatial relations stay stable over subsequent frames. In this study we adopt a symbolic description of a spatial relations constructed from 12 predicates: 3 topology, 4 directional, 3 distance and 2 size. These sets of 12 predicates are chosen as they tend to be the most common spatial relations in linguistic, as well as being suggested as sufficient to describe the objects (Freeman, 1975). A more in-depth explanation of individual aspects of qualitative spatial and temporal relations identified among interacting objects' tracks is presented below.

5.3.1 Topology

Topology, a field of mathematics which comes under geometry, contributes a coarsely granular approach to encoding an object region structure by looking at connectedness. 'Connectedness' is commonly defined as the property of two regions being in contact with each other. While the qualitative spatial reasoning (QSR) uses certain aspects of mathematical topology, both point-set and algebraic topology are too ambiguous to be used in practical situations with QSR, since they concentrate mostly on representation and do not involve the 'common-sense' aspects of human spatial reasoning (Cohn and Hazarika, 2001).

Topology in QSR is employed as the background for defining the overlap of pairs of regions A and B . Cognitive studies show that topological relationships are crucial for spatial cognition, and have been identified as the key basis for certain tasks such as classification. Topology is noteworthy as it is a very simple depiction of space, while being able to denote important spatial distinctions.

The RCC-8 is adopted (Randell et al., 1992) for calculating the intersection of bounding boxes. The spatial logic of RCC-8 individually processes the visually occupied area for objects. It is able to identify their spatial relations, for example the existence of interconnected (*e.g.* overlapped) regions. The use of RCC-8 is not as complicated as it might appear because the fine granularity, which RCC-8 can offer, is not needed for this task; additionally the number of RCC-8 relations can be reduced without much loss of essential spatial information. As a consequence, the following relations from RCC-8:

externally connected (EC),
partially overlapping (PO),
tangential proper part (TPP),
equal (EQ),
tangential proper part inverse (TPPi)

may simply be reduced to 'TOUCH' (if one object comes into contact with another object from any side, or if there is an overlapping between them), while

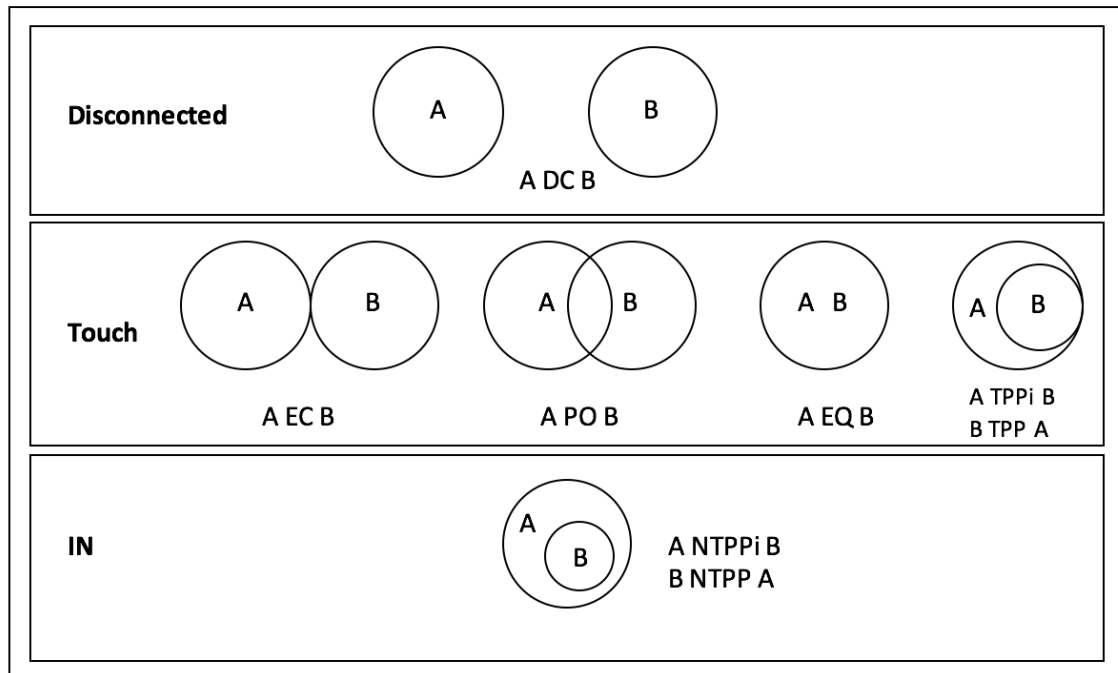


Fig. 5.6 Abstracting RCC-8 to RCC-3.

non-tangential proper part (NTPP),
non-tangential proper part inverse (NTPPi)

can be brought together, creating a so-called ‘INSIDE’ relation (if one object is completely inside the other). There also exist

disconnected (DC)

cases, resulting in a set of three basic relations as shown in Figure 5.6, ‘DC’, ‘TOUCH’ and ‘INSIDE’, that should be considered.

In practice, given two reference axes x and y and two bounding boxes of objects A and B , these rectangles will be projected to x and y given two intervals s_A, e_A and s_B, e_B on the x axis and two intervals s'_A, e'_A and s'_B, e'_B on the y axis as shown earlier in Figure 5.1, where s and e stand for start point and end point of each object on the x axis and s' and e' represent the same on the y axis. The region of interest for objects A and B is bounded by these intervals and formulate nine cores, written as $core_{i,j}(A,B)$ for $1 \leq i, j \leq 3$. The topology $state_{i,j}(A,B)$ in two-dimensional space can take one of the following values:

- $core_{i,j}(A,B) = AB, \iff core_{i,j}(A,B) \subseteq A \cap B$
- $core_{i,j}(A,B) = A, \iff core_{i,j}(A,B) \subseteq A \wedge core_{i,j}(A,B) \not\subseteq B$
- $core_{i,j}(A,B) = B, \iff core_{i,j}(A,B) \subseteq B \wedge core_{i,j}(A,B) \not\subseteq A$

- $core_{i,j}(A,B) = \emptyset, \iff core_{i,j}(A,B) \not\subseteq A \wedge core_{i,j}(A,B) \not\subseteq B$

Using the notion of CORE-9 representation, the topological relation between two objects A and B can be represented as a 9-tuple as follows:

$$Topology(A,B) = [core_{1,1}(A,B), core_{1,2}(A,B), core_{1,3}(A,B), core_{2,1}(A,B), core_{2,2}(A,B), core_{2,3}(A,B), core_{3,1}(A,B), core_{3,2}(A,B), core_{3,3}(A,B)]$$

Consequently, the three main relations in our task, which can be referred to as RCC-3, based on CORE-9 representation can be defined as follows:

- ‘DISCONNECT’ exists between objects A and $B \iff \nexists core_{i,j}(A,B) = AB \subseteq Topology(A,B)$
- ‘TOUCH’ exists between objects A and $B \iff \exists core_{i,j}(A,B) = AB \subseteq Topology(A,B)$
- ‘INSIDE’ exists between objects A and $B \iff \exists core_{i,j}(A,B) = AB \subseteq Topology(A,B) \wedge \nexists core_{i,j}(A,B) = \emptyset \subseteq Topology(A,B)$

5.3.2 Size

When examining the relations and interactions between objects in a video frame, their sizes are of particular importance. For example, suppose that some variation or alteration of the object size is observed over time. It can be estimated that the object is either approaching the camera or going further away from it. Size can be better evaluated based on the OBB, associated with the spatio-temporal volumes of the objects by evaluating the set of cores, which constitute the object, at a specific time point against the same set of cores at the following time point. By comparing the relative object sizes in two adjacent frames, given their bounding boxes and associated CORE-9 representation, we can discover if the size is getting larger or smaller. In this study, the size relation can be assigned to one of the following values, *small*, *same* or *big*.

For the purpose of video analysis, the exact values of the width, height, and area of cores is generally not important. Instead, we are mainly interested in relative size measures either for single or pairs of objects. By comparing the size of one core with another, we can infer information such as relative closeness of rectangles. By comparing how the size of a core at one time point compares to its size at the next time point we can infer how objects move relative to each other.

In practice, the original CORE-9 representation estimates the size relation of nine cores by keeping track of $9 \times 8/2 = 36$ different size relationships. However, in our task we simplified this calculation by comparing two sets of cores which basically belong to objects A and B . As a result, one comparison is performed for each pairwise object track. The ‘small’ relationship is assigned when object A is smaller than B , ‘big’ relationship is assigned when object A is bigger than B and otherwise ‘same’ relationship is assigned.

The identification of this attribute based on image statistics. For each object class, a mean object size $\bar{a}_{objectclass}$ is calculated by averaging the detected-box areas over the number of all tracks belong to that object class. An optional size adjective for a track is generated by comparing the detected-box area a of object track A to $\bar{a}_{objectclass}$:

- $Size(A) = BIG \iff a \gg \bar{a}_{objectclass}$
- $Size(A) = SMALL \iff a \ll \bar{a}_{objectclass}$

5.3.3 Direction

A direction calculus is one of the basic components of qualitative reasoning calculus. It describes relative positions of objects in large-scale spaces. The direction of the object is a spatially significant feature. It is a binary relation where the direction between target objects is measured relative to a reference object; for example, ‘object A is to the left of object B’. Standard CORE-9 uses a global direction fixed by the video frame. The Cardinal Direction Calculus (CDC) (Ligozat, 1998) is employed to calculate the direction.

In this study, CORE-9 representation is used to extract the direction relations between two objects A and B by analysing interval information. Using the notion of intervals as shown in Figure 5.7, the value of $dir(A, B)$ can be assigned to one of the following five states:

- $Direction(A, B) = SAME \iff s_B \leq s_A \wedge e_A \leq e_B \wedge s'_B \leq s'_A \wedge e'_A \leq e'_B$
- $Direction(A, B) = UP \iff e_A \leq e_B \wedge e'_B \leq s'_A \wedge s_B \leq s_A$
- $Direction(A, B) = DOWN \iff s_B \leq s_A \wedge e_A \leq e_B \wedge e'_A \leq s'_B$
- $Direction(A, B) = RIGHT \iff e_B \leq s_A \wedge s'_B \leq s'_A \wedge e'_A \leq e'_B$
- $Direction(A, B) = LEFT \iff e_A \leq s_B \wedge s'_B \leq s'_A \wedge e'_A \leq e'_B$

5.3.4 Distance

For temporal change of an object, or temporal development of a certain event, the position of the object in the video frame or the position relative to the other object may be traced over a span of time. To this end qualitative trajectory calculus (QTC) (Van de Weghe et al., 2005) is employed for primitives. In our system the distance between two objects A and B is calculated by evaluating $|e_A - s_B|$ value. Based on this value, the QTC relation consists of three basic relations *approach* (two objects moving towards each other), *depart* (two objects moving away from each other) or *static* (represents an object at rest). The following illustrates distance relations between objects A and B at subsequent frames t and $t + 1$:

- $Distance(A, B) = APPROACH \iff |e_A - s_B|_t > |e_A - s_B|_{t+1}$
- $Distance(A, B) = DEPART \iff |e_A - s_B|_t < |e_A - s_B|_{t+1}$
- $Distance(A, B) = STATIC \iff |e_A - s_B|_t = |e_A - s_B|_{t+1}$

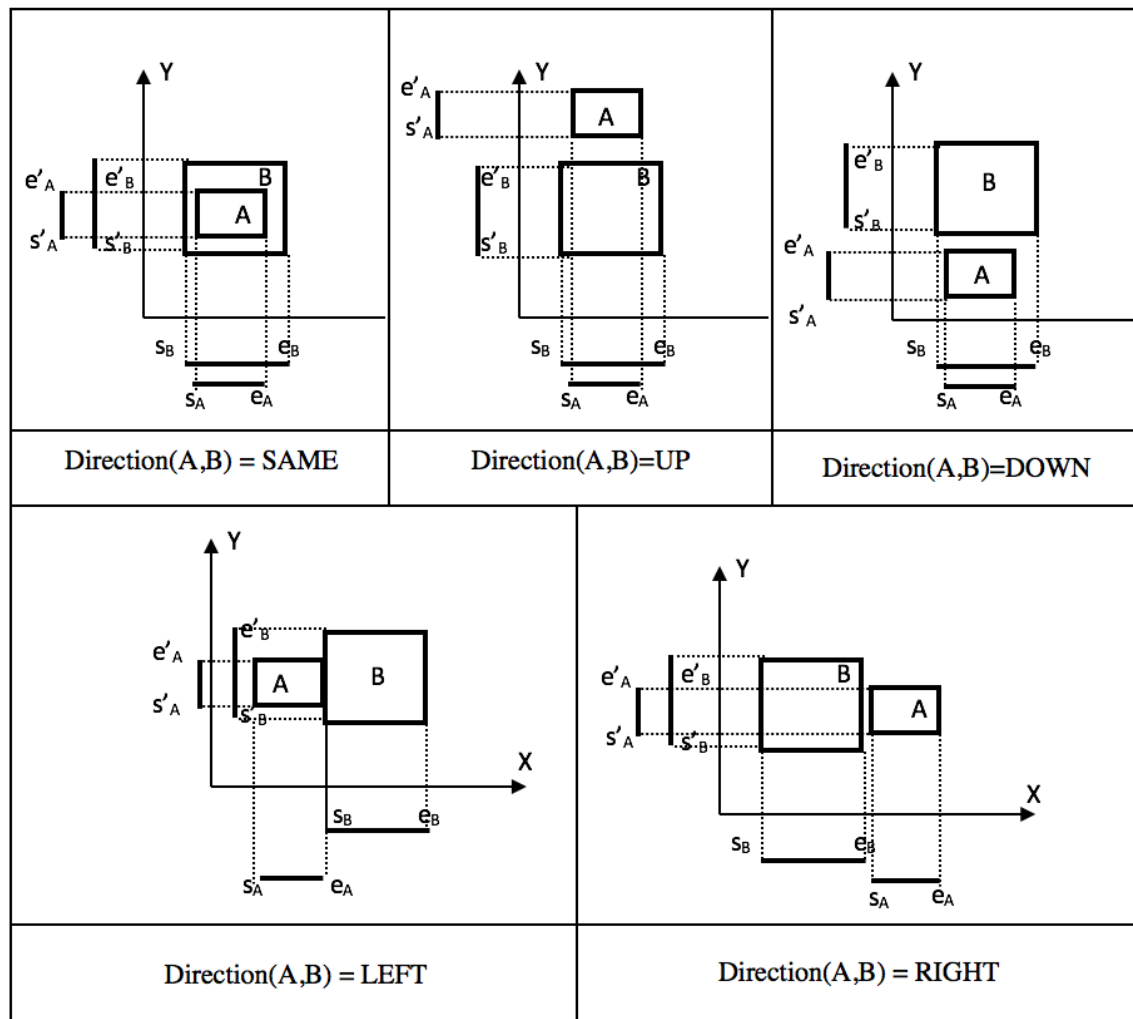


Fig. 5.7 Different types of direction relations between two objects *A* and *B*.

QTC relations can be of particular use in the visual demonstration of common activities, especially when topological relations such as RCC-3 cannot fully account for some of interconnections present in a video frame. It can be illustrated by the following examples: the topological relation is 'DC' if one person is chasing another person while maintaining the distance gap between them. In this case the relation 'DC' for a pair of moving objects is exist over all frame sequence while there is crucial changes in their spatial relations over time.

5.3.5 Temporal Relations

By themselves, spatial changes of bounding boxes and the associated CORE-9 do not show the notion of time order in an activity which, in certain instances, could be crucial. Capturing



Fig. 5.8 Sample frames from the two datasets used for the experiments: the TV Human Interaction dataset (first row), and the Mind's Eye video dataset (second row).

the time dependencies of spatial changes of performed actions can be helpful in semantic understanding of such scenes. A widely used approach describing temporal relations is Allen's Interval Algebra (Allen, 1983), which consists of 13 bases capturing the possible relations between two intervals (Figure 5.2).

To calculate the temporal order of two event A and B with their time intervals $[X_1, Y_1]$ and $[X_2, Y_2]$ respectively over sequence of frames, the temporal relations can be estimated as follows:

- $Temporal(A, B) = Equal \iff (X_1 = X_2 \wedge Y_1 = Y_2)$
- $Temporal(A, B) = Finishes \iff (X_1 \neq X_2 \wedge Y_1 = Y_2)$
- $Temporal(A, B) = Starts \iff (X_1 = X_2 \wedge Y_1 \neq Y_2)$
- $Temporal(A, B) = During \iff (X_1 > X_2 \wedge Y_1 < Y_2) \vee (X_1 < X_2 \wedge Y_1 > Y_2)$
- $Temporal(A, B) = Overlap \iff (X_1 < X_2 \wedge Y_1 < Y_2 \wedge Y_1 > X_2) \vee (X_2 < X_1 \wedge Y_2 < Y_1 \wedge Y_2 > X_1)$
- $Temporal(A, B) = Meets \iff (X_1 < X_2 \wedge Y_2 > Y_1 \wedge Y_1 = X_2) \vee (X_2 < X_1 \wedge Y_1 > Y_2 \wedge Y_2 = X_1)$
- $Temporal(A, B) = Before \iff (X_1 < X_2 \wedge Y_1 < X_2) \vee (X_2 < X_1 \wedge Y_2 < X_1)$

5.4 Experiments

We carried out two sets of experiments, the temporal change identification task and the human action classification task. The purpose was to make comparison between the extended CORE-9 with spatio-temporal volumes and hybrid OBBs, proposed in this chapter, and AngledCORE-9 with approximated regions and PCA (Sokeh et al., 2013) and prove the efficiency of extracted spatial and temporal relations.

5.4.1 Video Data

Two datasets were used in the experiments. One was the TV Human Interactions dataset,² which is a collection of real-life videos with complex settings. There are 300 video clips, drawn from more than twenty programmes screened on television. They have four different human interaction classes, consisting of ‘*hand shakes*’, ‘*high fives*’, ‘*hugs*’ and ‘*kisses*’, each containing 50 videos. There are 100 clips that did not feature any of the above four human interactions, forming a fifth action class, ‘*negative*’. This dataset was used for the human action classification task.

The second dataset was the same as the one used by Sokeh et al. (2013). It holds 50 clips selected from the Mind’s Eye video dataset,³ consisting of five action classes: ‘*approach*’, ‘*collide*’, ‘*drop*’, ‘*carry*’ and ‘*catch*’, each containing ten videos. This dataset was used for the temporal change identification task and the human action classification task. Sample frames of both datasets are presented in Figure 5.8.

5.4.2 Feature Selection and Clustering

Spatial information between objects in the videos was extracted and used as input features for Latent Dirichlet Allocation (LDA),⁴ an unsupervised clustering method able to cluster collections of discrete data efficiently. This learning algorithm was previously used for grouping words into topics from a block of text. More recently it has been employed in the computer vision field, for example in action classification (Sokeh et al., 2013).

Every video was seen as a document and the entire dataset as a corpus. Video frames showing two objects interacting with each other were seen as a single word in that document ‘video’. In this study it was our habitual practice to employ five features relating to space and time: topology, size, direction, distance and temporal relations. In frame f of video v where objects A and B interacted, we would have five-dimensional words, with each component representing these spatial and temporal relations individually:

$$W_{vf} = \{\theta(A, B), \alpha(A, B), \beta(A, B), \delta(A, B), \gamma(A, B)\}$$

where θ , α , β , δ , γ correspond to topology, size, direction, distance and temporal relations respectively. The k -means clustering algorithm was employed in order to discretise each word with five-dimensional values. Each video was associated with a latent variable z_k , which represents its cluster: the aim is to cluster similar activities together.

² Available from: www.robots.ox.ac.uk/~vgg/data/tv_human_interactions/

³ Available from: www.visint.org/datasets

⁴ LDA code is available from <https://github.com/kyamagu/lda-matlab>

	approach	carry	catch	collide	drop
approach	8	1	0	1	0
carry	0	9	0	0	1
catch	1	0	8	1	0
collide	1	0	2	7	0
drop	0	1	0	0	9

Table 5.1 Confusion matrix for the action classification task with the Mind’s Eye video dataset achieved by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.

	approach	carry	catch	collide	drop
approach	6	0	0	2	2
carry	0	8	1	1	0
catch	1	1	6	2	0
collide	3	1	1	5	0
drop	0	2	1	0	7

Table 5.2 Confusion matrix for the action classification task with the Mind’s Eye video dataset achieved by AngledCORE-9 (Sokeh et al., 2013).

5.4.3 Temporal Change Identification

This task was concerned with the identification of temporal change between two objects over video frames. More specifically, 20 video clips were selected from the Mind’s Eye video dataset and, for each video, the ground-truth was annotated with one of the following three relations:

- two objects moving toward each other (10 videos);
- two objects moving away from each other (3 videos);
- one object moving while the other at rest (7 videos).

Our approach (CORE-9 with spatio-temporal volumes plus hybrid OBBs) was able to identify 76% of relations successfully, which was a good improvement over 52% achieved by AngledCORE-9 (approximated regions with PCA OBBs). We observed that the weaker result from the latter was caused because PCA often failed to detect the correct direction and approximate the region shape. Moreover, extraction of spatio-temporal objects volumes improved the accuracy of video representation in terms of precise shape of interacting objects.

action	precision	recall	F1
approach	0.80	0.80	0.80
carry	0.90	0.82	0.86
catch	0.80	0.80	0.80
collide	0.70	0.78	0.74
drop	0.90	0.90	0.90

Table 5.3 Precision and recall scores for the human action classification task using the Mind’s Eye video dataset by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.

action	precision	recall	F1
approach	0.60	0.60	0.60
carry	0.80	0.67	0.73
catch	0.60	0.67	0.63
collide	0.50	0.50	0.50
drop	0.70	0.78	0.74

Table 5.4 Precision and recall scores for the human action classification task using the Mind’s Eye video dataset by AngledCORE-9 (Sokeh et al., 2013).

5.4.4 Human Action Classification

To compare the effectiveness of CORE-9 with spatio-temporal volumes and hybrid OBBs against AngledCORE-9 with approximated regions, experiments were undertaken with the human action classification task, follow the same setting in Sokeh et al. (2013).

Classification of the Mind’s Eye videos. For the classification of human activities, LDA was used to cluster the videos into five categories. The LDA parameters were set as $\alpha = 10$ for topic distribution (per document) and $\beta = 0.01$ for word distribution (per topic). Tables 5.1 (for extended CORE-9) and 5.2 (for AngledCORE-9) present the confusion matrices for the classification task using the Mind’s Eye video dataset. Their precision and recall scores, as well as the balanced F1-score, for each class are calculated in Tables 5.3 and 5.4. The figures indicate that the extension of CORE-9 with spatio-temporal volumes was able to perform better than AngledCORE-9 by a clear margin. This improvement resulted from more accurate extraction of shapes for regions of interest. Moreover, by incorporating Allen’s temporal relationships, the approach was able to capture the time dependencies between spatial changes. For example both ‘drop’ and ‘carry’ activities held the same topology, *DC* and *TOUCH*, in spatial relations over video frames. By capturing the order and duration of these spatial relations using Allen relations it was possible to efficiently differentiate between them.

action	precision	recall	F1
handShake	0.65	0.64	0.64
highFive	0.59	0.58	0.58
hug	0.73	0.72	0.72
kiss	0.77	0.69	0.73
negative	0.79	0.66	0.72

Table 5.5 Precision and recall scores for the human action classification task using the TV Human Interactions dataset by the extended CORE-9 with spatio-temporal volumes and hybrid OBBs.

Classification of the TV Human Interactions videos. Table 5.5 shows the precision and recall scores for classification of the TV Human Interactions data. As a comparison the average precision scores achieved by [Patron-Perez et al. \(2010\)](#), who developed this dataset, was in the range of 0.25 to 0.4 for automatic annotation when including a ‘negative’ class. Additionally they used a supervised classifier, while we made unsupervised, k -means based clustering. Direct comparison may not be readily available, but the extension of CORE-9 with spatio-temporal volumes and incorporation of Allen’s temporal relations presented clear and promising results for various types of human actions; it was particularly so when the action could be fully identified using mainly human body regions, such as ‘Kiss’ and ‘HugPerson’ actions.

5.5 Mapping from QSTR into Natural Language Terms

The primary objective of this chapter is to formulate the relations between interacting objects in video clips and then define a collection of rules that will facilitate the mapping process from qualitative spatial and temporal relations to natural language terminologies. To the best of our knowledge there has been no work done previously in the area of mapping the spatial and temporal relations between interacting objects in video clips into the corresponding natural language terms. Quite the contrary: previous research focuses on parsing natural language commands and mapping them into the corresponding spatial language for different purposes, such as for robotic navigation ([Matuszek et al., 2013](#)).

The only work we are aware of which tackles the problem of mapping spatial relations into natural language term is proposed by [Leopold et al. \(2015\)](#). They define metrics appropriate for describing topological relations of simple 3D regions which include the notions of terms such as splitting, closeness, and alongness. The association of the collection of metrics, 3D connectivity relationships, and many natural language spatial terms was manually examined

in a human subject study. As spatial queries usually tend to be in natural language format, they provide preliminary study into how 3D topological relations and natural language terms are correlated.

The results from the experiments conducted in this chapter prove the efficiency of the proposed model in determining the main aspects of spatial and temporal relations. To make use of extracted relations there is urgent need for practical approach to map between spatial reasoning and natural language terms. Given two interacting objects A , B and set of extracted spatial and temporal relations between them, R , the following set of natural language terms can be said to take place if the specified spatial relations between them are as follows:

- Term(A,B)= in, \iff Topology(A,B)= INSIDE
- Term(A,B)= away from \iff Distance(A,B)= DEPART
- Term(A,B)= next to \iff Topology(A,B)= TOUCH
- Term(A,B)= on \iff Topology(A,B)=TOUCH \wedge Direction(A,B)=UP
- Term(A,B)= to the left of \iff Direction(A,B)= LEFT
- Term(A,B)= to the right of \iff Direction(A,B)=RIGHT
- Term(A,B)= under \iff Topology(A,B)=TOUCH \wedge Direction(A,B)=DOWN
- Term(A,B)= toward \iff Distance(A,B)=APPROACH
- Term(A,B)= beside \iff Topology(A,B)= TOUCH
- Term(A,B)= insides \iff Topology(A,B)= INSIDE
- Term(A,B)= above \iff Direction(A,B)=UP
- Term(A) = big \iff Size(A) = BIG
- Term(A) = small \iff Size(A) = SMALL

On the other hand, the temporal relation between two intervals of action can be mapped directly to the same natural language term using Allen's notion, as described in in Section 5.3.5. However, synonyms of these temporal relation included to encourage sentence variation in the generation stage. The following synonyms are defined for the major Allen temporal relations:

- after: later, next, then, after
- during: while, at the same time, throughout
- before: previous, prior to, since
- start: at the beginning, at the start
- finish: at the end, finally

5.6 Conclusion

In this chapter we have presented an approach to formalising the spatial and temporal relations between interacting objects using the object volume-based extension of CORE-9. The algorithm was efficient; it used the segmented object's volume to form an OBB, which was then used for extracting spatial and temporal information from a video stream. The approach was able to detect spatial changes that occurred over time, promoting good semantic understanding of video content. Next, a set of rules was defined to transfer the qualitative spatial and temporal relations into meaningful natural language terms, which will be used later for generating semantic video description. The obtained results indicate that the object volume-based extension was complementary to AngledCORE-9, as they both use OBBs. However, utilisation of spatio-temporal segments of interacting objects has established its superiority. For video processing tasks it is very important that there is a satisfactory means of representing the object interaction in space and time. In this chapter we centred our approach around the qualitative representation of relations between objects' volumes. It was more precise in terms of shape regions compared to AngledCORE-9 and, as a result, it was able to provide a more accurate representation. Finally, one of the main challenges is to find a practical way to map between spatial reasoning and natural language, and to this end we presented a set of rules that facilitates the mapping process. The work in this chapter is preliminary to an investigation of the basic spatial and temporal relations. However, we believe this work will be an interesting starting point in a world which is saturated with video data and thus is in need of automated extraction of spatial and temporal relations data between interacting objects and their mapping into natural language terms, which can then be used for many applications, such as summarisation and retrieval.

Chapter 6

Generation of Textual Video Descriptions

Previous chapters have addressed the extraction of visual HLFs that include human objects among their visual attributes (action, age, gender and emotion), non-human objects, scene setting, and spatial and temporal relations between interacting entities. In this chapter, we present a framework that produces textual descriptions of video, based on the semantic video content extracted in previous chapters. Detected action classes will be rendered as verbs, participant objects will be converted to noun phrases (mainly subject for human and object for other), visual properties of detected objects will be rendered as adjectives and spatial relations between objects will be rendered as prepositions. Organising a video as a sequence of shots eliminates the redundancy problem caused by using frame-based description approaches and enriches the description by including temporal information. Further, in cases where no verb is assigned for a given track (called zero-shot action recognition) as the action recognition system is unable to identify the performed action that has not previously appeared in the training data, a language model is used to infer a missing verb, aided by the detection of objects and scene settings. These HLFs are converted into textual descriptions using a template-based approach. Paraphrasing of the resulting shot-based descriptions is introduced to create compact and coherent video descriptions. The proposed video descriptions framework will be evaluated on the NLDHA dataset introduced in Chapter 2, which contains video segments depicting a variety of human activities aligned with human annotations. ROUGE scores are used to evaluate the automatic video descriptions with respect to human annotations. Moreover, human judgment evaluation is introduced to provide qualitative evaluation of the automatic video descriptions.

The chapter is structured as follows. Section 6.1 introduces the natural language generation framework, along with the motivation for this work and its contributions. Section 6.2 reviews previous work related to the generation video descriptions task based on semantic visual content. The implementation of the proposed natural language generation framework

is presented in Section 6.3, while Section 6.4 presents a summary of the results obtained from the automatic evaluation and human judgments and a comparison with the baseline approach implemented recently. Finally, Section 6.5 provides a concluding discussion.

6.1 Introduction

The field of computer vision has advanced to detect humans, identify their activities, or to discriminate between a large number of object classes and assign them attributes. The outcome is usually a compact semantic representation that encodes activities associated with object categories. Such representations could be easily processed and interpreted by automatic systems. However, the natural way to convey this kind of information to humans is through natural language. Thus, this chapter addresses the issue of producing textual descriptions for human activities in videos. This task has a range of applications, such as human-computer/robot interaction, video summarising, indexing and retrieval. Furthermore, translation between video content and language provides a solid foundation for understanding relations between vision and linguistics, as they are the closest modalities to interact with humans.

Generating textual descriptions of visual content is an intriguing task that requires a combination of two major research aspects: visual recognition approaches and natural language generation (NLG) techniques. To generate descriptions for videos and images, a template-based approach is a powerful tool though one which needs to be manually identified (Barbu et al., 2012; Gygli et al., 2014a; Khan et al., 2015; Kulkarni et al., 2011). An alternative approach is to retrieve descriptive sentences from a training corpus based on visual similarity, or to utilise externally textual-based corpora to help rank the visual detections (Das et al., 2013b; Farhadi et al., 2010; Hanckmann et al., 2012; Kuznetsova et al., 2012; Mitchell et al., 2012).

The most relevant researches to us are the Khan et al. (2015) and Barbu et al. (2012). Both of these approaches identify high-level concepts such as humans, chairs, and so forth, and generate textual descriptions using a template-based approach. Khan et al. (2015) propose a method that relies on treating a video as a sequence of frames, and performs image detection for each frame independently, to identify HLFs without exploiting the temporal domain. Alternatively, Barbu et al. (2012) have used a dataset with simple video settings where only one action is performed. Consequently, their natural language descriptions consist of one sentence.

In contrast, this study focuses on generating descriptions of human activities in videos sequences at a shot-based level, relying mainly on visual detections. Specifically, objects'

tracks and their visual attributions are extracted from each shot, along with their spatial and temporal relations. In cases of zero-shot action recognition, where no verb (action class) is assigned for a given track, the detected objects' classes are used to mine the relative verb from web-scale textual corpora via incorporated text-mined likelihoods.

6.1.1 Motivations

With the increasing spread of mobile phones and other devices that use digital cameras, more and more videos are being captured and stored worldwide. Searching and retrieving related videos quickly turns out to be a challenging task. The automatic generation of semantic textual video descriptions that capture and summarise the main aspects of the video has become an urgent need.

The problem of generating textual descriptions for video data is necessary for two key reasons. Firstly, translating visual content into textual content will develop well-understood mechanisms for text-based retrieval, basically for free. Secondly, fine grained region labelling provides a rich semantic level for multimedia retrieval systems.

To this end, we propose a framework to generate textual description for human activities in videos at the shot-based level. Structuring videos at shot-level enables us to utilise the temporal information associated with video data. Moreover, for zero-shot cases the detected subject (which is usually human), object and place will be used to mine the relevant verb from web-based textual corpora. Finally, the set of detected HLFs will be used to generate the final description for the video using a template-based approach.

6.1.2 Generate Textual Descriptions for Video Content: Overview

We utilise the outcomes of the previous three chapters, meaning each shot of a set of participant subjects and objects is detected, along with their visual attributions and action class. Spatial and temporal relations between interacted tracks are formulated using their oriented bounding box and CORE-9 representation. Then, these detections are processed to produce a textual video description. The detected action class will be rendered as a verb, participant objects will be converted to noun phrases (mainly subject for human and object for other), the visual properties of detected objects will be rendered as adjectives and spatial relations between objects will be rendered as prepositions.

In zero-shot recognition cases, where no action class is detected, the detected objects' tracks will be used to mine the relevant verb from web-scale textual corpora via incorporated text-mined likelihoods. Next, a template-based approach will be employed to generate a shot-level description utilising extracted HLFs. As the system generates sentences independently

for each video shot, the generated multi-sentence descriptions tend to be a ‘list of sentences’ rather than a complete description ‘text’ for readers. Consequently, we automatically apply a set of rules to post-process the descriptions and generate more cohesive final video descriptions.

6.1.3 Generating Textual Descriptions for Video Content: Contributions

The contributions of the proposed work in this chapter can be summarised as follows:

- enhances visual detections for zero-shot cases by utilising detected triplets (Subject, Object, Place) and the web-based textual corpora;
- proposes a set of rules to generate coherent video descriptions by combining lists of generated shot-based sentence descriptions;
- the natural language generation experiment is carried out on the NLDHA dataset. The automatic descriptions at shot-based level are compared and evaluated against the frame-based baseline of [Khan et al. \(2015\)](#) using two different methods. ROUGE scores and human judgment evaluation are revealed the efficiency of automatic video descriptions.

6.2 Related Work

The issue of generating textual descriptions of images and videos has recently been gaining prominence in the field of computer vision. The purpose of the current section is to provide a review of previous research by examining two main aspects: textual image description and textual video description.

6.2.1 Textual Image Descriptions

Thus far, the most widely use approach through which images have been textually described is through modelling joint distributions over low-level image features and linguistics, specifically nouns. Early work in this area using multimodal topics was proposed by [Blei and Jordan \(2003\)](#) and the following extensions methods: [Cao and Fei-Fei \(2007\)](#), [Das et al. \(2013a\)](#), [Feng and Lapata \(2010\)](#), [Putthividhy et al. \(2010\)](#) and [Chong et al. \(2009\)](#). All of them jointly model image features (*e.g.* : SIFT and HOG) and language words over a latent topic model.

Other non-parametric approaches utilise nearest-neighbour as well as label transfer, as presented by [Makadia et al. \(2008\)](#) and [Guillaumin et al. \(2009\)](#). They depend on very big annotated sets in order to produce descriptions from similar samples. These categories of approaches have revealed the degree to which lingual descriptions of images on a variety of different levels can be generated. However, they possess two main shortcomings linked to low-level features and similarity measures. First, this method is not proven to scale up as the semantic richness space is increased. Second, text which is produced is mainly in word-list formats that do not possess any semantic validation.

In contrast, another class of linguistic descriptions of images mainly relies on the extraction of high-level concepts, for objects and scene categories, for example. Prominent and well-known image object detectors are the deformable parts model (DPM) ([Felzenszwalb et al., 2010](#)) and its associated visual phrases ([Sadeghi and Farhadi, 2011](#)) which have been shown to be successful in the field of image annotations. Despite being able to generate more credible semantic descriptions, these methods have been found to be very limited in reality due to the difficulty of enumerating all relevant concepts and learning their associated detector. [Kuznetsova et al. \(2012\)](#) implemented a system to create image descriptions through phrases obtained from a large image-caption database, with the help of visual similarity measures. However, this line of approach is not applicable for complex scenes where multiple objects interact.

[Li et al. \(2011\)](#) produce sentential descriptions by utilising visual objects detection associated with their visual attributes and spatial relations; however, action recognition is not covered. [Farhadi et al. \(2010\)](#) introduce a system which converts images and the associated descriptions to a ‘meaning’ space consisting of three components: object, action and scene. Next, for any test image the description will be retrieved from the training set based on visual similarity. However, this operates on the assumption of using one object per image. [Yang et al. \(2011b\)](#) utilise text-mined knowledge in order to produce sentential descriptions of static images after detecting objects and scenes. However, activity recognition is not carried out nor is text-mining used in order to choose the optimal verbs. [Mitchell et al. \(2012\)](#) and [Kulkarni et al. \(2011\)](#) describe a comparable method through which automatic language descriptions can be generated from static images, taking advantage of language statistics drawn from parsing large text-based corpora and visual recognition techniques.

6.2.2 Textual Video Descriptions

Video data introduces the additional dimension of time, with an associated set of challenges, such as temporal continuity. The majority of the literature pertaining to video descriptions

has centred around two fundamental themes: deriving the description from semantic visual content and/or mining the relevant description from text-based corpora.

Barbu et al. (2012) demonstrate a method whereby a single sentential description of a short video is generated by visual recognition techniques to render the language entities; specifically an event recognition approach is utilised to identify object tracks, role assignment and body posture variability. Finally, generation is achieved by pre-defined templates for each event class, in the form of subject-action-object. **Khan et al. (2015)** and **Hanckmann et al. (2012)** introduce a video description framework which starts with the extraction of the set of HLFs by the implementation of conventional image processing techniques. Context-free grammar (CFG) is used next to convert the extracted concepts into natural language descriptions. The drawback of these techniques is that they rely on only a limited set of high-level concepts, without exploiting text mined from text-based corpora. Moreover, videos are manipulated as sequences of images; hence no interaction between objects is considered over the time domain.

Guadarrama et al. (2013) introduce a new framework that addresses the challenges associated with describing activities ‘in-the-wild’. The method encompasses a wide range of verbs, objects and functions in an out-of-domain manner that does not necessitate videos consisting of the precise activity. If it is unable to provide a precise prediction by using the pre-trained model, it will generate a more concise and credible answer. The semantic hierarchies are learned from web-based corpora in order to decide upon the most suitable degree of generalisation. However, this work focuses on short videos clips that depict one activity; hence the resulting descriptions consist of single sentences, without investigation of any temporal associations between objects.

Gygli et al. (2014b) describe a novel way to carry out video summarisation, the process of which is initiated by segmenting the video via the use of a ‘super-frame’. Then, the degree to which the visuals are appealing is approximated for every super-frame with the use of low-, mid- and high-level characteristics. On the basis of this scoring method, an ideal subset of super-frames is chosen to produce an informative summary. However, this approach concentrates mainly on subject, verb, object (SVO) triples, without taking into account the spatial and temporal associations between objects.

Thomason et al. (2014) integrate the use of linguistics and computer vision techniques in order to enhance the description of objects in real-life videos. They propose a method through which textual descriptions of videos could be generated by combining visual detections with language statistics, via the use of a factor graph model. A conventional visual detection system was used to detect and score objects, activities and scenes involved in the video. Then, the factor graph model combines these detection confidences with probabilistic knowledge

mined from text corpora to estimate the most likely subject, verb, object, and place. Again, this study targets videos with single activity without identification of spatial and temporal relations.

Sharma (2016) proposes that action detection in videos and formulation of natural language descriptions of videos can be achieved via models based on soft attention. Presenting both spatial and temporal depth, Multi-layered Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units are employed. After learning with selective focus on portions of the video frames, this model creates a categorisation of the videos based on the couple of glimpses it captured. Additionally, it can create sentences that draw on spatio-temporal glimpses over videos to provide video descriptions. Basically, the model prioritises the portions in the frames it has learned to have higher relevance for the task in question.

Edke and Kagalkar (2016) integrate the results of cutting-edge object and activity markers with ‘real-world’ information to select the triplet of subject, verb and object for video representation. The standard data-driven approach that is used can describe short video content in a one sentence description in the Hindi language. Uncomplicated words and sentence formulations are generated by the preliminary and basic text description in Hindi that is suggested by the study. However, extraction of Hindi text information related to video content that is expressive and free of grammatical errors is the major difficulty facing this study. To enhance the outcome of testing video explanations based on activity and object detection, the triplet choice method can be employed, allowing the trainer to tag a video, especially subject, verb and object (SVO), followed by screening of the data.

Laokulrat et al. (2017) produce short video descriptions by using a sequence-to-sequence model with semantic attention and based on LSTM. Integration of external fine-grained visual information identified in every video frame is proposed as an approach for addressing the issue of overlooking correct objects that show up in videos. According to the study findings, objects in videos are not only accurately specified by the system, but the quality of video description is improved as well, through the use of semantic attention for selective focus on external fine-grained visual information.

In contrast to earlier researches, through which individual presences have been determined through the use of the DPM model (**Felzenszwalb et al., 2010**) at a frame-based level, our approach is different in several important ways. We consider the video as 3D (x, y, t) , and consequently individual detection is achieved by the novel human body segmentation approach introduced earlier in this thesis. This approach is designed for video data, to alleviate the shortcomings of the DPM model, such as partial occlusion, background noise and temporal variation. As a result it provides reliable physical interpretations. Visual

attributes for regions of detected salience are extracted, along with their spatial and temporal relations, to avoid generating long, complex and unnatural textual descriptions. The video in this approach is structured as a sequence of shots, to preserve the order of activities, combining the sentence description of each shot to generate a coherent multi-sentence video description at the required level of detail. Additionally, our work utilises a language model trained on text-based corpora only in cases of zero-shot action recognition, where no action class is detected, drawing on detected object tracks and scene setting information.

6.3 Framework for Generating Textual Video Description

Figure 6.1 shows the overall approach for the video description task, while Table 6.1 illustrates the set of vocabulary used to generate textual descriptions of video. The generating of video descriptions task basically includes two main modules: content planning and a surface realizer. In our system, the content planning is mainly accomplished by visual recognition techniques, with the exception of the case of zero-shot action recognition, where language statistics are utilised to infer the verb class, given the detected subject and object classes. For the surface realizer stage, the template-based approach is used to generate a single sentence shot-based description. As the system generates sentences independently for each video shot, the generated descriptions tend to be a ‘list of sentences’ rather than a complete description ‘text’. Consequently, a set of rules is automatically applied to post-process the descriptions and generate a compact and cohesive final video description. The following describes each of these components in turn.

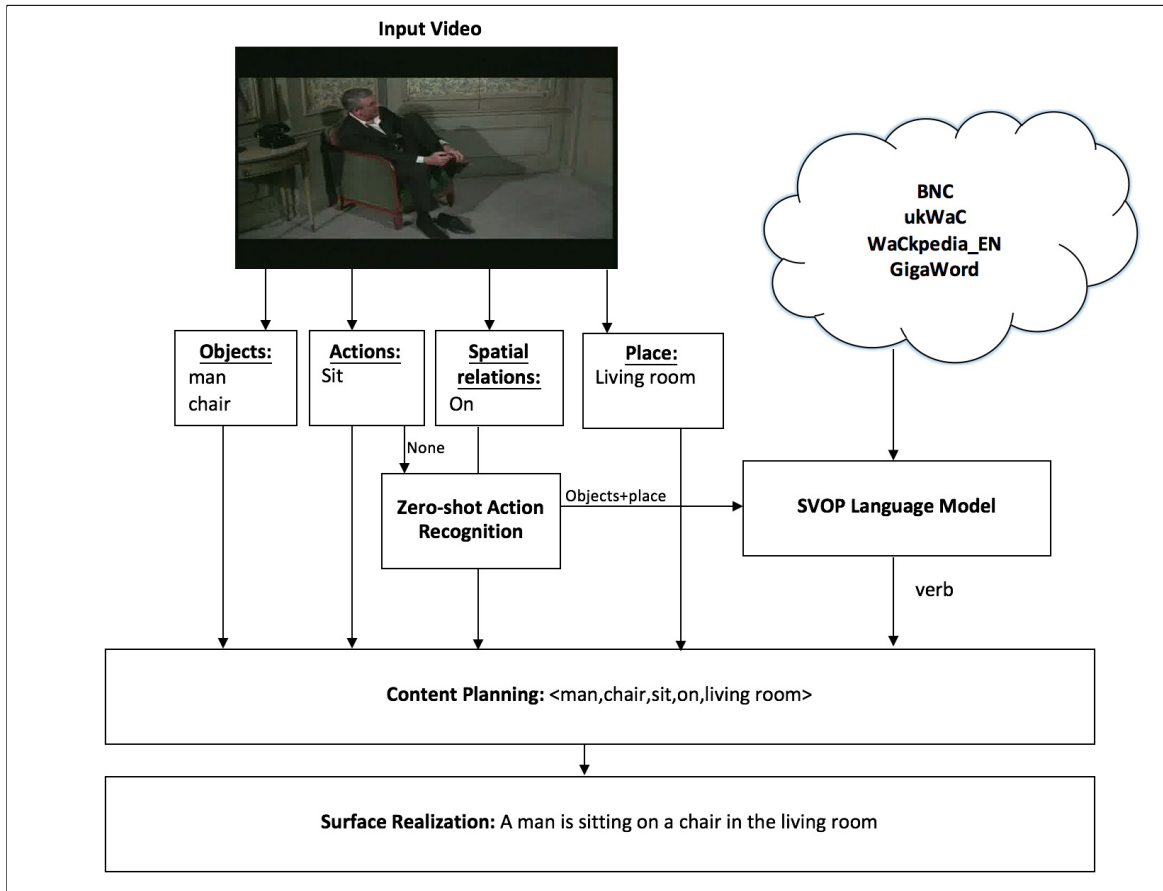


Fig. 6.1 Summary of proposed framework of generation of video description. The framework starts with content planning stage that identify the HLFs such as people, objects, actions, spatial and temporal relations. In the case of zero-shot action recognition, language statistics from large text-based English corpora will be used to predict the missing verb (action class), given detected objects' classes and recognised scene settings. Finally, a template-based approach is used as surface realizer to convert these HLFs into textual descriptions.

6.3.1 Visual recognition of Subjects

As humans are the main participants in the video activities, in this study the role of subject is assigned to human objects if they are present. A novel model is proposed that detects and segments human body regions across video frames (see Chapter 3), rather than using the human detector of Felzenszwalb et al. (2010), which is used by all previous works in generating video descriptions. This approach improves visual detection by focusing only on human regions rather than on holistic features (*e.g.* dense trajectories). We develop a robust segmentation approach involving two principle stages.

verbs:	clap, wave, jog, run, walk, dive, kick, lift, ride, skate, swing, answer phone, drive, eat, fight, kiss, hug, sit down, sit up, stand up, get out, hand shake, approach, carry, catch, collide, drop, high five, depart and touch
nouns:	man, women, baby, child , person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, phone, TV/monitor, home, road, bedroom, park, hotel , kitchen, living room , office, restaurant and shop
prepositions:	in, on, next to, to the left, to the right, under, beside, above and inside
conjunctions:	and, after, before, while, later, then, next, finally
adverbs:	away and toward
adjectives:	small, big, young, old, angry, happy, sad, surprised, serious and disgust
pronouns:	he, she, they, him and her
articles:	a, an, the
auxiliary:	is

Table 6.1 The set of vocabulary used to produce textual descriptions of video.

The first stage includes the identification of human objects at frame level that combines information from low-level image cues, such as colour and textures, with high-level part detector information. The pixels and parts represent nodes in a graph, whereas affinity and ordering relationships are encoded by edges. The second stage is to track the detected regions over the video stream by building a foreground model that combines colour and shape estimated models. As a result, a list of human objects' tracks is extracted which will be used for further processing to identify their adjective attributes, such as gender, age and emotion, using conventional image processing techniques (see Appendix A).

6.3.2 Visual recognition of Objects

We used the discriminatively trained part-based models from [Felzenszwalb et al. \(2010\)](#) in order to detect the non-human objects present in each video, creating a store of twenty object classes: bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, phone and TV/monitor. As these object detectors are mainly designed for images, they are applied to each keyframe, in order to obtain the maximum scores allocated to each objects, and top two objects are chosen per frame to reduce the false positive detections.

6.3.3 Visual recognition of Verbs

We aim to process and represent complex actions that are difficult to track efficiently using conventional descriptors. To this end a new model for action representation that relies on

extracted human regions has been introduced (see Chapter 4). It formulates a descriptor that encompasses the static and dynamic features of detected segments. The approach consists of two principle stages. The first involves extraction of motion and appearance features from the detected spatio-temporal regions using HOG/HOF descriptors (Laptev et al., 2008). In the second stage, the LLC is applied to the extracted descriptor in order to encode the local descriptors with similar bases from a codebook. 100,000 features are randomly selected and clustered using K -means for initialisation the codebook that is 4,000 words in size. Descriptors are assigned to their nearest vocabulary word using Euclidean distance. Finally, a support vector machine (SVM) classifier is used to learn a model from the feature vectors for each action (Fan et al., 2008). After several trials the classifier is applied every ten frames, to assign a human object's track with the appropriate action class. In our experiment, 30 different action classes are used to train the model, with an extra negative class that is assign to any action that doesn't appear in the training data.

6.3.4 Visual recognition of Prepositions

Generating elaborate textual descriptions demands more than simply applying object detection and event recognition. Producing a sentence with the embedding of spatial relations as a prepositional phrase requires the extraction of spatial relations between the detected interacting objects. To efficiently and accurately represent the relationships between the interacting objects present in a video stream, the AngledCORE-9 is adopted (Sokeh et al., 2013) (see Chapter 5). Firstly, an approximated region of OBB is replaced with a space-time volume for detected objects and for each extracted region a tight OBB is drawn. Finally, the compact CORE-9 representation is used to extract the spatial and temporal aspects for multiple inter-related object bodies by analysing the nine cores and six intervals in each binary relation. Compared to the commonly used representation CORE-9, the object-volume based method has a higher chance of generating reliable results regarding the direction of objects, topologies, size, distances and temporal changes. Symmetric relations are not allowed between any pairs, to eliminate the redundancy. In this study, the following prepositions are identified, including in, on, away from, next to, to the left, to the right, under, toward, beside, above and inside.

6.3.5 Visual recognition of Scene Settings

The state-of-the-art scene recognition system is utilised for this task. Zhou et al. (2014) proposes a scene-based system called 'Places 205' based on Convolutional Neural Networks (CNNs). It provides over 5,000 labelled images of environments and a collection of 205

Corpora name	Size of text
British National Corpus (BNC)	1.5GB
WaCkypedia EN	2.6GB
ukWaC	5.5GB
Gigaword	26GB

Table 6.2 The English corpora used to mine the SVOP tuples.

scene classes. When combined with CNNs, the system is capable of deep features learning for a broader range of scene identification tasks, and can offer valuable data in line with a number of scene-based datasets. In order to accurately identify the scene featured in the corpus for this study, the environment recognition method suggested by [Zhou et al. \(2014\)](#) was employed. The method was used to identify the scene setting of the first frame in each shot – whether it was an indoor or an outdoor scene, with a ranked list of the five most likely place categories. For this experiment, 12 different places are recognised for both indoor and outdoor settings (*e.g.* road, bedroom, park), for each of which the associated preposition is assigned manually.

6.3.6 Zero-shot Language Statistics

Our approach to generating a textual video description relies mainly on visual semantic content. However, there is a case called zero-shot action recognition where the action recognition system is unable to identify the performed action, as the action has not previously appeared in the training data; in this case a negative class is assigned. Subsequently, language statistics will be used to predict the missing verb (action class), given a detected objects' classes and recognised scene settings.

Language statistics are mined from four large text-based English corpora. As in [Thomason et al. \(2014\)](#) the dependency parser¹ is used to parse text from the following corpora: English Gigaword, British National Corpus (BNC), ukWac and WaCkypedia-EN. Table 6.2 shows their sizes after pre-processing. The quadruple of SVOP (subject, verb, object, place) are extracted using the dependency parser. The subject-verb relations are extracted on the basis of nsubj dependencies, while the verb-object relations are identified by dobj and prep dependencies (prep dependencies are used in order to account for intransitive verbs that occur with prepositional objects). Object-place relations are extracted by utilising the prep dependencies where the noun affected by the preposition belong to the recognisable places list.

¹The spacy's API: <https://spacy.io>

The quadruple frequency of SVOP are maintained and if no object or place is present in the sentence, their values in the quadruple are None. For the best performance, the frequency counts are a python dictionary with verbs as keys, and for each verb we keep the count of each context (subject, object, place) that co-occurs with that verb. To propose the best verb for a given context, the conditional probability $P(V|S, O, P)$ is calculated by maximum likelihood estimate (MLE) as follows:

$$P(V|S, O, P) = \frac{P(V, S, O, P)}{P(S, O, P)} = \frac{Count(V, S, O, P)}{Count(S, O, P)} \quad (6.1)$$

The verb with high probability given the context of subject, object and place is chosen to generate the sentence.

6.3.7 Sentence Generation

Finally, the extracted information from previous stages will be used to generate informative descriptions for each shot. For this purpose the template-based approach will be used. The same template will be used to create a description for each human track present in the video shot, if no human track is detected the object is considered as a subject and described in term of it motion. The list of generated sentences will be further processed to generate a coherent description. Like [Thomason et al. \(2014\)](#), the following template will be used for the generation task:

‘Determiner (A, The) - Adjective (optional)- Subject - Verb (Present Continuous) - Preposition (optional) - Determiner (A, The) - Adjective (optional) - Object (optional) - Preposition (optional) - Determiner (A,The) - Place (optional)’.

For implementation purposes, the surface realizer simpleNLG is utilised ([Gatt and Reiter, 2009](#)). This package also provides some extra processing applied automatically to the generated sentence: (1) the first letter is capitalised for each sentence; (2) ‘-ing’ is attached to the verb if the progressive tense is chosen; (3) the words are assembled in the correct grammatical order; (4) white spaces are automatically added to separate words; and (5) at the end of each sentence a full stop is inserted.

6.3.8 Creating Cohesive Descriptions

Our system independently describes each video shot. The generated multi-sentence descriptions for the video as a whole tend to be a ‘list of sentences’ rather than a coherent ‘text’. Generating coherent natural language descriptions requires linking sentences at a surface level without any need for deep understanding of the text produced. Hence, the generated

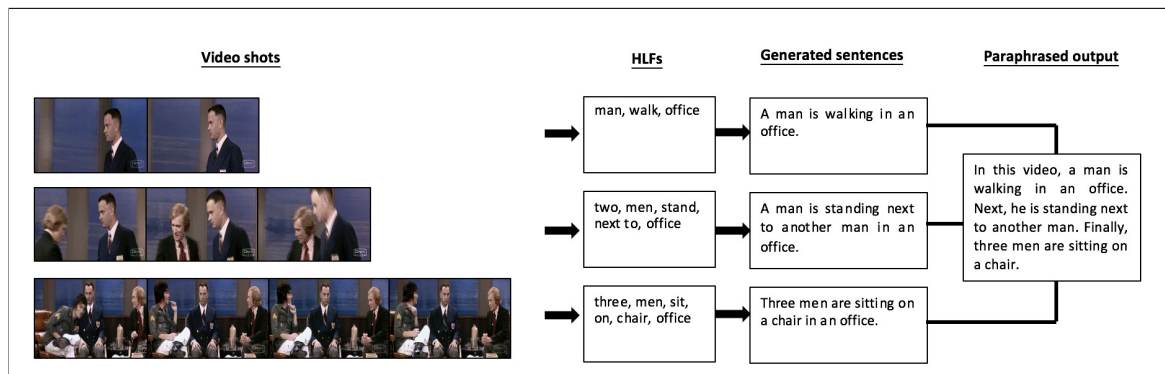


Fig. 6.2 Generating video description stages for the ‘actionclipautoautotrain00428’ video from the SitDown class, with three shots; the process start with the extraction of the HLFs list, followed by shot-based sentences generation. Finally, the paraphrasing stage is applied.

list of sentences for each video is automatically post-processed at two levels – shot-level and video-level – in order to create more cohesive and informative descriptions, as shown in Figure 6.2. First, each human track in each shot will be described independently in a complete sentence, which results in a list of sentences describing a given shot. The following set of rules is applied in order to generate compact and coherent sentence:

1. When multiple subjects perform the same action at the same time, the subjects of these sentences are combined by ‘and’. (e.g. If (i) ‘A man is eating.’ and (ii) ‘A woman is eating.’ these are combined to become (iii) ‘A man and woman are eating.’)
2. If multiple subjects perform different action simultaneously, they will be combined using ‘while’. (e.g. in Figure 6.3 (a)(b), (i) ‘A man is standing.’ and (ii) ‘A woman is walking.’ these are combined to become (iii) ‘A man is standing, while a woman is walking.’)
3. In the case where multiple subjects interact to create certain common actions (e.g. hug or fight), one is considered as the subject while the other(s) serve as objects in the sentence. (e.g. If (i) ‘A man is fighting.’ and (ii) ‘A man is fighting.’ these are combined to become (iii) ‘A man is fighting another man.’)
4. Proper pronouns (co-reference) are added if multiple verbs are allocated to the same subject during the same video shot. In this case, when a subject is mentioned again after its debut, a proper pronoun is used to improve the sentence’s concision. (e.g. in Figure 6.3 (c)(d), ‘A man is walking toward a phone.’ and (ii) ‘A man is answering a phone.’ these are combined to become (iii) ‘A man is walking toward a phone. Later, he is answering the phone.’)

Secondly, shot-based descriptions are combined to produce the final video description. For this purpose the following rules are applied:

**Shot-based descriptions:**

(a) A man is standing to the right of a woman in a living room.
 (b) She is walking toward him in a living room.

(c) A man is walking toward a phone in a living room.
 (d) He is answering the phone in a living room.

Post-processed description:

In this video, a man is standing to the right of a woman, while she is walking toward him in a living room. Next, a man is walking toward a phone. Later, he is answering the phone.

Fig. 6.3 Example of applying post-processing rules to the system-generated description of ‘actionclipautoautotrain00463’ video from the AnswerPhone category, with two shots.

1. Temporal adverbials (*e.g.* next, then and finally) are incorporated between subsequent sentences as a powerful device for conserving the logical order of events performed over different shots.
2. Scene-setting information is added only to the leading sentence and discarded from subsequent sentences if the event take place in the same setting to eliminate redundancy.
3. The phrase ‘In this video,’ is added to the leading sentence of each video description.

6.4 Experiments and results

This section presents the evaluation procedure of our video description framework on the NLDHA dataset introduced in Chapter 2. First, a brief overview of the baseline approach used to provide a comparison with our system is presented. Next, the results of quantitative evaluation with the ROUGE Metric, along with qualitative human judgements, are discussed.

6.4.1 Frame-based Video Description Baseline

To put our performance in perspective, we compare our proposed approach against the baseline video description framework of [Khan et al. \(2015\)](#). This approach is chosen as the baseline as it augments the sentence components largely on the basis of semantic video content by applying conventional image processing techniques. Additionally, in order to make a fair comparison, the same set of detected objects are used for both systems. However, we advanced the detection to accommodate temporal information from the videos.

Hand annotations of characterisations of video segments created on the basis of a TREC Video set of data are made initially. To determine the video contents that attracted individuals’

attention, the data are analysed. Subsequently, to describe the video streams smoothly and coherently, a framework is created, drawing on traditional methods of image processing that derive high-level features (e.g. humans and their activities) from each video frame.

These features are afterwards used to generate a description in natural language. A limited set of spatial relations are calculated between the extracted HLFs' geometric features, though no temporal information is considered. These HLFs are translated into sentential descriptions utilising the SimpleNLG, a template-based approach. Evaluation involves the determination of ROUGE scores between descriptions annotated by hand and those produced by machine. The generated descriptions are also qualitatively assessed through a human-controlled task-based assessment.

6.4.2 Evaluation with ROUGE Metric

The complexity of evaluating video textual descriptions comes from the fact that defining the criteria is a challenging task. To evaluate our method, we examine the metrics commonly used for this purpose in machine translation. These metrics include the BLEU (bilingual evaluation understudy) (Papineni et al., 2002) and ROUGE (Recall Oriented Understudy for Gisting Evaluation) (Lin, 2004) metrics, among others. The BLEU score calculates precision on n-grams level as it measures how many n-grams of the automatic summaries appear in the human annotation references, and for this reason is not suitable for our task of lingual video description, as has already been suggested by (Mitchell et al., 2012) and (Das et al., 2013a).

By contrast, ROUGE score is an n-gram recall oriented measure of the information coverage of human annotation references compared to automatic summaries produced by a system. ROUGE score measures how many n-grams of the human annotation references appear in the automatic summaries. A higher ROUGE score denotes a higher degree of match between them. In general, a score of '1' indicates a perfect match, whereas a score close to '0' means the match occurs in only a small portion of the data. Four different ROUGE scores are used in this experiment, ROUGE-1 (unigram) recall is the perfect option to compare descriptions based on predicted keywords only (Das et al., 2013a). ROUGE-2 (bigram) and ROUGE-SU4 (skip-4 bi-gram) scores are best to evaluate lingual video descriptions for coherence and fluency, whereas ROUGE-L scores depend on the longest common subsequence. ROUGE metrics are chosen for this study following Das et al. (2013a), who used it to evaluate lingual video summarisation.

Table 6.3 present the average ROUGE scores achieved between the automatic descriptions produced by the baseline and our system, averaged over all twelve different human action categories, with respect to manual annotations. Manual annotations tend to be subjective as they depend on the annotator's perception and understanding. Moreover, this subjectivity

		Baseline	Our approach
ROUGE-1	R	0.2480	0.3513
	P	0.3443	0.2474
	F	0.2749	0.2806
ROUGE-2	R	0.0532	0.0737
	P	0.0801	0.0500
	F	0.0592	0.0577
ROUGE-L	R	0.2353	0.33365
	P	0.3275	0.2354
	F	0.2609	0.26689
ROUGE-SU4	R	0.0939	0.1526
	P	0.1745	0.0951
	F	0.1064	0.1098

Table 6.3 ROUGE scores calculated for the baseline and our approach, with respect to hand annotations. Four different ROUGE metrics are measured: ROUGE-1 (unigram), ROUGE-2 (bigram), ROUGE-L (longest common subsequence) and ROUGE-SU4 (skip-4 bi-gram). For each ROUGE metric, the recall (R), precision (P), and F-measure (F) are averaged over all twelve categories from the NLDHA dataset.

might be affected by personal education level, interests, background and experiences. As a result, the ROUGE metric inevitably penalises many automatically generated sentences where these do not match the manual annotations, despite being technically correct.

The ROUGE metric measures three aspects – recall, precision and F-measure – for any automatic summary and reference summary. Generally, recall and precision can be defined as follows: ‘a system with high precision indicates that an algorithm retrieved substantially more relevant than irrelevant data’ whereas ‘a system with high recall means that an algorithm return most of the relevant results’.² In our experiment ROUGE metrics will be analysed in terms of recall, as we are interested in relevant results. It can be observed that our proposed system significantly outperforms the baseline for all the ROUGE metrics.

Clearly, the best results were obtained by ROUGE-1, as our method involves an extended language vocabulary compared to the baseline. This richness comes from a varied set of verbs included along with their scene setting, especially when the language model is involved for the case of zero-shot action recognition. (*e.g.* When ‘person’ and ‘TV’ are detected in the scene without a connected verb, the language model will infer the verb ‘watch’ to complete the sentence.) Additionally, ROUGE-L results confirm the efficiency of our approach as it captures similarity at sentence-level between the automatic generated descriptions and hand annotations. There is also an observable improvement for ROUGE-2

²https://en.wikipedia.org/wiki/Precision_and_recall

	Grammar	Correctness	Relevance
Baseline	3.40	3.40	2.25
Our approach	3.54	3.75	3.74

Table 6.4 Human evaluation for the baseline and our approach, with respect to three aspects: grammatical correctness, cognitive correctness, and relevance.

and ROUGE-SU4. This is not surprising since attributes (such as adjectives and prepositions) and co-reference enhance the quality of description by generating richer and less verbose descriptions. However, this kind of improvement in quality does not usually contribute considerably to the ROUGE score, which is based on n-gram comparisons.

6.4.3 Human Evaluation

The ROUGE metrics produce only a rough estimate of the informativeness of an automatically produced summary, as it does not consider other significant aspects, such as readability or overall responsiveness. To evaluate these types of aspects there is an urgent need for manual evaluation. For this task Amazon Mechanical Turk was used to collect human judgements of automatic video descriptions. We follow [Kuznetsova et al. \(2012\)](#) and asked 10 Turk workers to rate video descriptions generated by the baseline and our description. Each worker watched each video and rated the description on a scale of 1 to 5, where 5 means ‘perfect description’, and 1 indicates ‘bad description’.

The description rating was based on three different criteria: grammar, correctness, and relevance. For both the correctness and relevance aspects, the video was displayed with its description. The correctness evaluates to what extent the textual description depicted the video semantic content, while the relevance rates if the sentence captures the most salient actions and objects. For the grammar correctness, only lingual descriptions were presented to the worker, without the video, to evaluate the sentence. Table 6.4 shows the results of human evaluation of both the baseline and our approach. It can be observed that our system improves on the baseline in all three aspects. However, the relevance score significantly outperforms the baseline with margin of 1.61. This indicates that our approach is able to describe much more semantic video content, especially in terms of activities, attributes and scene setting.

6.4.4 Discussion

Describing video semantics automatically in natural language is a compelling task and challenging as well. This chapter has described the video semantics based on the spatio-

temporal volume of entities and their attributes. To this end, a human object segmentation model is utilised, to extract the human subject at the shot-based level, as videos provide useful information such as motion and temporal continuity. Consequently, a robust descriptor is formalised for the extracted regions and embedded into an action recognition framework, in order to assign the appropriate action to each track. Using segmentation substantially contributes to an enhancement of the accuracy of visual attribution identification and activities recognition.

The extracted HLFs are plugged into an NLG template-based system to generate a coherent description. In an extensive experimental evaluation we show the improvements that our framework provides compared to the baseline frame-based video description system. The improvements are consistent among both automatic evaluation with ROUGE metrics and manual human evaluation of correctness and relevance. This improvement by the proposed system stems from the fact that the main sentence components are extracted after visually parsing the video semantic content with respect to temporal information, See Figure 6.4 for some examples of automatic video descriptions.

The majority of previous works, including the baseline system, rely on the image-based detector deformable parts model (DPM) (Felzenszwalb et al., 2010) which is applied to each frame to augment a store of detected objects, without preserving any temporal dependency between video frames. As a result, the descriptions generated using this detector suffer from several weaknesses, mainly redundancy and lack of coherence. The redundancy issue basically results from applying the object detector at each frame without maintaining any temporal correlation; hence if the object changes its position gradually between frames it will be considered as a new detection.

Moreover, consistent co-reference of pronouns to visual objects across multiple sentences cannot be reliably identified for image-based detections, as prior information is required from the preceding frames to prove the previous detection. As a result, the generated description will be verbose, unnatural and contain irrelevancies. The Figure 6.4(c) shows an example of co-reference identification achieved successfully by the proposed system in ‘*she* is sitting next to *him*’, while the baseline was unable to identify such information as its detection based on individual frames rather than tracking the detection over video frames and exploiting the temporal continuity.

Generating elaborate textual descriptions demands more than action recognition and object detection. Identifying spatial and temporal relations between entities allows them to be mapped onto prepositions and adverbs in the output description. Figure 6.4(b) shows an example of improvement over the baseline as the proposed system was able to identify the scene layout by formalising spatial relations in ‘a man is standing *next to* a car; while a




<p>(a) Clip name: actionclipautotrain00230 Class: SitDown, 2 shots</p>	
	<p>Hand annotation: This scene starts with a conversation between a couple. Later, the person sits on a chair and starts removing his shoes Baseline system: An old man and a young woman are talking. The old man sits down Our system: In this video, an old man is standing next to a woman in an office. Later, he is walking away from her. Next, an old man is sitting on a chair.</p>
<p>(b) Clip name: actionclipautotrain00153 Class: DriveCar, 4 shots</p>	
	<p>Hand annotation: A woman is driving a car. Next, the husband who is wearing a brown coat and wife who is wearing a black dress are standing in front of the car. Later, they are arriving to a memorable house. A mother who is wearing black dress is visiting her daughter in the college. Baseline system: A man and woman sit in a car and talk. A woman talks to a young woman. Our system: In this video, a happy woman is driving a car on the road. Later, a man is standing next to a car; while a woman is standing to the right of him. Next, a car is moving on the road. Finally, a woman is walking toward a young woman in the park.</p>
<p>(c) Clip name: actioncliptrain00676 Class: HandShake, 1 shot</p>	
	<p>Hand annotation: A man is on a car. They are in a country place. The woman walks toward the man. He helps her sit on the car. They begin to talk and smile. Baseline system: A woman talks to a man. The woman sits down. Our system: In this video, a happy woman is walking toward a man in the park. Next, she is shaking him. Finally, she is sitting next to him.</p>

Fig. 6.4 Sample of textual video descriptions along with their video shots from different categories from the NLDHA dataset.

woman is standing *to the right of* him’. Additionally, temporal relations are captured by the system in Figure 6.4(c) ‘a woman is walking *toward* a man’ and in Figure 6.4(a) ‘a man is walking away from her’ as this relation is calculated by comparing the distance between two objects over sequence of frames.

We are currently unaware of any prior work that generates such information from video data, apart from [Kulkarni et al. \(2011\)](#), who identify basic topology relations for images; all previous works estimate prepositions by retrieving relevant prepositions from text-based



Fig. 6.5 Computer vision techniques can result in errors: (a) although the man is detected, his action is misclassified as walking rather than as answering the phone; (b) only a man is identified while the woman next to him not detected; (c) the car is not detected and only two sitting men are identified; (d) two persons who are eating are detected correctly from this scene but the rest are missed.

corpora based on types of detected objects. In this work we instead make these relations the foundation stone of the analysis, and identify a primary set of spatial and temporal relations that could be extended in future work.

The proposed framework is applicable to any video genre with human actions and even if no human is detected, the video will be described based on detected non-human objects and scene setting. Although this framework produces a syntactically and grammatically correct description, the current immaturity of computer vision techniques can lead to false positive detections or missing information; refer to Figure 6.5 for some examples. As a result, the generated description can be inaccurate and mismatch the real action performed in the video sequences. There is a room for improvement, especially in object detections and their associated attributes, such as actions, colour and dress, which can significantly enhance the accuracy and quality of automatically generated description.

6.5 Conclusion

This chapter has introduced a framework that produces textual descriptions of video based on semantic video content extracted in the previous chapters. Detected action classes will be rendered as verbs, participant objects will be converted to noun phrases (mainly subject for humans and object for others), visual properties of detected objects will be rendered as adjectives, spatial relations between objects will be rendered as prepositions, and event characteristics as adverbs. Further, a language model is utilised to infer the best probable fit in case of no verb being assigned for a given track, aided by the objects and scene settings detected. These detections are converted into lingual descriptions using a template-based approach.

In an extensive experimental evaluation we show the improvements of our framework compared to the baseline frame-based video description system. The improvements are consistent among both automatic evaluation with ROUGE metrics and manual human evaluations of correctness and relevance. This improvement offered by the proposed system stems from the fact that the main sentence components are extracted by visually parsing the video content with respect to temporal information.

Chapter 7

Human Action Retrieval via Textual Descriptions

Generating natural language descriptions for videos offers a multimedia repository that facilitates video summarisation, indexing and retrieval. The number of videos clips uploaded to the world wide web continues to increase dramatically. Hence, content-based video retrieval is becoming ever more important to meet the users' need and retrieve their desired data from huge video databases. One of the major challenges associated with big data is memorising video details and the sophisticated relations between video objects. The majority of video data relates to humans, and hence human action retrieval has become a specific new topic in the big data field, in which searches for videos are based on the human actions performed (Ramezani and Yaghmaee, 2016). Human action retrieval can be used for human-computer interaction, human behaviour analysis and video surveillance.

However, most current video retrieval systems still rely only on user metadata text-matching. Unfortunately, such solutions can fail because of the large number of videos that have small amounts of user metadata or none at all. The need for intelligent video search systems, which could bridge the semantic gap between the users' needs and the video content, has become pressing. In this chapter, we introduce a new video retrieval system based on textual video descriptions. The prime focus of this thesis is describing human activities within video streams, as has been evaluated in previous chapters. To verify the efficiency of the proposed framework, the human action retrieval task is selected as an application.

In using the framework developed throughout this thesis, the video dataset is indexed both by video clips and by their automatic natural language descriptions. The user provides a query as a text formula, as this is the most convenient way for humans and is computationally efficient as well; next the vector space model (VSM) and the semantic indexing (SI) are used to search the video database for a user query, and the text similarity between the query

and automatic video descriptions is measured. Consequently, for the top five most similar automatic descriptions, the corresponding video clips are retrieved. Average precision and average recall scores are used to evaluate the human action video retrieval on the NLDHA dataset.

The chapter is structured as follows. Section 7.1 introduces the human action video retrieval framework. Section 7.2 reviews previous work related to content-based video retrieval. The implementation of the proposed video retrieval framework is presented in Section 7.3, while Section 7.4 presents a summary of the results obtained from the automatic evaluation with precision and recall scores. Finally, Section 7.5 provides a concluding discussion.

7.1 Introduction

Today's significant growth in digital technology is enabling users to share their multimedia data, such as images, audio and videos, much more easily than before. However, managing these huge amounts of data to search for desired objects has become a challenging task (Wang et al., 2015). To overcome this challenge, multimedia retrieval systems are being proposed that are able to retrieve subsets of these multimedia data that match user enquiries.

Unlike the progress of multimedia data growth, video retrieval techniques have not received the desired attention and have dropped behind. The majority of search engine systems retrieve multimedia data based on the textual data assigned by the user. Such methods search multimedia repositories via metadata, such as titles, tags or descriptions uploaded by the user (Arman et al., 1994; Zhang et al., 1997). However, this method can be easily flawed or manipulated, and irrelevant objects are retrieved when the associated metadata is faulty, incorrect or irrelevant to the video content (Zhai et al., 2013).

An alternative method for metadata video retrieval systems exists, in which the actual content of objects is utilised as a similarity measure (Lew et al., 2006). The visual objects are considered as a query in order to find similar objects that match the user's query. Hence, representative features should be extracted from objects' content in order to describe them efficiently. However, the main drawback of such techniques is that their on-line processing is computationally expensive and the user is not always has a visual example of the query. Moreover, some search methods use both types of model – content and metadata information together. These approaches are called 'concept annotation' and can be used in the image domain (Wu et al., 2013) and video domain (Wang et al., 2012).

Most retrieval systems seek to retrieve one type of media, with the query and retrieved results sharing the same media type. However, a new generation of retrieval applications has

emerged called cross-media retrieval systems; these are able to retrieve a variety of media contents (Zhai et al., 2013). In this retrieval system the query can be a text and the retrieved results on different modality form, such as images or videos, as it is the most convenient way for the end user. Existing video retrieval systems mainly rely on retrieval by example (the query can be image or video) and can be categorised based on the video content into scene retrieval (Sun et al., 2008), action retrieval (Paez et al., 2013), or actor retrieval (Gao et al., 2007) types.

Providing an alternative, content-based video search methods have emerged to alleviate the shortcomings of metadata search systems. This category of search approach relies on the semantic content of videos, such as people, actions, objects and scenes that are automatically detected using video processing techniques (Ramezani and Yaghmaee, 2016). In this chapter, we are motivated by the challenge of content-based video searches and propose a new video retrieval system based on semantic natural language video descriptions. The objects detected in a video are translated into convenient textual forms as this is more human-friendly and facilitates the search process by creating a multimedia repository.

7.2 Related Work

With the explosion in multimedia data levels, there is an urgent need for efficient multimedia retrieval systems. Textual forms of query provide a natural interface for humans, and can claim to be the most convenient approach for retrieving a query from video databases. Hence, existing video retrieval approaches can be categorised into three major groups: retrieve by metadata, retrieve by low-level features or retrieve by video semantic content where the query can be keywords (*e.g.* ‘bike’ or ‘horse’) or complete text (*e.g.* ‘a man is driving a car’).

7.2.1 Metadata retrieval

Users can freely add textual descriptions when a video is uploaded. Such metadata has no pre-defined structure and the user can use any combination of words and sentences; for example, when a user uploads a YouTube video, any textual description can be added as metadata (Abburu, 2010). Most existing search engine systems retrieve video data based on textual data assigned by users. Such methods search multimedia repositories via metadata such as titles, tags or descriptions uploaded by the user (Arman et al., 1994; Zhang et al., 1997).

Further, existing systems, such as the one used by YouTube, are based on matching user query words against the textual metadata generated by the uploader (Davidson et al.,

2010). Such systems can prove to be faulty and unreliable when metadata data are incorrect or irrelevant to the actual video content. Although the annotation process seems very easy and speedy, the retrieval task tends to be challenging when the annotation is less relevant to the video content, and impossible when the annotation is completely irrelevant to the video content (Zhai et al., 2013).

7.2.2 Low-level features retrieval

Early approaches to video retrieval (Marques and Furht, 2002; Smith et al., 2002) modelled video data as a group of low-level features produced from various modalities. Low-level features are extracted automatically from video data, such as histograms in colour space, wavelets, texture features and structure through edge direction. However, due to the semantic gap, this type of system cannot express the video content. As a result they have very limited success with semantic queries for video retrieval.

Several researchers have demonstrated the difficulty of retrieving relevant data using low-level features (Markkula and Sormunen, 2000; Rodden et al., 2001). Recently, graphs have received more attention due to their effective representation power in many computer vision applications. Regions of interest present in images or video data are represented as a graph of connected nodes and edges, where nodes usually correspond to pixels and edges encode their affinity. The task of performing video clip comparisons is formulated as seeking the consistent relation of pairs of features over their graphs. Alignment of two graphs is achieved by matching their nodes while conserving the edges belonging to the matched nodes (Zaslavskiy et al., 2010).

Video retrieval techniques based on direct visual matching between video features are categorised based on the targeted video content into scene retrieval (Sun et al., 2008), action retrieval (Paez et al., 2013), or actor retrieval (Gao et al., 2007). In addition to the semantics of video content, methods for recognising human actions are being explored. Li et al. (2010) propose a video analysis system for sports videos. The approach starts by segmenting the player's body in jump and diving videos, with action recognition subsequently achieved using Hidden Markov Models. However, using on-line visual matching of query against the video database is computational expensive and time consuming process which is not desired by the end user.

Searching through these repositories of multimedia indexed by low-level features such as colour and texture usually fails to fulfil the user's needs due to the persistent semantic gap resulting from the information extracted from the video data and the interpretation of the user's query (Smeulders et al., 2000). Despite the success of such approaches in many applications of video data, they cannot be easily applied to the video retrieval task.

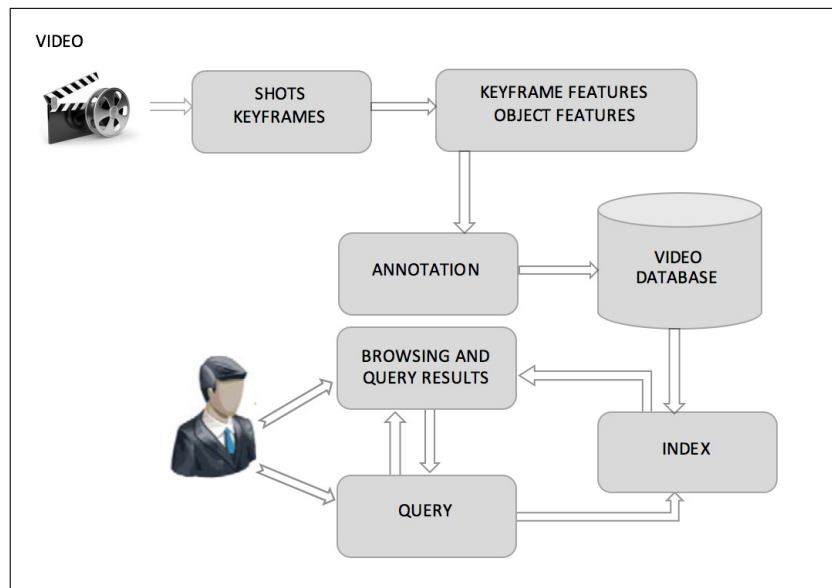


Fig. 7.1 The content-based video retrieval pipeline (Bhat et al., 2014).

7.2.3 Semantic content retrieval

The vision community struggles to bridge the semantic gap between low-level features such as colour and texture, and high-level video semantic representation, especially for real-life video where the content is varied. An alternative substitution approach for metadata and low-level features video retrieval systems is content-based video retrieval (CBVR) (Lew et al., 2006) which utilises the actual content of the video. In CBVR, multiple objects can be considered as a user query in order to seek for similar objects to be retrieved. Hence, representative features can be extracted from a video's content in order to robustly describe them. Further, many video retrieval systems use semantic content as keywords annotation to efficiently retrieve relevant video. This type of method is called concept annotation. Concept annotation has been applied successfully in both the image (Wu et al., 2013) and video domains (Wang et al., 2012).

Concept annotation approaches basically utilise a variety of descriptors that can be reliably detected and recognised (*e.g.* : faces, cars, aeroplane). Many automatic semantic classifiers have already been developed, such as concepts related to human figures (face, body parts), acoustic concepts (music, speech) and scene settings (indoor, outdoor) (Chang et al., 2005). The semantic concept detection task has been studied by many researchers and they conclude that training these classifiers with enough data could help them to reach the mature level needed for their use by video retrieval systems (Natsev et al., 2005).

Generally, CBVR systems can be divided into four main parts, as shown in Figure 7.1: (a) video segmentation (how to partition the video and extract the representative features),

(b) video annotation (specify the technique by which video content is described, *e.g.* : visual signatures, keywords or complete text), (c) video indexing (a process for extracting from the annotation a feature and its value), and (d) video database (how to organise the data in a way that allows a computer to quickly retrieve the desired content). The effectiveness of each component in the system is directly reflected in the accuracy and efficiency of the video retrieval system (Bhat et al., 2014).

Semantic searches for video content are considered a new and challenging task in multimedia retrieval, specifically when textual query is used. Existing systems are largely restricted to exact text matching, where the query is mainly matched to the textual annotation attached by the uploader. Instead, and following the format of concept annotation, we propose a novel system where this annotation is automatically generated based on the semantic video content. For raw video clips the low- and high-level features are leveraged together to detect and extract a list of semantic concepts. For these concepts, action recognition and spatial and temporal relations are extracted and converted into textual form, using a rule-based approach. We are interested in repositories of video data associated with automatic semantic annotation. The textual form of the query provides a natural interface for humans and can be treated as the most convenient and speedy approach for retrieving a query from video databases. Hence, the query is received as a text and the results will be a list of relevant videos for which the annotations are the most similar to the query, based on using the classical information retrieval approaches VSM and SI.

7.3 Human Action Text-based Retrieval Framework

The semantic video search task will be modelled as a classical text-based retrieval system, in which a user submits a text query and the result is a list of relevant videos. The system presented in this section comprises two main components, namely off-line processing involving video segmentation, annotation and indexing, and on-line processing, which involves the user query, similarity measurement and retrieval of relevant videos. The following sections discuss this in further detail.

7.3.1 Off-line Video Segmentation, Annotation and Indexing

The collection of video clips in the dataset is segmented into meaningful shots. For each shot, a fine-grained segmentation approach is applied in order to detect the human participants over a sequence of frames. Consequently, the support vector machine is adopted to train the actions classifiers. Once a list of HLFs has been extracted, a template-based approach is used

to translate these concepts into readable annotations. Finally, paraphrasing of the resulting shot-based descriptions is introduced to create compact and coherent video descriptions. The automatic annotations along with their corresponding video clips are stored together for further processing in the next stage. The Vector Space Model (VSM) and Semantic indexing (SI) techniques are commonly used in the text retrieval field (Parameswaran et al., 2012). However, the use of such approaches in image and video retrieval is somewhat limited. In this chapter, these methods are utilised to retrieve videos content based on their automatic natural language descriptions. The majority of text retrieval systems function in a similar manner, by indexing a database of documents and user queries as a set of terms' weight. Then, the similarity between the query and all indexed document are measured to retrieve the most similar documents. Below is a brief description of these two models.

A- The Vector Space Model (VSM)

Once the automatic description of each video has been generated, the well-known information retrieval model named vector space model (VSM) is used for indexing and retrieval purposes (Salton and Yang, 1973). This model transforms the given text into a vector in a very high-dimensional vector space. The main power of this model comes from its ability to measure the proximity between any two vectors, which is to say the 'closeness' between any two texts.

In this model, if m distinct terms from a collection of documents are present, a document d will be represented as an m dimensional vector $d = (w_1, w_2, \dots, w_m)$ where w_i represents the terms weight calculated for the i -th term. Generally, the document can be represented by this model following three main stages. First, that the document corresponds to a video clips' annotations is index by extracting representative terms. Second, the extracted terms are weighted to create an inverted index, which facilitates the retrieval process. Finally, the user query is ranked against the video annotations documents, based on a cosine similarity measure, and the most relevant video clips are retrieved, as for text-based video retrieval systems (Yan and Hauptmann, 2007).

Document Normalisation The documents are prepared for indexing by applying two types of linguistic pre-processing. First, the document text is split into a list of words, while punctuation and spaces are removed during this process. The second stage involves stemming, which is mapping words with different endings onto a single word using the Porter stemmer (Porter, 1980). The prevalent current thinking is that morphological variations of words with the same root/stem can be regarded as a single word because thematically they are the same.

Term weighting and indexing. During this stage the document will be indexed according to each term that occurs to create a term–document matrix X . In this matrix each dimension corresponds to a unique term and its weight in each document. In the classical

VSM the term is weighted using local and global parameters (Baeza-Yates et al., 1999). The normalised weight vector for document d of m terms is calculated as follows:

$$tfidf(w_{t,d}) = \frac{tf_{t,d}}{\max(tf_{t,d})} \cdot \log \frac{D}{df_t} \quad (7.1)$$

where, $tf_{t,d}$ is a term frequency local parameter that represents the number of times that term t occurs in document d . The df_t is document frequency and a global parameter representing the total number of documents containing term t , and D stands for the total number of documents in the database. The $\max(tf_{t,d})$ is a normalisation scaler used to penalise larger documents, so that all documents share equal significance. Finally, a term-document matrix X is created with m terms in the rows and D documents in the columns, as follows.

$$X = \begin{bmatrix} tfidf(t_1, d_1) & \dots & tfidf(t_1, d_D) \\ \vdots & & \vdots \\ tfidf(t_m, d_1) & \dots & tfidf(t_m, d_D) \end{bmatrix} \quad (7.2)$$

B- The Semantic Indexing (SI) The VSM model represents the video description texts as bag of words without preserving the sentence structure. Inspired by Lin et al. (2014) who proposed a videos retrieval system by complex textual query where the textual query is represented as a graph using dependency parser, the classical inverted index is replaced by a semantic index (SI). The dependency parser is utilised for this task to capture the semantic structure and sentence context of a video description.

The dependency parser is applied to parse the video description documents and store the syntactic information rather than single term, where each token represent meaningful role such as subject, verb, object, *etc.* Initially, the documents' contents are tokenised and then a lemmatisation process is applied for each token, where the vocabulary and morphological analysis of individual words are used in order to get rid of inflectional endings only and return base form, called lemma. Next, the semantic inverted index is creating by using spacy's dependency parser.¹ The single term of inverted index is replaced with the extracted dependency triples which are simply of the form (lemma, relation, lemma of the head), as shown in Figure 7.2. The frequency of each triple is counted for each document, and those values are stored in the index.

Additionally, to account for cases where the user query is not a complete sentence (*e.g.* phrase or keyword), another index with dependency labels, which are simply doubles of the form (lemma, relation), but no head of the relation is created. This is to capture a case, such

¹The spacy's API: <https://spacy.io>

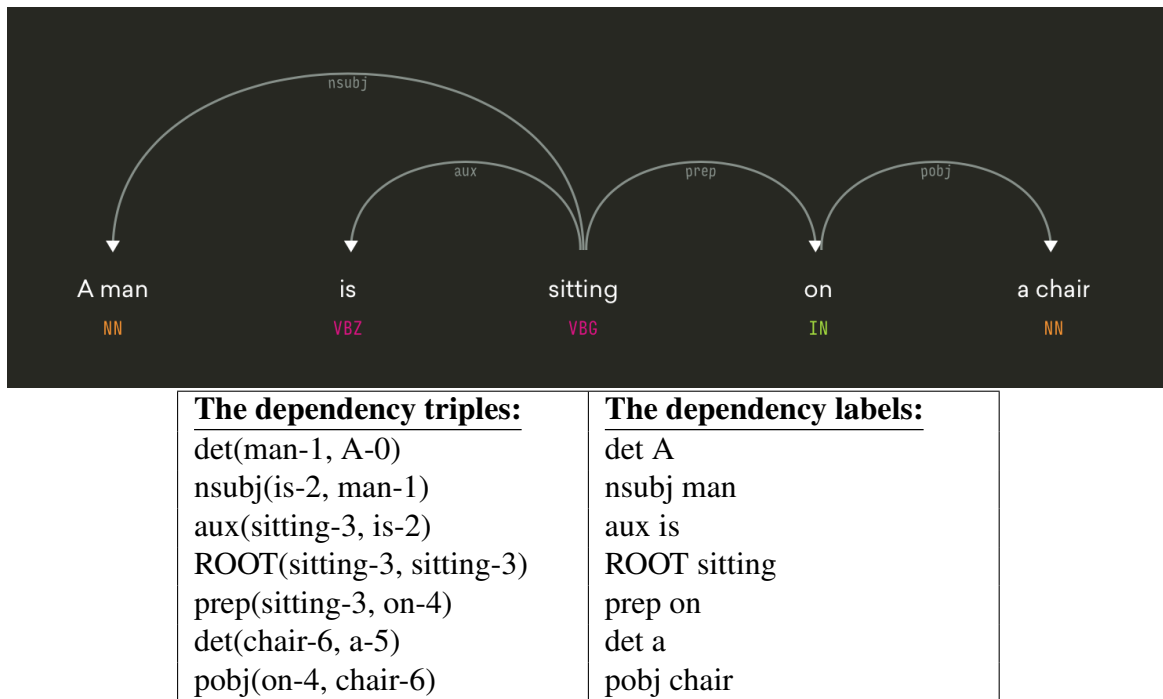


Fig. 7.2 A visualisation of dependency parse tree for the sentence ‘A man is sitting on a chair’ associated with syntactic relations between tokens using Spacy’s dependency parser.

as if a query is ‘woman is walking’ and there is no document in which satisfies this query; in this case all documents where woman is doing something will be retrieved. The key idea for a semantic inverted index is that, by using dependency triples and labels, we are capturing context and semantics structure of documents and queries, and ranking is basically done by frequencies of that context, rather than frequencies of individual words.

7.3.2 On-line Searching and Ranking the Results

Given a system query, it will be processed by either VSM or SI. For VSM the query is converted into a weighted terms vector of *tf.idf*, as in Equation 7.1. Once the query vector is created, it will be compared against all vectors in the database, using cosine similarity, which measures the angle between the query vector and all vectors in the database (Huang, 2008). Given two vectors \vec{d}_j and \vec{q} , their cosine similarity calculated by:

$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^N w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^N w_{i,j}^2} \sqrt{\sum_{i=1}^N w_{i,q}^2}}, \quad (7.3)$$

The cosine similarity is non-negative and bounded within $[0, 1]$. Where the documents are completely different the result is 0, and where two documents are identical the result is 1.

For semantic indexing (SI), the given query is parsed to extract the syntactic entries (dependency triples and labels). Then, following the *tf-idf* weighting approach to retrieve relevant documents, the *tf-idf* scores are calculated for each indexed document where the query entry is exists using Equation 7.1 (Ramos et al., 2003). The final document score is estimated by taking the weighted sum of both dependency triples and dependency labels as follows:

$$sum = triples_{score} + 0.25 * labels_{score} \quad (7.4)$$

The label's weight here are chosen arbitrarily, and the goal was to ensure that the labels score don't have more effect than the triples score. Finally, the documents are sorted in decreasing order by their weighted sum, the higher the weighted sum; the higher the relevance to the user query.

7.4 Experiments and results

This section presents an evaluation of the text-based human action video retrieval system proposed in this chapter. As there is no public dataset available for this task, we tested the framework on our dataset, the NLDHA dataset introduced in Chapter 2. A direct comparison between the performance of our framework and the work proposed in Khan et al. (2015) was not possible due to the limited set of actions identified in the latter study, which included only five actions (stand, walk, sit, run, wave). To assess the performance, two systems were implemented: the baseline used the vector space model, and the second system implemented the semantic indexing. For this experiment, the dataset contained 120 video clips coupled with their automatic textual descriptions. Due to the limited size of the dataset, for each class two queries are designed, forming a total of 24 different textual queries. For each action class the queries were designed to be a phrase (depict single action, *e.g.* man is running, woman is walking) and a complete sentences (which might include actions with either spatial or temporal relations between interacting objects, *e.g.* a man is sitting on a chair, a woman is driving a car on the road).

7.4.1 Evaluation scheme

As the relevant video clips for every query have already been defined from existing classes, following the classic information retrieval system, the list of relevant results is evaluated according to average precision and average recall. Given a textual query Q_i from set of two

queries $N_q = 2(i = 1, 2)$, while the dataset contains 12 different action classes $c = 12(j = 1, \dots, c)$, to evaluate the performance the following quantity should be calculated:

- TP_i : representing the number of retrieved videos that belong to the same human action class j as textual query Q_i
- FP_i : representing the number of retrieved videos that do not belong to the same human action class j as textual query Q_i
- TN_i : represent the number of non-retrieved videos that do not belong to the same human action class j as textual query Q_i
- FN_i : represent the number of non-retrieved videos that belong to the same human same action class j as textual query Q_i .

The average precision and recall are calculated over total number of queries as follows:

$$AP_j = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{TP_i}{TP_i + FP_i} \quad (7.5)$$

$$AR_j = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{TP_i}{TP_i + FN_i} \quad (7.6)$$

Finally, using the results of Equation 7.5 and Equation 7.6, the overall average precision and recall are calculated over the total number of human action classes:

$$\text{overall average precision} = \frac{1}{c} \sum_{j=1}^c AP_j \quad (7.7)$$

$$\text{overall average recall} = \frac{1}{c} \sum_{j=1}^c AR_j \quad (7.8)$$

7.4.2 Results

Table 7.1 shows the average precision and recall measures computed from all 12 action classes for two standard text retrieval systems (VSM and SI) with term-weighting techniques *tf.idf*. It can be observed that the SI outperforms the classical VSM for many classes. This superiority stems from the fact that similarity matching between the query and indexed documents considers the semantic correlations between terms and sentence structure. The efficiency of semantic matching is clearly noticeable for some classes, such as DriveCar, AnswerPhone and Eat, whereas the co-occurrence of salient objects ('phone', 'car' and 'dining table') with specific scene setting ('road', 'living room' and 'restaurant') enhances the retrieval of relevant video clips. However, the VSM retrieve the videos based on individual

Query category	VSM-tf.idf		SI-tf.idf	
	AP	AR	AP	AR
AnswerPhone	0.80	0.40	1.00	0.50
DriveCar	0.40	0.20	0.90	0.45
Eat	0.60	0.30	0.90	0.45
FightPerson	0.40	0.20	0.60	0.30
GetOutCar	0.70	0.35	0.80	0.40
HandShake	0.60	0.30	0.70	0.35
HugPerson	0.70	0.35	0.60	0.30
Kiss	0.60	0.30	0.60	0.30
Run	0.80	0.40	0.80	0.40
SitDown	0.40	0.20	0.60	0.30
SitUp	0.40	0.20	0.60	0.30
StandUp	0.30	0.15	0.40	0.20
Overall average	0.56	0.28	0.71	0.35

Table 7.1 The text-based video retrieval results using two systems the VSM and SI using set of two queries for each of 12 human action classes from the NLDHA dataset.

keywords regardless of their role in the sentence which was the main reason for failed retrieval.

Furthermore, the SI retrieval system is successfully capture the dependency between terms in query and in the indexed documents and assign specific meaning to each word, for example, nouns (car, person), verbs (walk, drive), prepositions (on, in) or adverbs (towards, away). For example, for query ‘a man is driving a car’ the classical VSM retrieved the documents that contain these individual terms ‘man’, ‘drive’ and ‘car’ and maximises the similarity scores based on $tf - idf$ regardless of their role in the sentence. However, the SI which is based on a parsing-based semantic match retrieved only the documents that semantically satisfy the query where the ‘man’ occurs as a subject doing a ‘drive’ action with a ‘car’ object.

The SI retrieval system was able to identify the relevant video for a user query that specified the actions in chronological order. In other words, incorporating temporal relations into video descriptions facilitated such retrieval of user queries and constructed a time dependency between actions performed in the video clip. Figure 7.3 shows an example of a user query and the top five retrieval results, with the user query ‘a man is driving a car. Next, he is getting out the car’. For this specific query the occurrence of dependency triple (‘next’, ‘advmod’, ‘getting’) improved the quality of retrieved results. The top two videos correctly depict both actions – ‘drive a car’ and ‘get out a car’ – in the same order, as required by the user. The other retrieved videos can be described as somewhat relevant to the user query as

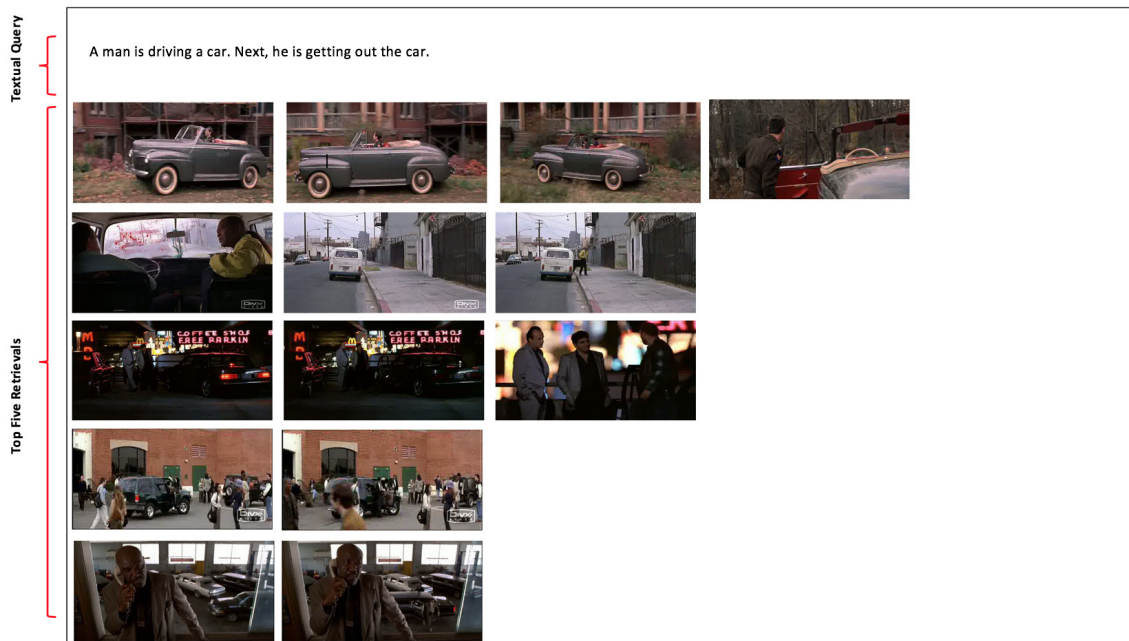


Fig. 7.3 The retrieved video clips samples from a ‘a man is driving a car. Next, he is getting out the car’ query. The top row is the user textual query. Followed by five top relevant retrievals that depict the main actions in the query (drive a car, get out a car) using SI.

they satisfy the role assigned to each word in the query, where they show either the drive action or the get out action.

Additionally, the presented retrieval system SI was able to satisfy the user query when simultaneous actions are required. This can be achieved by utilising the temporal relation of actions performed at the same time. For example, ‘a man is sitting, while a woman is standing’. Incorporating spatial relations into automatic video descriptions provides a user with a wider range of options, enabling them to enquire about the spatial layout between interacting objects in video clips. The user query can be accompanied with a variety of spatial relations, such as topology relations (*e.g.* ‘a person is sitting on a chair’), direction relations (*e.g.* ‘a man is standing to the right of a woman’) or distance relations (*e.g.* ‘a man is walking towards a car’). Figure 7.4 shows some examples of a user query with spatial relations and the retrieved videos.

Searching for individual concepts such as action class or object class is done efficiently by both VSM and SI systems. However, searching for phrases or new terms not previously present in the database documents requires more advanced representation that takes into account the semantic correlation between terms and concepts. For instance, the occurrence of some patterns of words provides a strong clue as to the likely occurrence of others. For example, if the query has ‘automobile moving’, the positive retrieved videos will show both

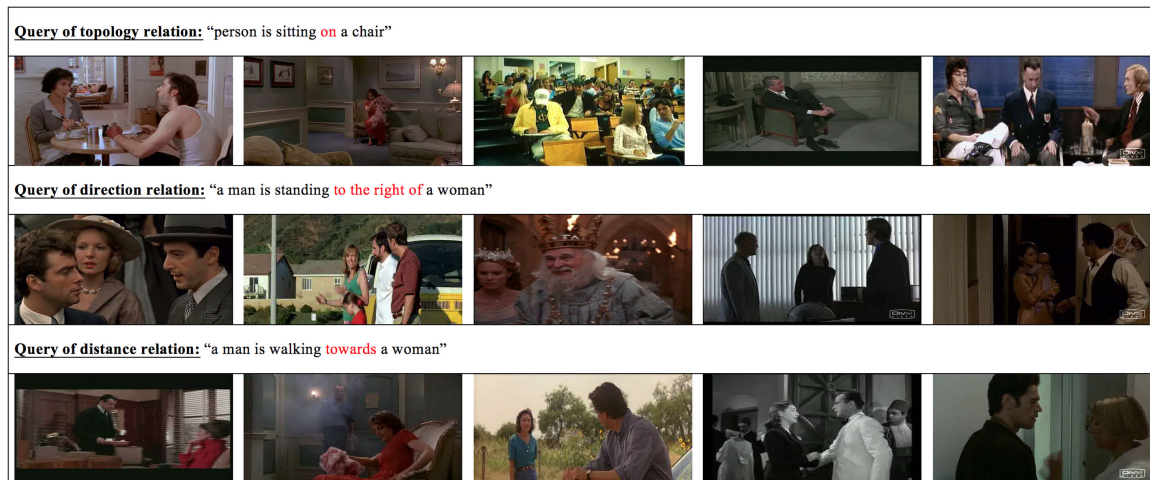


Fig. 7.4 The retrieved video clip samples from different queries that involved spatial relations.

car and road, as they are semantically correlated, providing a clue to the main action theme; simple term retrieval by exact lexical comparison would fail as neither 'automobile' or 'moving' are present in the database. To achieve such performance synonymy and polysemy should be considered and one possible solution is incorporating latent semantic analysis (LSA).

7.5 Conclusion

This chapter has presented a preliminary experiment of a text-based video retrieval framework that utilises the textual descriptions generated in previous chapters. The retrieval framework was implemented using two models, the space vector model and semantic indexing. Once the automatic description has been indexed, the similarity between the user query and all indexed documents is calculated, and for the top five similar video descriptions the corresponding video clips will be retrieved as relevant. Results from experimental evaluation using average precision and recall show that the system with semantic indexing was able to efficiently retrieve relevant human action video clips using their natural language descriptions, whether the query was a concept or complete text. However, intensive evaluation and analysis was difficult due to the limited size of the dataset. Although the proposed system shows promising results, especially when spatial and temporal relations are included in the user query, there is much more room for improvement especially when dealing with words that were not observed before or not in the list (*e.g.* synonymy and polysemy). For this purpose the WordNET (Ahsae et al., 2014), Latent semantic analysis (LSA) (Shen et al., 2014), Latent

dirichlet allocation (LDA) (Biggers et al., 2014) or SenticNet (Dashtipour et al., 2016; Tran et al., 2016) can be utilised.

Chapter 8

Conclusion

Describing human activities of video stream in natural language is in demand for various video analysis applications. The generation of textual description of human activities task seeks to translate prominent visual information from video into textual descriptions, specifically focusing on parsing semantic video content that can capture and summarise the main aspects of the human activities shown that demand detecting, tracking the interacting objects. It is a fundamental task that helps with understanding the spatial and temporal relations between objects over video sequence that cannot be achieved by processing individual frame using image-processing approaches. Accurate content video representation is a core step for numerous applications including generation of textual video description. The semantic video content was parsed in this thesis by extraction of spatio-temporal entities' segments, the identification of their visual attributions, and the formulation of their spatial and temporal relations. Finally, the extracted information was expressed using natural language generation techniques.

This thesis is concerned with the generation of natural language description of human activities in video stream, which can be used for video indexing, retrieval and summarisation applications. Initially, hand annotations were generated for relatively long videos which consisted of 120 manually selected from Hollywood2 dataset (see Chapter 2). Analysis of this corpus presents insights into human interests and thoughts while watching videos. For automatic video descriptions, initially, a spatio-temporal human body segmentation was proposed to identify the representative features and extract the meaningful content of the video content (see Chapter 3). Consequently, an action recognition framework that is able to identify the action performed by human subjects during video sequences was developed (see Chapter 4). The spatial and temporal relations were formalised between extracted segments to describe the changes of relations between two objects over the time domain (see Chapter 5). This visual information was translated into textual descriptions using a template-based

approach. The automatic descriptions are compared and evaluated using ROUGE scores and human judgment evaluation (see Chapter 6). Finally, to verify the efficiency of automatic descriptions, the human action retrieval task was selected as an application, using classical text-based retrieval approaches (see Chapter 7).

The main motivation of this research stems from two parts. First, a wide range of video data have been recorded every day without processing, which causes potential problems for content management systems that serve user needs. In order to solve this problem, many approaches have been introduced to represent, index and retrieve video data. However, a big challenge still exists to access semantic video content efficiently and to aid the end user to gain relevant information to their interests. Second, representing video data over the time domain provides a unique clues for analyses. The majority of video activities can be identified through the sequence of frames rather than processing individual frames. Object movement, non-rigid motion changes, spatial and temporal relations between objects in the scene are all examples of information that cannot be captured by processing video as individual frames. All these observations inspired this thesis: to study the accurate video content representation using meaningful segments and identify the spatial and temporal relations between these segments, and the translation of these visual information into textual form to facilitate the indexing and retrieval systems.

However, the biggest barrier to video description as a research area is the large numbers of actions and objects that can occur within and across the video frames, and the scarcity of training data. Moreover, the complexity associated with video recordings, such as lighting, occlusion and camera movement, contribute to making the video processing a more challenging task. In order to adapt such video descriptions approach to real-world technology, powerful object detectors and action classifiers need to be proposed that can be used to enhance the quality of video description. Additionally, a variety of pre-processing techniques should be applied to video content to facilitate detection, segmentation and description tasks, such as video stabilisation.

8.1 Original Contributions

1. **Corpus Generation and Analysis:** In order to identify the visual features that are usually verbalised by humans to describe human activities in video data, a new corpus was generated. The corpus consist of human annotations for 120 video clips that are manually selected from 12 different classes from the Hollywood2 dataset. These videos were chosen based on two main criteria – the number of camera shots and the variety of human actions performed in the video – in order to explore spatial and

temporal relationships between individual shot components, and to produce a story for the complete video clip. Annotations were made manually by 12 participants in two ways: a title, which consists of a single phrase or sentence (where a specific topic or primary concept is outlined), and a full textual description using a number of sentences (providing an extensive description of the visual scene).

2. **Spatio-temporal Human Body Segmentation:** To obtain an accurate and compact description of an input video, this description should be based on semantic visual content, taking into account the temporal dimension. Based on the fact that a human figure is the most important video component over a temporal dimension, extracting and describing their body volume over the duration of the video sequence leads to a comprehensive and compact representation. In this thesis, a novel approach was presented where human body regions are extracted from a video sequence. The approach detects and segments human body volumes by using joint embedding of parts and pixels, which utilises the advances of low-level image cues and high-level part detectors information. The appearance and shape models are learned for extracted segments to track them automatically across frame sequence and identify the foreground objects.
3. **Human Action Recognition Framework:** Once the human body segments are determined, the performed actions must be identified. For this task a new action recognition framework was implemented, where the video representation is improved using extracted spatio-temporal human regions combined with the extended spatio-temporal locality-constrained linear coding (LLC) technique. The using of spatio-temporal human regions improves the feature extraction by focusing on accurate position of actors and enables the action recognition to be applied for multiple regions simultaneously. Additionally, the LLC coding technique successfully represents video content with less coding than the original extracted set of features, which assists in reducing processing time and storage space.
4. **Extraction of Qualitative Spatial and Temporal Relations:** Spatial and temporal relations of prominent objects play a vital role in describing videos' semantic content. To the extent, an approach was designed for formalising the spatial and temporal relations between interacting objects using their spatio-temporal object bounding boxes and the intervals that result from aligning them to the video frames and applying AngledCORE-9 representation. By measuring nine cores, spatial information about two objects can be obtained, such as their topology, direction, relative size and the distance between them. Temporal changes can also be extracted from these cores and their associated intervals by processing a sequence of video frames. The approach

was able to detect spatial changes that occurred over time, promoting good semantic understanding of the video content.

5. **Generation of Textual Video Descriptions:** A framework that produces textual descriptions of video, based on the semantic video content at shot-level, was implemented. Detected action classes are rendered as verbs, participant objects are converted to noun phrases (mainly subject for humans and object for other), visual properties of detected objects are rendered as adjectives and spatial relations between objects are rendered as prepositions. Further, in cases where no verb is assigned for a given track, a language model is used to infer a missing verb, aided by the detection of objects and scene settings. These HLFs are converted into textual descriptions using a template-based approach. Paraphrasing of the resulting multi-sentences shot-based descriptions is introduced to create compact and coherent video descriptions.

8.2 Future Work

The previous section describes the main contributions towards the problem of describing human activities in video stream. However, there are some problems which still need to be addressed to improve the overall performance of the proposed framework. Future directions for this work are explored below.

1. The spatio-temporal human body segmentation approach was based on leverage of low-level image cues and high-level part detectors information (poselets detector) aiming for robustness from the image processing field to the video signal domain. However, although this approach is able to cope with partial occlusion thanks to the power of the poselets detector, full occlusion handling when the tracked object is completely not visible in the scene still needs to be addressed (Zhang et al., 2017). Further, to overcome the camera motion effect, the videos need to be stabilised so that as many high and low level features as possible are extracted. The optical flow of the surrounding entities within the flow field is not estimated simply when the background is not static and the camera is moving; therefore it is necessary to stabilise the camera prior to extracting those characteristics. For this purpose, a dominant motion compensation procedure can be used (Walha et al., 2015).
2. The coding stage used in action recognition framework can be improved by several ways. The k -means data structure can be replaced with different techniques such as kd -tree or cover-tree to enhance the search speed. The codebook learning can be optimised using other techniques, such as supervised or semi-supervised dictionary learning to enhance the performance. Moreover, the Euclidean distance used to measure the

- distance between the spatio-temporal descriptor and the basis codebook at coding stage can be replaced with the geodesic distance that is able to define the shortest path in the video signal (Wang et al., 2017).
3. The HLFs set can be extended to accommodate more visual attribution such as object colour and dress information. Further, spatial and temporal relations can be reliably estimated using bounding boxes and their intervals. The set of extracted relations in this thesis can be extended to accommodate more relations such as speed (Barbu et al., 2012), which can be estimated by comparing the magnitude of track velocity over frame sequence.
 4. The template-based approach is a robust NLG technique and was used to generate textual descriptions defined by rules and templates. However, we believe that learning these rules from data itself is a much more attractive approach. For any sufficiently rich domain, the statistical models can be used to learn the rules and template inspired by the language translation, where statistical machine translation has replaced rule-based approaches (Rohrbach et al., 2013).
 5. Video content includes variety of data such as visual, audio, and sometimes contains closed captions. Integration of these data can lead to improvement of retrieval accuracy. In practice, there are various methods that realise the improvement of retrieval accuracy by introducing a multimodal approach. Additionally, there is much more room for improving text-based video retrieval especially when query dealing with words that were not observed before or not in the list (*e.g.* synonymy and polysemy). For this purpose the WordNET (Ahsae et al., 2014), Latent semantic analysis (LSA) (Shen et al., 2014), Latent dirichlet allocation (LDA) (Biggers et al., 2014) or SenticNet (Dashtipour et al., 2016; Tran et al., 2016) can be utilised

References

- Abburu, S. (2010). Multi level semantic extraction for cricket video by text processing. *International Journal of Engineering Science and Technology*, 1(2):5377–5384.
- Ahsae, M. G., Naghibzadeh, M., and Naeini, S. E. Y. (2014). Semantic similarity assessment of words using weighted wordnet. *International Journal of Machine Learning and Cybernetics*, 5(3):479–490.
- Al Ghamdi, M., Al Harbi, N., and Gotoh, Y. (2012). Spatio-temporal video representation with locality-constrained linear coding. In *European Conference on Computer Vision*, pages 101–110. Springer.
- Al Harbi, N. and Gotoh, Y. (2013a). Action recognition: spatio-temporal human body region tracking approach. In *Proceedings of the Second Workshop on Recognition and Action for Scene Understanding, CAIP*.
- Al Harbi, N. and Gotoh, Y. (2013b). Spatio-temporal human body segmentation from video stream. In *International Conference on Computer Analysis of Images and Patterns*, pages 78–85. Springer.
- Al Harbi, N. and Gotoh, Y. (2015a). Describing spatio-temporal relations between object volumes in video streams. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Al Harbi, N. and Gotoh, Y. (2015b). A unified spatio-temporal human body region tracking approach to action recognition. *Neurocomputing*, 161:56–64.
- Al Harbi, N. and Gotoh, Y. (2016). Natural language descriptions of human activities scenes: Corpus generation and analysis. pages 39–47.
- Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.
- Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579.
- Arman, F., Depommier, R., Hsu, A., and Chiu, M.-Y. (1994). Content-based browsing of video sequences. In *Proceedings of the second ACM international conference on Multimedia*, pages 97–103. ACM.
- Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

- Bai, X., Wang, J., Simons, D., and Sapiro, G. (2009). Video snapcut: robust video object cutout using localized classifiers. *ACM Transactions on Graphics (TOG)*, 28(3):70.
- Barbu, A., Bridge, A., Burchill, Z., Coroian, D., Dickinson, S., Fidler, S., Michaux, A., Mussman, S., Narayanaswamy, S., Salvi, D., et al. (2012). Video in sentences out. *arXiv preprint arXiv:1204.2742*.
- Barbu, T. (2014). Pedestrian detection and tracking using temporal differencing and hog features. *Computers & Electrical Engineering*, 40(4):1072–1079.
- Bekios-Calfa, J., Buenaposada, J. M., and Baumela, L. (2011). Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):858–864.
- Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*.
- Bhat, S. A., Sardesai, O. V., Kunde, P. P., and Shirodkar, S. S. (2014). Overview of existing content based video retrieval systems. *International Journal of Advanced Engineering and Global Technology*, 2.
- Biggers, L. R., Bocovich, C., Capshaw, R., Eddy, B. P., Etkorn, L. H., and Kraft, N. A. (2014). Configuring latent dirichlet allocation based feature location. *Empirical Software Engineering*, 19(3):465–500.
- Bilen, H., Namboodiri, V. P., and Van Gool, L. (2011). Action recognition: A region based approach. In *Proceedings of IEEE Workshop on Applications of Computer Vision*, pages 294–300.
- Bin, Y., Yang, Y., Shen, F., Xu, X., and Shen, H. T. (2016). Bidirectional long-short term memory for video description. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 436–440. ACM.
- Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE.
- Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM.
- Bolle, R. M., Yeo, B.-L., and Yeung, M. M. (2010). Video query: Research directions. *IBM Journal of Research and Development*, 42(2):233–252.
- Bouchrika, I., Carter, J. N., Nixon, M. S., Morzinger, R., and Thallinger, G. (2010). Using gait features for improving walking people detection. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3097–3100. IEEE.
- Bourdev, L., Maji, S., Brox, T., and Malik, J. (2010). Detecting people using mutually consistent poselet activations. In *European conference on computer vision*, pages 168–181. Springer.

- Bourdev, L. and Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372. IEEE.
- Bouwman, T. (2011). Recent advanced statistical background modeling for foreground detection—a systematic survey. *Recent Patents on Computer Science*, 4(3):147–176.
- Bouwman, T. and Zahzah, E. H. (2014). Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34.
- Brendel, W., Fern, A., and Todorovic, S. (2011). Probabilistic event logic for interval-based event recognition. In *Proceedings of CVPR*, pages 3329–3336.
- Brox, T. and Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In *Computer Vision—ECCV 2010*, pages 282–295. Springer.
- Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Cai, X., Nie, F., and Huang, H. (2013). Multi-view k-means clustering on big data. In *IJCAI*, pages 2598–2604.
- Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.
- Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The ami meeting corpus: a pre-announcement. In *Proceedings of the Second international conference on Machine Learning for Multimodal Interaction*, MLMI’05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.
- Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.
- Chang, C., Gorissen, B., and Melchior, S. (2011). Fast oriented bounding box optimization on the rotation group $SO(3, \mathbb{R})$. *ACM Transactions on Graphics*, 30(5).
- Chang, S.-F., Ma, W.-Y., and Smeulders, A. (2007). Recent advances and challenges of semantic image/video search. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1205. IEEE.
- Chang, S.-F., Manmatha, R., and Chua, T.-S. (2005). Combining text and audio-visual features in video indexing. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP’05). IEEE International Conference on*, volume 5, pages v–1005. IEEE.
- Cheung, S.-S. and Zakhor, A. (2003). Efficient video similarity measurement with video signature. *IEEE Transactions on Circuits and Systems for video Technology*, 13(1):59–74.

- Cho, M., Sun, J., Duchenne, O., and Ponce, J. (2014). Finding matches in a haystack: A max-pooling strategy for graph matching in the presence of outliers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2083–2090.
- Choi, S. E., Lee, Y. J., Lee, S. J., Park, K. R., and Kim, J. (2011). Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6):1262–1281.
- Chong, W., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.
- Cohn, A. and Hazarika, S. (2001). Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1):1–29.
- Cohn, A., Renz, J., and Sridhar, M. (2012). Thinking inside the box: A comprehensive spatial representation for video analysis.
- Cohn, A. G., Magee, D. R., Galata, A., Hogg, D. C., and Hazarika, S. M. (2002). Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. In *International Conference on Spatial Cognition*, pages 232–248. Springer.
- Cohn, A. G., Renz, J., et al. (2008). Qualitative spatial representation and reasoning. *Handbook of knowledge representation*, 3:551–596.
- Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.
- Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.
- Dalal, N. (2006). *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.
- Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer.
- Das, P., Srihari, R. K., and Corso, J. J. (2013a). Translating related words to videos and back through latent topics. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 485–494. ACM.

- Das, P., Xu, C., Doell, R. F., and Corso, J. J. (2013b). A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641.
- Dashtipour, K., Hussain, A., Zhou, Q., Gelbukh, A., Hawalah, A. Y., and Cambria, E. (2016). Persent: A freely available persian sentiment lexicon. In *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings 8*, pages 310–320. Springer.
- Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM.
- Devi, O. R., Reddy, L., Prasad, E., Lasya, V. S., and Siddartha, V. S. (2015). Robust rule based local binary pattern for face recognition. *International Journal of Advanced Research in Computer Science*, 6(3).
- Dhamecha, T. I., Sharma, P., Singh, R., and Vatsa, M. (2014). On effectiveness of histogram of oriented gradient features for visible to near infrared face matching. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1788–1793. IEEE.
- Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72.
- Dubba, K., Cohn, A., and Hogg, D. (2010). Event model learning from complex videos using LP. In *Proceedings of ECAI*, volume 215, pages 93–98.
- Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM.
- Edke, V. D. and Kagalkar, R. M. (2016). Video object description of short videos in hindi text language. *International Journal of Computational Intelligence Research*, 12(2):103–116.
- Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.
- Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163.
- Elhabian, S. Y., El-Sayed, K. M., and Ahmed, S. H. (2008). Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent patents on computer science*, 1(1):32–54.

- Elkerdawi, S. M., Sayed, R., and ElHelw, M. (2014). Real-time vehicle detection and tracking using haar-like features and compressive tracking. In *ROBOT2013: First Iberian Robotics Conference*, pages 381–390. Springer.
- Ensafi, S., Lu, S., Kassim, A. A., and Tan, C. L. (2014). A bag of words based approach for classification of hep-2 cell images. In *Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 2014 1st Workshop on*, pages 29–32. IEEE.
- Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.
- Fejes, S. and Davis, L. S. (1998). What can projections of flow fields tell us about the visual motion. In *Computer Vision, 1998. Sixth International Conference on*, pages 979–986. IEEE.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.
- Feng, Y. and Lapata, M. (2010). Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839. Association for Computational Linguistics.
- Fernyhough, J., Cohn, A. G., and Hogg, D. C. (2000). Constructing qualitative event models automatically from video input. *Image and Vision Computing*, 18(2):81–103.
- Flood, B. J. (1999). Historical note: the start of a stop list at biological abstracts. *Journal of the American Society for Information Science*, 50(12):1066–1066.
- Forbus, K. (2008). Qualitative modeling. *Handbook of Knowledge Representation*, 3:361–393.
- Fowlkes, C., Martin, D., and Malik, J. (2003). Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–54. IEEE.
- Freeman, J. (1975). The modelling of spatial relations. *Computer graphics and image processing*, 4(2):156–171.
- Galata, A., Cohn, A., Magee, D., and Hogg, D. (2002). Modeling interaction using learnt qualitative spatio-temporal relations and variable length markov models. In *ECAI*, pages 741–745.
- Galata, A., Johnson, N., and Hogg, D. (1999). Learning behaviour models of human activities.

- Galata, A., Johnson, N., and Hogg, D. (2001). Learning variable-length Markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413.
- Gao, Y., Wang, T., Li, J., Du, Y., Hu, W., Zhang, Y., and Ai, H. (2007). Cast indexing for videos by ncuts and page ranking. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 441–447. ACM.
- Garg, A. and Choudhary, V. (2012). Facial expression recognition using principal component analysis. *Int. J. Sci. Eng. Res. Technol.*
- Gatt, A. and Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.
- Gavrila, D. M. and Munder, S. (2007). Multi-cue pedestrian detection and tracking from a moving vehicle. *International journal of computer vision*, 73(1):41–59.
- Ge, W., Collins, R. T., and Ruback, R. B. (2012). Vision-based analysis of small groups in pedestrian crowds. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):1003–1016.
- Geronimo, D., Lopez, A. M., Sappa, A. D., and Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1239–1258.
- Gilbert, A., Illingworth, J., and Bowden, R. (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proceedings of ICCV*, pages 925–931.
- Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., and Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: theory and practice. *Archives of General Psychiatry*, 38(4):408–413.
- Grundmann, M., Kwatra, V., Han, M., and Essa, I. (2010). Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2141–2148. IEEE.
- Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., and Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 14th International Conference on Computer Vision (ICCV-2013)*, pages 2712–2719, Sydney, Australia.
- Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE.
- Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014a). Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.
- Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014b). Creating Summaries from User Videos. In *ECCV*.

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.
- Hanckmann, P., Schutte, K., and Burghouts, G. J. (2012). Automated textual descriptions for a wide range of video events with 48 human actions. In *European Conference on Computer Vision*, pages 372–380. Springer.
- Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.-M., Hicks, S. L., and Torr, P. H. (2016). Struck: Structured output tracking with kernels. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2096–2109.
- Haritaoglu, I., Harwood, D., and Davis, L. S. (2000). W 4: Real-time surveillance of people and their activities. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):809–830.
- Hayashi, J., Yasumoto, M., Ito, H., and Koshimizu, H. (2001). Method for estimating and modeling age and gender using facial image processing. In *Virtual Systems and Multimedia, 2001. Proceedings. Seventh International Conference on*, pages 439–448. IEEE.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916.
- Heikkilä, J. and Silvén, O. (2004). A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, 22(7):563–570.
- Hong, W.-B., Lee, C.-P., and Chen, C.-W. (2001). Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering*, 4(3):183–192.
- Hossen, M. K. and Tuli, S. H. (2016). A surveillance system based on motion detection and motion estimation using optical flow. In *Informatics, Electronics and Vision (ICIEV), 2016 5th International Conference on*, pages 646–651. IEEE.
- Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.
- Huang, Y., Wu, Z., Wang, L., and Tan, T. (2014). Feature coding in image classification: A comprehensive study. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):493–506.
- Isard, M. and MacCormick, J. (2001). Bramble: A bayesian multiple-blob tracker. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 34–41. IEEE.
- Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493.
- Jiang, W., Chan, K. L., Li, M., and Zhang, H. (2005). Mapping low-level features to high-level semantic concepts in region-based image retrieval. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 244–249. IEEE.

- Jiang, Z., Zhang, G., and Davis, L. S. (2012). Submodular dictionary learning for sparse coding. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3418–3425. IEEE.
- Jones, M. J. and Snow, D. (2008). Pedestrian detection using boosted features over many frames. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.
- Khan, M., Nawab, R., and Gotoh, Y. (2012). Natural language descriptions of visual scenes - corpus generation and analysis. In *EACL 2012 workshop, Joint workshop of ESIRMT and HYTRA*.
- Khan, M. U. G., Al Harbi, N., and Gotoh, Y. (2015). A framework for creating natural language descriptions of video streams. *Information Sciences*, 303:61–82.
- Khan, M. U. G. and Gotoh, Y. (2012). Describing video contents in natural language. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 27–35. Association for Computational Linguistics.
- Kim, J. and Grauman, K. (2011). Boundary preserving dense local regions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1553–1560. IEEE.
- Kovashka, A. and Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proceedings of CVPR*, pages 2046–2053.
- Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.
- Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics.
- Kwon, Y. H. and da Vitoria Lobo, N. (1999). Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21.
- Laokulrat, N., Okazaki, N., and Nakayama, H. (2017). Generating video description using rnn with semantic attention.
- Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of CVPR*, pages 1–8.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

- Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of CVPR*, pages 3361–3368.
- Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2006). Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808.
- Lee, Y. J., Kim, J., and Grauman, K. (2011). Key-segments for video object segmentation. In *Proceedings of ICCV*.
- Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885. IEEE.
- Leopold, J. L., Sabharwal, C. L., and Ward, K. J. (2015). Spatial relations between 3d objects: The association between natural language, topology, and metrics. *Journal of Visual Languages & Computing*, 27:29–37.
- Lew, M. S., Sebe, N., Djeraba, C., and Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19.
- Li, H., Tang, J., Wu, S., Zhang, Y., and Lin, S. (2010). Automatic detection and analysis of player action in moving background sports video sequences. *IEEE transactions on circuits and systems for video technology*, 20(3):351–364.
- Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics.
- Ligozat, G. (1998). Reasoning about cardinal directions. *Journal of Visual Languages & Computing*, 9(1):23–44.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.
- Lin, D., Fidler, S., Kong, C., and Urtasun, R. (2014). Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2657–2664.
- Lindquist, K. A. and Barrett, L. F. (2008). Constructing emotion the experience of fear as a conceptual act. *Psychological science*, 19(9):898–903.
- Lipton, A. J., Fujiyoshi, H., and Patil, R. S. (2005). Moving target classification and tracking from real-time video. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pages 8–14. IEEE.
- Liu, C., Wu, X., and Jia, Y. (2016). A hierarchical video description for complex activity understanding. *International Journal of Computer Vision*, 118(2):240–255.

- Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos “in the wild”. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE.
- Liu, L., Wang, P., Shen, C., Wang, L., van den Hengel, A., Wang, C., and Shen, H. T. (2017). Compositional model based fisher vector coding for image classification. *IEEE transactions on pattern analysis and machine intelligence*.
- Liu, Y., Liu, Y., and Chan, K. C. (2008). Multiple video trajectories representation using double-layer isometric feature mapping. In *2008 IEEE International Conference on Multimedia and Expo*, pages 129–132. IEEE.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110.
- Maire, M., Yu, S. X., and Perona, P. (2011). Object detection and segmentation from joint embedding of parts and pixels. In *Proceedings of ICCV*.
- Makadia, A., Pavlovic, V., and Kumar, S. (2008). A new baseline for image annotation. In *European conference on computer vision*, pages 316–329. Springer.
- Mandellos, N. A., Keramitsoglou, I., and Kiranoudis, C. T. (2011). A background subtraction algorithm for detecting and tracking vehicles. *Expert Systems with Applications*, 38(3):1619–1631.
- Markkula, M. and Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1(4):259–285.
- Marques, O. and Furht, B. (2002). *Content-based image and video retrieval*, volume 21. Springer Science & Business Media.
- Marszalek, M., Laptev, I., and Schmid, C. (2009a). Actions in context. In *Proceedings of CVPR*.
- Marszalek, M., Laptev, I., and Schmid, C. (2009b). Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE.
- Matuszek, C., Herbst, E., Zettlemoyer, L., and Fox, D. (2013). Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer.
- Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., and Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.
- Mondal, A., Ghosh, S., and Ghosh, A. (2016). Efficient silhouette-based contour tracking using local information. *Soft Computing*, 20(2):785–805.

- Morariu, V. I. and Davis, L. S. (2011). Multi-agent event recognition in structured scenarios. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3289–3296. IEEE.
- Natsev, A. P., Naphade, M. R., and Tešić, J. (2005). Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 598–607. ACM.
- Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.
- Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *IEEE transactions on pattern analysis and machine intelligence*, 22(8):831–843.
- Over, P., Smeaton, A., and Kelly, P. (2007). The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the international workshop on TRECVID video summarization*, page 15. ACM.
- Packer, B., Saenko, K., and Koller, D. (2012). A combined pose, object, and feature model for action understanding. In *Proceedings of CVPR*, pages 1378–1385.
- Paez, F., Vanegas, J. A., and Gonzalez, F. A. (2013). An evaluation of nmf algorithm on human action video retrieval. In *Image, Signal Processing, and Artificial Vision (STSIVA), 2013 XVIII Symposium of*, pages 1–4. IEEE.
- Paisitkriangkrai, S., Shen, C., and Zhang, J. (2008). Fast pedestrian detection using a cascade of boosted covariance features. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1140–1151.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Paragios, N. and Deriche, R. (2000). Geodesic active contours and level sets for the detection and tracking of moving objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(3):266–280.
- Parameswaran, A., Kaushik, R., and Arasu, A. (2012). Efficient parsing-based keyword search over databases. Technical report, Stanford InfoLab.
- Patron-Perez, A., Marszalek, M., Zisserman, A., and Reid, I. (2010). High five: Recognising human interactions in TV shows. In *Proceedings of BMVC*.
- Paul, N., Singh, A., Midya, A., Roy, P. P., and Dogra, D. P. (2017). Moving object detection using modified temporal differencing and local fuzzy thresholding. *The Journal of Supercomputing*, 73(3):1120–1139.
- Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732.

- Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.
- Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.
- Price, B. L., Morse, B. S., and Cohen, S. (2009). Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 779–786. IEEE.
- Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.
- Putthividhy, D., Attias, H. T., and Nagarajan, S. S. (2010). Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3408–3415. IEEE.
- Ramakrishnan, S. K., Ravindran, S. K., and Mittal, A. (2017). Comal tracking: Tracking points at the object boundaries. *arXiv preprint arXiv:1706.02331*.
- Ramanathan, N., Chellappa, R., and Biswas, S. (2009). Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144.
- Ramezani, M. and Yaghmaee, F. (2016). A review on human action analysis in videos for retrieval applications. *Artificial Intelligence Review*, 46(4):485–514.
- Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.
- Randell, D., Cui, Z., and Cohn, A. (1992). A spatial logic based on regions and connection. *Proceedings of KR*, 92:165–176.
- Rautaray, S. S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54.
- Ravindran, S. K. and Mittal, A. (2016). Comal: Good features to match on object boundaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345.
- Reddy, K. K. and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981.
- Regneri, M., Rohrbach, M., Wetzell, D., Thater, S., Schiele, B., and Pinkal, M. (2013). Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.
- Ren, X. and Malik, J. (2007). Tracking as repeated figure/ground segmentation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

- Rodden, K., Basalaj, W., Sinclair, D., and Wood, K. (2001). Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 190–197. ACM.
- Rodriguez, M. D., Ahmed, J., and Shah, M. (2008a). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.
- Rodriguez, M. D., Ahmed, J., and Shah, M. (2008b). Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *Proceedings of CVPR*.
- Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., and Schiele, B. (2013). Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 433–440.
- Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., and Schiele, B. (2012). Script data for attribute-based recognition of composite activities. In *Computer Vision—ECCV 2012*, pages 144–157. Springer.
- Rother, C., Kolmogorov, V., and Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM.
- Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *Proceedings of CVPR*, pages 1234–1241.
- Sadeghi, M. A. and Farhadi, A. (2011). Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE.
- Salton, G. and Yang, C.-S. (1973). On the specification of term values in automatic indexing. *Journal of documentation*, 29(4):351–372.
- Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local SVM approach. In *Proceedings of ICPR*, volume 3, pages 32–36.
- Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019.
- Sharma, S. (2016). *Action Recognition and Video Description using Visual Attention*. PhD thesis, Master’s thesis, University of Toronto.
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 101–110. ACM.
- Shi, Y., Wan, Y., Wu, K., and Chen, X. (2017). Non-negativity and locality constrained laplacian sparse coding for image classification. *Expert Systems with Applications*, 72:121–129.

- Singh, S., Ren, W., and Singh, M. (2009). A novel approach to spatio-temporal video analysis and retrieval. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 106–115. Springer.
- Smeaton, A., Over, P., and Kraaij, W. (2009). High-level feature detection from video in TRECVID: a 5-year retrospective of achievements. *Multimedia Content Analysis*, pages 1–24.
- Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380.
- Smith, J. R., Basu, S., Lin, C.-Y., Naphade, M., and Tseng, B. (2002). Interactive content-based retrieval of video. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–I. IEEE.
- Sokeh, H., Gould, S., and Renz, J. (2013). Efficient extraction and representation of spatial information from video data. In *Proceedings of AAAI*, pages 1076–1082.
- Southey, T. and Little, J. J. (2007). Learning qualitative spatial relations for object classification. In *IROS 2007 Workshop: From Sensors to Human Spatial Concepts*. Citeseer.
- Sridhar, M., Cohn, A., and Hogg, D. (2008). Learning functional object categories from a relational spatio-temporal representation. In *Proceedings of ECAI*, pages 606–610.
- Sridhar, M., Cohn, A., and Hogg, D. (2010a). Discovering an event taxonomy from video using qualitative spatio-temporal graphs. In *Proceedings of ECAI*, pages 1103–1104.
- Sridhar, M., Cohn, A., and Hogg, D. (2011). From video to RCC8: exploiting a distance based semantics to stabilise the interpretation of mereotopological relations. In *International Conference on Spatial Information Theory*, pages 110–125.
- Sridhar, M., Cohn, A. G., and Hogg, D. C. (2010b). Relational graph mining for learning events from video. *STAIRS 2010*, pages 315–327.
- Sridhar, M., Cohn, A. G., and Hogg, D. C. (2010c). Unsupervised learning of event classes from video. In *Proceedings of AAAI*, pages 1631–1638.
- Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on*, volume 2. IEEE.
- Stauffer, C. and Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):747–757.
- Stella, X. Y. (2009). Angular embedding: from jarring intensity differences to perceived luminance. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2302–2309. IEEE.
- Sun, X., Ji, R., Yao, H., Xu, P., Liu, T., and Liu, X. (2008). Place retrieval with graph-based place-view model. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 268–275. ACM.

- Sun, X., Yao, H., Zhang, S., and Li, D. (2015). Non-rigid object contour tracking via a novel supervised level set model. *IEEE Transactions on Image Processing*, 24(11):3386–3399.
- Susan, S., Jain, A., Sharma, A., Verma, S., and Jain, S. (2015). Fuzzy match index for scale-invariant feature transform (sift) features with application to face recognition with weak supervision. *IET Image Processing*, 9(11):951–958.
- Tah, A., Roy, S., Das, P., and Mitra, A. (2017). Moving object detection and segmentation using background subtraction by kalman filter. *Indian Journal of Science and Technology*, 10(19).
- Tavanai, A., Sridhar, M., Gu, F., Cohn, A. G., and Hogg, D. C. (2013). Carried object detection and tracking using geometric shape models and spatio-temporal consistency. In *International Conference on Computer Vision Systems*, pages 223–233. Springer.
- Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., and Mooney, R. J. (2014). Integrating language and vision to generate natural language descriptions of videos in the wild. In *Coling*, volume 2, page 9.
- Tran, H.-N., Cambria, E., and Hussain, A. (2016). Towards gpu-based common-sense reasoning: using fast subgraph matching. *Cognitive Computation*, 8(6):1074–1086.
- Tsai, D., Flagg, M., Nakazawa, A., and Rehg, J. M. (2012). Motion coherent tracking using multi-label mrf optimization. *International journal of computer vision*, 100(2):190–202.
- Tsai, D., Flagg, M., and Rehg, J. M. (2010). Motion coherent tracking with multi-label MRF optimization. In *Proceedings of BMVC*.
- Txia, J.-D. and Huang, C.-L. (2009). Age estimation using aam and local facial features. In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IHH-MSP'09. Fifth International Conference on*, pages 885–888. IEEE.
- Van de Weghe, N., Kuijpers, B., Bogaert, P., and De Maeyer, P. (2005). A qualitative trajectory calculus and the composition of its relations. In *International Conference on GeoSpatial Semantics*, pages 60–76.
- Van Gemert, J. C., Geusebroek, J.-M., Veenman, C. J., and Smeulders, A. W. (2008). Kernel codebooks for scene categorization. In *European conference on computer vision*, pages 696–709. Springer.
- Vazquez-Reina, A., Avidan, S., Pfister, H., and Miller, E. (2010). Multiple hypothesis video segmentation from superpixel flows. In *Computer Vision–ECCV 2010*, pages 268–281. Springer.
- Veenman, C. J., Reinders, M. J., and Backer, E. (2001). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):54–72.
- Veltkamp, R., Burkhardt, H., and Kriegel, H.-P. (2013). *State-of-the-art in content-based image and video retrieval*, volume 22. Springer Science & Business Media.

- Viola, M., Jones, M. J., and Viola, P. (2003). Fast multi-view face detection. In *Proc. of Computer Vision and Pattern Recognition*.
- Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE.
- Viola, P., Jones, M. J., and Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161.
- Vondrick, C., Ramanan, D., and Patterson, D. (2010). Efficiently scaling up video annotation with crowdsourced marketplaces. In *European Conference on Computer Vision*, pages 610–623. Springer.
- Walha, A., Wali, A., and Alimi, A. M. (2015). Video stabilization with moving object detecting and tracking for aerial video surveillance. *Multimedia Tools and Applications*, 74(17):6745–6767.
- Wang, H., Ullah, M. M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *Proceedings of BMVC*.
- Wang, H., Zheng, X., and Xiao, B. (2015). Large-scale human action recognition with spark. In *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*, pages 1–6. IEEE.
- Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *Proceedings of CVPR*, pages 3360–3367.
- Wang, M., Hong, R., Li, G., Zha, Z.-J., Yan, S., and Chua, T.-S. (2012). Event driven web video summarization by tag localization and key-shot identification. *IEEE Transactions on Multimedia*, 14(4):975–985.
- Wang, W., Shen, J., Yang, R., and Porikli, F. (2017). A unified spatiotemporal prior based on geodesic distance for video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wang, Y. and Mori, G. (2009). Human action recognition by semilattent topic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(10):1762–1774.
- Weinland, D., Boyer, E., and Ronfard, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–7. IEEE.
- Wixson, L. (2000). Detecting salient motion by accumulating directionally-consistent flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):774–780.
- Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfnder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):780–785.
- Wu, L., Jin, R., and Jain, A. K. (2013). Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727.

- Wu, X., Xu, D., Duan, L., and Luo, J. (2011). Action recognition using context and appearance distribution features. In *Proceedings of CVPR*, pages 489–496.
- Xia, L. and Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of CVPR*, pages 2834–2841.
- Yadav, R. K., Sharma, S., and Verma, J. S. (2011). Deformation and improvement of video segmentation based on morphology using ssd technique. *International Journal of Computer Technology and Applications*, 2(5).
- Yan, R. and Hauptmann, A. G. (2007). A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484.
- Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE.
- Yang, J., Yu, K., and Huang, T. (2010). Supervised translation-invariant sparse coding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3517–3524. IEEE.
- Yang, M., Zhang, L., Yang, J., and Zhang, D. (2011a). Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 625–632. IEEE.
- Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011b). Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics.
- Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13.
- Yu, H. and Siskind, J. M. (2013). Grounded language learning from video described with sentences. In *ACL (1)*, pages 53–63.
- Yu, S., Tan, T., Huang, K., Jia, K., and Wu, X. (2009). A study on gait-based gender classification. *IEEE Transactions on image processing*, 18(8):1905–1910.
- Yu, S. X. and Shi, J. (2004). Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183.
- Zaslavskiy, M., Bach, F., and Vert, J.-P. (2010). Many-to-many graph matching: a continuous relaxation approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer.
- Zhai, X., Peng, Y., and Xiao, J. (2013). Cross-media retrieval by intra-media and inter-media correlation mining. *Multimedia systems*, 19(5):395–406.
- Zhang, C., Liu, J., Liang, C., Xue, Z., Pang, J., and Huang, Q. (2014). Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition. *Computer Vision and Image Understanding*, 123:14–22.

- Zhang, H. and Parker, L. (2011). 4-dimensional local spatio-temporal features for human activity recognition. In *International Conference on Intelligent Robots and Systems*, pages 2044–2049.
- Zhang, H. J., Wu, J., Zhong, D., and Smoliar, S. W. (1997). An integrated system for content-based video retrieval and browsing. *Pattern recognition*, 30(4):643–658.
- Zhang, X., Shiqiang, H., Zhang, H., and Xing, H. (2017). Full occlusion handling for pedestrian tracking via hybrid system. *Turkish Journal of Electrical Engineering & Computer Sciences*, 25(2).
- Zhang, Y., Liu, X., Chang, M., Ge, W., and Chen, T. (2012). Spatio-temporal phrases for activity recognition. In *Proceedings of ECCV*, pages 707–721.
- Zhao, T. and Nevatia, R. (2003). Bayesian human segmentation in crowded situations. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–459. IEEE.
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.

Appendix A

Identification of Additional Visual Attributions

As the interactions of ‘humans’ are usually the most essential and intriguing element of video content, a new approach is proposed to extract human body regions and their associated actions. In fact, the overarching objective of this thesis is to create a system of natural language description for human activity in video stream. First, the most valuable HLFs within the video content are identified by analysing and evaluating the annotations. Then, these HLFs are efficiently extracted via the use of video and image processing techniques. As a way to further enhance and develop the system, a number of extra HLFs will be highlighted and identified via off-the-shelf software packages. This section provides an overview of the methods used to identify age, gender, emotion and scene setting. In the case of gender, age, and emotion, the process is repeated for each human body track in each shot, whereas the scene setting can be identified by testing the first frame of each shot. Moreover, the face detector developed by [Viola et al. \(2003\)](#) will be employed during pre-processing stage as a way to highlight facial region within body tracks. Finally, the area being identified will be carefully cropped to make it easier to pick out key facial features. The following section outlines the identification steps for each HLF.

A.1 Gender Identification

In most cases, it is very easy to identify gender from a human face alone. The framework used to achieve this begins with scrutiny of the individual facial region. The facial features which can be used to discriminate male and female can take the form of many different facial features such as the shape and size of the nose, cheeks, chin, jaw, forehead, eyebrows,

lips, and more. Currently, there are two key methods used to determine gender within video content. The first analyses facial features and takes measurements of each one, in order to build up a picture of their position and size. For instance, the space between the eyes and mouth or the nose and mouth are calculated. The downside to this technique is that detecting facial parts is still challenging due to the variety of face poses. The second method employs low-level features from image pixel values. These low-level data include textural aspects, the coefficients of wavelet transformations, raw gray-scale pixel values, and histograms of facial gradients. Crucially, studies have shown that using low-level features is much more reliable, in this context, than using facial measurements. Refer to [Yu et al. \(2009\)](#) for an in-depth outline of these earlier studies on gender estimation.

This part will discuss the approach introduced by [Bekios-Calfa et al. \(2011\)](#) to estimate human gender. Refer to Figure A.1 to see that the process begins with a face detection framework. This is directly based on the framework constructed by Viola and Jones. Following this, an Oval Mask is combined with the input feed as a way to stop the background content from being a distraction. Next, Linear Discriminant Analysis (LDA) and Principal Components Analysis (PCA) features are identified from the training dataset. Lastly, trained image characteristics are linked to a Bayesian classifier associated with their true labels. Then, a preliminary test is conducted by feeding test image features through the Bayesian framework. The aim is to determine whether they match up with the trained image features previously linked to them and to produce a decision on what gender has been identified.

Table A.2 shows a confusion matrix of the gender identification process. The frontal faces were present in 369 of detected human body tracks in our corpus. It tends to be that estimating female gender is trickier than the male. This is due to a variety of factors such as the presence of cosmetics, complex hairstyles, and accessories like scarves, hats, and jewellery.

A.2 Age Identification

Over the years, there has been much debate over the best ways to successfully approximate age using only the faces in videos. The most popular method, however, focuses on two key phases – feature identification and feature classification. For instance, the facial features presented in earlier studies are grouped into one of three classifications: local features, global features and hybrid features. The local features include things like skin colour, hair colour, wrinkles, and geometric characteristics. For earlier investigations, these were a common way to determine relevant age groups. For example, the work of [Kwon and da Vitoria Lobo](#)

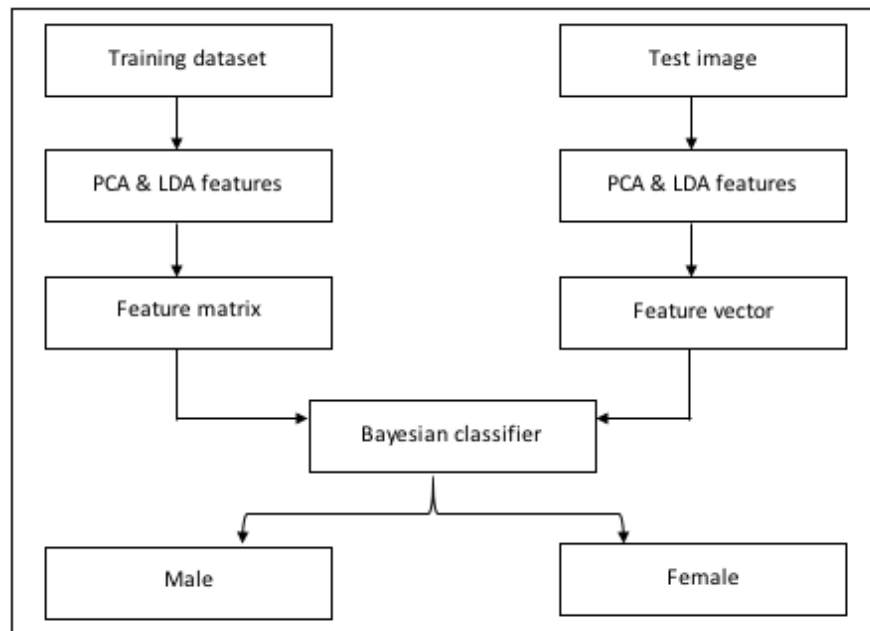


Fig. A.1 Gender classification approach proposed by Bekios-Calfa et al. (2011).

	Male	Female
Male	212	9
Female	23	125

Fig. A.2 Confusion matrix for gender identification. Columns show the ground truth, and rows indicate the classification results.

(1999) categorises faces as either senior adults, young adults, or babies. This is achieved by analysing the faces, in line with a distance ratio and any existing wrinkle features.

The approach proposed by Horng et al. (2001) scrutinises geometric features and the presence of wrinkles as a way to identify four key age categories. The researcher uses a Sobel filter to calculate the abundance and severity of wrinkles. This involves defining each individual wrinkle feature by measuring both density and depth. The Sobel edge magnitude calculation is later employed to determine the degree of inconsistency on the surface of the skin. According to Txia and Huang (2009), another technique is an age categorisation technique which combines wrinkle elements analysed by the Sobel filter with specific hair tone features. The areas targeted for extraction are chosen according to facial ‘highlights’ identified by the Active Appearance Model (AAM). In the Hayashi et al. (2001) study, a Digital Template Hough Transform (DTHT) technique is used to identify the wrinkle features. Once the targeted skin areas have been extracted according to skin tone, the wrinkle features

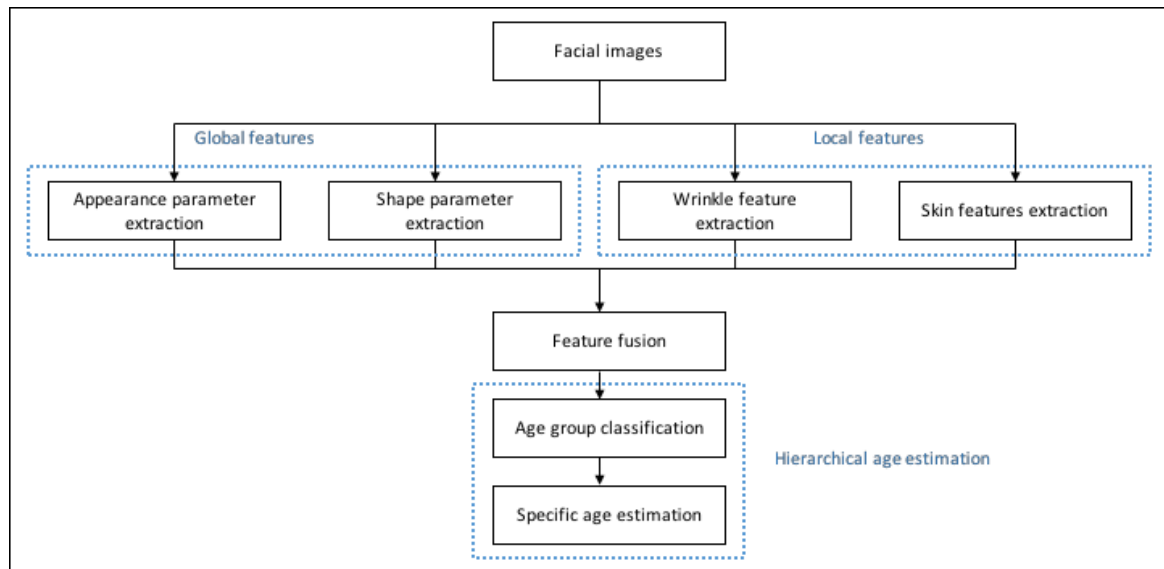


Fig. A.3 Age classification approach proposed by Choi et al. (2011).

are modelled via the DTHT method. Refer to Ramanathan et al. (2009) for a detailed review of all of the techniques outlined.

This section will describe the Hyper Features technique proposed by Choi et al. (2011) as a way to approximate age. Figure A.3 demonstrates that this system can be split into two key stages – feature extraction and age estimation. The feature extraction stage involves both global and local characteristics. This method employs the AAM as a way to approximate the age of global facial features, as it is a generative parametric framework which considers the appearance and shape of a face. Some local characteristics are specific skin or wrinkle features and they are handled as valuable biological quantities. A Gabor filter is used to investigate facial wrinkles, according to their directional properties. This necessitates the use of several different constraints, depending on the direction and density of the targeted area of the face.

While wrinkles are distinct and can be easily measured, the age of skin is much vaguer. It emerges in a random fashion and can appear more pronounced in some areas than in others. For most people, there is no even distribution of age across their features. This is why age approximation methods must employ very sophisticated tools to scrutinise and evaluate the condition of skin. For this study, skin features were detected using the Local Binary Pattern (LBP). As Ojala et al. (2002) explain, the technique identifies microstructures within the skin; things like smoother spots, ridges, lines, and blemishes. Lastly, the categorisation phase necessitates the use of a Support Vector Machine (SVM) for age group classification.

	Baby	Child	Young	Adult	Old
Baby	2	0	0	0	0
Child	0	13	2	0	0
Young	0	0	10	33	4
Adult	0	0	16	176	58
Old	0	0	0	10	45

Fig. A.4 Confusion matrix for age identification. Columns correspond to the ground truth, whereas the rows represent the classification results.

Thus far, the SVM process is the only one that has been carried out in this experiment as we aim to classify the age into five groups. It has successfully identified five age categories. They are baby (less than 2), child (2-10), young (11-24), adult (25-65), and old (over 65). Refer to Table A.4 for a confusion matrix for age identification in our corpus.

A.3 Facial Emotions Recognition

For human beings, facial expressions are an important part of conveying mood. They are our first response to the world around us and they help us to communicate. The biggest obstacle to facial expression research, however, is the fact that there is such a huge variety of potential expressions. Nevertheless, behavioural specialists believe that humans are born with the capacity to create five key expressions. These are seriousness, anger, surprise, sadness, and happiness. All other expressions are learned responses that we pick up as we grow and develop (Lindquist and Barrett, 2008). Bettadapura (2012) brings together some of the most important studies on this topic and uses the results to build up a picture of the way in which the field has evolved. For instance, the creation of new expression databases and processes, the construction of automatic face expression recognisers, and the move towards a formalised system of identification.

In this section we discuss a facial recognition system grounded in the use of PCA features proposed by Garg and Choudhary (2012). The employment of PCA techniques allows researchers to identify a subset of primary directions (or primary elements) within a series of targeted training faces. If faces are combined with these primary elements, the associated feature vectors are produced. Ordinarily, facial imagery is compared and contrasted by determining the Euclidean measurement between targeted vectors. Therefore, comparisons are conducted by measuring the space between the vectors.

The identified expressions are those highlighted by behavioural science – happiness, sadness, surprise, anger and seriousness. Within this particular data ‘happy’, ‘sad’, and

	Angry	Happy	Sad	Surprised	Serious	Neutral
Angry	45	0	15	6	0	0
Happy	0	87	4	10	0	0
Sad	0	0	59	4	0	9
Surprised	12	0	0	14	3	0
Serious	0	0	0	19	34	19
Neutral	0	0	1	0	1	27

Fig. A.5 Confusion matrix for human emotion recognition. Columns show the ground truth, and rows indicate the automatic recognition results.

‘angry’ were the most frequently occurring expressions. The happy expression was the most easily identified and, therefore, the most accurately estimated each time. In contrast, the emotion of ‘surprised’ was much harder to identify and wasn’t always picked up. Refer to Table A.5 for the confusion matrix for human emotion recognition results.

A.4 Scene Setting Identification

Scene identification is a key aspect of digital vision, as it enables a machine to make estimations based on contextual information. Over the years, there has been a remarkable degree of development across the field of object recognition. This is largely due to the increased accessibility of rich datasets such as ImageNet. Recently, the work of [Zhou et al. \(2014\)](#) proposes a contemporary scene-based system called ‘Places 205’ based on the Convolutional Neural Networks (CNNs) and. It provides over 5,000 labelled images of environments and a collection of 205 scene classes. When combined with CNN, the system is able to gain deep features learning for a broader range of scene identification tasks and offer valuable data, in line with a number of scene-based datasets.

In order to accurately identify the scene featured in the corpus for this study, the environment recognition method suggested by [Zhou et al. \(2014\)](#) was employed.¹ The method was used to analyse the first frame in every shot. Almost all the indoor shots were correctly classified with 96.75%, whereas only 66.66% of outdoor scene were correctly identified due to variety of illumination and background noise. Figure A.6 shows an example of a scene recognition result obtained from our corpus.

¹ The dataset and the source code available at: <http://places.csail.mit.edu/index.html>.

**Predictions:**

- **Type of environment:** outdoor
- **Semantic categories:** forest_road:0.19, campsite:0.15, orchard:0.13, forest_path:0.09, cemetery:0.08
- **SUN scene attributes:** trees, foliage, naturallight, nohorizon, vegetation, leaves, openarea, man-made, natural, camping
- **Informative region for the category *forest_road* is:**

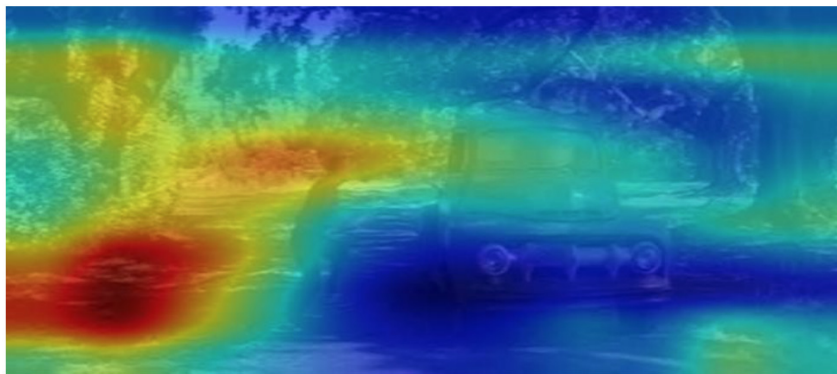


Fig. A.6 Scene recognition result for video clip named ‘actioncliptrain00366’ from ‘DriveCar’ class using system proposed by [Zhou et al. \(2014\)](#).

