



The
University
Of
Sheffield.

Access to Electronic Thesis

Author: Muhammad Usman Ghani Khan
Thesis title: Natural Language Descriptions for Video Streams
Qualification: PhD

This electronic thesis is protected by the Copyright, Designs and Patents Act 1988. No reproduction is permitted without consent of the author. It is also protected by the Creative Commons Licence allowing Attributions-Non-commercial-No derivatives.

If this electronic thesis has been edited by the author it will be indicated as such on the title page and in the text.

Natural Language Descriptions for Video Streams



Muhammad Usman Ghani Khan

Supervisor: Dr. Yoshihiko Gotoh

Department of Computer Science

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

12-09-2012

Dedication

I would like to dedicate this thesis to my father, Sardar Muhammad Arshad Khan, who has been the biggest and continuous source of motivation throughout my life.

Acknowledgements

Thanks to Allah who is the source of all the knowledge in this world, and imparts as much as He wishes to any one He finds suitable.

I offer my sincere gratitude to my supervisor, Yoshihiko Gotoh, who has supported me throughout my research work with his patience and knowledge. Without his guidance, motivation and support, this dissertation would not have been completed. One could not wish for more kind, accessible and friendlier supervisor than him. Special thanks to my panel members, Thomas Hain and Steve Maddock for their useful suggestions and being so kind to me. I am thankful to my examiners Phil Green and Natalia Grabar for useful discussion during the final viva of my defence.

I am thankful to my parents for their support, prayers, love and care throughout my life and they have played a vital role in achieving this milestone. My wife has always been a wonderful being to me and extended her whole-hearted support especially during my PhD studies, which I could not have completed without her. To my sons Taha and Fahad; you have played your part very well as your naughtiness always made me relaxed. I extend my thanks to my sisters, brother (Muhammad Kamran Khan), in-laws, relatives and friends for their continuous support and prayers.

I thank all my colleagues and group fellows who never let me feel that I am away from my homeland: Mauro, Amy, Siripinyo Chantamunee, Lei, Sarah, Manal, Nouf, Tim, Davide, Jan, Jamy, Jon Barker, Guy Brown, Roger Moore, Georg Struth, Mick Humble and Phil Green.

Also many thanks to my elders and friends in Pakistan; Pervaiz Iqbal Khan Niazi, Rao Muhammad Adeel, Hafiz Shahzad Asif, Ilyas Sarwar, Ali Khan, Muhammad Ayub, Moazzam Naveed, Muzammil Shahbaz, Asad Sandhu, Tahir Aziz, Rashid Sharif, Mirza Moazzam.

I am much obliged to University of Engineering and Technology, Lahore, Pakistan for funding this work under the Faculty Development Program. Special thanks to Dr. Shaiq Haq since he is the one who introduced me to this domain.

Last but not the least I thank all my teachers, who shared their knowledge with me to the best of their efforts and raised me to the position where I am.

Abstract

This thesis is concerned with the automatic generation of natural language descriptions that can be used for video indexing, retrieval and summarization applications. It is a step ahead of keyword based tagging as it captures relations between keywords associated with videos, thus clarifying the context between them. Initially, we prepare hand annotations consisting of descriptions for video segments crafted from a TREC Video dataset. Analysis of this data presents insights into humans interests on video contents. For machine generated descriptions, conventional image processing techniques are applied to extract high level features (HLFs) from individual video frames. Natural language description is then produced based on these HLFs. Although feature extraction processes are erroneous at various levels, approaches are explored to put them together for producing coherent descriptions. For scalability purpose, application of framework to several different video genres is also discussed. For complete video sequences, a scheme to generate coherent and compact descriptions for video streams is presented which makes use of spatial and temporal relations between HLFs and individual frames respectively. Calculating overlap between machine generated and human annotated descriptions concludes that machine generated descriptions capture context information and are in accordance with human's watching capabilities. Further, a task based evaluation shows improvement in video identification task as compared to keywords alone. Finally, application of generated natural language descriptions, for video scene classification is discussed.

Declaration

I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been acknowledged.

Muhammad Usman Ghani Khan

Contents

Contents	v
Nomenclature	viii
1 Introduction	1
1.1 Background	1
1.2 Motivation	1
1.3 Scope and Aim	2
1.4 Thesis Contributions	3
1.5 Justification For the Research	6
1.5.1 Application Areas	6
1.5.2 Future Research Areas	7
1.6 Thesis Overview	7
1.7 Published Work	9
2 Related Work	11
2.1 Introduction	11
2.2 Video Corpus and Annotation Tools	12
2.3 Visual Contents Identification:	14
2.4 High Level Features Extraction	15
2.4.1 Face Detection	16
2.4.2 Age Identification	18
2.4.3 Gender Detection	19
2.4.4 Facial Emotion Recognition	19
2.4.5 Action recognition	21
2.4.6 Object Recognition	23
2.4.7 Scene Settings- Indoor / outdoor Scene Classification	25
2.4.8 Speaking Person Identification	27
2.5 Natural Language Generation	28
2.5.1 Stages of Natural Language Generation	28
2.6 Natural language Description of Videos	30

CONTENTS

2.7	Video scene classification	33
2.7.1	Feature Extraction	33
2.7.2	LSA (Latent Semantic Analysis)	35
2.8	Summary	37
3	Corpus Generation and Analysis	39
3.1	Introduction	39
3.2	NLDV - Corpus 1	40
3.2.1	Annotation Process	40
3.2.2	Corpus Analysis	41
3.2.3	Human Related Features	43
3.2.4	Objects and Scene Settings	44
3.2.5	Spatial Relations	45
3.2.6	Temporal Relations	46
3.2.7	Similarity between Descriptions	47
3.2.8	Sequence of Events Matching	48
3.3	NLDV - Corpus 2: Dataset with Lengthy Videos	49
3.3.1	Annotation Process	50
3.3.2	Corpus Analysis	50
3.4	NLDV - Corpus 3: Evaluation Dataset	54
3.5	Findings from the Corpora Analysis	54
3.6	Summary	55
4	Image Processing Methods and Evaluations	57
4.1	Overview of HLF Extraction Procedure	57
4.2	Human Identification	58
4.3	Human Age/ Gender Detection	60
4.4	Emotion Recognition	62
4.4.1	Localization of Facial Features	62
4.5	Action Recognition	64
4.6	Objects Recognition	66
4.7	Indoor / Outdoor Scene Identification	67
4.8	Speaking Person Identification	69
4.9	Formalising Spatial and Temporal Relations	72
4.10	Summary	75
5	Natural Language Descriptions for Visual Images	77
5.1	Introduction	77
5.2	Framework Overview	78
5.3	Natural Language Generation	79
5.3.1	Summary of Generation Procedure	86
5.4	Experiments	88

5.4.1	Machine Generated Annotation Samples	88
5.4.2	Evaluation with ROUGE	93
5.4.3	Task Based Evaluation	94
5.5	Summary	97
6	Dealing with Missing and Erroneous Data	99
6.1	Introduction	99
6.2	Application Scenarios	100
6.3	Scalability of Work for other Video Categories	102
6.4	Evaluation of Generated Descriptions	111
6.4.1	Evaluation with ROUGE	111
6.4.2	Task based Evaluation	112
6.5	Discussion of Framework Shortcomings	112
6.6	Summary	114
7	Natural Language Descriptions for Video Streams	115
7.1	Introduction	115
7.2	Identifying Units for Description	115
7.3	Paraphrasing Unit-based Descriptions	118
7.4	Incorporating Temporal Information	121
7.5	Experiments	122
7.5.1	Evaluation with ROUGE	122
7.5.2	Task Based Evaluation	122
7.6	Summary	128
8	Use of Natural Language Descriptions for Video Scene Classification	129
8.1	Introduction	129
8.2	The Approach	130
8.3	Feature Extraction	132
8.4	LSA (Latent Semantic Analysis)	133
8.5	Pseudo Semantic Tree	134
8.6	Experiments	136
8.7	Summary	140
9	Conclusions	141
9.1	Original Contributions	142
9.2	Future Work	143
9.2.1	RST Based Video Summarization Framework.	144
	References	149
A	Video Annotation Tool	165

CONTENTS

B Templates for Sentence Generation	169
--	------------

Chapter 1

Introduction

This thesis is concerned with the task of automatically generating descriptions for video streams. Mostly, previous work has focused on identification of individual keywords that may be present in video sequences. Recently, research is becoming focussed towards generation of descriptions for video streams based on these individual keywords. In this chapter, we provide brief background, motivation, scope and aim of this work. We further present main contributions and some application areas of this research work. Finally, the structure of the thesis is presented.

1.1 Background

Digital images and videos collection has increased exponentially during the past few years as more and more data is available in the form of personal photo albums, handheld camera videos, feature films and multi-lingual broadcast news videos which presents visual data ranging from unstructured to highly structured. Today video traffic accounts for 80 percent of all video traffic.¹ Videos consist of audio and visual contents and are often provided with textual information resulting in the increase of data in all the three dimensions. Because of this huge increase in data, there is a need for qualitative filtering to differentiate between relevant and irrelevant information according to user requirements. In addition, time constraints enforce a limit on how much time one can spend watching videos. Therefore, one has to be selective when accessing appropriate information. Such a distillation process requires comprehensive information processing including categorization, description and explanation about various videos.

1.2 Motivation

Since, there is an exponential increase in video data these days, there is need for formalizing video semantics to help users gain useful information relevant to their interests. One approach to explain video semantics and contents is to convert it into some other modality such as text.

¹<http://techcrunchies.com/what-percentage-of-internet-traffic-is-video/>

1. INTRODUCTION

Human language is a natural way of communication. Useful entities extracted from videos and their inter-relations can be presented by natural language in a syntactically and semantically correct formulation.

Humans can intuitively describe a video in their natural language based on natural capabilities of visual scene understanding. They describe a scene using visual contents and their domain knowledge. On the other hand, computers can only identify and recognize some objects and certain activities. Most of the previous studies were related to semantic indexing of video using keywords [Chang et al., 2007; Naphade and Huang, 2001]. However it is often difficult with keywords alone to represent relations between various entities and events in video. An interesting extension to a keyword based scheme is natural language textual description of video streams. They are more human friendly. They can clarify context between keywords by capturing their relations.

Natural language description of videos is fundamental for the conventional video retrieval tasks. The problem is that, given the sheer volume of multimedia produced publicly and privately, only a small fraction of them are annotated, and the large majority cannot be retrieved using natural language queries. Natural language description is one way to represent a video, saving storage space and processing time for retrieval. It can be as granular as a set of keywords, as verbose as a paragraph, or even a full length story. The latter has an advantage over keywords based description in that relation between entities is clearer. As scenes in videos are often lengthy, it is sometimes difficult to describe them using a set of keywords only. Descriptions can guide generation of video summaries by converting a video to natural language. They can provide basis for creating a multimedia repository for video analysis, retrieval and summarization tasks.

1.3 Scope and Aim

This work addresses generation of natural language descriptions for human actions, behaviour and their relations with other objects observed in video streams. We start the work with manual development of a dataset, consisting of natural language descriptions of video segments crafted from a small subset of TREC¹ Video data. In a broad sense, the task may be considered one form of machine translation as it translates video streams into textual descriptions. To date, the number of studies in this field is relatively small partially because of lack of appropriate dataset for such a task. Another obstacle may be inherently larger variation for descriptions that can be produced for videos than a conventional translation from one language to another. Indeed humans are very subjective when annotating video streams, *e.g.*, two humans may produce quite different descriptions for the same video. Based on these descriptions, we are interested to identify the most important and frequent high level features (HLFs); they may be keywords, such as a particular object and its position/move, used for a semantic indexing task in video retrieval.

¹<http://trecvid.nist.gov/>

Machine generated descriptions for videos sequences require investigation in two disciplines, namely image processing and natural language processing. Image processing techniques lead to identification of HLFs such as humans, objects, their moves and properties (*e.g.*, gender, emotion and action) [Smeaton et al., 2009] from individual frames. Natural language processing deals with merging these HLFs into syntactically and semantically correct textual presentations. The frame based natural language generation procedure results in many identical descriptions produced from adjacent frames. Hence simple concatenation of descriptions may lead to redundancy, lacking coherency. Additionally visual feature extraction processes are erroneous at various levels. Use of spatial and temporal information comes to rescue for these problems. Finally, generated descriptions are used for visual scene classification application.

1.4 Thesis Contributions

This section enlists thesis contributions together with their background where background presents motivation and need of the contribution.

Contribution 1: Corpus Generation and Analysis

Background. As video contents continue to expand, it is increasingly important to properly annotate videos for effective search, mining and retrieval purposes. While the idea of annotating images with keywords is relatively well explored, work is still needed for annotating videos with natural language to improve the quality of video search. This contribution is based on generating a video dataset with natural language descriptions as annotations.

Further, there is a need to determine important contents for the description of a video stream. Though content selection can be much subjective to human’s perception, still some matching and overlap contents can be enlisted to provide an overview of human’s interest while watching videos. Generally, visual contents determination encompasses two fields, (i) objects determination and (ii) finding objects activities and interaction between them. Most important objects, activities and interactions should be identified for proper semantic understanding of visual scenes. Further these contents should be presented in suitable words from natural language.

Contribution. A video corpus of short video clips ranging from 10 to 30 seconds in length is manually created from TRECVID provided videos. It consisted of 140 segments of videos — 20 segments for each of the following seven categories: human actions, human close up, news, meeting, grouping, traffic, and indoor / outdoor videos. 13 human participants manually annotated this dataset in three flavors, identification of important concepts (keywords), a very short summary (title) comprising of one phrase or sentence (presenting theme or main idea) and a complete natural language description comprising of several sentences (detailed description of the visual scene). Analysis of this corpus presents insights of human behavior and interest while watching videos. Such resource can also be used to evaluate automatic natural language generation systems for videos.

1. INTRODUCTION

Contribution 2: Framework for Natural Language Description of Visual Images

Background. Visual scene interpretation is a task which is still in its infancy. Though important concepts in a visual scene can be presented by keywords, they lack context information which is needed to explain the detailed semantics of video. Natural language descriptions of visual scenes are needed as they are more human friendly and capture relationships among keywords; clarifying the context related to separate keywords. Thus they help in explaining scenario and situation depicted in the video sequence. Descriptions can guide generation of video summaries by converting a video to natural language. They can provide basis for creating a multimedia repository for video analysis, retrieval and summarisation tasks.

Contribution. A framework is presented for generation of natural language descriptions for human actions, behaviour and their relations with other objects observed in video streams. The work starts with implementation of conventional image processing techniques to extract high level features from video. These features are converted into natural language descriptions using context free grammar.

Contribution 3: Dealing with Missing and Erroneous Data

Background. This question is further decomposed into two directions, *i.e.*, (i) application of research into different video genres and (ii) dealing with missing and erroneous outputs resulting due to limitedness of image processing techniques.

Any video processing paradigm must incorporate the strengths and weaknesses of the media by which it is conveyed. Different media sources have specific attributes and genres; therefore, have their own audiences and application areas. For instance, handheld video camera movies are mostly family oriented, feature films are often entertainment, lesson and recreation providers, historical videos are informative and television broadcasts are better at tracking developing news stories. Video structure plays significant role in the identification of video genre. For example, personal videos are unstructured; on the other hand, broadcast videos are highly structured.

The quality of description needs to be investigated when the limited number of features are available. It is anticipated that the larger number of features leads to the higher quality in description. However it is not feasible to produce a full list of features with the help of current technologies. Further the processing time may become an issue when the feature size is very large. A balance between the quality of description and quantity of features needs to be explored. Finally, there may be many possible outputs for a single frame by image processing. Strategies to combine best available outputs to generate the coherent description need to be explored.

Contribution. Of all the various sources available, this work focuses on the processing of broadcast news and Rushes¹ videos. Different scenarios taken from Rushes and news videos are

¹Rushes is a raw material which will be further used to produce a video stream such as movies or television

discussed and evaluated. Application scenarios of this framework include dealing with missing information, absence of human subjects and dealing with erroneous HLF extraction outputs.

Contribution 4: Evolving a Framework for Coherent Natural Language Descriptions of Complete Video Stream

Background. Video domain can be described at two levels; frame based and full video based descriptions. Initially, description is generated for individual frames. For describing a complete video sequence, simply joining frame based descriptions will have several shortcomings. Descriptions of individual frames are crude, repeated and in some cases missing useful information due to sparseness of HLFs that can be produced by current technologies. Image processing errors can be accumulated. Lack of temporal information may cause a further problem.

Contribution. We consider a sequence of video frames from which some HLFs (*e.g.*, human, objects or their moves) can be identified. For example in a scene where a man walks out from a room after desk work, the following actions may be identified: ‘*sitting*’ (in front of a desk), ‘*standing up*’ (from a chair), and ‘*exiting*’ (from a room), each of which can span over multiple frames. It may also contain another identifiable features such as facial expressions (*e.g.*, serious in the beginning and smiling later) and some objects in the background. In this work we refer to a sequence of frames with an identical set of visual HLFs as a unit. Using this definition, the length of individual units may be affected by the availability, as well as the quality, of HLF extraction techniques. Textual descriptions of individual frames are used for ‘Unit’ generation. Merging descriptions of similar frames, paraphrasing in cases where there is some similarity in descriptions and introducing temporal information are main building blocks for this Unit generation framework.

Contribution 5: Use of Natural Language Descriptions for Video Scene Classification

Background. Natural language descriptions produced for video sequences based solely on visual contents are compact¹. When the amount of (meta)data is small, it is often difficult to achieve a good classification performance based only on information derived from the data in a conventional fashion. There is a need to devise a classification method which works fine for such scenarios.

Contribution. Incorporating complementarity knowledge extracted from individual video scene classes comes to rescue for classification task. Finding co-occurrence terms between documents further improves classification problem. For classification of visual scenes based solely on machine generated descriptions, human annotations are used as a complement to the limited size of machine generated descriptions, and out of vocabulary terms are also handled.

programmes. The material contains natural sound and highly repetitive frame sequences.

¹roughly 3 to 4 sentences for hand and machine generated annotations

1.5 Justification For the Research

One of most important goals of computer vision research is to provide computers with human like perception capabilities. Natural language is a mechanism used by humans for communicating outputs based on mixtures of natural senses such as sight, hearing and taste. For presenting these abilities, humans mostly use natural language descriptions rather than individual words alone, since descriptions are more human friendly and provide detailed explanation of the visual scenes. Although, most of the video indexing and retrieval methods are based on keywords alone, there are good reasons for using more linguistically meaningful descriptions. Attaching list of keywords with a video may lead to ambiguous understanding of the video stream. A video sequence annotated with the words blue, sky and car could depict a blue car or a blue sky. On the other hand, descriptive annotation such as ‘A car is moving under the blue sky’ would make the relations between the words explicit.

1.5.1 Application Areas

Conversion of video streams into natural language descriptions leads to variety of useful applications. Video retrieval search engines can be generated which make use of longer and more semantic queries. Installed for monitoring human activities in an old house environment, they can help old people in their daily life activities. Human behavior identification, monitoring and understanding to minimize risks of miss happenings in security and surveillance videos. Robots which communicate using natural language can be built and employed in various environments(*e.g.*, dangerous mining work, helping people with limited mobility and blindness in doing daily life tasks).

- **Video Retrieval:** A video searching tool based on description of location, objects, people and their activities in a vast amount of visual data. This tool can be used in movies, broadcast news, personal videos, and internet videos. A new user interface allowing fast or reverse forwarding to show shots containing a particular person or an object at a specific location with a specific action. Suppose a user is interested in ‘*Barack Obama*’ and ‘*Ahmed Nazad*’ meeting where both are smiling, presenting some cheerful scenario for many users. Users may be interested to go directly to this scene instead of watching the full news.
- **Video Mining and Summarization:** Video summarization plays significant role in reducing the storage space without losing meanings of the contents and helps in efficient and quick video retrieval according to user needs. The main problem in video summarization is how to detect a crucial moment in video sequence. If the behavior of a human is recognized, the moment of an action is likely to be a candidate of a key frame. Therefore, video summarization has a close relation to text generation. From this point of view, extraction of key frames or shots corresponding to notable textual descriptions in the process of text summarization becomes useful.

1.5.2 Future Research Areas

- Video semantics and video understanding: A semantic analyzer for discovering semantics information according to users' requirements. Video semantics identifies meanings of video in the form of a theme or an underlying idea.
- Video indexer and retrieval system: A video retrieval system need to be built which can answer queries which consist of complete sentences rather than words alone.

1.6 Thesis Overview

- **Chapter 1: Introduction** — This chapter gives background, research questions, aims and application of this research.
- **Chapter 2: Related work** — It provides review of work directly related to this research *i.e.*, natural language description generation for images and videos. This chapter explains following areas, *i.e.*, literature of visual scene description, image processing methods for HLFs extraction, natural language generation and description generation for images and videos.
- **Chapter 3: Corpus Generation and Analysis** — This chapter provides details about the generated corpora and analysis of hand annotations for these corpora. This chapter is based on the paper [Khan et al., 2012a]. It justifies the contribution 1.
- **Chapter 4: Image Processing Methods and Evaluations** — This chapter presents details about the implementation methods based on image processing used in this research. Results of these methods and discussion of results is presented in this chapter.
- **Chapter 5: Natural Language Descriptions for Visual Images** — This chapter presents the framework for generation of language descriptions for visual images. Contents of this chapter are based on the papers [Khan and Gotoh, 2012a; Khan et al., 2011a, 2012b]. It justifies contribution 2.
- **Chapter 6: Dealing with Missing and Erroneous Data** — This chapter continues the discussion on the framework for language descriptions generation. Application of the framework on different video genres and scenarios for dealing with missing and erroneous data are discussed in this chapter. This chapter is based on the papers [Khan and Gotoh, 2012a,b]. It justifies contribution 3.
- **Chapter 7: Natural Language Descriptions of Video Streams** — This chapter further extends framework for generating descriptions of complete video streams. This chapter is based on the papers [Khan and Gotoh, 2012b; Khan et al., 2011b]. It justifies the contribution 4.

1. INTRODUCTION

- **Chapter 8: Video Scene Classification Based on Machine Generated Natural Language Descriptions.** — This chapter provides application of machine generated language descriptions for visual scene classification task. This chapter is based on the papers [Zhang et al., 2011, 2012]. It justifies contribution 5.
- **Chapter 9: Conclusion** — This chapter provides conclusion of this research by providing summary, recommendations and future directions. For future work we present our ongoing work on the paper [Khan and Gotoh, 2012c] related to video summarization which is based on natural language description of video streams.

1.7 Published Work

This thesis is mainly based on following publications:

1. M. U. G. Khan and Y. Gotoh. Generating Natural Language Tags for Video Information Management. Submitted to the International Journal of Information Processing and Management.
2. M. U. G. Khan and Y. Gotoh. Natural language descriptions of visual scenes - corpus generation and analysis. In joint workshop of ESIRMT and HYTRA, EACL 2012.
3. M. U. G. Khan and Y. Gotoh. Describing video contents in natural language. In Workshop on Innovative hybrid approaches to the processing of textual data, EACL 2012,
4. M. U. G. Khan, L. Zhang, and Y. Gotoh. Natural language description of video streams. In International conference on Image processing, 2012.
5. M. U. G. Khan, L. Zhang, and Y. Gotoh. Human focused video description. In The 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications, ICCV, 2011.
6. M. U. G. Khan, L. Zhang, and Y. Gotoh. Towards coherent natural language description of video streams. In 2nd IEEE International Workshop on Stochastic Image Grammars, ICCV, 2011.
7. L. Zhang, M. U. G. Khan, and Y. Gotoh. Video scene classification based on natural language description. In 2nd IEEE Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, ICCV, 2011.
8. L. Zhang, M. U. G. Khan, and Y. Gotoh. Video scene classification based on limited dataset of natural language descriptions. Submitted to the Journal of Zhejiang University.
9. M. U. G. Khan and Y. Gotoh. Framework for video indexing and retrieval based on natural language descriptions. (In preparation for Journal of Information Sciences)
10. M. U. G. Khan and Y. Gotoh. Summarizing Video Contents using Natural Language Descriptions. (In preparation for International Conference on Computational Linguistics, COLING, 2012.)

1. INTRODUCTION

Chapter 2

Related Work

2.1 Introduction

This chapter provides survey of previous studies related to our research work. The contents of this chapter is not summary of this topic in the world, but it points only subjects with direct relationship with this work.

Background. Recently, textual document processing, *i.e.*, storage, indexing and retrieval has attained a certain mature level. On the contrary, processing of multi-media documents such as images, videos and speech is still in its infancy. For example, finding a video which depicts ‘*war scenes in Gaza*’ is rather impossible to achieve. It requires video classification based on video semantics. Video semantics help in understanding theme or idea presented by the video instead of sole identification of objects and activities happening in the video. Video description is a necessary step towards video semantics as description provides more information than individual features in the video.

More recent works start video content analysis by integrating both acoustics and visual features, which are the two inseparable parts of a video. Visual scene understanding is heavily dependent on low-level features like color, texture, energy and pitch *etc.*, which leads to the well-known semantic gap problem. ‘*Semantic gap*’ describes the difference between video provided information and user understanding. User is always interested in abstract information which correspond to high level features rather than low level features. The semantic gap problem has been recognized as the biggest challenge that the multimedia community is facing for multimedia data retrieval.

Chapter Structure. This thesis is focussed towards visual scene description based on visual information only. Roughly, this work can be decomposed into two parts, *i.e.*, visual contents identification (HLFs) and natural language generation based on these visual contents. Analysis of metadata for a video corpus is helpful in filtering useful HLFs in video sequences (Section 2.2). Image processing leads to automatic identification of these HLFs (Section 2.3). Methods used for extraction of HLFs are discussed in Section 2.4. Second part is related to natural

2. RELATED WORK

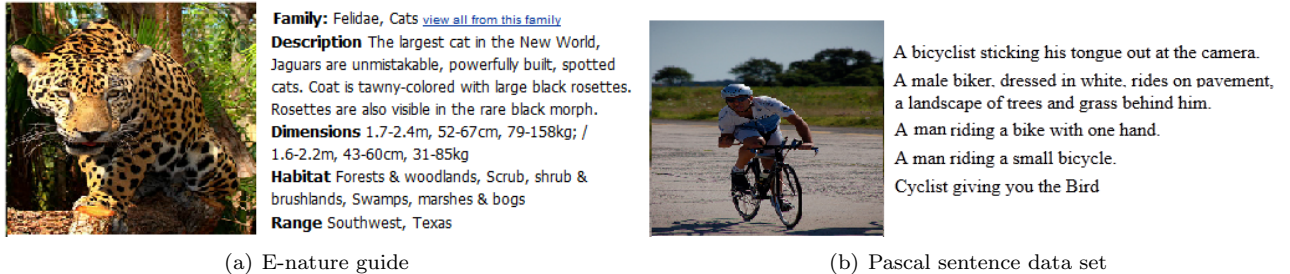


Figure 2.1: Sample images from image databases with metadata in the form of natural language descriptions.

language generation domain where HLFs are combined to generate well phrased, smooth and coherent descriptions (Section 2.5). Some previous studies are related to generation of natural language descriptions for images and videos (Section 2.6). Finally, visual contents could be useful for visual scene classification task (Section 2.7).

2.2 Video Corpus and Annotation Tools

Many corpora have been developed for video processing studies, ranging from a simple object identification to more complex scene analysis. Performance evaluation of tracking and surveillance (PETS) is a series of workshops providing datasets for object detection, human tracking and surveillance purposes [Young and Ferryman, 2005]. Approaches to human action tracking and recognition in different conditions (*e.g.*, indoor/outdoor, shopping center) can be explored using corpora such as the KTH human action dataset [Schuldt et al.] and the Hollywood action dataset [Marszalek et al., 2009]. CAVIAR [Fisher et al., 2005] and the Terrascope [Jaynes et al., 2005] are publicly available datasets for video surveillance. Annotations are available in the form of keywords (*e.g.*, actions — sit, stand, walk). Most datasets have been developed for keyword based searching, object recognition and event identification tasks. Recently, trend is shifted towards generating images / video corpora with natural language annotations.

Online Nature Guide. Enature guide¹ provides description of animal and plant species with their attributes. Figure 2.1(a) shows an example image of Jaguar with the metadata in the form of family, description, habitant *etc.* These descriptions are based on appearance of visual contents only, *i.e.*, there is no information about semantic properties or activities of these species. Information about the structure of these species such as color, texture and shape are quite obvious in these descriptions.

PASCAL Sentence Dataset. UIUC dataset² [Farhadi et al., 2010] contains 1000 images taken from a subset of the Pascal-VOC 2008 challenge image dataset, and metadata with sentences that describe the image by paid human annotators using Amazon Mechanical Turk.

¹<http://www.enature.com/fieldguides/>

²<http://vision.cs.uiuc.edu/pascal-sentences/>

Figure 2.1(b) shows an example image with the five sets of hand annotations. There are 5 annotations for each image, resulting in 5000 hand annotations in total. Mostly, these annotations are short – around 10 words in length, and covering foreground objects in more details.

Both of the above mentioned datasets are related to individual images, *i.e.*, they provide natural language descriptions for static image. On the contrary, video sequences are composed of temporal information which explains the relationship between individual static images. TRECVID¹ provides a range of video corpora with varying sets of annotations.

TREC Video provided Datasets. The TREC video evaluation consists of on-going series of annual workshops focusing on a list of information retrieval (IR) tasks. The TREC video promotes research activities by providing a large test collection, uniform scoring procedures, and a forum for research teams interested in presenting their results. The high level feature extraction task aims to identify presence or absence of high level semantic features in a given video sequence [Smeaton et al., 2009]. Approaches to video summarisation have been explored using rushes video² [Over et al., 2007].

TREC video also provides a variety of meta data annotations for video datasets. For the HLF task, speech recognition transcripts, a list of master shot references, and shot IDs having HLFs in them are provided. Annotations are created for shots (*i.e.*, one camera take) for the summarisation task. Multiple humans performing multiple actions in different backgrounds can be shown in one shot. Annotations typically consist of a few phrases with several words per phrase. Human related features (*e.g.*, their presence, gender, age, action) are often described. Additionally, camera motion and camera angle, ethnicity information and human’s dressing are often stated. On the other hand, details relating to events and objects are usually missing. Human emotion is another missing information in many of such annotations.

Video Annotation Tools. There exist several freely available video annotation tools. One of the popular video annotation tool is *Simple Video Annotation tool*³. It allows to place a simple tag or annotation on a specified part of the screen at a particular time. The approach is similar to the one used by *YouTube*⁴. Another well-known video annotation tool is *Video Annotation Tool*⁵. A video can be scrolled for a certain time period and place annotations for that part of the video. In addition, an annotator can view a video clip, mark a time segment, attach a note to the time segment on a video timeline, or play back the segment. ‘*Elan*’ annotation tool allows to create annotations for both audio and visual data using temporal information [Wittenburg et al., 2006]. During that annotation process, a user selects a section of video using the timeline capability and writes annotation for the specific time.

¹<http://trecvid.nist.gov/>

²Rushes are the unedited video footage, sometimes referred to as a pre-production video.

³videoannotation.codeplex.com/

⁴www.youtube.com/t/annotations_about

⁵dewey.at.northwestern.edu/ppad2/documents/help/video.html

2.3 Visual Contents Identification:

Any visual scene can be partitioned into foreground and background, where foreground is the main ‘*area of interest*’ and background can help in better understanding of the foreground. Important visual contents such as objects and their activities are often considered part of foreground. In the following we first provide review of low level features used for object detection research. Then a discussion of classifiers is provided. Finally, concept based object detection is discussed. In the literature, there are mainly three types of low-level image features, *i.e.*, color-based features, texture-based features and shape-based features [Antani et al., 2005].

Color Based Features have been shown to be the most widely-used features in video processing domain, as they maintain strong cues that capture human perception in a low dimensional space and they can be generated with less computational effort than other advanced features. Most of them are independent of variations of view and resolution, and thus possess the power to locate the target images robustly. They have also been demonstrated to be the most effective image features in the TRECVID evaluation [Amir et al., 2005]. Many color spaces have been suggested in previous studies such as *RGB*, *YUV*, *HSV*, *HVC*, *L*u*v* and *L*a*b* [Del Bimbo, 1999]. The simplest representation of color-based features are color histograms. Each component in color histograms is the percentage of pixels that are most similar to the represented color in the underlying color space [Smith and Chang, 1997]. Another type of color-based image features are called color moments, which only compute the first two or three central moments of color distributions [Smith and Chang, 1996b]. They aim to create a compact and effective representation in image retrieval.

Texture Based Features aim to capture the visual characteristics of homogeneous regions which do not come from a single color or intensity [Smith and Chang, 1996a]. These regions may have unique visual patterns or spatial pixel arrangements, which gray level or color features in a region may not sufficiently describe. The process of extracting texture based features often begins with passing images into a number of Gabor or Haar wavelet filters [Amir et al., 2005]. The feature vector can then be either constructed by concatenating central moments from multiple scales and orientations into a long vector [Ngo et al., 2002] or extracting statistics from image distributions directly. Effectiveness of texture features is also an important research direction. For instance, Ohanian and Dubes [1992] compared four types of texture representations and observed that the co-occurrence matrix representation work the best in their test collections.

Shape Based Features can be generated either from boundary-based approaches that use only the outer contour of shape segments, or from region-based approaches that considers the entire shape regions [Rui et al., 1996]. One of the simplest approaches to extract shape features is to detect visible edges in query images and then match their edge distribution or histogram against those of the target images. Another approach to extract shape features is to use implicit polynomials to effectively represent the geometric shape structures [Chuang and Kuo, 1996], which is robust and stable to general image transformation. Mehtre et al. [1997] presented a comprehensive comparison of shape features for retrieval by evaluating them on a 500 element

trademark dataset.

Local Image Features. To derive local features from images, a natural idea is to partition the entire image into a set of regular image grids and extract image features (especially color features) from image grids [Amir et al., 2005]. Wilkins et al. [2008] used a local color descriptor based on the average color components on an 8×8 block partition of images. Hauptmann et al. [2004] studied color layout features on a 5×5 regular image grid. Extended from regular image grids, quad-tree based layout approaches first split the image into a quad-tree structure and construct color histogram for each tree branch so as to describe its local image content.

Image Segmentation. In order to locate specific objects in images, it would be advantageous to extract image features (*e.g.*, color or shape) from segmented image regions. Image segmentation is defined as a division of the image data into regions in such a way that one region contains the pixels of the silhouette of the given object without anything else [Smeulders et al., 2000]. Most segmentation algorithms proceed by automatically clustering the pixels into groups. Normalized cut [Malik et al., 2001] has been widely applied in visual retrieval, object recognition and motion segmentation. Numerous alternative criteria have also been suggested for segmentation. The use of image segments has been widely studied in the context of content based video retrieval.

Classifiers for Object Detection. Initial step of any object detection paradigm is to define a meaningful and manageable list of semantic concepts based on human prior knowledge. Mostly, object detection is treated as a supervised learning problem that attempts to discriminate positive and negative annotated examples. Initially, low-level features are extracted from several modalities, *e.g.*, text, audio, and visual. For each concept, separate mono-modal classifiers are built using the corresponding labeled data and low-level features. For classification, one of the most common learning algorithms is called support vector machines (SVMs) [Burges, 1998] which has been applied in image processing domain for variety of tasks such as 2-D and 3-D object recognition, human action classification, activity detection *etc.* Apart from SVMs, there are a large variety of other classifiers [Naphade and Smith, 2004] that have been investigated, including gaussian mixture models (GMM), hidden Markov models (HMM), k nearest neighbor (kNN) *etc.*

2.4 High Level Features Extraction

Since, ‘human’ is often the most important and also interesting feature in any video, a survey of methods related to human’s structure and activities is presented. Presence of human face or body (upper or lower, complete) can help in detection of humans in the video. Identification based on human face / body leads to other interesting applications. Facial features play an important role in identifying age, gender and emotion information. Body information helps in identification of human actions and their interactions with other objects. We briefly present literature related to human detection using face and body information. A further review of HLFs which directly rely on face or body information is also presented. Secondly, literature

2. RELATED WORK

about other HLFs which may be objects in video streams is provided. Third scene settings classification methods are reviewed. Finally speaker identification techniques based on visual information are discussed.

2.4.1 Face Detection

Detection of human face can guide presence of human in a video stream. However, presence of different variations in brightness, lightings, contrast levels, poses, and backgrounds make this task much complicated. Literature related to face detection started appearing in early 70's and has matured a lot since then. Several survey studies have been performed based on different categories [Hjelmås and Low, 2001; Li et al., 2004]. Roughly face detection algorithms can be categorized into four main classes *i.e.* (i) knowledge based (ii) facial features based (iii) appearance based and (iv) template matching based approaches. First two approaches are based on extraction of facial features from an image and manipulate its parameters such as angles, size, and distances. Last two approaches rely on training and learning set of examples for objects of interest. However, dealing with video introduces other approaches for face detection such as motion based approaches. A brief description of the most common approaches and examples of algorithms used in each of them is given in the rest of this section.

Knowledge Based Approaches. Also known as top down approaches, these approaches translate human knowledge to well-defined rules. Firstly facial features are extracted from provided images. Then, relationship between facial features is captured to represent the contents of a face and encode it as a set of rules. Rules are encoded to capture and describe the relation between the features of a face. As an example of facial feature relationship, eyes should be symmetric to each other and should have a relative distance to the nose and mouth. There should be a balance in nature of rules *i.e.*, if the rules are more general, they may fail to detect the right face and there will be many false detections. However, if the rules are more specific, they may fail to detect some faces and will miss most of the faces. [Berbar et al.; Mohamed et al., 2007]

Yang and Huang [1994] proposed a three level hierarchical knowledge-based method to detect human faces in a complex background. The higher two levels are based on mosaic images at different resolutions. The lower level made use of an edge detection method. Berbar et al. presented a method for faces and facial features detection for colored images. Skin locations were identified using skin detection algorithm, which become the region of interest. This skin location is searched for face features such as eyes, mouth and nose. A verification step is further applied to ensure that the extracted features are facial features. Although knowledge-based methods work well for face localization in clear background, it is often difficult to transform human knowledge into correct rules and also to detect faces in different positions.

Facial Features Based Methods. Feature based approaches are techniques that aim to find the features of the face that do not change when the position, viewpoint or lighting conditions varies and using these features, faces are located. Phimoltares et al. [2007] proposed their work

for detection of face from images using a mean face template and a Canny edge detector. The mean face template was generated by averaging intensity of image faces of same size. A Canny edge detector is selected to find the edge image from a colour, grey or binary image. Their method differs from other edge detection methods, in that, it includes the weak edges in the output only if they are connected to the strong edges. The edge image is further compared with their introduced face template. Face template is converted to black-and-white version and edge version by using the Canny edge detector. However, regions are discarded if there are no features detected. [Mondal et al.](#) proposed a method for face detection by using a geometric definition of human face. They computed the mean and variance of the input image to obtain a median filter that was used to reduce the noise effect within the image. The face region was detected based on the technique of human face geometry and center of gravity template matching.

Template Matching Methods. The template matching method is used to detect an individual face in images. The idea of template matching is to compute the correlation between given images and a template face to determine if given images have faces or no faces. [Manoria et al. \[2007\]](#) used a template face to determine if the segmented region is a face or not. The template face is resized according to the region size based on the measurement of height and width of this region. The angular position and the centroid region are first determined. Then the template is rotated to fit the region. The correlation between them is computed and if it is more than a defined threshold, then the region is matched and classified as a face. Otherwise it is classified as non-face.

Appearance Based Methods. This method is based on the use of statistical analysis and machine learning to compute facial features in order to determine if there is a face or not. It uses the models learned from a set of training images where the key method of face detection is based on the image intensity. Therefore, the method is not sufficient to detect the face with images that have poor quality intensity and occlusions. The technique has a high detection rate but slower as compared to facial feature-based techniques. In addition, it has the advantage of being simple to implement, but it cannot efficiently handle differences in scale, pose and shape [[Mohamed et al., 2007](#)]. [Turk and Pentland \[1991\]](#) developed a near real-time Eigen faces systems motivated by PCA for face recognition using Eigen faces and the Euclidean distance measure. Their method computes the Eigenvectors of the covariance matrix of the set of face images. These Eigenvectors can be considered as a set of features that combine to characterize the variation between facial images.

Small Movements Adjustments. Tracking or searching for moving objects needs separating the foreground from background to avoid detection of false areas, especially if the background has same areas which have same features as the wanted objects. Unlike still images, video sequences hold more details about the history of moving objects (foreground), which help in isolating the foreground from the background. Generally, the moving areas are detected by finding the changes that happen among the sequences of images [[Elgammal et al., 2002](#); [Radke](#)

2. RELATED WORK

et al., 2005]. Most of the research done in movement detection applied preprocessing steps before applying the change detection algorithms [Radke et al., 2005]. Such preprocessing steps involve geometric and intensity adjustments. In geometric adjustment, the frames are adjusted to have same coordinates. Affine transformation is an example of such approaches used in this adjustment. It is mainly used for small camera movements adjustments. The problem of variation in light intensity is solved by intensity adjustment in which illumination effect is reduced to some degrees based on the method used.

2.4.2 Age Identification

Age can be identified based on facial information only. Ramanathan et al. [2009] and Fu et al. [2010] presented a detailed review of age identification methods using facial information. One of the first attempts to develop facial age estimation algorithms was reported by [Kwon and Lobo, 1999]. They used two main types of features: Geometrical ratios calculated based on the distance and the size of certain facial characteristics and an estimation of the amount of wrinkles detected by deformable contours (snakes) in facial areas where wrinkles are usually encountered. Based on these features faces were classified as babies, adults and seniors. Lanitis et al. [2002] used an Active Appearance Model based coding scheme for projecting faces into a low dimensional space. Aging functions in the form of quadratic equations are used for relating the coded representation of faces to the actual age allowing in the way the estimation of the age of a subject. According to the results the use of person specific aging functions produced improved age estimation results when compared to the use of a common aging function for all subjects.

Geng et al. [2007] generate aging patterns for each person in a dataset consisting of face images showing each subject at different ages. Each collection of temporal face images is considered as a single sample, which can then be projected to a low dimensional space. Given a previously unseen face, the face is substituted at different positions in a pattern and the position that minimizes the reconstruction error indicates the age of the subject. Fu and Huang [2008] represent aging patterns using manifold learning. A discriminative subspace learning based on manifold criterion is developed for low-dimensional representations of the aging manifold. Regression is generally applied on the aging manifold patterns, which shows significant improvements on age estimation.

Most age estimation methodologies described in the literature use information from the overall face. As an alternative [Suo et al., 2008] used a three-level hierarchical face model as the basis for age estimation. The first level is the global face representation; the second level refers to multiple local facial regions corresponding to different features and the third level involves the use of fine details such as wrinkles and hairline information. Experimental results indicate that the use of local features is important for achieving improved performance. Instead of estimating the age of a subject, Ramanathan and Chellappa [2006] estimated the age-difference between a pair of faces belonging to the same individual. The problem was treated as a classification task where difference vectors between pairs of age-separated faces were used for establishing

the statistical distributions for different age range separations which were subsequently used during the age-separation classification problem.

2.4.3 Gender Detection

Human's gender can be identified using face information only. Once again it starts with processing on facial parts. The system takes the video as input and extracts necessary information and classifies the human being as male or female. Different facial features *i.e.*, forehead, eyebrows, nose, cheek, top lips length, chin jaw and Adam's apple help in identification of human gender. There are mainly two approaches related to gender identification task. The first one is based on facial features and calculating relations between these features. Mostly distances between nose, eyes and mouth, areas of different face parts are employed. The main drawback of these methods is that automatic detection of face parts for different face poses and views is still very difficult to achieve. Another approach is to use low-level information about face image areas based on image pixels values. Among low-level features the most popular are various texture features, histograms of gradients, coefficients of wavelet transformation of image or even raw gray-scale pixel values. Classification methods based on low-level features outperform methods based on high-level ones in accuracy.

The earliest attempt to use computer vision techniques for gender classification was based on neural networks. [Golomb et al. \[1991\]](#) trained a fully connected two-layer network, called SEXNET, to identify gender from facial images. [Tamura et al. \[1996\]](#) used a multi layered neural network to identify gender from face images of different resolutions. [Gutta and Wechsler \[1999\]](#) proposed a hybrid approach that consists of a collection of neural networks and decision trees. A PCA based image representation was used along with radial basis functions and perceptron networks by [\[Abdi et al., 1995\]](#). [OToole et al. \[1993\]](#) have also used PCA and neural networks and have reported good performance. [Moghaddam and Yang \[2000\]](#) investigated the use of SVMs for gender classification.

[Brunelli and Poggio \[1992\]](#) computed 16 geometric features (like pupil to eye brow separation, eye brow thickness, *etc.*.) from the frontal images of a face. These features were used for identifying the gender. [Sun et al. \[2002\]](#) used genetic feature subset selection from frontal images. A multi-modal gender classification approach using images and voice is proposed in [\[Walawalkar et al., 2002\]](#). [Jain et al. \[2005\]](#) worked on the problem of gender identification using frontal faces. They solved this problem using different classifiers like SVM, LDA and ICA. Their method worked fine for female faces but for male faces its performance is not much satisfactory. The performance degrades due to presence of moustaches, beards and glasses as we are focussing on geometric features of human face.

2.4.4 Facial Emotion Recognition

Facial expressions are the most natural means by which human express their emotions. The fundamental issue about the facial expression classification is to define a set of categories to deal with. Psychologists have stated that people are born with the ability to produce six facial

2. RELATED WORK

expressions such as joy, anger, disgust, fear, surprise and sadness. The other expressions are learned from the environments [Lindquist and Barrett, 2008]. Emotion recognition from facial expression can be done either by using machine learning technique or rule based approaches.

Emotion Recognition using Machine Learning Algorithms. Emotion detection from facial expression using HMM were explored by [Cohen et al., 2000]. Each facial expression was modelled using specific HMM trained for that expression. There were a total of six HMMs, each one for modelling the expressions happy, sad, surprise, disgust, fear and angry. Support vector machine (SVMs) classifier [Bartlett et al., 2004] was used to train sample images of faces, each labelled as belonging to a particular category of emotion. The samples used for training are represented as points in space mapped into distinct categories based on their labels. The trained model of SVMs can then be used to predict whether a given test input falls into one category or the other.

Tong et al. [2007] proposed a hierarchical framework based on DBNs to represent to represent probabilistic relationships among various AUs and to account for the temporal changes in facial action development. Instead of recognising each AU or AU combinations individually or statically as other people do, they proposed a more reliable and consistent approach that did not ignore the semantic relationships between the AUs and their dynamics. Lee et al. [2008a] used SVM to recognise facial expressions from a sequence of images. The facial feature displacements tracked by the optical flow are used as input parameters to SVM in order to classify the facial expression. Optical flow is a method used to estimate the motion of the brightness pattern in the image for motion detection. Zhou et al. [2004] proposed an embedded HMM (EHMM) based on AdaBoost for real-time facial expression recognition. AdaBoost is a learning algorithm that produces a strong classifier as a linear combination of a number of weak classifiers. Embedded HMM consists of a set of super states that are used to model a set of HMM. The transition between the super states in the EHMM is different from transition within each HMM. Their proposed system used five basic expressions: normal, laugh, anger, sleep and surprise.

Emotion Recognition using Heuristics or Rule Based Models. Heuristics are based on biological theories which describe human emotions using facial expressions. They help us to understand the emotional state of human from expressive features and categorize them based on the emotion [de Lera and Garreta-Domingo, 2007]. Ekman and Friesen [1977] developed the Facial Action Coding System (FACS) to code facial expressions where movements on the face are described by a set of action units (AUs). It codes the facial action using 44 facial Action Units (AUs) from psychology point of view. Each AU has some related muscular basis. FACS divides the face into upper and lower face action and splits each sub-face into action units (AUs). FACS was not applied in Computer Vision for facial behaviour until the 1990s. According to the rules of FACS, a trained computer program automatically decomposes the expression into a specific set of AUs that describe these expressions. Tian et al. [2001] extended this method to recognize 17 human emotions using facial images. Rizon et al. [2007] used genetic algorithms for extraction of facial features like the eyes, nose, mouth, etc. Ellipse fitting algorithm which

is a genetic algorithm can be used to compute the emotion. The shape of human lip can be considered as a combination of two ellipses. Furthermore, the ellipse is termed irregular, since it has two minor axis of different lengths and a fixed major axis.

Images of the crying faces of babies can be analyzed and the reason for the cry can be classified as sad, anger, hunger, pain and fear [Pal et al., 2006]. Input to the emotion detection system includes static image of the crying face of the infant. The changes that occur to facial features like mouth, eyes and eyebrows were noted. The different states of these facial features includes open eyes, closed eyes, raised eyebrows, lowered eyebrows, open mouth and closed mouth. Combination of mouth state, eye state and the position of eyebrows were used to determine the reason for the baby's cry.

2.4.5 Action recognition

Recognition procedures generally consist of three steps; detect the target object, generate useful representation, recognize and analyze the object movement. The first two steps create an object representation that applies various image processing techniques, while the last step identifies the action pattern and produces semantic description. Representing the target object movement through a sequence of images facilitates the recognition process. It provides a basic model that can be further processed by different recognition approaches. There are many representations that can illustrate the movements from an image sequence. Local features, skeletonisation and temporal templates have all been investigated in combinations with different recognition methods. The following reviews will summarize these approaches.

Local Features. Local features can be represented using a histogram representation that identifies spatio-temporal features from a sequence of images by extracting the interest points around various events [Laptev and Lindeberg, 2006]. The interest points represent locations in the image that have crucial variations in both spatial and temporal domains [Laptev, 2005]. Spatial variations represent changes within the frame, while temporal variations represent changes between frames in the video sequence. Various events can be captured depending on the spatial and temporal scales. This approach creates a histogram with brightness surfaces representing the background and dark ellipsoids representing the interest points that hold significant events such as stopping and starting a feet movement [Laptev and Lindeberg, 2006]. Interest points can be detected by finding the regions with significant eigenvalues using the Gaussian kernel and Harris corner equations. The approach has been successfully applied to the recognition of human activity.

Schuldt et al. extended this approach for finding descriptors that represent the image structure around these interest points. It can be achieved using spatio-temporal gradients and optic flow by using histograms, N-jets (a set of derivatives to the order N) or the principal components analysis (PCA) in different combinations. These descriptors achieved good performance when they were used in conjunction with histograms. They do not require image segmentations or pre-processing while some other approaches rely on complementary segmentations or special image pre-processing.

2. RELATED WORK

Skeletonisation. This approach transfers a temporal image sequence to a feature vector sequence by representing the object with a skeleton. A 2D stick figure, consists of 13 to 18 feature points representing the highlighted points on body joints, referred to as the moving light displays (MLD) [Cutting and Kozlowski, 1977]. Another example is a 2D labelling, provided by [Leung and Yang, 1995] that represents a tree of connected parts at certain body joints. Chen et al. [2006] proposed a method using star skeleton. It works by identifying a human object in the scene using the background subtraction approach that isolates the required pixels from the background [Fujiyoshi and Lipton, 1998]. Isolated foreground objects should be processed in order to enhance its quality by a morphological dilation followed by erosion; this cleans up the extracted objects and smoothes their lines. The border points are extracted to determine the external outline of the target. A star skeleton is drawn from the five extreme points in the object outline. Five points normally represent a head, two hands and two legs. Lastly, five points with the maximum distances are chosen to represent the target object skeleton in the star fashion. It is the effective approach to extract object features in two-dimensional environments using a five-dimensional vector.

Temporal Templates. Temporal templates representation has been used by [Wren et al., 1997] to construct a view-specific representation of human motions within an image sequence. It is based on two concepts: where the motion is and how the motion occurs. In order to achieve this, two components must be constructed: Motion Energy Image (MEI) and Motion History Image (MHI). The MEI represents the area of the image that contains motion; it is used to confirm that there is motion activity and to show the angle of view. On the other hand, the MHI represents the direction of the movement; the most recent moves will be shown by brighter pixels. In order to apply this approach, an image sequence should contain one person and the tracking movement should be isolated from the rest of the objects present in the scene.

Action Recognition. In monitoring human activity, each action may have a unique HMM with certain parameters to describe it [Yamato et al., 1992]. Typical model parameters consist of the state transition probabilities, the output probabilities and the initial state probability. In the learning phase, each category or action parameters are optimized to best describe the action. The recognition phase finds the HMM that produces the symbol sequence. Chen et al. [2006] used HMM in combination with skeleton representation to recognize different actions. They extracted the feature vectors from the images, and assigned a symbol to each feature vector using a codebook created by Vector Quantization (VQ) to store these symbols. Symbols were used for training the models and recognizing the best equivalent action.

Learning and matching algorithms consists of two steps: learning and recognition. The learning phase is responsible for creating a set of categories, one for each movement, by computing parameters from a large number of training data. On the other hand the recognition phase compares the input with each category to find the best matches. Bobick and Davis [2001] used learning and matching algorithm to recognize the temporal template that has been described previously in this chapter. Their algorithm worked by collecting large sample of data to represent each movement with different views. They computed statistical descriptions from

the temporal templates using moment-based features. These descriptors are used as training data. To recognize the test data, the Mahalanobis distance of the input and each movement descriptors was calculated. Finally, the movement with approximately similar distance will be the candidate movement object. If more than one candidate were found, then the one with the smallest distance is the matching movement.

There are many other approaches used as action recognition such as detecting the cyclic motion for specific body segments. Using this approach legs and torso are defined by the autocorrelation and Fourier transformation techniques. Another approach invented by [Polana and Nelson, 1994] is used to recognize repetitive motion represented by low-level features for the entire body. They classified the motions using a nearest centroid algorithm, which is a Kernel based classifier for assigning the testing data to a specific class attained by the training data.

2.4.6 Object Recognition

The image/video analysis community has long struggled in bridging the semantic gap from low-level features to high-level semantic content. To overcome this gap, recent years have seen the emergence of a new retrieval approach, called concept-based retrieval, which aims to design and utilize a set of intermediate semantic concepts [Naphade and Smith, 2004] to describe frequent visual content in video collections and improve the retrieval performance. These concepts cover a wide range of topics [Chang et al., 2008] such as those related to people (face, anchor, etc), acoustics (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports) and production (camera motion, blank frames).

Semantic Concept Detection. The task of semantic concept detection has been investigated by many studies. Their successes have demonstrated that a large number of high-level semantic concepts are able to be inferred from low-level multi-modal features. Typically, the first step of developing semantic concept detection systems is to define a meaningful and manageable list of semantic concepts based on human prior knowledge. Most previous work approached concept detection as a supervised learning problem that attempts to discriminate positive and negative annotated examples through automatically extracted low-level features. As the first step, a variety of low-level features are extracted from several modalities, e.g., text, audio, motion and visual modality. For each concept, separate uni-modal classifiers are built using the corresponding labelled data and low-level features. One of the most common learning algorithms is called support vector machines (SVMs) [Burges, 1998], which have been proposed with sound theoretical justifications in order to provide good generalization performance. SVMs have been applied in image processing domain for variety of tasks such as 2-D and 3-D object recognition, human action classification, activity detection. Apart from SVMs, there are a large variety of other classifiers [Naphade and Smith, 2004] that have been investigated, including Gaussian mixture models (GMM), hidden markov models (HMM), k nearest neighbour (kNN), logistic regression, and adaboost. Basic area of application for these classifiers is object and

2. RELATED WORK

object categories recognition, image annotation, segmentation and classification.

To further refine the detection results, it is beneficial to combine prediction outputs from multiple modalities that provide complementary information with each other. Generally speaking, there are two families of multi-modal fusion approaches, i.e., early fusion and late fusion. The early fusion method begins with merging multi-modal features into a longer feature vector and takes as the input of learning algorithms. In contrast, the late fusion method directly fuses the detection outputs from multiple uni-modal classifiers. Both fusion methods have their own strengths and weaknesses [Snoek et al., 2005], but late fusion appears to be more popular and more extensively studied than early fusion in the literature. Finally, since the detection results of semantic concepts are not related to any query topic, they can be indexed offline without consuming any online computation resources. Such detection approaches have been applied in most existing video semantic concept extraction systems.

Recognition of Object Instances. Object recognition is not the ultimate goal of this research. A brief discussion about the object recognition methods is provided. The goal of visual object recognition is to detect the presence of an object in an image, and possibly localize the object in the image and estimate its pose. This usually involves designing an object representation that can model the imaged appearance of an object under a broad class of imaging conditions, such as varying object and camera pose, scene lighting, partial occlusion, and possibly deformation. Such representation should also be robust enough to deal with large amounts of background clutter. Early approaches to object recognition made strong simplification assumptions about the real world. An example is the ‘Blocks world’ where objects were made of combinations of polyhedral on a uniform background [Roberts, 1973].

Later, objects were represented by combinations of generalized cylinders, which enabled modelling of fairly complex curved objects. Object instances can be found in the video using one of these methods. Early works were based on 3D geometric object models and geometric invariants. Then global appearance methods that essentially represent objects by storing images taken from different viewpoints. Finally, approaches representing objects by local image regions. Object recognition methods can be classified into one of the following. Geometry based methods, Global appearance methods, local appearance methods, affine covariant regions, local descriptors, representing 3D objects using local patches [Sivic and Zisserman, 2006].

Recognition of Object Categories. In contrast to object instance recognition, which is reaching some maturity, object class recognition is still a very active research area. There are some competing, and sometimes complementary, approaches. Object categories are important as they model the objects in hierarchal form. This hierarchal structure is also beneficial in the task of topic detection in the visual domain. Hierarchies in visual scenes can be constructed using bottom up and top down approaches and decisions are postponed in the presence of uncertainty [Marszalek and Schmid, 2007]. Topic models with spatial transformations and geometric constraints are used to produce transformed Dirichlet process (TDP) in which Monte Carlo algorithms recognize objects in street and office scenes [Sudderth et al., 2008]. A method to extract multiple object categories and their locations given a set of images containing, without

supervision using probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA) was proposed [Sivic et al., 2005].

There are other hierarchical probabilistic models for the detection and recognition of objects in cluttered, natural scenes using low level features [Sudderth et al., 2008]. A biologically inspired model of visual object recognition was applied to the multiclass object categorization problem [Mutch and Lowe, 2006]. Simultaneous recognition and localization of multiple object classes using a generative model is also possible [Mikolajczyk et al., 2006]. There are some methods where visual dictionary is generated for objects categorization. Visual dictionary is learned by pair-wise merging of visual words from an initially large dictionary. The final visual words are described by GMMs [Winn et al., 2005]. Object class recognition by unsupervised scale-invariant learning is also possible in which case objects are modelled as constellation of parts [Fergus et al., 2005]. Some work related to semantic hierarchies is also on the go like we can get semantic hierarchies for visual object recognition by incorporating semantics of image labels to integrate prior knowledge about inter-class relationships into the visual appearance learning [Marszalek and Schmid, 2007].

There are methods where visual data can be divided into layers and outer layer are composed of subsequent inner layers [Fidler and Leonardis, 2007]. An unsupervised learning of categories from sets of partially matching image features is present in which case images are divided into local features and put in space where clustering and affinities are used for merging [Grauman and Darrell, 2006]. Scene interpretation task was also applied to movies using the motion content shot length and colour properties of shots as the features. Backward shot coherence (BSC) and scene dynamics (SD) are the major technologies used by [Rasheed and Shah, 2003].

Object Detection using Haar Classifier. Viola and Jones [2001] presented method of object detection using rectangular Haar feature. Out of all the rectangular features in an image only the ones that best separate between positive and negative image were selected by the classifier for training. An optimum threshold was obtained for each feature so minimum misclassification takes place. They used adaboost classifier to boost the efficiency of weak learners. Different classifiers are used in cascade to build a classifier where positive response from one classifier triggers the next classifier and if negative response is received that subwindow is rejected. Lienhart and Maydt [2002] used same approach as [Viola and Jones, 2001] and also added rectangular feature at 45 degree to the initial Haar feature set. The results achieved showed that even though the computation time of the classifier increases there was reasonable increase in object detection task.

2.4.7 Scene Settings- Indoor / outdoor Scene Classification

Classifying an image into indoor/outdoor image category is computationally a very challenging task. It is difficult due to vast range of variations in both of these scene categories. An outdoor image can be of beach, urban scene, houses, road traffic etc and an indoor image can have scene from office, bedrooms, bakery, hotels etc. Therefore indoor/outdoor image classification is highly researched topic. A lot of previous work on indoor/outdoor image classification has

2. RELATED WORK

been done using the low level features of an image such as colour and texture. The global or local approach was used for classifying and improving performance of the classification. Some used combination of different features to improve performance of classification. This chapter discusses such previous work on indoor/outdoor image classification and results achieved by them.

The early research on indoor/outdoor image classification was done by [Szummer and Picard, 1998] who used 3 features of the image for classification. They were colour, texture and frequency. These features were calculated for the entire image as well as for subblocks of the image. First tests were performed using RGB colour space. Then the colour space was changed to OHTA [Ohta et al., 1980] and using histogram intersection the performance of the classifier increased significantly. They also calculated the frequency feature using 2D DFT (2 Dimensional Discrete Fourier Transform) and then taking 2D DCT (2-Dimensional Discrete Cosine transform) of an image. Finally combining all this features a k-nn classifier was trained and best performance was achieved by this combination. In their research the results obtained by using entire image were better than using subblocks of image.

Luo and Savakis [2001] their method along with low level features used semantic features like sky and grass in a image for classification. For colour and texture information they used techniques similar to [Szummer and Picard, 1998]. Mid level semantic feature extraction was done using the ground truth information of a image such that sky and grass are always true. Serrano et al. [2002] based their approach on low level features of image like colour and texture just like [Szummer and Picard, 1998] but using half the dimensions. They used LST colour space to generate a 16-bin histogram for each colour channel and used it to generate a 48 dimension feature vector to train the SVM classifier. The texture feature was obtained by using two level wavelet decomposition. A two stage classifier was then used in which the first stage was used to extract the colour and texture features from an image, from the first stage the resulting distances were calculated and these distance values were used to produce a new distance feature. A new SVM was trained with this distance feature which yielded best results on testing dataset.

Payne and Singh [2005] provided a novel way of classifying the images. Their method was based on the hypothesis that organic objects had a large amount of small erratic edges whereas synthetic objects had straight and less erratic edges. They used canny edge detector to detect edges in images. After which they calculated the straightness value of edge between start and end point of edge. The image was then divided in 16 equal size blocks and edge straightness value for each block was calculated. The results were obtained by two methods, in one method a threshold value was obtained and used to classify the image and in other method a k-nn classifier was trained to classify the image. Their method failed when images contained some objects prevalent in both indoor and outdoor environments. Da Deng [2005] in their experiment they extracted features from local and global part of the image. The global features were the ones obtained from the entire image and the local features were obtained from segmented parts of the image. They used the LUV colour space from which each channel was quantized into 5 bins and a colour histogram was calculated as a global colour feature. To calculate the edge

histogram descriptors (EHD) edge descriptors were applied to the image to detect edges of the 2x2 pixels block and quantised into six edge features like horizontal, vertical, 45, 135 degree, non-directed and no edge. Related to this features a 6-bin histogram was obtained. EHD were then calculated from all 4x4 blocks of image creating a 96-dimension histogram feature. For the local features image segmentation was done based on the colour, texture information in which similar regions were merged through region growing. Classifier for all 4 features (2 local, 2 global) were built and applied. Kim et al. [2010a] calculated the Edge Oriented Histogram (EOH) and Colour Oriented Histogram (COH) from the image and used a combination of these two features to train a SVM classifier. This classifier was then used for predicting class of image and results shown by them are satisfactory.

2.4.8 Speaking Person Identification

Speakers in a video sequence can be identified from their visual features, such as head movements and mouth shape changes. The person who makes more head and hand movements is most likely to be the person speaking. This claim is backed up by [Rose and Clarke, 2009], who conducted an experiment that supported the hypothesis that speakers moved more than listeners. Bull and Connelly [1985] also showed that body movements can be linked to speech and aimed to show a significant relationship between body movement and phonemic clause structure. A phonemic clause consists of about five words, with changes in pitch, loudness and rhythm indicating just one main stress. The clause stops at a juncture, when the pitch, loudness and rhythm average out again. The next phonemic clause then starts after the juncture. The second test in this paper investigates the relationship of body movement and vocal stress, using phonemic clauses. The results of this test showed a strong relationship between the two. They found the tonic stresses were accompanied by body movements, to be more precise over 90% of the tonic stresses had body movements.

Cootes et al. [1995] have come up with the Active Shape Models (ASM) which is capable of capturing the variability of image structures which belong to the same class. ASM maintains a training set where images are marked with set of landmark points which will exist in almost all the similar images. A Point Distribution Model (PDM) is derived out of the figures of the variations for the labelled points in the training set. The shapes are allowed to be varied within this model. A new image is identified with the model if it exists within the deformity of the set of images the model contains. Iwano et al. [2001] have used the Optical Flow Analysis Method for tracking the lip movement. By this method we can determine the object movement. The images of the lip region are extracted from the video and then it takes a pair of adjacent images for calculating the optical flow velocities. The horizontal and vertical features of the images are computed which tells you whether the mouth is moving or not. It basically helps when the mouth pauses or mouth is shut.

Kaucic and Blake [1998] discussed lip tracking and that most approaches utilise Kass's snake approach which track the outer lip despite the impressive performance that comes from this approach it is not distinctive enough. Another tracker is instead of relying on prior models it

2. RELATED WORK

can detect the nostrils, and then colour thresholds are used to identify the black area in the inner mouth region, contour is then grown around the area identified as the inner mouth. [Kaucic and Blake \[1998\]](#) describes a possible way of tracking the lips by using dynamic contour tracking through sparse representation of the lip contours, via Bsplines, combined with a Kalman filter utilising prior shape and motion models of deforming lips. For image feature detection, edge detection is used for the lips but it can be difficult as lip colour is similar to the skin colour. Using Bayesian classification, which uses probability for whether the pixel belongs to the lips by its classification of colour. Afterwards Fishers linear discriminate analysis is used to determine the boundary between the lips and facial skin. The next step is the inner-outer lip contour tracking and inner contour tracking enables additional reasoning to be made about the presence of the tongue and teeth.

[Yuille et al. \[1992\]](#) used shape templates with snakes in order to extract the lip contours from an image face. The lips can be described using a parameterized shape template. The shape template models an object within the image. By adjusting the parameters the model can be made to deform to fit the object in the image. The shape of this template is based on prior knowledge of the shape of the lips. One of the most successful lip reading systems to the present date is developed by ([Bregler and Omohundro \[1995\]](#))

2.5 Natural Language Generation

Natural Language Generation (NLG) is the process of constructing natural language outputs from non-linguistic inputs [[Reiter and Dale, 2000](#)]. Its goal is to generate readable and user understandable natural language text from a machine representation system such as a knowledge base or a logical form. In video processing domain, NLG system can be considered as ‘*machine translator*’ that converts a video stream into a natural language representation. There are two widely adopted approaches to NLG, the ‘*deep-linguistic*’ and the ‘*template-based*’ [[Gagne et al., 2005](#)]. The deep-linguistic approach attempts to build the sentences up from a logical representation. The template-based approach make use of templates that contain a predefined structure with slots for filling those structure. Deep linguistic approach is flexible and can generate variety of sentences while template based approach is commonly specific to defined domains.

2.5.1 Stages of Natural Language Generation

Approaches for natural language generation vary a lot based on their respective goals. Sometimes, simple copying and pasting text with function words may produce satisfactory results such as for horoscope machines or personalized business letters. However, a sophisticated NLG system needs to include stages of planning and merging of information to enable the generation of text that looks natural and does not become repetitive. [Dalianis \[1996\]](#) introduced a two-stage framework for NLG systems containing (1) ‘*Discourse Planner*’ which chooses the appropriate content to express the communicative goal, and gathers information from a knowledge base to

transform the content into text and, (2) ‘*Surface Realizer*’ which generates sentences for the discourse specification based on its lexical and grammatical resources. Reiter and Dale [1997] proposed three staged framework for NLG systems, where they composed discourse planner into two components, *i.e.*, text planner and sentence planner. Text planner deals with the contents, sentence planner decides structure of sentence and linguistic realization is concerned with morphological and syntactic processing.

Text Planning. This task is further divided into two subtasks, *i.e.*, content determination and discourse planning. Content determination deals with selection of contents for NLG. It includes the information which needs to be communicated. Discourse planning is related to structure of the contents. Structure of the information and order of that information presentation is decided by this subtask. Main concern is to keep the rhetorical structure of the produced text intact. These tasks can be done at different levels of sophistication. One approach is use of simple hard coded text planner written in some programming language like C or Java. It may be less flexible, but can be effective if the produced text has standardized content. Other approaches include rule-based systems and schema or text planning languages based systems. Text plans are represented as transition networks of one sort or another, with the nodes giving information content and arcs giving the rhetorical structure.

Sentence Planning. Sentence planning is related to structure of sentences as sentences are useful for producing user understandable outputs. This task is further divided into three subtasks. Sentence aggregation, lexicalization and referring expression generation. Sentence aggregation is used for deciding structure of individual sentences. Decision of information/messages that should be combined to form a single sentence is made by this subtask. Finally, range of information coverage by a single sentence is decided by this subtask. Linguistic means to combine messages (clauses) are investigated for the purpose of sentence aggregation. There are different type of clauses such as relative clauses, coordination, subordination and lists¹. Lexicalization is related to presentation of domain concepts and relations, *i.e.*, selecting words and phrases to express domain concepts. Referring expression generation deals with entities to be referred. The common use of this step is not to change the contents of the NLG system. Rather focus is to improve the fluency and readability of the produced text. Another aim of this task is to present text such as it is human written.

Linguistic Realization. This step corresponds to well structured, well formed and grammatically correct produced text. The main concern is that the produced text is orthographically well formed. A realizer generates individual sentences with correct English grammar and rules. Some of the rules are point absorption and other punctuation rules, morphology, agreement and reflexives. There are some general purpose engines which help in application of linguistic rules such as FUF [Elhadad, 1992].

Reiter and Dale [2000] further elaborated stages for natural language generation systems to make distinction between subtasks more clearer. ‘*Content determination*’ is related to selection of contents which need to be expressed. ‘*Lexicalization*’ encompasses selection of words to

¹<http://www.cs.uiuc.edu/class/fa08/cs498jh/Slides/Lecture23HO.pdf>

2. RELATED WORK

express the concepts. ‘*Document structuring*’ deals with overall organization of the information to convey. ‘*Syntactic and morphological realization*’ is focussed on producing the surface document or text by using syntactic and morphological rules. ‘*Aggregation*’ decides merging similar sentences into one sentence. ‘*Referring expression generation*’ deals with creation of referring expressions that identify objects and regions. It further identifies pronouns to replace repeated noun phrases. ‘*Realization*’ is related to creation of the actual text, which should be correct according to the rules of syntax, morphology, and orthography. Resolving matters such as formats, casing, and punctuation.

SimpleNLG. There exists several freely available NLG systems. One of them is SimpleNLG, which can be used to write a program which generates grammatically correct English sentences [Gatt and Reiter, 2009]. It is a library (not an application) written in Java, that performs the basic tasks necessary for natural language generation; it assembles parts of a sentence into grammatical form and outputs the result. For example, it capitalizes the first letter of the sentence, can add an auxiliary verb and make it agree with the subject, or simply add ‘*ing*’ to the end of the verb if the progressive aspect of the verb is desired. Working of SimpleNLG can be divided into three tasks.

- **Grammar:** SimpleNLG puts sentences into grammatical form (enforces noun-verb agreement, creates well formed verb groups, etc).
- **Morphology:** SimpleNLG handles in inflection (modifies words to reflect information such as gender, tense, number or person).
- **Orthography:** SimpleNLG inserts the appropriate whitespace between the words of the sentence, puts a period at the end of the sentence, and formats lists such as apples, pears, and oranges.

2.6 Natural language Description of Videos

It is becoming popular to introduce natural language concepts into a vision system. While literature relating to object recognition [Galleguillos and Belongie, 2010], human action recognition [Torralba et al., 2008], and emotion detection [Zheng et al., 2010] are moving towards maturity, automatic description of visual scenes is still in its infancy. To a certain extent machines are able to identify human activities in videos [Torralba et al., 2008] but only a small number of works exist towards automatic description of visual scenes. Most studies in video retrieval have been based on keywords [Bolle et al., 2010].

On the whole, previous approaches follow the similar strategy for converting images to natural language descriptions. To begin with, the image is represented by image features, which are then replaced by an abstract representation, essentially a set of description words, according to a visual-to-textual representation dictionary. The features used to represent the image content mainly include color information [Héde et al., 2004; Kojima et al., 2002; Yao et al., 2010], textual features [Héde et al., 2004; Yao et al., 2010], detected edges [Kojima et al.,

2002], and so on. For certain applications, some objects are detected and recognized with prior knowledge to supply higher level features [Abella et al., 1995; Héde et al., 2004; Kojima et al., 2002; Yao et al., 2010].

Héde et al. [2004] presented description of objects based on semantic relations such as ‘an orange ball’ indicated a single word based on the relationship between orange and ball while isolated words orange, ball presented two separate objects, an orange and a ball, rather than single object (an orange ball). Initially a dictionary of objects was created based on image signature which itself consisted of features such as color and texture and object’s name and its category. Secondly, images are segmented into regions and corresponding signatures are fetched from the database by comparing the region features with entries in the dictionary. Finally, description is generated based on the retrieved signature keywords using manually defined templates.

Kojima *et al.* presented a method for describing human activities based on a concept hierarchy for actions [Kojima et al., 2002]. They described head, hands and body movements using natural language texts. Firstly, human poses and head movements along with their trajectories are identified in the form of numerical values. Secondly, these numerical values are converted into actions such as enter, carry, turn and exit *etc.*, based on a manually created concept dictionary. Finally, these actions are combined to generate natural language descriptions using predefined grammars. Kojima *et al.* further improved their method by incorporating more objects and interaction between human and non human objects. They further extended the concept hierarchy of actions related to human body and their interaction with other objects in office environments [Kojima et al., 2008].

For a traffic control application, Nagel investigated automatic visual surveillance systems where human behaviour was represented by scenarios, consisting of predefined sequences of events [Nagel, 2004]. The scenario was automatically translated into text by analysing the contents of the image over time, and deciding on the most suitable event. Lee *et al.* introduced a framework for semantic annotation of visual events in three steps; image parsing, event inference and language generation [Lee et al., 2008b]. Instead of humans and their activities, they focused on object detection, their inter-relations and events in videos. Baiget *et al.* performed human identification and scene modelling manually and focused on human behaviour description for crosswalk scenes [Baiget et al., 2007].

Yao *et al.* introduced a framework for video to text description which is dependent on the significant amount of annotated data [Yao et al., 2010]. Main building blocks of this framework are image parser, visual knowledge representation, the semantic web and a text generation module. Firstly, images are hierarchically decomposed into their constituent visual patterns into an And-Or Graph. Secondly, these graphs are converted into structured representations with specified semantic relations including categorical, spatial and functional relations using a visual knowledge database. Finally, natural language descriptions are generated with the help of the semantic web using templates or grammars which are manually defined for specific applications such as video surveillance. However, both the parser and visual semantic representation are built based on a large-scale ground truth image database which is manually annotated.

2. RELATED WORK

The approaches discussed above generate grammatical natural language sentences for images or videos by analyzing the image content. However, note that all of them rely on large amounts of manually created resources. This includes the annotation of the image database for the training purpose, the construction of a visual-textual correspondence dictionary or ontology, and the engineering of application-specific sentence templates or grammars for generation. And most of this data can not be reused cross domains or applications and thus manual effort must be invested for a new one, which is obviously costly and time consuming.

Recently, research community has shifted its focus towards exploiting textual data and natural language processing techniques for generating image descriptions. Main interest while combining vision and natural language is to investigate any significant improvement towards producing readable and descriptive sentences compared to naive strategies that use vision alone.

Li *et al.* proposed a three steps framework for generating textual descriptions of images [Li *et al.*, 2011]. First step is related to image processing and identification of objects, their visual attributes and spatial relationships between them. These three types of visual output are presented in the form of tuples (one triple for every pair of detected objects). Finally, smooth and well phrased sentences are generated using web-scale n-grams which provide frequency count of each possible n-gram sequence for $1 \leq n \leq 5$. Yang *et al.* presented a framework for static images to textual descriptions by predicting nouns, verbs, scenes and prepositions that make up the core sentence structure. Initially, nouns and scenes were detected using image processing methods. Secondly, a language model was trained from the English Gigaword corpus to estimate the verbs and prepositions. These estimates are used as parameters of a Hidden Markov Model that models the sentence generation process, with hidden nodes as sentence components and image detections as the emissions. Both of the above mentioned works are structured for static images instead of video streams. Video streams have temporal information attached which can provide some useful information such as movements and actions being performed by the objects. Yang *et al.* restricted their work to images containing maximum of two objects [Yang *et al.*, 2011].

This work contrasts to most previous approaches in several aspects:

- Descriptions are generated from scratch, instead of retrieving [Farhadi *et al.*, 2009], or summarizing existing text fragments associated with an image [Aker and Gaizauskas, 2010; Feng and Lapata, 2010].
- Descriptions are based on the real contents of the video sequences rather than using news related information [Feng and Lapata, 2010] or encyclopedic text [Aker and Gaizauskas, 2010] which is not related to the real visual contents.
- A generic framework for descriptions generation is provided instead of domain specific hand-written grammar rules [Yao *et al.*, 2010].
- Finally, descriptions are generated for sets of images which can include more useful high level features such as human action and motion information as compared to approaches which are based on single images only [Li *et al.*, 2011; Yang *et al.*].

2.7 Video scene classification

There are a large number of research studies in natural language processing for information extraction, retrieval and summarisation. Their ideas and approaches have also been successfully extended to spoken document and video retrieval, classification tasks [Chen, 2006; Liu and Chen, 2007; Xu et al., 2008; Zeng et al., 2010]. Most of the text retrieval techniques are based on the vector space model, with which features are built into a term-document matrix. Various classifiers, such as the k-nearest neighbour (kNN) and the support vector machine (SVM), can be used with the feature matrix. Because there are many synonyms (*i.e.*, words with the same meaning) in text documents, latent semantic analysis (LSA) [Papadimitriou et al., 1998, 2000] and probabilistic LSA (pLSA) [Hofmann, 1999] can be applied. They are able to find the hidden semantic space, projecting terms having the similar meaning to the close location in that space. Latent Dirichlet allocation (LDA) [Blei et al., 2003] can be viewed as an extension of pLSA with a Dirichlet prior; it builds a three-level hierarchical Bayesian network. It has been reported that all these techniques are able to achieve a sound performance not only for text processing but also for spoken document and video signal processing [Bosch et al., 2006; Lazebnik et al., 2006; Niebles et al., 2008].

Many approaches have been developed for information extraction from text documents; they include document frequency, chi-square statistic, term strength, mutual information and information gain. Their performances were compared by Yang and Pedersen using a text classification task [Yang and Pedersen, 1997]. Comparison was also made by Hazen *et al.* between information gain, chi-square statistic and maximum a posteriori (MAP) probability estimates [Hazen et al., 2007]. Their analysis indicated that document frequency, information gain and chi-square statistic scores were strongly correlated and that MAP estimates can achieve better in topic identification.

2.7.1 Feature Extraction

The traditional *tf-idf* feature characterised the relation between a term t and a document d . The chi-square statistic and the MAP estimate are able to reflect the relation between a term t and a video scene classification c . ***Tf-idf* term-document matrix.** Suppose that we have D documents and a list of T vocabulary terms that occur in any of these documents. *Tf-idf* can give a measure of the importance for a term t in a particular document d [Salton and Buckley, 1988]. It is calculated by

$$tfidf(t, d) = tf(t, d) \cdot idf(t) \quad (2.1)$$

$tf(t, d)$ and $idf(t)$ are referred to as the term frequency and the inverse document frequency respectively. The term frequency calculated as follows:

$$tf(t, d) = \frac{N_{t,d}}{\sum_{\tau} N_{\tau,d}} \quad (2.2)$$

2. RELATED WORK

where $N_{t,d}$ is the number of occurrences of a term t in a document d , and the denominator is the sum of occurrences of all terms in that document d . The inverse document frequency is

$$idf(t) = \log \frac{D}{W(t)} \quad (2.3)$$

where $W(t)$ is the number of documents containing the term t .

Finally a term-document matrix X is defined as a $T \times D$ matrix of the form

$$X = \begin{bmatrix} tfidf(t_1, d_1) & \dots & tfidf(t_1, d_D) \\ \vdots & & \vdots \\ tfidf(t_T, d_1) & \dots & tfidf(t_T, d_D) \end{bmatrix} \quad (2.4)$$

with T terms in the rows and D documents in the columns.

Chi-Square Statistic. A chi-square test is used to evaluate the independence between two events. The relevance of a term t in a scene class c can be estimated by the following formula [Manning et al., 2008]:

$$\chi^2(t, c) = \frac{(F_{11} + F_{10} + F_{01} + F_{00}) \times (F_{11}F_{00} - F_{10}F_{01})^2}{(F_{11} + F_{01})(F_{11} + F_{10})(F_{10} + F_{00})(F_{01} + F_{00})} \quad (2.5)$$

where

F_{11} : #documents belonging to c and containing t ;

F_{10} : #documents which are not in c but containing t ;

F_{01} : #documents belonging to c but not containing t ;

F_{00} : #documents which are not in c and not containing t .

MAP Estimates. A typical use of a MAP estimator is for a classification task, *i.e.*, to decide which class to choose. It is used to measure the importance of a term t for a scene class c . The larger the posterior probability $P(c | t)$ is, the more important the term t is. Conversely if t does not occur in c , the probability can be zero. The MAP estimates may be calculated in the following equation, which can handle the zero probability [Hazen et al., 2007]:

$$map(t, c) \equiv P_{map}(c | t) = \frac{N_{t|c} + \alpha_1 N_c P_{map}(c)}{N_t + \alpha_1 N_c} \quad (2.6)$$

where $N_{t|c}$ and N_t are the numbers of term t in the scene class c and in the entire corpus, respectively. N_c is the number of distinct scene classes. Further,

$$P_{map}(c) = \frac{N_{d|c} + \alpha_2}{N_d + \alpha_2 N_c} \quad (2.7)$$

where $N_{d|c}$ is the number of documents in the scene class c , and N_d is the entire number of documents. Note that α_1 and α_2 are the smoothing parameters that are typically determined empirically.

Combination. In this section we consider combination of the *tf-idf* term-document matrix with the chi-square statistics and/or the MAP estimates for scene classes. To this end the latter two can be normalised for each class c as follows:

$$\widehat{\chi^2}(t, c) = \frac{\chi^2(t, c)}{\sum_{\tau} \chi^2(\tau, c)} \quad (2.8)$$

$$\widehat{map}(t, c) = \frac{map(t, c)}{\sum_{\tau} map(\tau, c)} \quad (2.9)$$

For each scene class, there is only one vector to represent each term weight in the vocabulary. The *tf-idf* scores can be combined with the normalised forms of the chi-square statistics and the MAP estimates:

$$\sigma(t, d) = w_t tfidf(t, d) + w_c \widehat{\chi^2}(t, c_d) + w_m \widehat{map}(t, c_d) \quad (2.10)$$

using weighting factors w_t , w_c and w_m that satisfy

$$w_t + w_c + w_m = 1$$

and c_d indicates the class to which a document d belongs. Using $\sigma(t, d)$, a revised term-document matrix X' is given by

$$X' = \begin{bmatrix} \sigma(t_1, d_1) & \dots & \sigma(t_1, d_D) \\ \vdots & & \vdots \\ \sigma(t_T, d_1) & \dots & \sigma(t_T, d_D) \end{bmatrix} \quad (2.11)$$

2.7.2 LSA (Latent Semantic Analysis)

LSA has been successfully applied for information extraction and retrieval tasks in the natural language processing area. The idea can be applied for a video classification task using their metadata annotation.

Basic Idea. LSA maps a document into the semantic space [Deerwester et al., 1990]. Synonyms can be projected close to each other in the reduced space. Technically, LSA is based on the singular value decomposition (SVD) of a $T \times D$ term-document matrix X :

$$X = U\Sigma V^* \quad (2.12)$$

2. RELATED WORK

where $*$ implies the transpose, and U and V are $T \times T$ and $D \times D$ unitary matrices satisfying

$$U^*U = I; \quad V^*V = I.$$

Further Σ is the $T \times D$ diagonal matrix containing the singular values $\{s_1, \dots, s_r\}$ of X where r represents the rank of matrix X . Note that singular values may be arranged in the descending order.

There are two ways to view the term-document matrix X . Firstly, X may be composed as follows:

$$X = \begin{pmatrix} t_1 \\ \vdots \\ t_T \end{pmatrix}$$

Note that t_i is a row vector of length D representing the i -th term spanning all D documents. Alternatively, X may be built up with

$$X = (d_1, \dots, d_D)$$

where d_j means a column vector of length T for the j -th document containing T terms.

To measure the documents' similarity, documents are projected onto the latent semantic space, then their cosine distance is calculated. The definition of the SVD (2.12) leads to the following expression for a document vector d_j with T terms:

$$d_j = U\Sigma v_j^* \tag{2.13}$$

and v_j is the j -th row of V with D elements. With a little manipulation of Equation (2.13) we obtain

$$v_j = d_j^* U \Sigma^{-1}. \tag{2.14}$$

If a test document q is given, it can also be projected onto the same space:

$$\hat{q}^* = q^* U \Sigma^{-1}.$$

The approach works well in many natural language processing tasks when the sufficient amount of data is provided. A potential problem is that a small size document (*e.g.*, short length annotation of a video clip) may not always carry sufficient information to make an accurate classification. Further, use of synonyms is less obvious than tasks with a large dataset, hence it is more difficult to improve the retrieval performance using the query expansion. The conventional use of LSA projects documents into the semantic space, effectively accumulating information into the fewer number of dimensions, which probably is '*not*' very helpful for classification.

Term co-occurrence by LSA projection. Projection to the latent semantic space by LSA has been well studied. There is an alternative interpretation for LSA; it is able to represent

co-occurrence patterns for terms and documents. Consider the following relation:

$$XX^* = (U\Sigma V^*)(U\Sigma V^*)^* = U(\Sigma\Sigma^*)U^* \quad (2.15)$$

Equation (2.15) represents the eigenvalue decomposition of XX^* whose individual elements are the inner product of t_i (row vector) and t_j^* (column vector):

$$(XX^*)_{ij} = t_i \cdot t_j^* \quad (2.16)$$

The above expression measures the similarity between the i -th and the j -th term vectors. Hence XX^* can be interpreted as the term similarity matrix.

The original term-document matrix X can be approximated using the K largest singular values. Typically K is chosen such that $K \ll T$ and $K \ll D$. The sparseness can be reduced in the K dimensional semantic space. This approximation is calculated by

$$\hat{X} \approx U\hat{\Sigma}V^* \quad (2.17)$$

where $\hat{\Sigma}$ is the $T \times D$ diagonal matrix containing the K largest singular values $\{s_1, \dots, s_K, 0, \dots, 0\}$ in its diagonal. Note that s_{K+1}, \dots, s_r in the original Σ are replaced with 0. We refer to this expression \hat{X} as *tfidf* term-document matrix.

In the similar fashion, Equation (2.13) is approximated as

$$\begin{aligned} \hat{d}_j &= (U\hat{\Sigma})v_j^* \\ &= \begin{bmatrix} u_{11}s_1 & \dots & u_{1K}s_K & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ u_{T1}s_1 & \dots & u_{TK}s_K & 0 & \dots & 0 \end{bmatrix} v_j^* \end{aligned}$$

The similarity of the i -th and the j -th term vectors depends on their frequency in each document, implying that matrix U , in the eigenvalue decomposition in Equation (2.15), can represent the co-occurrences of terms. In the approximation procedure, matrix U can influence the document vector \hat{d}_j .

2.8 Summary

This chapter presented review of literature related to natural language descriptions of video streams. Since development and testing corpus is heart and core of any research work, this chapter started with the discussion of video corpora and their metadata. Review of some of the annotation tools with their limitations was also provided. This research work is based on the extraction of useful HLFs from the video sequences, review of which was provided next. A brief introduction about NLG and its stages was presented next. Discussion of work which is similar to our proposed work is provided in the next section. The same section discussed pros and cons of each approach and highlighted differences between our approach and other approaches.

2. RELATED WORK

Finally a brief review about visual scene classification task was provided.

Chapter 3

Corpus Generation and Analysis

3.1 Introduction

We are exploring approaches to natural language descriptions of video data. The step one of the study is to create a dataset that can be used for development and evaluation because we could not identify a suitable dataset.¹ Creation of such corpora will help in limiting this research work to tight and manageable domains. It further leads to identification of high level features to be extracted by image processing methods. Finally this resource can be used for test data and ground truths for evaluation. Human-authored descriptions of the same visual scene vary greatly due to the individual influence of the people involved in description generation. In spite of these variations, there exists some similarities among descriptions derived from the same video. It can be argued that the dissimilarity is in the words used and the similarity is in the facts that are included in the description. Human authored descriptions can be used as a reference to measure the information content of machine generated descriptions.

To this end, three video datasets are manually generated and hand annotated with natural language descriptions. First dataset is named '*NLDV - Corpus 1*' and consists of short video segments. Each video segment contains a single camera shot and purpose of this dataset is to filter out important contents for description generation. These contents are then extracted automatically using image processing methodologies. Second dataset is named '*NLDV - Corpus 2*'² and consists of relatively lengthier videos compared to Corpus 1, which consist of video taken in multiple camera shots. This resource is needed for exploring relationships between individual shots and generating a story for complete video sequence. Third dataset is named '*NLDV - Corpus 3*' and used as an evaluation set for the language description framework. Section 3.2 presents details about the corpus 1, Section 3.3 explains corpus 2 and Section 3.4 explains corpus 3. Finally, findings based on corpora analysis are presented in Section 3.5.

¹At current time, there is no such publicly available resource.

²NLDV stands for Natural language descriptions for videos

3. CORPUS GENERATION AND ANALYSIS

3.2 NLDV - Corpus 1

In this study we select video clips from TREC Video benchmark. They include categories such as news, meeting, crowd, grouping, indoor/outdoor scene settings, traffic, costume, documentary, identity, music, sports and animals. The most important and probably the most frequent content in these videos appears to be a human (or humans), showing their activities, emotions and interactions with other objects. We do not intend to derive a dataset with a full coverage of all the video categories, which is beyond the scope of our work. Instead, to keep the task manageable, we aim to create a compact dataset that can be used for developing approaches to translate video contents to natural language descriptions.

Annotations were manually created for a small subset of data prepared for the rushes video summarisation task and the HLF extraction task for the 2007 and 2008 TREC Video evaluations. It consisted of 140 segments of videos — 20 segments for each of the following seven categories:

Action videos: Human posture is visible and human can be seen performing some action such as ‘*sitting, standing, walking and running*’.

Close-up: Human face is visible. Facial expressions and emotions usually define mood of the video (*e.g.*, happy, sad).

News: Presence of an anchor or reporter. Characterised by scene settings such as weather boards at the background.

Meeting: Multiple humans are sitting and communicating. Presence of objects such as chairs and a table.

Grouping: Multiple humans interaction scenes that do not belong to a meeting scenario. A table or chairs may not be present.

Traffic: Presence of vehicles such as cars, buses and trucks. Traffic signals.

Indoor/Outdoor: Scene settings are more obvious than human activities. Examples may be park scenes and office scenes (where computers and files are visible).

Each segment contained a single camera shot, spanning between 10 and 30 seconds in length. Two categories, ‘*Close-up*’ and ‘*Action*’, are mainly related to humans’ activities, expressions and emotions. ‘*Grouping*’ and ‘*Meeting*’ depict relation and interaction between multiple humans. ‘*News*’ videos explain human activities in a constrained environment such as a broadcast studio. Last two categories, ‘*Indoor/Outdoor*’ and ‘*Traffic*’, are often observed in surveillance videos. They show humans’ interaction with other objects in indoor and outdoor settings.

3.2.1 Annotation Process

A total of 13 annotators were recruited to create texts for the video corpus. They were undergraduate or postgraduate students and fluent in English. It was expected that they could

produce descriptions of good quality without detailed instructions or further training. A simple instruction set was given, leaving a wide room for individual interpretation about what might be included in the description. For quality reasons each annotator was given one week to complete the full set of videos.

Each annotator was presented with a complete set of 140 video segments on the annotation tool. For each video annotators were instructed to provide

- selection of high level features (*e.g.*, male, female, walk, smile, table);
- a title of one sentence long, indicating the main theme of the video;
- description of four to six sentences, related to what are shown in the video.

The annotations are made with open vocabulary — that is, they can use any English words as long as they contain only standard (ASCII) characters. They should avoid using any symbols or computer codes. Annotators were further guided not to use proper nouns (*e.g.*, do not state the person name) and information obtained from audio. They were also instructed to select all HLFs appeared in the video.

3.2.2 Corpus Analysis

13 annotations were created for 140 videos, resulting in 1820 documents in the corpus. They are referred to as **hand annotations** in the rest of this paper. The total number of words is 30954, hence the average length of one document is 17 words. We counted 1823 unique words and 1643 keywords (nouns and verbs).

Figure 3.1 shows a video segment for a meeting scene, sampled at 1 fps (frame per second), and three examples for hand annotations. For ‘*keywords*’, almost all the three annotators choose the similar HLFs from predefined HLFs, although there is some mismatch in human’s age and emotion information. For open ended keywords, annotator (1) and (2) supplied some keywords shown in the second against the heading of keywords in figure 3.1. For ‘*titles*’, hand annotations are usually very short, 1 sentence or phrase containing 3 to 6 words. Some annotators try to provide main theme of the video based on the semantic interpretation of the scene (annotators (1) and (2) in figure 3.1), while others just use the visual information without any context and try to present a very short title (annotator (3) in figure 3.1).

‘*Descriptions*’ typically contain two to five phrases or sentences. Most sentences are short, ranging between two to six words. Descriptions for human, gender, emotion and action are commonly observed. Occasionally minor details for objects and events are also stated. Descriptions for the background are often associated with objects rather than humans. Again, as for titles, descriptions vary on semantic and contextual basis. For example, (annotators (1) and (2) in figure 3.1) noticed semantic information in the video such as TV host, guests and interview scene while annotator (3) did not make any semantic boundaries and just provided plain descriptions. Finally, it is interesting to observe the subjectivity with the task; the variety of words were selected by individual annotators to express the same video contents, *e.g.*, annotator (1) used ‘*TV presenter*’ while annotator (2) used ‘*host*’ word for the same person.

3. CORPUS GENERATION AND ANALYSIS



Hand annotation 1

(**keywords**) male, adult, old, sit, serious, table, chair, indoor, tv presenter, interview, papers, formal clothes

(**title**) interview in the studio;

(**description**) three people are sitting on a red table; a tv presenter is interviewing his guests; he is talking to the guests; he is reading from papers in front of him; they are wearing a formal suit;

Hand annotation 2

(**keywords**) male, old, sit, happy, serious, table, chair, indoor, tv presenter, host, guests

(**title**) tv presenter and guests;

(**description**) there are three persons; the one is host; others are guests; they are all men;

Hand annotation 3

(**keywords**) male, old, adult, sit, serious, table, chair, indoor

(**title**) three men are talking;

(**description**) three people are sitting around the table and talking each other;

Figure 3.1: A montage showing a meeting scene in a news video and three sets of hand annotations. In this video segment, three persons are shown sitting on chairs around a table — extracted from TREC Video ‘20041116_150100_CCTV4_DAILY_NEWS.CHN33050028’.

Figure 3.2 shows another example of a video segment for a human activity and three sets of hand annotations. Again, annotators had difference of opinions about HLFs in the feature, theme of the video and description styles. For ‘keywords’, selected HLFs from predefined list are quite similar, still there is difference in open ended keywords such as annotator (1) selected park as the location of the scene while annotator (2) described it as a street scene. Further, annotator (2) and (3) both noticed dressing information as important keywords. ‘Titles’ for this video segment looks quite similar as all the annotator stated agreed that this is ‘*talking scene of a couple*’.

For ‘description’, annotator (1) tended to create a concise description, annotator (2) explained everything in detail, while annotator (3) is in the middle of these two extremes. Sometimes, annotators discarded background information altogether, *e.g.*, (annotator (1) and (3)) just described the video based on foreground information alone while annotator (2) noticed background movements and objects too.¹

After removing function words, the frequency for each word was counted in hand annotations

¹ Although annotations were also provided by TREC Video for these two video segments they were not used for this study. TREC Video annotations differ from our hand annotations to some extent; they are shot based, created for one camera take. Multiple humans performing multiple actions in different backgrounds can be shown in one shot. Descriptions for human, gender and action are observed. Additionally camera motion and angle, ethnicity information and human’s dressing are frequently stated, however there are not much details for events or objects.



<p>Hand annotation 1 (keywords) male, female, adult, young, sit, stand, sad, serious, chair, outdoor, park (title) outdoor talking scene of a man and woman; (description) young woman is sitting on chair in park and talking to man who is standing next to her;</p> <p>Hand annotation 2 (keywords) male, female, adult, sit, walk, stand, serious, chair, bus, outdoor, formal suit, people, street, taxi (title) a couple is talking; (description) two person are talking; a lady is sitting and a man is standing; a man is wearing a black formal suit; a red bus is moving in the street; people are walking in the street; a yellow taxi is moving in the street;</p> <p>Hand annotation 3 (keywords) male, female, sit, stand, serious, chair, outdoor, dark clothes, talking (title) talk of two persons; (description) a man is wearing dark clothes; he is standing there; a woman is sitting in front of him; they are saying to each other;</p>
--

Figure 3.2: A montage of video showing a human activity in an outdoor scene and three sets of hand annotations. In this video segment, a man is standing while a woman is sitting in outdoor — from TREC Video ‘20041101_160000.CCTV4_DAILY_NEWS.CHN_41504210’.

(full descriptions). Following two classes are manually defined:

1. A class, relating directly to humans, their body structure, identity, action and interaction with other humans;
2. Another class, representing artificial and natural objects and scene settings (*i.e.*, all the words that are not directly related to humans, although they are important for semantic understanding of the visual scene).

Note that some related words (*e.g.*, ‘*woman*’ and ‘*lady*’) were replaced with a single concept (‘*female*’); concepts were then built up into a hierarchical structure.

3.2.3 Human Related Features

Figure 3.3 presents human related information observed in hand annotations. Annotators paid full attention to human gender information as the number of occurrences for ‘*female*’ and ‘*male*’ is the highest among HLFs. This supported our prediction that most interesting and important HLF was humans when they appeared in a video. On the other hand age information (*e.g.*, ‘*old*’, ‘*young*’, ‘*child*’) was not identified very often. Names for human body parts had mixed occurrences ranging from high (‘*hand*’) to low (‘*moustache*’). Six basic emotions — anger,

3. CORPUS GENERATION AND ANALYSIS

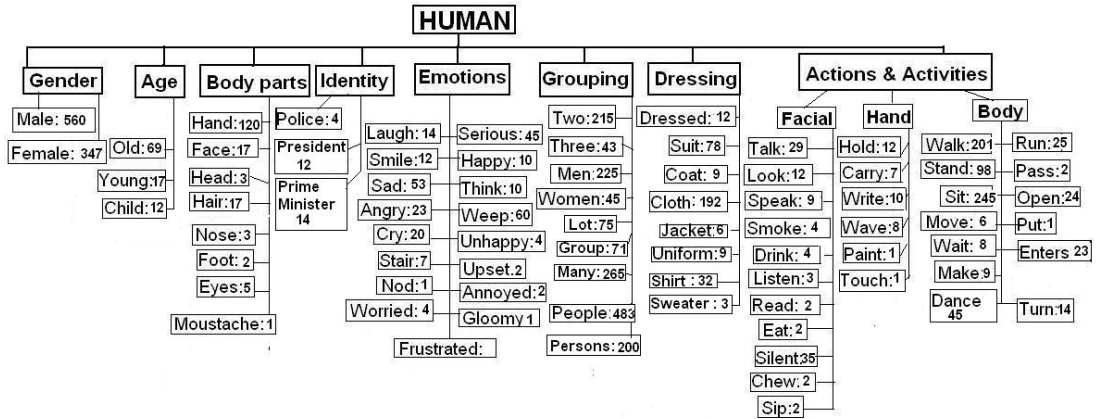


Figure 3.3: Human related information found in 13 hand annotations. Information is divided into structures (gender, age, identity, emotion, dressing, grouping and body parts) and activities (facial, hand and body). Each box contains a high level concept (e.g., ‘woman’ and ‘lady’ are both merged into ‘female’) and the number of its occurrences.

disgust, fear, happiness, sadness, and surprise as discussed by Paul Ekman¹ — covered most of facial expressions.

Dressing became an interesting feature when a human was in a unique dress such as a formal suit, a coloured jacket, an army or police uniform. Videos with multiple humans were common, and thus human grouping information was frequently recognised. Human body parts were involved in identification of human activities; they included actions such as standing, sitting, walking, moving, holding and carrying. Actions related to human body and posture were frequently identified. It was rare that unique human identities, such as police, president and prime minister, were described. This may indicate that a viewer might want to know a specific type of an object to describe a particular situation instead of generalised concepts.

3.2.4 Objects and Scene Settings

Figure 3.4 shows the hierarchy created for HLFs that did not appear in Figure 3.3. Most of the words were related to artificial objects. Humans interacted with these objects to complete activities — ‘man is sitting on a chair’, ‘she is talking on the phone’, ‘he is wearing a hat’. Natural objects were usually in the background, providing the additional context of a visual scene — ‘human is standing in the jungle’, ‘sky is clear today’. Place and location information (e.g., room, office, hospital, cafeteria) were important as they showed the position of humans or other objects in the scene — ‘there is a car on the road’, ‘people are walking in the park’.

Colour information often played an important part in identifying separate HLFs — e.g., ‘a man in black shirt is walking with a woman with green jacket’, ‘she is wearing a white uniform’. The large number of occurrences for colours indicated human’s interest in observing not only objects but also their colour scheme in a visual scene. Some hand descriptions reflected anno-

¹en.wikipedia.org/wiki/Paul_Ekman

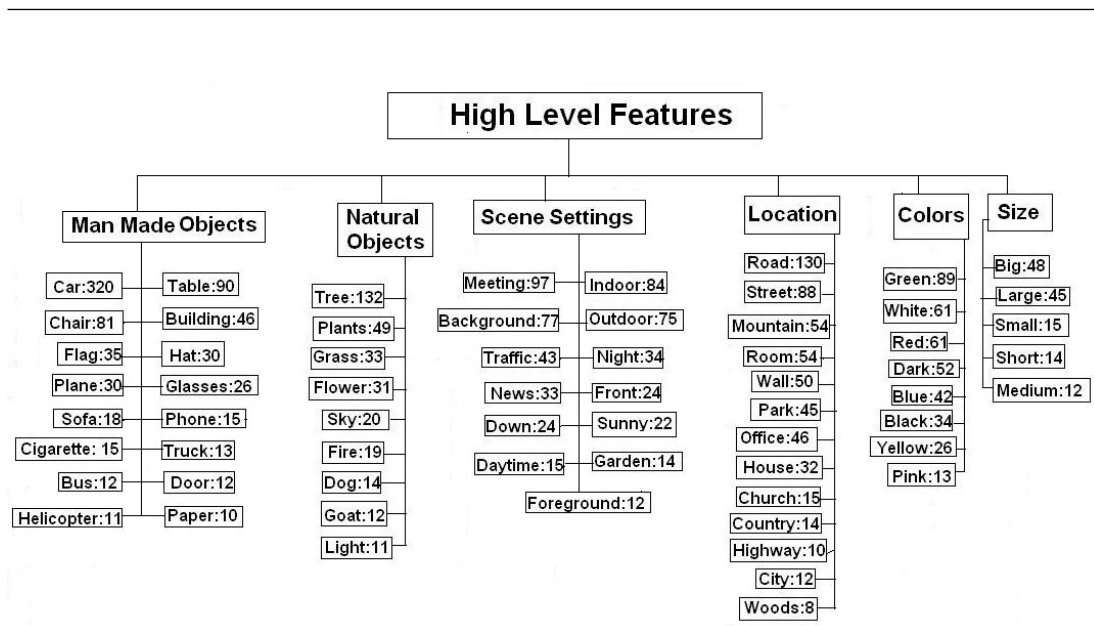


Figure 3.4: Artificial and natural objects and scene settings were summarised into six groups.

tator’s interest in scene settings shown in the foreground or in the background. Indoor/outdoor scene settings were also interested in by some annotators. These observations demonstrated that viewers were interested in high level details of a video and relationships between different prominent objects in a visual scene.

3.2.5 Spatial Relations

Spatial relation specifies how some object is spatially located in relation to some reference object. A reference object is usually a part of foreground in a video stream. Spatial relations are important when explaining visual scenes. Prepositions (*e.g.*, ‘on’, ‘at’, ‘inside’, ‘above’) can present the spatial relations between objects. Their effective use helps in generating smooth and clear descriptions, *e.g.*, ‘man is sitting on the chair’ is more descriptive than ‘man is sitting’ and ‘there is a chair’. Spatial relations can be categorised into

static: relations between not moving objects;

dynamic: direction and path of moving objects;

inter-static and dynamic: relations between moving and not moving objects.

Static relations can establish the scene settings (*e.g.*, ‘chairs around a table’ may imply an indoor scene). Dynamic relations are used for finding activities of moving objects present in the video (*e.g.*, ‘a man is running with a dog’). Inter-static and dynamic relations are a mixture of stationary and non stationary objects; they explain semantics of the complete scene (*e.g.*, ‘persons are sitting on the chairs around the table’ indicates a meeting scene). For this study videos containing humans are considered candidates for dynamic, inter-static and dynamic relations. Videos having little motion information are candidates for static relations.

3. CORPUS GENERATION AND ANALYSIS

on: 237; in: 121; around: 53; with: 44; near: 43; at: 41; on the left: 35; in front of: 24; together: 24; behind: 22; between: 18; beside: 16; on the right: 16; on the left: 12, in the middle: 10; inside: 7; middle: 7; under: 7

Figure 3.5: List of frequent spatial relations with their frequency counts.

in: 121; with: 44; on: 237; near: 43; around: 53; at: 41; on the left: 35; in front of : 24; together: 24; beside: 16; on the right: 16; between: 18; in the middle: 10; inside: 7; middle: 7; under: 7; on the left: 12; behind: 22

Figure 3.6: Manually counted frequently occurring spatial relations.

Figure 3.5 presents the list of most frequent words in the corpus related to spatial relations. High number of these words prove the fact that they are commonly used by humans to describe the visual scenes. Furthermore, they are helpful as they capture relationship between different HLFs which leads to clear understanding of the semantics of the visual scene.

Most of words shown as spatial relations in Figure 3.5 have multiple uses instead of just spatial relations. For example ‘in’ can be used for several different purposes; a sentence, such as ‘a man is sitting in the car’, indicates the spatial relation, while ‘there is a man in the video’ improves the readability of description, or ‘the man in the previous video’ explains a link between various scenes. To remedy this short coming, Figure 3.6 shows list of manually counted spatial relations.

3.2.6 Temporal Relations

Video is a class of time series data formed with highly complex multi dimensional contents, involving not only spatial but also temporal relations. Individual frames are connected together to form a complete and coherent video sequence. To generate a full description of video contents, annotators can use temporal information to join descriptions for sequential frames. The following example uses two temporal relations, *i.e.*, ‘after’ and ‘later on’, for connecting descriptions of three individual frames:

a man is walking; **after** sometime he enters the room; **later on** he is sitting on the chair.

Allen and Ferguson [1994] suggested that it was more common to describe scenarios by time intervals rather than by time points, and listed thirteen relations formulating a temporal logic (*before, after, meets, meet-by, overlaps, overlapped-by, starts, started-by, finishes, finished-by, during, contains, equals*). Temporal relations play a major role in identifying activities in

Single human: then: 25; end: 24; before: 22; after: 16; next: 12; later on: 12; start: 11; previous: 11; throughout: 10; finish: 8; afterwards: 6; prior to: 4; since: 4;
Multiple humans: meeting:114; while: 37; during: 27; at the same time: 19; overlap: 12; meanwhile: 12; throughout:7; equals: 4,

Figure 3.7: List of frequent temporal relations with their frequency counts.

videos. According to Allen’s temporal logic, most common relations in video sequences are ‘before’, ‘after’, ‘start’ and ‘finish’ for single humans. For this corpus, ‘overlap’ and ‘during’ are also frequently observed.

Based on analysis of the corpus, we describe temporal information in two flavors; (1) temporal information extracted from activities by a single human, and (2) interactions between multiple humans. Figure 3.7 presents a list of most frequent words in the corpus related to temporal relations. As can be seen, annotators put much focus on keywords related to activities of multiple humans. Keyword ‘meeting’ had the highest frequency because annotators usually considered most scenes involving multiple humans as the meeting scene. Keyword ‘while’ was typically used for presenting separate activities by multiple humans such as ‘a man is walking while a woman is sitting.’

3.2.7 Similarity between Descriptions

A well-established approach to calculating human inter-annotator agreement is kappa statistics [Eugenio and Glass, 2004]. However in the current task it is not possible to compute inter-annotator agreement using this approach because of the subjectivity in hand annotations. Further, the description length for each video can vary among annotators. Another approach that has also been effective is based upon n-gram overlap. The source and target documents are converted into fixed length n-grams (either characters or words) and the proportion of common n-grams is used to determine the similarity between documents. The similarity between two documents can be computed by counting the number of n-grams they have in common. An effective and commonly used measure to find the similarity between a pair of documents is the overlap similarity coefficient [Manning and Schütze, 1999]:

$$Sim_{overlap}(X, Y) = \frac{|S(X, n) \cap S(Y, n)|}{\min(|S(X, n)|, |S(Y, n)|)}$$

where $S(X, n)$ and $S(Y, n)$ are the set of distinct n -grams in documents X and Y respectively. It is a similarity measure related to the Jaccard index Tan et al. [2006]. Note that when a set X is a subset of Y or the converse, the overlap coefficient is equal to one. Values for the overlap coefficient range between 0 and 1, where 0 presents the situation where documents are completely different and 1 describes the case where two documents are exactly the same.

Table 7.1 shows the average overlap similarity scores for seven scene categories within 13 hand annotations. The average was calculated from scores for individual description, that was compared with the rest of descriptions in the same category. The outcome demonstrate the fact that humans have different observations and interests while watching videos. Calculation was repeated with two conditions, one with stop words removed and Porter stemmer [Porter, 1993] applied, but synonyms NOT replaced, and the other with stop words NOT removed, but Porter stemmer applied and synonyms replaced. It was found the later combination of preprocessing techniques resulted in better scores. Not surprisingly synonym replacement led to increased performance, indicating that humans do express the same concept using different terms.

3. CORPUS GENERATION AND ANALYSIS

	Action	Close-up	News	Meeting	Grouping	Traffic	Indoor/Outdoor
unigram (A)	0.3827	0.3913	0.4378	0.3968	0.3809	0.4687	0.4217
(B)	0.4135	0.4269	0.4635	0.4271	0.4067	0.5174	0.4544
bigram (A)	0.1483	0.1572	0.1872	0.1649	0.1605	0.1765	0.1870
(B)	0.2490	0.2616	0.2890	0.2651	0.2619	0.2825	0.2877
trigram (A)	0.0136	0.0153	0.0279	0.0219	0.0227	0.0261	0.0301
(B)	0.1138	0.1163	0.1279	0.1214	0.1229	0.1298	0.1302

Table 3.1: Average overlapping similarity scores within 13 hand annotations. For each of unigram, bigram and trigram, scores are calculated for seven categories in two conditions: (A) stop words removed and Porter stemmer applied, but synonyms NOT replaced; (B) stop words NOT removed, but Porter stemmer applied and synonyms replaced.

	raw	synonym	keywords
Action	0.3782	0.3934	0.3955
Close-up	0.4181	0.4332	0.4257
Indoor	0.4248	0.4386	0.4338
Grouping	0.3941	0.4104	0.3832
Meeting	0.3939	0.4107	0.4124
News	0.4382	0.4587	0.4531
Traffic	0.4036	0.4222	0.4093

Table 3.2: Similarity scores based on the longest common subsequence (LCS) in three conditions: scores without any preprocessing (raw), scores after synonym replacement (synonym), and scores by keyword comparison (keywords). For keyword comparison, verbs and nouns were presented as keywords after stemming and removing stop words.

The average overlap similarity score was higher for ‘Traffic videos’ in comparison to rest of the categories. Presence of vehicles as the major entity in traffic videos, rather than humans and their actions, resulted in generating uniform hand annotations for these videos. Scores for some other categories were lower. It probably means that there are more aspects to pay attention when watching videos in, *e.g.*, ‘Grouping’ category, hence resulting in the wider range of natural language expressions produced.

3.2.8 Sequence of Events Matching

Since frames in a video sequence are tied together temporally, it will be beneficial to know how the annotators captured the temporal information present in a video. As the order is preserved in a sequence of events, a suitable measure to quantify temporal information contained in the description is the longest common subsequence (LCS). This approach computes the similarity between a pair of tokens (*i.e.*, words) sequences by simply counting the number of ‘edit’ operations (insertions and deletions) required to transform one sequence into the other. The output is a sequence of common elements such that no other longer string is available. In the experiments, the LCS score between word sequences is normalised by the length of the shorter sequence.

Table 3.2 presents results for identifying sequences of events in hand descriptions using the LCS similarity score. Individual descriptions were compared with the rest of descriptions in the

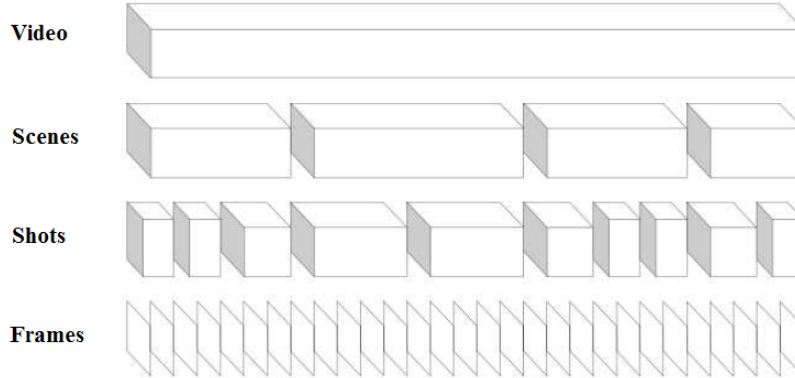


Figure 3.8: Hierarchical structure of video sequence.

same category and the average score was calculated. Relatively low scores in the table indicate the great variation in annotators’ attention on the sequence of events, or temporal information, in a video. Events described by one annotator may not have been listed by another annotator. The ‘News’ videos category resulted in the highest similarity score, confirming the fact that videos in this category are highly structured.

3.3 NLDV - Corpus 2: Dataset with Lengthy Videos

Video is digitized as a continuous stream that joins a number of different content portions. It detects the boundary of each video unit, usually defined by frame, shot or scene. Frame is a still picture which is sequentially segmented by a certain time duration, called a frame rate. The standard frame rates are 25 and 30 frames/second [Oh et al., 2005]. Although it is not difficult to extract frame, there is too wide range of detail captured from all frames. Therefore, it is inefficient to handle such a huge data to represent a video. Instead, consecutive frames are assembled to create a semantic segment called shot and scene. While a shot is a single and continuous camera action, a scene is a story unit which gives more semantic notion concerning an object, a person, and time [Lienhart, 1999]. The relation between frame, shot, scene and video is illustrated in Figure 3.8.

NLDV - Corpus 1, explained in section 3.3 consisted of short videos, which usually contained only one camera shot. To extend this work for lengthy videos, another video dataset was created. This dataset contained manually crafted videos from the rushes video summarization task for the 2007 and 2008 TREC video evaluations. It consisted of 10 segments of videos — Each segment contained multiple camera shots, combined together to present a coherent and complete video story, where length of complete video was spanning between 3 and 5 minutes. The idea here was to select shots which have semantic relation between them and can generate a story for the complete video sequence. For this corpus we selected videos containing humans and their activities only. Although there is no categorization like NLDV - Corpus 1, still these

3. CORPUS GENERATION AND ANALYSIS

video segments contained shots related to humans close ups, actions, grouping and meeting sceneries. Finally, it is worth mentioning that purpose of this dataset is to find a story line in the video segment which may consist of several actors performing several actions in different shots.

3.3.1 Annotation Process

Five human subjects prepared annotations for these video segments, consisting of a summary and a full description with multiple sentences¹, where summary is very short, presenting the theme and overall idea of the video segment and description provides detailed explanation of objects and activities happening in the video stream. In addition to instructions provided for the previous dataset they were asked to provide detailed description about the complete video, where each video was a combination of several shots, roughly 4-6 shots for each video. They were asked to write 5-6 sentences for each shot. They were further asked to find a connection between each shot and describe the relationship between each shot. For summary generation they were asked to write a short and compact description of the video. Finally, they were motivated to find a story of the video segment.

3.3.2 Corpus Analysis

Total number of documents for this corpus was 50 (5 annotators created descriptions for 10 videos each). The total number of words was 5653, hence the average length of one document was roughly 113 words. We counted 429 unique words and 299 keywords (nouns and verbs). There were 731 sentences in total which roughly corresponded to 7 to 8 words per sentence.

Figure 3.9 presents human related information observed in hand annotations. It is interesting to see that most of the presented information is similar to that of Figure 3.3. Some missing information include, human identity information such as president and prime minister, dressing information such as suit, coat, jacket and uniform. Since this dataset is much smaller than the previous one, frequency of most of the these features is much lower in this figure. Finally, it is interesting to note that both figures have similar ordering of features with respect to frequency for most of human related features.

Further analysis was done to point out HLFs other than humans. In comparison to Figure 3.4, this list of HLFs is very short. Firstly, there are only four broad categories, *i.e.*, objects, scene settings, location and color, in comparison to 6 presented in Figure 3.4. HLFs related to natural objects and size are completely absent, whereas color category has only one HLF *i.e.*, black color. Secondly, HLFs present in the remaining three categories are very few. For example, list of HLFs in objects category is very limited as compared to man made and natural objects in Figure 3.4. Similarly, for scene settings there are only four HLFs, *i.e.*, scene, indoor, outdoor and ceremony, whereas Figure 3.4 contains 13 HLFs for this category. Finally for

¹All the five annotators were fellow PhD students in Computer Science Department at The University of Sheffield. They were well familiar with video processing concepts and descriptions of good quality were expected from them.

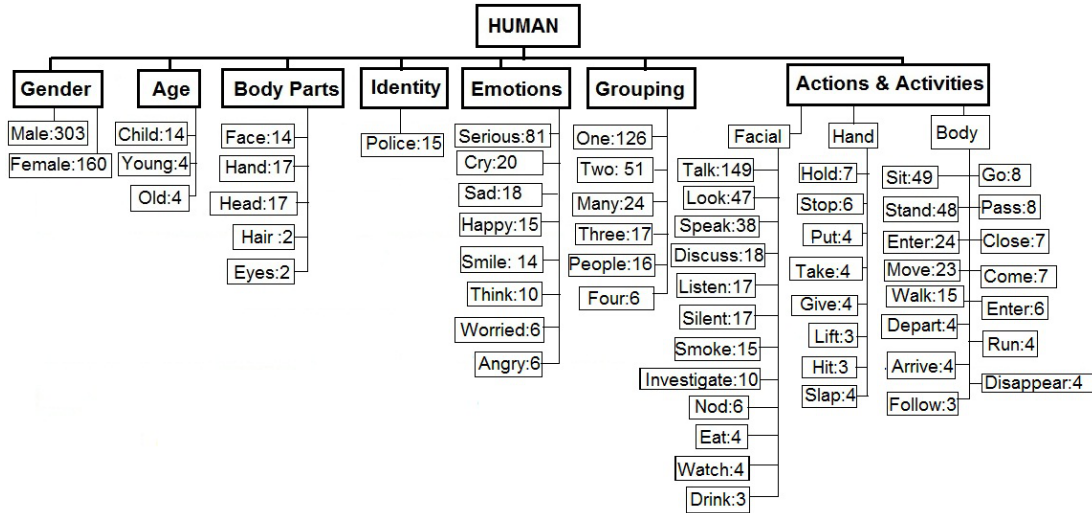


Figure 3.9: List of keywords related to humans in the dataset 2.

location category, there are only six HLFs, *i.e.*, office, room, restaurant, church, funeral and building, which is less than half of the same category in Figure 3.4.

Sample Video and Annotations. Figure 3.10 presents an example video from this dataset. This video segment depicts activities of a couple sitting in a restaurant. In the beginning, they are sitting with another couple and having dinner with them. After sometime, the other couple leaves, and the original couple is sitting and chatting in the restaurant. Later on, there are shots of another man and woman talking to each other. Table 3.3 presents three sets of hand annotations for the video sequence shown in Figure 3.10. In comparison to previous dataset, these descriptions are quite lengthy, more than 120 words typically.

Some annotators were able to differentiate between different scenes within the video segments such as annotator (1) noticed that there are six scenes in the video. Mostly, human related information is noticed by the participants such as human age, gender, emotions and their activities. Presence of objects such as table, chairs, cup *etc.* was also noticed by most of the annotators. In addition to recognizing general actions such as sitting, walking and standing, annotators were also able to identify specific actions such as smoking and drinking. Interaction between humans and their background information was also successfully stated by the annotators. On the other hand, summaries are shorter in length, usually ranging between 40 to 60 words. Summaries tend to unleash some sort of story in the video contents. Further, annotators tried to cover most important events and actors while generating summaries. Some annotators generated compact summaries providing an overview of the video contents such as annotator (1). On the other hand, annotator (1) and (2) tried to focus on each actor and generated a summary which explained details of their activities.

3. CORPUS GENERATION AND ANALYSIS



Figure 3.10: Montage of a 5 minutes video segment taken from TREC video 'MRS042538'.

Hand Annotation 1:

(summary) The video shows many people gathering in a party and there are some discussion take place between couple of man and women.

(description) There are 6 scenes in this video. In first one, there are 2 men and 2 women sitting together talking an laughing and one man is smoking. There is fifth person stand in front of camera cooking and mixing something during that a sixth person pass quickly. The second scene shows many people who arriving and handshaking together and the couple of them sitting together and talk. The women face has been zoomed in and she seems surprised. the man is smoking and talking. Then a scene shows a man who seems nervous. the last scene shows man and women sitting together and talking.

Hand Annotation 2:

(summary) This is a video of a man and women who are present in a restaurant. In the beginning, two mans and two women are sitting on chairs and talking together. Then, one man and woman are talking. Later on, there is another man and woman are talking. In the end, a pretty woman is speaking.

(description) This is a video of a man and women who are present in a restaurant. In the beginning, two mans and two women are sitting on chairs and talking together. They are talking and doing dinner together. After sometime, one man and one woman get ready for leaving. All of them are standing. They meet each other and then one man and one woman leave the restaurant. Remaining one man and woman are sitting around the table and talking. Man is smoking cigarette and woman is drinking something. After sometime, in another scene, there is serious man present who is looking into camera and talking. There is another man behind him. Later on, in another scene, a woman is shown talking to someone. Woman is very pretty and she is serious.

Hand Annotation 3:

(summary) Two men, one woman and a human are sitting around a table in an indoor scene. Four persons are standing in an indoor scene. A man and woman are sitting around a table. There are cups on the table. They are happy and talking to each other. A serious woman is talking to someone in an indoor scene.

(description) Four persons are sitting in a restaurant. There is a table between humans. There is a man, woman and two other humans. There are two men, one woman and one other human. Humans are sitting. Four humans are standing. There is one woman and three humans. A woman is sitting with a human. There is a cup between woman and human. A happy woman is present while there is a woman in the background. There is a man in the background. Woman is speaking. Woman is happy. A man is happy. A man is gesturing. A man is happy and speaking with a human. A man is happy and gesturing. A man is there while there are other humans in the background. The man is happy and speaking. The man is serious. Two humans are present. A serious woman is present. A serious woman is speaking. There is another human with the serious woman. A serious woman is silent.

Table 3.3: Three sets of hand annotations for the video montage shown in Figure 3.10.

3.4 NLDV - Corpus 3: Evaluation Dataset

TRECVID provided videos can be roughly divided into categories such as news, meeting, crowd, grouping, indoor/outdoor scene settings, traffic, costume, documentary, identity, music, sports and animals videos. Seven out of these categories are used for NLDV - Corpus 1 (section 3.2). For evaluation of description generation framework, a new dataset was generated. This dataset consisted of five different categories, *i.e.*, costume, crowd, sports, violence, and animal videos. In relation to ‘closeness’ of these five video categories with the previously mentioned seven categories, first four out of these five categories usually contain scenes related to humans and their activities. Costume videos usually differ from other categories based on humans’ dressing such as old Roman or Islamic dresses. Crowd videos contain large number of humans and variety of activities simultaneously such as people in a procession raising slogans and holding banners. Sports videos have special scene settings and human’s dressing information such as woman in a white skirt in a tennis court. Violence videos have special objects such as guns, army trucks and color information such as colors of fire or smoke. Animal videos contain scenes of animals and usually specific scene settings such as park and fences. Four short video segments were manually created for each of these videos from TRECVID provided data, mostly from HLF extraction task of year, 2004 and BBC rushes videos.

3.5 Findings from the Corpora Analysis

This section presented analysis of video annotation dataset¹. The corpus is important for the following reasons:

1. limiting this study to a manageable and defined domain.
2. decision of HLFs that should be extracted by image processing.
3. preparation for development/test data and ground truths for evaluation.

Concerning 2. above, several conclusions can be drawn based on the analysis of hand annotations. Annotators are most interested in human emotions, actions and their interaction with other humans and objects. Natural objects play important role for identification of scene settings (*e.g.*, presence of trees indicates ‘*park scene*’, presence of sky generates ‘*outdoor scene*’). Artificial objects are mostly attached with humans and their activities (*e.g.*, ‘*man is sitting on the chair*’). Colour information is important to distinguish one object from others. Humans are normally considered a part of foreground while other objects constitute the background. Based on these observations, we derive a wish list for HLFs for automatic extraction. Roughly, these HLFs include, ‘*human, age, gender, human counting, human emotion, human action, objects, scene setting and temporal relations between HLFs.*’ Natural language description of video streams starts with identification of these HLFs.

¹We plan to make this dataset public with the following structure, video ID, start time, end time, set of keywords, title, description and annotator ID.

3.6 Summary

This chapter discussed generation of metadata in the form of natural language descriptions for three video corpora. These corpora are used as training and evaluation datasets for all the experiments presented in this thesis. For first corpus, 13 annotators produced titles, descriptions and specific high level features for 140 short segments. Analysis of this corpus presents insights into humans interests and thoughts while watching videos. This analysis, further provides a list of important visual contents which are referred as high level features. Chapter 5 discusses generation of automatic language descriptions of these 140 video segments based on HLFs which are extracted using image processing methods. Same resource, *i.e.*, corpus 1 is used for evaluation of machine generated descriptions. Second corpus, which consists of 10 video segments is generated for evaluation of lengthy videos (Chapter 7).

3. CORPUS GENERATION AND ANALYSIS

Chapter 4

Image Processing Methods and Evaluations

Section 2.4 in Chapter 2 presented literature survey of feature extraction task used for this research work. This chapter further discusses implementation details, experimental setup and evaluation results of feature extraction task which is based on image processing methods. In the current implementations, the results are evaluated using confusion matrix. It is a specific table layout that allows visualization of the performance of an algorithm based on false positives, false negatives, true positives, and true negatives. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (*i.e.*, commonly mislabeling one as another).

Frame extraction was performed using `ffmpeg`¹. It is a computer program that can record, convert and stream digital audio and video in numerous formats. `Ffmpeg` is a command line tool that is composed of a collection of free software / open source libraries. It includes `libavcodec`, an audio/video codec library used by several other projects, and `libavformat`, an audio/video container mux and demux library. 25 frames/second is the default conversion rate. Multiple frame rates were tested and used for implementation purposes.

It is important to note that all the HLF extraction methods discussed in this chapter are based on the idea of supervised learning, *i.e.*, classifiers are trained using feature vectors with attached labels. For each of the HLF extraction method, fixed labels are defined as output options, such as for human detection two labels are yes are no.

4.1 Overview of HLF Extraction Procedure

This work starts with the extraction of HLFs from video sequences. Figure 4.1 provides a list of HLFs together with the features used for their identification. Closed rectangles present the

¹<http://www.ffmpeg.org/>

4. IMAGE PROCESSING METHODS AND EVALUATIONS

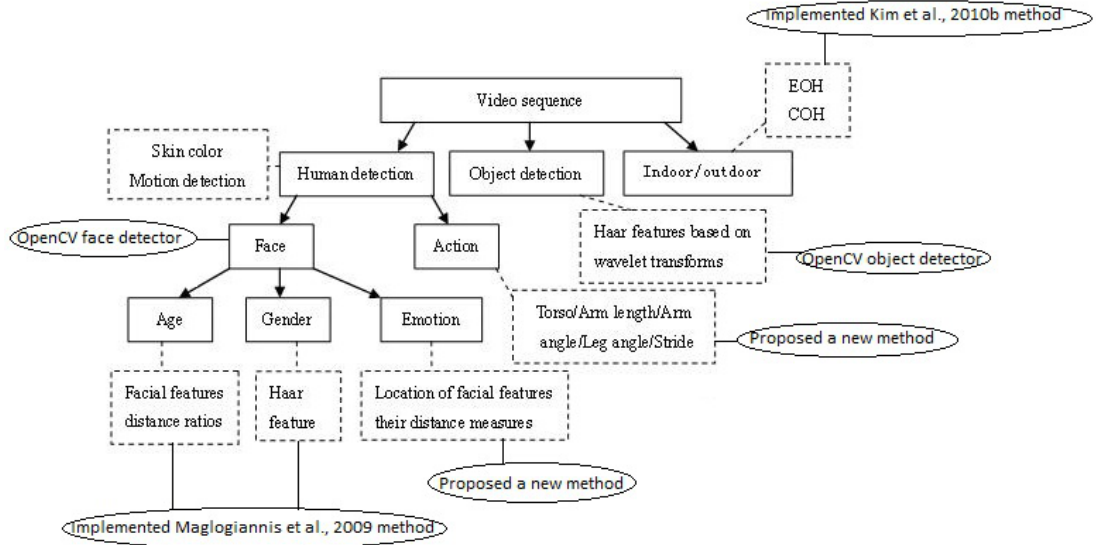


Figure 4.1: HLF identification from video sequences using conventional image processing techniques. Closed rectangles present the extracted HLFs, while dotted rectangles represent features used for extraction of HLFs.

extracted HLFs, dotted rectangles represent features used for extraction of HLFs and circles present implementation methods used for this extraction.

4.2 Human Identification

Identification of human face or body can prove the presence of human in a video. Based on this assumption, human face detection was first performed to find the humans in the video sequences. Method proposed by [Viola and Jones, 2001] was implemented to detect faces in real-time and with very high detection rate. It is essentially a feature-based approach in which a classifier is trained for Haar-like rectangular features selected by AdaBoost. The test image is scanned at different scales and positions using a rectangular window, and the regions which pass the classifier are declared as faces. Concept of Integral Image was proposed which enabled the detection in real-time. Additionally, instead of learning a single classifier and computing all the features for all the scanning windows in the image, a number of classifiers are learnt which are put together in a series to form a cascade.

Features. The features used for face detection are simple Haar-like rectangular features as shown in left panel of Figure 4.2. Example rectangle features shown relative to the enclosing detection window. The sum of the pixels which lie within the white rectangles are subtracted from the sum of pixels in the grey rectangles. Three versions of these features are used: two two-rectangle features, and one three-rectangle feature and four-rectangle feature each.

A major advantage of using these rectangular features is that they can be computed very

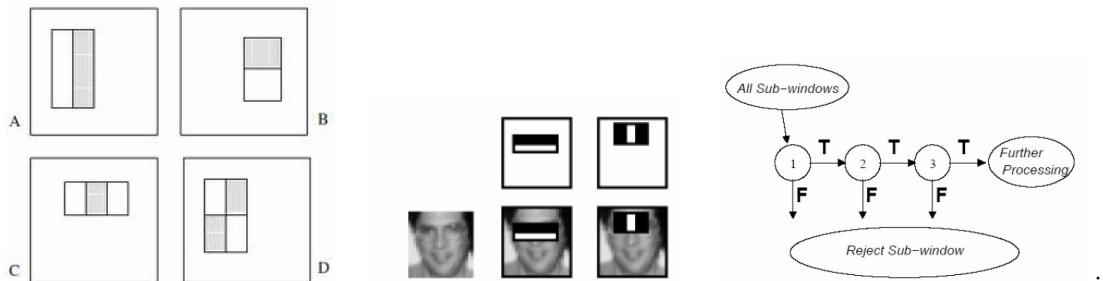


Figure 4.2: (left panel) (A) and (B) are the two Two-Rectangle features, (C) shows the Three-Rectangle feature and (D) the Four-Rectangle feature. (center panel) The first and second features selected by AdaBoost. (right panel) schematic depiction of a the detection cascade.

quickly using the concept of integral image. The value in the integral image at the pixel (x, y) is the sum of all the pixels to the left and above (x, y) in the original test image:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (4.1)$$

where $ii(x, y)$ is the integral image and $i(x, y)$ is the actual image.

The integral image can be computed in one pass for an image and thereafter, the sum of pixels for any rectangular region can be computed with just four array references. **Learning**

Classifier Functions. This is the main learning stage of the system which accomplishes two things simultaneously. Firstly, it selects ‘good’ discriminative features out of a pool of thousands of possible candidates. Secondly, it learns a classifier using these features that decides whether a region is a face or not. Both the objectives are achieved using a well known learning algorithm called Adaboost [Freund and Schapire, 1995]. Two of the first features selected by AdaBoost are shown in center panel of Figure 4.2. The two features are shown in the top row and then overlaid on a typical training face in the bottom row. The first feature measures the difference in intensity between the region of the eyes and a region across the upper cheeks. The feature capitalizes on the observation that the eye region is often darker than the cheeks. The second feature compares the intensities in the eye regions to the intensity across the bridge of the nose. These features make sense since eyes, nose and cheeks are the most discriminate parts of a face.

The Detection Cascade. In practice, no single strong classifier is used. Instead, a series of many such classifiers are learnt to form a cascade of classifiers. The simpler classifiers come earlier in the As shown in right panel of Figure 4.2, a series of classifiers are applied to every sub-window. The initial classifier eliminates a large number of negative examples with very little processing. Subsequent layers eliminate additional negatives but require additional computation.

Implementation and Evaluation. OpenCV Bradski and Kaehler [2008] is an open source computer vision library written in C/C++ programming language. It is optimized and intended

4. IMAGE PROCESSING METHODS AND EVALUATIONS

	(ground truth)	
	exist	not exist
exist	1795	29
not exist	95	601

Table 4.1: Confusion matrix for human detection. Columns show the ground truth, and rows indicate the automatic recognition results. The human detection task is biased towards existence of humans.

for real-time applications. It provides image processing functionalities ranging from very basic (filtering, edge detection, corner detection, sampling and interpolation) to very complex like object recognition and motion analysis. OpenCV provides already built face detector built on the idea of Haar cascade classifiers. For this work, we used this face detector without any further modifications. Table 4.1 presents a confusion matrix for human detection for corpus 1 of this research work. It was a heavily biased dataset where human(s) were present in 1890 out of 2520 frames. Of these 1890, misclassification occurred on 95 occasions.

4.3 Human Age/ Gender Detection

Human face is of fundamental importance for guessing age of a person. Facial features play an important role in identifying age, gender and emotion information [Maglogiannis et al. \[2009\]](#). Human emotion can be estimated using eyes, lips and their measures (gradient, distance of eyelids or lips). The same set of facial features and measures can be used to identify a human gender¹. Based on this idea, facial parts information was used for this age prediction application. For this task implementation of the idea proposed by ([Hornig et al. \[2001\]](#)) was performed. At this time system can classify humans in three age groups; baby, young and old aged.

The process of the system development is divided into three phases: location of facial parts, feature extraction, and age classification (Figure 4.9). Sobel edge operator and region labelling were used for finding the positions of eyes, nose, and mouth based on the symmetry of human faces and the variation of gray levels. Two geometric features and three wrinkle features from a facial image were calculated. Finally, Linde-Buzo-Gray (LBG) algorithm [Lin and Tai \[1998\]](#) was used to train system for classification. The LBG takes input vectors consisting of feature parameters. Figure 4.9 presents steps for face based human age identification. In the location phase, the symmetry of human faces helps find vertical central lines of faces. Since eyes, nose, and mouth have significant brightness changes, the Sobel edge operator and region labelling were applied to locate them. Both geometric and wrinkle features were employed in the system for classification. In the feature extraction phase, two geometric features are evaluated as the ratios of the distances between eyes, nose, and mouth.

At the moment three age groups can be successfully identified; *i.e.* baby, young and old. In video corpus 1, there was no baby present, all the humans are either young or old aged. Table 4.2(b) shows a confusion matrix for age identification. Young human identification seems to be easier and better achieved as compared to old humans identification. Table 4.2(b) shows a

¹ www.virtualffs.co.uk/In_a_Nutshell.html

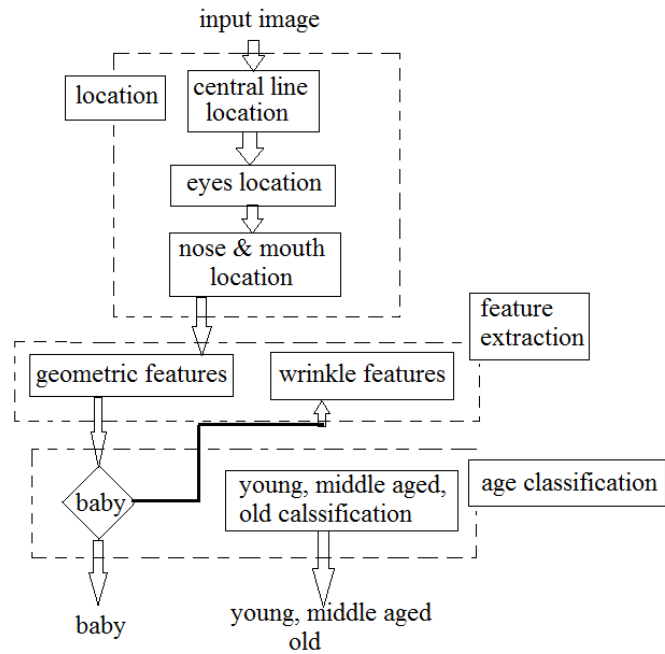


Figure 4.3: (Age prediction using facial features: Algorithm is divided into three steps; Location, feature extraction and age classification. Diagram taken from *Hornig et al. [2001]*)

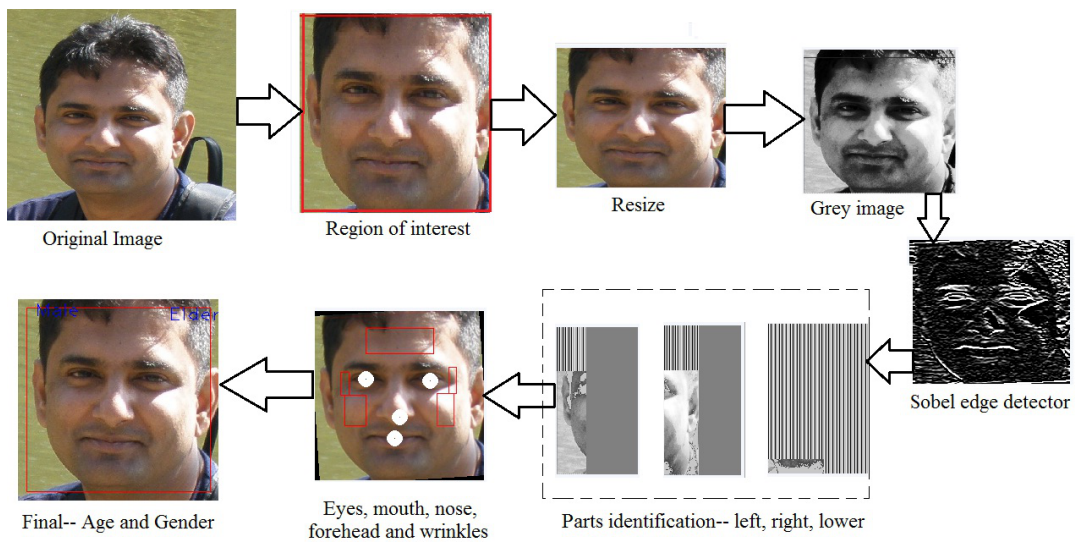


Figure 4.4: Results of age and gender identification: Extracted region of interest, resizing, greyscale conversion, greyscale, histogram, applying Sobel filter, dividing face into three parts as left, right and down, finding position of facial parts: nose, eyes, lips, cheeks and forehead, finally classifying face as male and adult

4. IMAGE PROCESSING METHODS AND EVALUATIONS

	(ground truth)			(ground truth)	
	young	old		male	female
young	589	233	male	636	175
old	229	298	female	346	192
(a) adult identification			(b) gender identification		

Table 4.2: Confusion tables for (a) age identification and (b) gender identification. Columns show the ground truth, and rows indicate the automatic recognition results. The age identification task is biased towards young humans, while in the gender identification presence of male and female are roughly balanced.

confusion matrix for gender identification. Female identification was often more difficult due to make ups, variety of hair styles and wearing hats, veils and scarfs. Out of 1890 frames in which human(s) were present, frontal faces were shown in 1349 images. The total of 3555 humans were present in 1890 frames (1168 frames contained multiple humans), however the table shows the results when at least one human w.r.t. age / gender is correctly identified.

4.4 Emotion Recognition

Emotion recognition can be performed in three steps; *i.e.*, face detection, facial features localization, and emotion classification. OpenCV face detector was used to separate face part from the given image. The two other steps are explained in the following.

4.4.1 Localization of Facial Features

The features that are commonly used to characterize and define a human face are the eyes and the mouth. Ioannou et al. [2007] presented a method for locating the eyes and mouth based on their feature maps derived from both the luminance and chrominance of an image. For this we add one extra step of skin detection to further enhance chances of mouth and eyes detection.

Constructing the Eye Map. Two separate eye maps are built, one from the chrominance components and the other from the luminance component. These two maps are then combined into a single eye map. The eye map from the chrominance is based on the fact that high C_b and low C_r values can be found around the eyes. This is easily constructed by the formula

$$EyeMap_{(chroma)} = \frac{1}{3} \left\{ (C_b)^2 + (255 - C_r)^2 + \left(\frac{C_b}{C_r} \right) \right\} \quad (4.2)$$

Eyes usually contain both dark and bright pixels in the luminance component, so grayscale operators can be designed to emphasize brighter and darker pixels in the luminance component around eye regions. Such operators are dilation and erosion. We use grayscale dilation and erosion with a spherical structuring element to construct the eye map from the luminance as follows:

$$EyeMap_{(Lum)} = \frac{Dilation(Y(x, y))}{Erosion(Y(x, y)) + 1} \quad (4.3)$$

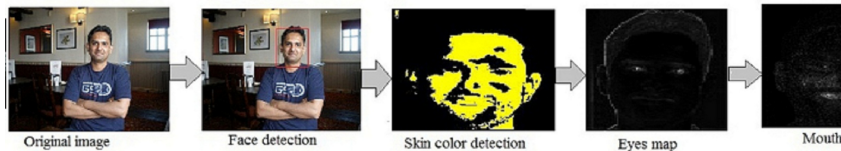


Figure 4.5: Framework for emotion recognition.

The eye map from the chrominance is then combined with the eye map from the luminance by an AND (multiplication) operation, $EyeMap = EyeMap_{(chroma)} AND EyeMap_{(Lum)}$. The resulting eye map is then dilated and normalized to brighten both the eyes and suppress other facial areas. From this eye maps, areas related to both eyes are bounded in rectangles for further processing.

Constructing the Mouth Map. The color of the mouth region contains stronger red component and weaker blue component than other facial regions. Then the chrominance component C_r is grater than C_b in the mouth region. So the mouth map is constructed as follows:

$$Mouth_{map} = C_r^2(C_r^2 - \eta \cdot \frac{C_r}{C_b}) \quad (4.4)$$

where η is calculated as the ration of the average of C_r^2 to $\frac{C_r}{C_b}$.

Emotion Detection and Recognition. The main branch of the proposed algorithm is the utilization using of an edge detection technique to determine the lines of the eyes and mouth, curves and gradients. The result of the edge detection is a binary image indicating the edges of the facial feature. Then the Hough transformation is applied on this binary image, for extracting the details of facial features (*i.e.*eyes, lips) and the details measured (*i.e.*gradient, distance of eyelids or lips, etc.), represented through some thresholds. A database of real images showing persons with different emotions was used for threshold's definition.¹

Performance of Emotion Recognition. Figure 4.5 shows an example of steps for emotion recognition procedure. Initially face is cropped from initial image. Skin color detection, construction of eyes map and then mouth map is performed. Based on eyes and mouth map, position of eyes and mouth is finalized. Using, features and distance measures as proposed by [Maglogiannis et al., 2009], which are based on eyes and mouth position, emotion of the human is identified. Table 4.3 shows the confusion matrix for human emotion recognition. ‘Serious’, ‘happy’ and ‘sad’ were most common emotions in this dataset, in particular ‘happy’ emotion was most correctly identified. On the other hand, emotions like angry and surprised were seldom noticed.

¹For emotion detection we used the same set of feature measures and distances between facial parts as proposed by Maglogiannis et al. [2009]

4. IMAGE PROCESSING METHODS AND EVALUATIONS

	(ground truth)				
	angry	serious	happy	sad	surprised
angry	59	0	0	15	16
serious	0	661	0	164	40
happy	0	35	427	27	8
sad	61	13	0	281	2
surprised	9	19	0	0	53

Table 4.3: Confusion table for human emotion recognition. Columns show the ground truth, and rows indicate the automatic recognition results.

4.5 Action Recognition

For action recognition, a mechanism is needed to recognize single and multiple human actions in video streams. Method proposed by [Chen et al., 2006] was implemented for this task, although human body was presented by stick figures [Haritaoglu et al., 1998] instead of star skeleton. Then low level features are extracted from these stick figures. These features are used for training Hidden Markov Models. Based on the most likely performance criterion, human action can be recognized through evaluating the trained HMMs. For multiple humans, a tracking algorithm is applied to first track individual humans. Three types of actions *i.e.* separate, transitive and parallel human actions are proposed.

System Overview. Preprocessing, feature extraction, training and recognition are main steps as shown in Figure 4.6. Initially, original video sequence is converted into individually visual frames. Preprocessing step includes grey scale conversion, object detection, finalization and noise removal. Each image sequence is converted to grey scale. Background image is subtracted from current image to find the foreground object. After object detection, image binarization is applied to convert image into two colors *i.e.*, black and white. Object is presented by white and background by black color. Noise produced in the image due to binarization is removed in the final stage of preprocessing.

Once object *i.e.*, human is separated from the background, it is further processed to find the feature vectors. Binary image is in the form of human silhouette. This silhouette is converted into human stick figures. Stick figures are twisted lines and its not possible to measure features from them. Hough enhancer is applied to this stick figure to straighten lines representing body parts. From this image, useful features such as torso, stride, arm length, arm angle, length angle and speed are calculated.

The action recognition module involves two phases: training of HMM and recognition. One HMM is trained for every unique action. Recognition is achieved by calculating the maximum probability of predefined HMM models against a testing video sequence.

Preprocessing and Feature extraction. Figure 4.7 presents steps of preprocessing and feature extraction. Image is converted to grey scale image to show only intensity information. From this grey scale image object is detected using background subtraction. Image binarization is performed to brighten the human silhouette. There is some noise in the image after finalization. Image is convolved with median filter Chan et al. [2005] to remove noise from the

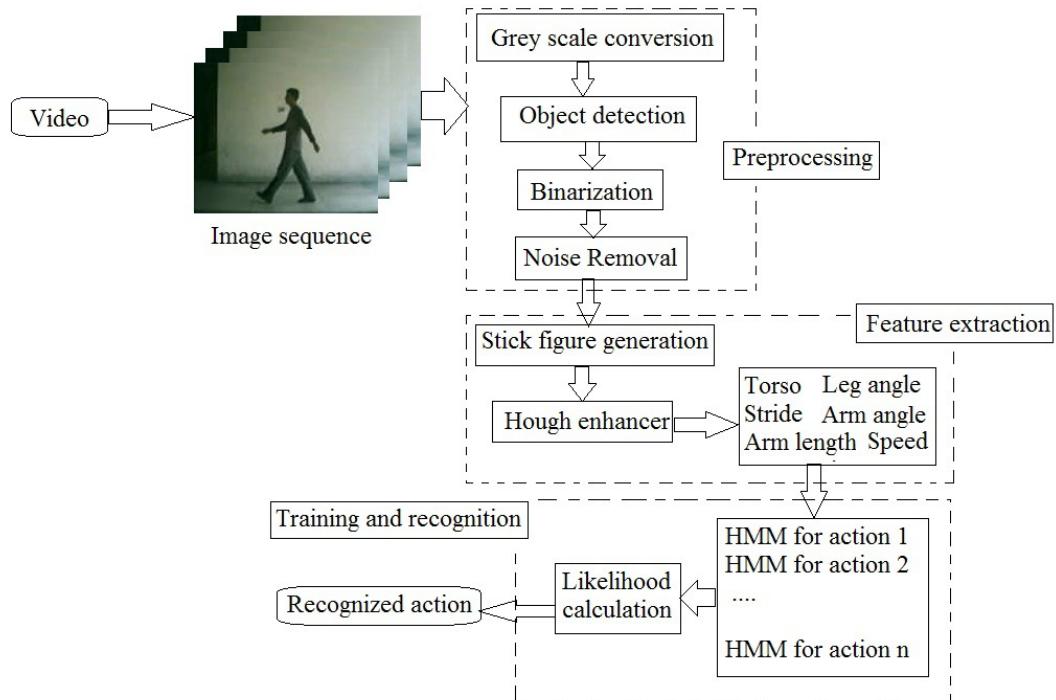


Figure 4.6: Framework for action recognition

image. Further, dilation and erosion are applied to remove small regions of noise pixels and holes which are not connected to human body.

To convert silhouette image to stick figures, thinning operation as proposed by Staunton [Staunton \[1996\]](#) was applied. The thinning operation is used as a skeletonisation operation; converting silhouette to human to single pixel thickness image *i.e.*, stick figure. This stick figure represents human body as configuration of 7 sticks and 6 joints combining those sticks. [Cunado et al. \[2003\]](#) For a body height H , an initial estimate of the vertical position of the neck, shoulder, waist, pelvis, knee and ankle was set by study of anatomical data to be $0.870H$, $0.818H$, $0.530H$, $0.480H$, $0.285H$, and $0.039H$, respectively [Dempster and Gaughran \[1967\]](#). The skeleton can be simply calculated by two border points of each body part p with a range constraint. The angles Θ_p of body part p from skeleton data can be approximated by using the slope of the lines in linear regression equation. Also, each body point (position) can be calculated by [Yoo et al. \[2005\]](#)

$$x_p, y_p = [x_i + L_p \cos(\phi + \theta_p) \quad y_i + L_p \sin(\phi + \theta_p)] \quad (4.5)$$

where ϕ is the phase shift, x_i and y_i are the coordinates of a previously established position, and L_p is the length of body segments. The 2D stick figures are combination of these points. The action signature is defined as a sequence of the stick figures obtained from the silhouette data. Torso, height, arm length, leg length, stride, arm waving speed, arm angle and leg angle

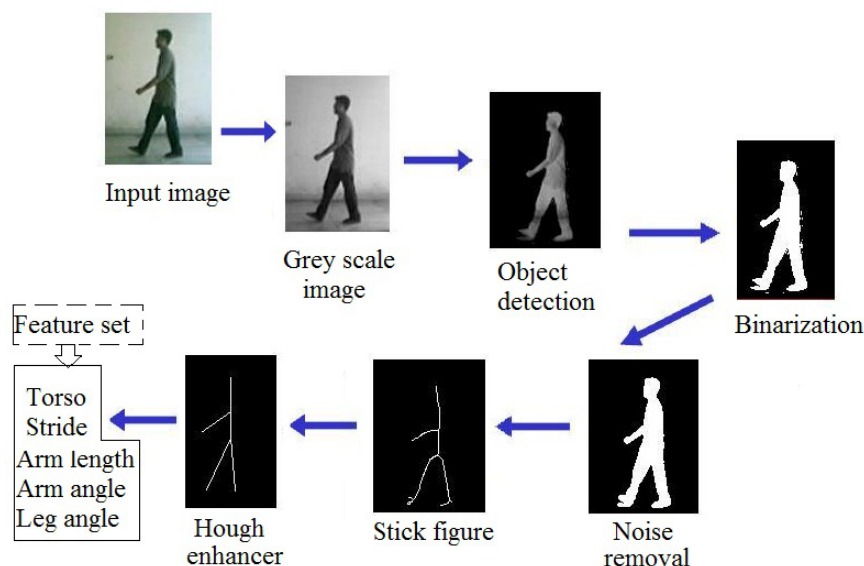


Figure 4.7: Steps for feature extraction

are the main features used for action recognition task. From binary image, extraction of the stick figures from the human silhouette is done. These stick figures generate these features related to action recognition. Human's action can be recognized by using information related to the motion of skeletal components. Therefore, it becomes necessary to find out which part of body (*e.g.*, head, hands, legs, *etc.*) is moving. Angles between two legs can be used to distinguish walking from running [Fujiyoshi et al. \[2004\]](#). Similarly upper limbs movement can help in identifying actions related to upper human body such as waving, boxing and clapping.

Table 4.4 shows the human action recognition performance tested for corpus 1 of this work. In total, there were 450 sets of actions. It was difficult to recognise 'sitting' actions, probably because HMMs were trained on postures of a complete human body, while a complete posture was often not available when a person was sitting. 'Hand waving' and 'clapping' were related to movements in upper body parts, and 'walking' and 'running' were based on lower body movements. In particular 'waving' appeared an easy action to identify because of its significant moves of upper body parts.

4.6 Objects Recognition

Haar features are extracted and classifiers are trained to identify non-human objects. First, a classifier (namely a cascade of boosted classifiers working with haar-like features) is trained with a few hundred sample views of a particular object (*i.e.*, a face or a car), called positive examples, that are scaled to the same size (say, 20x20), and negative examples - arbitrary images of the same size. After a classifier is trained, it can be applied to a region of interest

	(ground truth)					
	stand	sit	walk	run	wave	clap
stand	98	12	19	3	0	0
sit	0	68	0	0	0	0
walk	22	9	105	8	0	0
run	4	0	18	27	0	0
wave	2	5	0	0	19	2
clap	0	0	0	0	4	9

Table 4.4: Confusion table for human action recognition. Columns show the ground truth, and rows indicate the automatic recognition results. Some actions (e.g., ‘standing’) were more commonly seen than others (e.g., ‘waving’).

(of the same size as used during the training) in an input image. The classifier outputs a ‘1’ if the region is likely to show the object (*i.e.*, car, bike, tree *etc.*), and ‘0’ otherwise. To search for the object in the whole image one can move the search window across the image and check every location using the classifier. The classifier is designed so that it can be easily ‘resized’ in order to be able to find the objects of interest at different sizes, which is more efficient than resizing the image itself. So, to find an object of an unknown size in the image the scan procedure should be done several times at different scales. The word ‘cascade’ in the classifier name means that the resultant classifier consists of several simpler classifiers (stages) that are applied subsequently to a region of interest until at some stage the candidate is rejected or all the stages are passed. The word ‘boosted’ means that the classifiers at every stage of the cascade are complex themselves and they are built out of basic classifiers using one of four different boosting techniques (weighted voting).

OPENCV¹ presents complete steps for training a classifier for rapid object detection [Lienhart and Maydt, 2002]. For corpus 1 of this research work, objects which can be successfully identified include car, bus, motor-bike, bicycle, building, tree, table, chair, cup, bottle and TV-monitor. We found their average precision² scores ranging between 44.8 (table) and 77.8 (car).

4.7 Indoor / Outdoor Scene Identification

Scene settings *i.e.*, indoor or outdoor scene is performed based on the edge oriented histogram (EOH) and the colour oriented histogram (COH) [Kim et al., 2010b]. Most images tend to contain the objects of interest at the center area of the image and it is easily noted that the objects hardly play a role in the case of indoor/outdoor image classification. For example, a human face in the image provides little information since it often appears regardless of the image class. Therefore, it is natural to observe that the more likely area to get clues for the classification is boundary areas rather than the center areas. A given image is first divided into boundary regions and center regions, which are further partitioned into three and two blocks,

¹<http://opencv.willowgarage.com/wiki/>

²defined by Everingham et al. [2010]

4. IMAGE PROCESSING METHODS AND EVALUATIONS

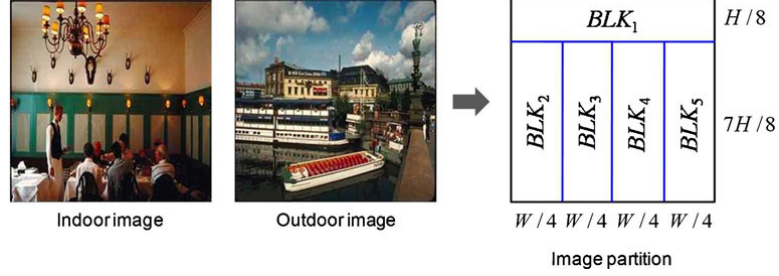


Figure 4.8: 16-bin ECOH descriptor generation. (figure taken from [Kim et al., 2010b])

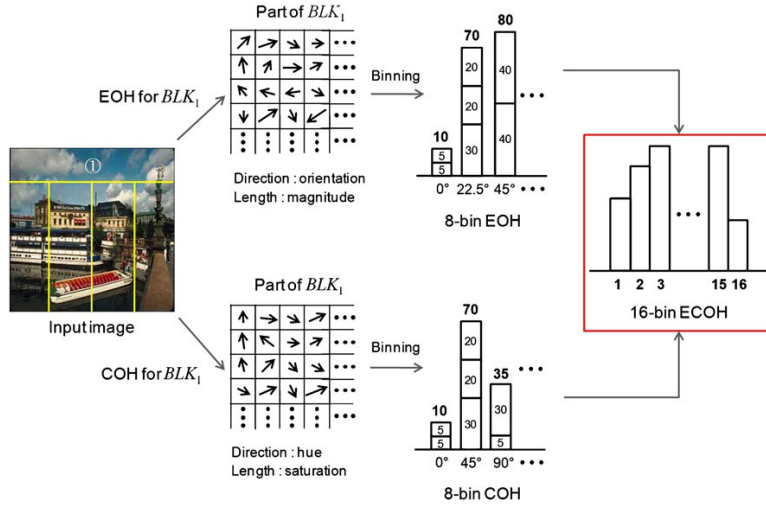


Figure 4.9: 16-bin ECOH descriptor generation. (figure taken from [Kim et al., 2010b])

respectively, as shown in Figure 4.8.

Let us define the pixel set R as $R = (x, y) \mid 1 \leq x \leq W, 1 \leq y \leq H$ where the image size is $W * H$. The pixel set R is divided into boundary and center regions. P , defined as the form of segmented R , can be represented by $P = BLK_1, \dots, BLK_5$ where BLK_1, BLK_2, BLK_5 belong to the boundary regions BR and BLK_3, BLK_4 to the center regions CR .

ECOH Descriptor Definition. The edge orientations are firstly computed over all pixels belonging to each block BLK_i (1..5) and then quantized into K angles $0^\circ, 180^\circ/K, 2 \times 180^\circ/K, \dots, (K-1) \times 180^\circ/K$. Next, the K bin histogram is generated by accumulating the corresponding edge magnitude belonging to each edge orientation over the pixels within each block

$$F_{i,n}^{EOH} = \frac{E_{i,n}}{\sqrt{\sum_{j=1}^k + \varepsilon}} \quad (4.6)$$

where $E_{i,n} = \sum_{(x,y) \in BLK_i, \Theta(x,y) \in n} m(x,y)$ and $1 \leq i \leq 5$. Here $m(x,y)$ and $\theta(x,y)$ denote the edge magnitude and quantized orientation at the pixel position (x,y) , respectively. $n, 1 \leq n \leq K$ denotes the histogram index. ε is a small positive constant.

Since the hue value represents the different color tone according to the angle ranged in the HSV color space, it is also quantized and used as the indices of the color orientation histogram bin. Also, the color magnitude can be represented by the saturation value. That is, purer colors have larger magnitudes for the corresponding color tone in the COH. Moreover, as mentioned, different illumination conditions between indoor and outdoor environments yields different saturation values, it is robust to the misclassification due to colors similar with sky and grass. In detail, the COH descriptor is generated by accumulating the saturation value to the corresponding hue value. Unlike the edge orientation, since the color tones are different through the range of $0^0 \sim 360^0$ the entire range needs to be quantized into K angles. The COH can be generated as follows:

$$F_{i,n}^{COH} = \frac{C_{i,n}}{\sqrt{\sum_{j=1}^K (C_{i,j})^2 + \varepsilon}} \quad (4.7)$$

where $C_{i,n} = \sum_{(x,y) \in BLK_i, \Theta(x,y) \in n} s(x,y)$ and $1 \leq i \leq 5$. Here $s(x,y)$ and $h(x,y)$ denote the saturation value and the quantized hue value at the pixel position (x,y) , respectively. $n(n = 1, \dots, K)$ denotes histogram index. The ECOH descriptor is finally defined by combining the EOH and COH descriptor for each block. The feature vector is generated by multiplying weights for each block to ECOH descriptor. The feature vector is shown in the following equation.

$$F = (w_1 F_1^{ECOH}, w_2 F_2^{ECOH}, \dots, w_5 F_5^{ECOH}) \quad (4.8)$$

$$F_i^{ECOH} = (F_{i1}^{EOH}, F_{i1}^{COH}, F_{i2}^{EOH}, F_{i2}^{COH}, \dots, F_{iK}^{EOH}, F_{iK}^{COH})$$

The ECOH descriptor makes possible to effectively classify the indoor and outdoor images since it is robust to the effect of sky and grass colors in both classes. In corpus 1 of this work, there were 15 videos where human or any other moving HLF (*e.g.*, car, bus) were absent. Out of these 15 videos, 12 were related to outdoor environments where trees, greenery, or buildings were present. Three videos showed indoor settings with objects such as chairs, tables and cups. All frames from outdoor scenes were correctly identified; for indoor scenes 80% of frames were correct. Presence of multiple objects seems to have caused negative impact on EOH and COH features, hence resulted in some erroneous classifications.

4.8 Speaking Person Identification

For speaking person identification, a method was implemented which combined two techniques, *i.e.*, lip movements and body movements detection. The approach was able to identify the speaker in different scenarios, *i.e.*, a single speaker or multiple speakers in the video or if the speaker's lips are not in view. Figure 4.10 presents steps for detection of a speaking person in a video stream. Head nodding, hand gesture or body movement can lead to identification of a speaking person in a video sequence.

4. IMAGE PROCESSING METHODS AND EVALUATIONS

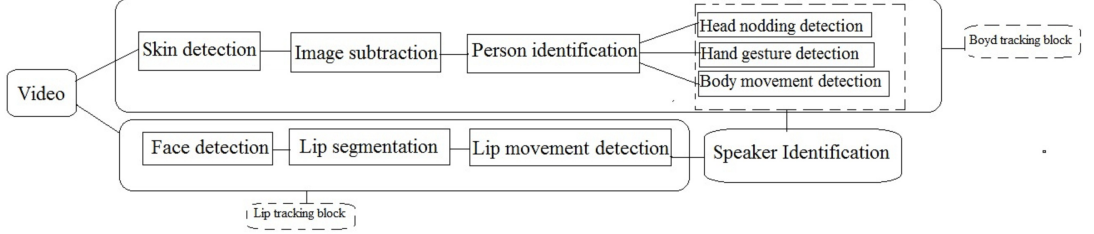


Figure 4.10: Speaker detection framework is decomposed into two steps; (i) lip tracking (ii) body movement detection based on head nodding, gesture identification and simple body movements.

Body based speaker identification.

$$\begin{aligned}
 &(R > 95) \text{ AND } (G > 40) \text{ AND } (B > 20) \\
 &\text{AND } (\max \{R, G, B\} - \min \{R, G, B\} > 15) \\
 &\text{AND } (|R - G| > 15) \text{ AND } (R > G) \text{ AND } (R > B)
 \end{aligned}$$

Initially areas related to human skin are selected and remaining portion of video sequences are truncated. Iwano et al. [2001] presented mathematical equation 4.9 for differentiating skin color from other colors using RGB values. This method is by no means perfect, *i.e.*, it would not work for all skin colours at once and it will more than likely detect background areas which are not a person. However major concern here is, body movement or moving skin areas detection, so background being detected as skin colour is not a problem as long as background is not moving. Next, image subtraction between the consecutive frames is used to detect any movement. Using grouping and connected component methods, clusters related to body movement are selected. Properties of these clusters such as ‘area’ were also stored. One observation made here was that identified clusters due to significant body movements were bigger than their counter parts. This observation lead to defining a threshold for keeping specific sized clusters. In particular when a person is far away from the camera, the skin detected areas become small enough to be the same size as non-skin detected areas. Figure 4.11 shows an example of a speaking person identification based on her body movements alone.

Lip tracking based speaker identification. Lip tracking starts from identification of human face which was achieved using OPENCV provided face detector. Similar to discussion of Section 4.4, mouth region from the given face was cropped. Later, lips from mouth region were separated and openness or closeness of lips guided for speaking person identification. Lips area detection was performed by using structuring elements [Da Silveira et al., 2003], which resulted in converting the lip area as ‘white colored’ and non-lip region as ‘black’. The openness of the mouth is then calculated for the image by calculating the area of the holes present in the image larger than the structuring elements size. For deciding a speaking person a manual threshold was defined, *i.e.*, if for three consecutive frames mouth region is identified as closed, then the speaker is classed as ‘not speaking’. Figure 4.12 shows an example of speaking person identification based on lip movements.

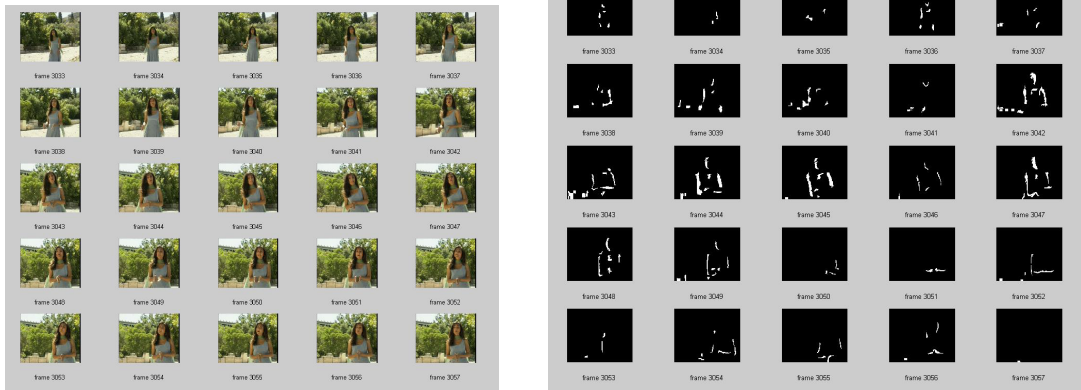


Figure 4.11: (left panel) Video sequence showing motion of one person. (right panel) Detection of motion related to one person movement in a video sequence.

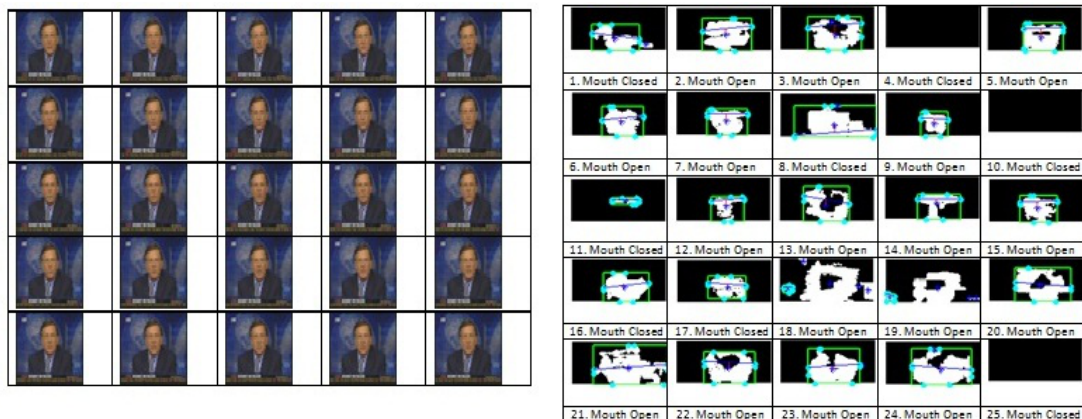


Figure 4.12: Example scenario for one person lip detection. For speaking person identification, mouth need to be opened in three consecutive frames, such as frames 12 to 14 on line 3.

	speak	not speak		speak	not speak
speak	312	127	speak	59	24
not speak	120	211	not speak	16	37
(a) Lip based speaker detection			(b) Body movement based speaker detection		

Table 4.5: Confusion table for (a) Lip based speaker detection (b) Body movement based speaker detection. Columns show the ground truth, and rows indicate the automatic recognition results. Note that the human detection task is biased towards existence of human, while in the gender identification presence of male and female are roughly balanced.

4. IMAGE PROCESSING METHODS AND EVALUATIONS

Table 4.5 presents results for speaking person identification in video corpus 1 of this work. Since human’s face was obvious in most of the frames, mostly, speaker identification was done based on the lip movements. Nearly about 80% lip movement based speaker identification performed as expected, whereas for body movement based speaker identification had much lower results like 60%. Since our assumption was that ‘body movement is directly linked to speaking’, these results are much satisfactory. Other aspects which affected the results were the content and quality of the videos. Content is related to colour of the background, moving objects in the background (cars, human *etc.*), the position of the humans in relation to the camera and the scene settings of the video *i.e.*, indoor and outdoor. For lip based speaker detection, results for situations, when the mouth was open were quite satisfactory. Whereas results for situations when the mouth was closed did not produce the results as expected. The reason for this is the the fact that, when the lip region is not correctly found, the default setting for this is to say the mouth region is closed.

4.9 Formalising Spatial and Temporal Relations

Spatial relations between human and other objects and within objects are extracted. Since, detected humans and other objects are present in bounding boxes, finding spatial relation between them becomes straight forward. Following relationships can be recognised between two or three objects: ‘in front of’, ‘behind’, ‘to the left’, ‘to the right’, ‘beside’, ‘at’, ‘on’, ‘in’, and ‘between’. Figure 4.13 illustrates steps for calculating the three-place relationship ‘between’. Schirra et al. [1987] explained the algorithm:

- Calculate the two tangents g_1 and g_2 between the reference objects using their closed-rectangle representation;
- If (1) both tangents cross the target (or its rectangle representation), or (2) the target is totally enclosed by the tangents and the references, the relationship ‘between’ is true.
- If only one tangent intersects the subject, the applicability depends on its penetration depth in the area between the tangents, thus calculate: $\max(a/(a+b), a/(a+c))$
- Otherwise ‘between’ relation does not hold.

The final step for unit based video description is related to induction of temporal information. *TimeML* specification language is a standard for presenting temporal information in a natural language text Pustejovsky et al. [2003]. This includes temporal expressions, events, and the relationships they share. The following four tag types are mainly used:

TIMEX3 tag is used for temporal expressions such as date, time, duration and sets. In general they are hard facts presented by specific keywords such as ‘12th June’, ‘12:10’, ‘20 mins’ and ‘two days every week’.

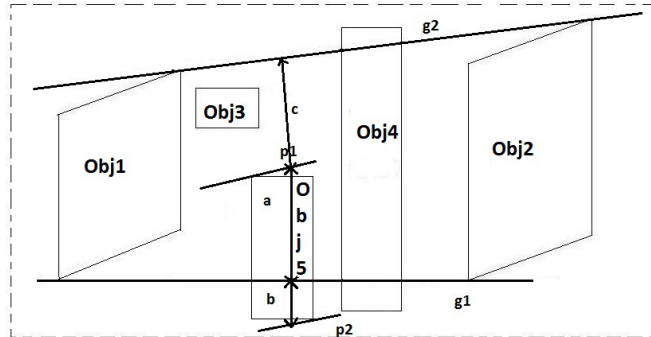


Figure 4.13: Procedure for calculating the ‘between’ relation. Obj 1 and 2 are the two reference objects, while Obj 3, 4 and 5 are the target objects.

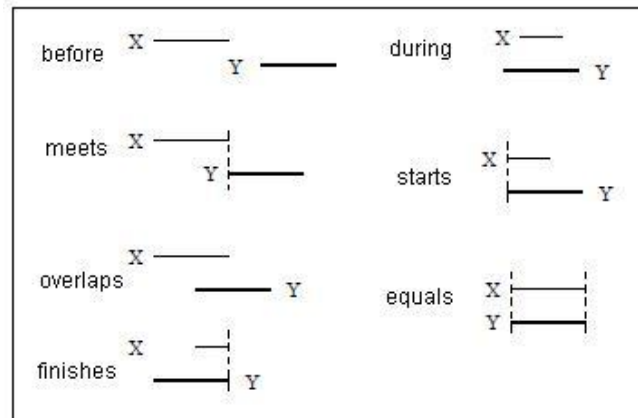


Figure 4.14: Temporal relations between two events X and Y. Figure from Muller and Reymonet [Muller and Reymonet \[2005\]](#).

EVENT tag usually presents verbs. Nouns, adjectives, and even some prepositions can also be described by this tag. There are seven event categories, *i.e.*, REPORTING, PERCEPTION, ASPECTUAL, I_ACTION, I_STATE, STATE and OCCURRENCE.

SIGNAL tag is used to relate temporal objects to each other by an additional word present, such as ‘at’, ‘on’, and ‘between’, whose function is to specify the nature of that relationship.

LINK tag presents relationship between times, events, or between times and events.

Details of *TimeML* and its tags can be found from Pustejovsky *et al.* [Pustejovsky et al. \[2004\]](#). Note that LINK tag in *TimeML* is based on Allen’s relation algebra [Allen \[1984\]](#). Relations between two time intervals are based on position of the interval endpoints (before, after or simultaneous). Combination of these intervals results in 13 relations; some of which are shown [Figure 4.14](#).

[Nevatia et al. \[2003\]](#) introduced Event Recognition Language based on Allen’s algebra for recognition of composite events. [Ryoo and Aggarwal \[2006\]](#) pre-

4. IMAGE PROCESSING METHODS AND EVALUATIONS

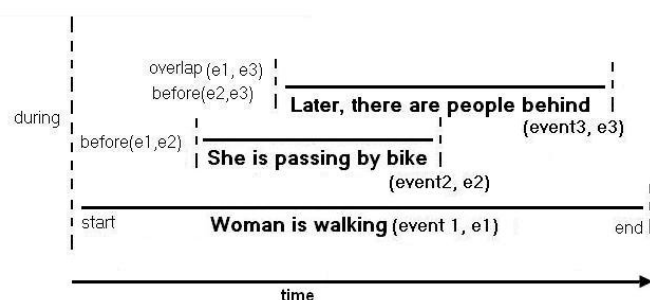


Figure 4.15: TimeML applying temporal relations to machine generated descriptions for Figure 7.1. Three events are identified and relations between events are shown as ‘start’, ‘before’, and ‘overlap’.

after	then, after, later, next, thereafter
during	while, at the same time, meanwhile, throughout
before	previous, afterwards, prior to, since
start	initiation, at the beginning, at the start

Figure 4.16: Keywords for temporal relations between sentences. These keywords are defined for explaining relations based on Allen algebra.

sented an approach for automatic composite actions recognition based on context free grammar. Generally both these approaches focus on recognizing human activities based on the relations among atomic-level actions. Our work is more concerned towards the detection of the relations between atomic-level actions and events. Use of these relations in our work is twofold; first, we aim to incorporate temporal information into description of activities by a single human, using expressions such as ‘before’ and ‘after’. Second, interactions between multiple humans can also be presented, using expressions such as ‘meets’, ‘overlaps’, ‘equals’ and ‘during’.

We use TARSQI toolkit Verhagen et al. [2005] to find these relations between sentences. Figure 4.15 shows temporal relations for the video sequence from Figure 7.1. There are three events identified — ‘walking’, ‘passing by bike’ and ‘walking while other people at background’. ‘Duration’ relations is common among these three events. First two events have ‘before’ and ‘after’ relation between them. Last event has two subjects involved; firstly a walking woman and secondly other humans. There is ‘overlap’ relation between woman walking and other people walking.

Most common relations in video sequences are ‘before’, ‘after’, ‘start’ and ‘finish’ for single humans, and ‘overlap’, ‘during’ and ‘meeting’ for multiple humans. Once these relations are identified between sentences, keywords are manually defined to present them. Figure 7.5 shows a list of keywords used for TimeML to represent relations. Finally they are put into templates to produce temporally coherent description of video sequences.

4.10 Summary

This chapter presented implementation details and evaluation results of feature extraction task which is based on image processing methods. Results related to human structure related information *i.e.*, face, age, gender and emotion and humans' action recognition are presented. Methods for detection of other objects such as car, table, chair and tv-monitors *etc.*, are presented and evaluated. Scene settings which are either indoor or outdoor are discussed. Results for identification of speaking person based on visual features on are also presented. Finally results for finding spatial relations between HLFs and temporal relations across individual frames are discussed. The results are evaluated using confusion matrix *i.e.*, in the form of false positives, false negatives, true positives, and true negatives. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

4. IMAGE PROCESSING METHODS AND EVALUATIONS

Chapter 5

Natural Language Descriptions for Visual Images

5.1 Introduction

The study presented in this chapter is concerned with production of natural language descriptions for visual images in a time series using a bottom-up approach. Initially high level features (HLFs) are identified in individual video frames. They may be ‘keywords’, such as a particular object and its position/moves, used for a semantic indexing task in video retrieval. Spatial relations between HLFs are calculated to improve the explanation of the semantics of visual scene. Extracted HLFs and spatial relations between them are then presented by syntactically and semantically correct expressions using context free grammar. The approach is evaluated using video corpus 1 which consists of video segments drafted manually from the TREC video dataset. Overlap between human annotated and machine generated descriptions is calculated using ROUGE score¹ which presents quantitative evaluation of this description generation task. Further, a task based evaluation is performed by human subjects, providing qualitative evaluation of generated descriptions.

Chapter 6 further extends this language description work towards a generic model for dealing with noisy data and to accommodate other video categories in addition to categories discussed in this chapter. Scenarios for dealing with missing and erroneous data to produce coherent descriptions for video sequences are presented. That chapter also explains application of framework to different video genres. Chapter 7 continues discussion of this chapter for complete video sequences. The frame based natural language generation procedure results in many identical descriptions produced from adjacent frames. Hence simple concatenation of descriptions may lead to redundancy, lacking coherency. Lack of temporal information may cause a further problem. To remedy these shortcomings, a scheme to generate coherent and compact descriptions for video streams is presented in that chapter.

¹standard evaluation measure for text summarization

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

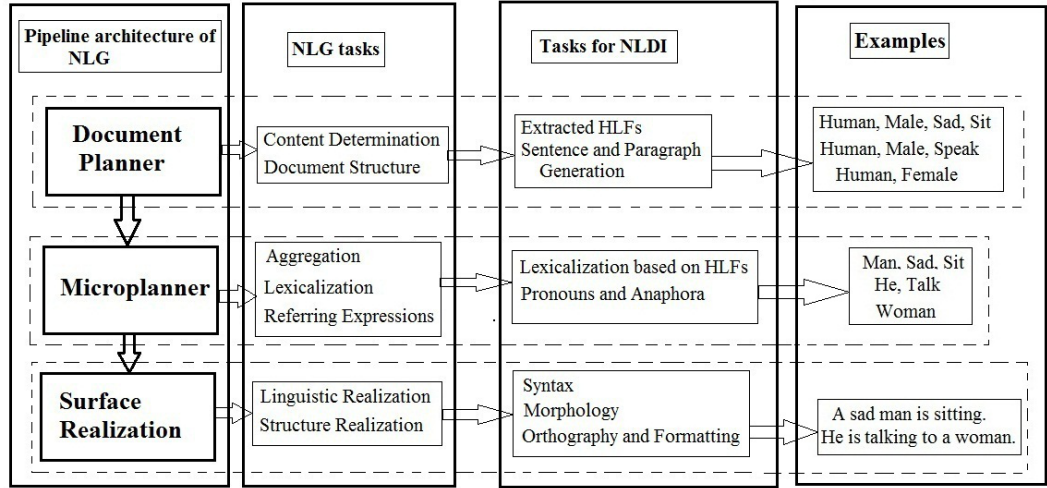


Figure 5.1: Pipeline architecture of NLG presented by [Reiter and Dale, 1997]. Tasks for NLG and relevant tasks for Natural language description of images (NLDI) are presented. Columns present steps for NLG and rows provides details about individual tasks for NLG and NLDI.

5.2 Framework Overview

The overall process of generation of natural language descriptions is based on the ‘Natural language generation (NLG)’ architecture proposed by Reiter and Dale [1997], which includes three modules; a document planner, a microplanner, and a surface realizer.

Document Planner. Image processing module provides the information to be described; this task is considered to be part of the document planner. Document planner is further responsible for deciding structure of the generated description and providing coherency to the description. Content determination and document structure are two subtasks performed by document planner. HLF extracted from image sequences are considered contents in this work. Chapter 4 provided details about the algorithms for extraction of HLFs from image sequences and procedure for extracting spatial relations between HLFs. In relation to document structure, HLFs are first combined to generate sentences. These sentences are then put together to generate full length paragraphs for generating a coherent description of the given images.

Microplanner. Microplanner is responsible for performing three subtask, (i) Lexicalization: choice of words/syntactic constructs and mark up annotations (ii) Aggregation: deciding how much information is communicated by each sentence (iii) Determination of referring expressions: determining what phrases should be used to identify entities to the user. HLFs extracted from image processing need to be assigned their proper semantic tags such as agents, patients, objects and events, for instance. Lexicalization is responsible for such a mapping of HLFs into semantic tags. Figure 5.4 presents a list of lexicons with their part of speech tags for each of the extracted HLF. Pronouns and anaphora are handled manually in this work.

Surface Realization. The final stage of generation is surface realization. This is a purely

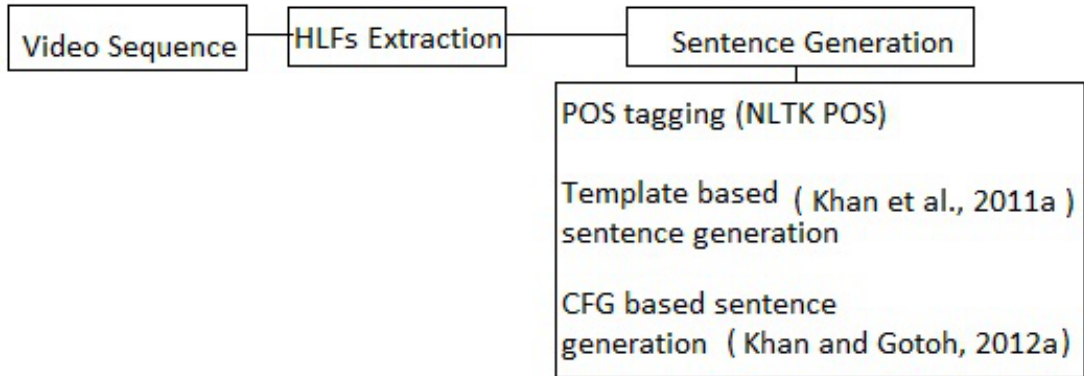


Figure 5.2: Steps for video to sentence generation.

linguistic level, which takes choices about words and syntactic structures made during sentence planning, and constructs a sentence using them. Surface realization is the final stage of this description generation process. Section 5.3 presents the proposed approach for surface realization which combines context free grammar with templates for syntactically and semantically correct text generation.

Example Scenario. Figure 5.1 presents pipeline architecture of natural language generation (NLG) [Reiter and Dale, 1997]. First column presents main tasks of NLG, while second column shows subtasks against each main task. Third column presents corresponding tasks for natural language description for visual images. Finally column four, presents as example scenario. Suppose a frame shows ‘a sad man and woman talking to each other’. For generating output for such a frame, initially, in document planner information related to HLF extraction is stored. For such a scenario extracted HLFs include, human, male, sad, sit, speak, human and female. Document planner further decides that extracted HLFs will be combined into sentences which further combine to generate a paragraph. Microplanner selects proper lexicons for the extracted HLFs such as ‘human + male’ is replaced with ‘man’, ‘human + female’ with ‘woman’, ‘speak’ with ‘talk’ *etc.* Surface realization is the final step of this generation process and generates sentence which are syntactically and semantically correct such as for lexicons ‘man, sad, sit, talk, woman’, generated sentences are ‘A sad man is sitting. He is talking to a woman’.

5.3 Natural Language Generation

Figure 5.2 provides steps for video to sentence generation framework. Two new approaches are proposed for the generation of sentences from HLFs. Extracted HLFs are combined using template based approach for generating sentences related to single human. CFG based approach is used for multiple humans scenarios.

HLFs acquired by image processing require abstraction and fine tuning for generating syntactically and semantically sound natural language expressions. Firstly, a part of speech (POS)

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

HLF	POS tag	
man	NN	noun, singular
faces	NNS	noun, plural
have	VB	verb
is	VBZ	verb, present tense
sitting	VBG	verb, present participle

Table 5.1: Examples of POS tags assigned for HLFs. Note that VBG (verb, present participle) also calls progressing verb ‘is’ in the expression (e.g., ‘person is walking’).

<p><u>Subject + Verb</u> Man is walking. A woman is standing.</p>
<p><u>Subject + Verb + Object</u> Person is smoking cigarette. A man is drinking tea.</p>
<p><u>Subject + Verb + Complement</u> He looks tired. Man is old.</p>
<p><u>Subject + Verb + Object + Complement</u> He left the door open. A man is kicking the ball with his right leg.</p>
<p><u>Tense: Present continuous tense</u> They are jogging. A man is walking.</p>

Figure 5.3: A partial list of templates for sentence generation. To fill in the template, POS tagger assigns labels for all HLFs, such as subject, verb, complement, object — direct and indirect object.

tag is assigned to each HLF using NLTK¹ POS tagger (see Table 5.1 for examples). Further humans and objects need to be assigned proper semantic roles. In this study, a human is treated as a subject, performing a certain action. Other HLFs are treated as objects, affected by human’s activities. These objects are usually helpful for description of background and scene settings.

A template filling approach based on context free grammar (CFG) is implemented for sentence generation. A template is a pre-defined structure with slots for user specified parameters. Each template requires three parts for proper functioning: lexicons, template rules and grammar. Lexicon is a vocabulary containing HLFs extracted from a video stream (Figure 5.4). Grammar assures syntactical correctness of the sentence. Template rules are defined for selection of proper lexicons with well defined grammar.

Given a frame, a sentence is generated for each of most important entities. Figure 5.3² presents a partial list of templates used for this work. For example, a simple template can be

¹www.nltk.org/

²Complete list is provided in the Appendix D.

Noun	→	man woman car cup table chair cycle head hand body
Verb	→	stand walk sit run wave
Adjective	→	happy sad serious surprise angry one two many young old middle-aged child baby
Pronoun	→	me i you it she he
Determiner	→	the a an this these that
Preposition	→	from on to near while
Conjunction	→	and or but

Figure 5.4: Lexicons and their POS tags.

```

If (gender == male) then man else woman
Select 1 (Action == walk, run, wave, clap, sit, stand)
Select n (Object == car, chair, table, bike)
Elaboration (If 'the car is moving' and 'man is  
inside the car') then 'man is driving the car'

```

Figure 5.5: Template rules applied for creating a sentence 'man is driving the car'.

subject (S) performs action (A) on object (O)
(*e.g.*, 'John (S) kicked (A) the ball (O)')

Template Rules. Template rules are employed for the selection of appropriate lexicons for sentence generation. Followings are the template rules used in this work:

Base returns a pre-defined string (*e.g.*, when no HLF is detected)

If same as an if-then statement of programming languages, returning a result when the antecedent of the rule is true

Select 1 same as a condition statement of programming languages, returning a result when one of antecedent conditions is true

Select n is used for returning a result while more than one antecedent conditions is true

Concatenation appends the the result of one template rule with the results of a second rule

Alternative is used for selecting the most specific template when multiple templates can be used

Elaboration evaluates the value of a template slot

Figure 5.5 presents an example to illustrate template rules selection procedure. This example assumes human presence in the video. **If-else** statements are used for fitting proper gender in the template. Human can be performing only one action at a time referred by **Select 1**. There can be multiple objects which are either part of background or interacting with humans. Objects are selected by **Select n** rule. These values can be directly attained from HLFs extraction step. **Elaboration** rule is used for generating new words by joining multiple HLFs. '*Driving*' is achieved by combing '*man is inside car*' and '*car is moving*'.

Grammar. Grammar is the body of rules that describe the structure of expressions in any language. We make use of context free grammar (CFG) for the sentence generation task. CFG

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

S → NP VP	man is walking
S → NP	man
NP → Pronoun	he
NP → Det Nominal	a man
Nominal → Noun	man
Nominal → Adjective nominal	old man
VP → Verb	wave
VP → Verb NP	wave hand
VP → Verb PP NP	sitting on chair
PP → Preposition NP	on chair

Figure 5.6: Grammar for lexicons shown in figure 5.4, with example phrases for each rule.

	Determiner +	Cardinal +	Adjective +
Subject			
Subject + Verb			
Subject + Verb + Object			

Table 5.2: Template selection procedure. Rows present blocks and columns present sub-blocks, while cells contain templates for sentence generation.

based formulation enables us to define a hierarchical presentation for sentence generation; *e.g.*, a description for multiple humans is comprised of single human actions. CFG is formalised by 4-tuple:

$$G = (T, N, S, R)$$

where T is set of terminals (lexicon) shown in Figure 5.4, N is a set of non-terminals (usually POS tags), S is a start symbol (one of non-terminals). Finally R is rules / productions (Figure 5.6) of the form $X \rightarrow \gamma$, where X is a non-terminal and γ is a sequence of terminals and non-terminals which may be empty.

Creating Candidate Sentences. Given the list of lexicons, grammar, and template rules, the sentence generation algorithm aims to produce a sentence without losing the original contents. Table 5.2 presents sentence generation algorithm. Rows present blocks, columns present sub-blocks, and cells contain templates for sentence generation. Based on part of speech tags, information such as subject, verb, object, determiner, cardinality and adjectives is decided for each of the extracted HLF *i.e.*, for each lexicon. Suppose that available information is about ‘subject’ only, it triggers focus on the first row and remaining two rows are discarded. Once row is selected, then the starting phrase of the sentence needs to be generated and put into the most matching column. For example, if only ‘one man’ is extracted by image processing then, templates from first column *i.e.*, ‘Determiner +’ column is selected, for HLFs such as two men, ‘Cardinal +’ is selected and for HLFs happy man, column ‘Adjective +’ is selected. Selecting column leads to finalizing the cell for template selection. Each cell contains multiple templates and selection of best template for paraphrase generation is based on language modeling scores.

Suppose given list of extracted HLFs for a frame consists of ‘Human + Male’, which is converted to ‘Man’ as the lexicon for sentence generation. Since man is considered as subject so

first block (row) from step 1 is selected which consists of subject information alone. Now in the second step, sub-block (column) presenting determiner as the starting phrase is selected which is ‘a’ for this particular example. Finally, in the third step best matching template is selected based on language modeling score. For this particular example, there are only two choices of templates, *i.e.*, (i) ‘A man is present.’ (ii) ‘There is a man.’ Suppose given list of templates is ‘Human + Male + Happy + Sit + Chair’. Since there is information about subject, verb and object in this information, block 1 is selected for the first step. For starting phrase, a choice needs to be made between ‘Determiner +’ and ‘Adjective +’ columns. If third column is selected which starts with the adjectives, generated sentence is ‘Happy man is sitting on the chair.’ If first column is selected then generated sentence is ‘A man is happy and sitting on the chair.’ Based on the language modeling scores for both of these sentences, final sentence is generated which is ‘Happy man is sitting on the chair.’ for this particular example.

Figure 5.7 (upper part) shows an example for sentence generation related to a single human. This mechanism is built with three blocks when only one subject¹ is present. The first block expresses a human subject with age, gender and emotion information. The second block contains a verb describing a human action, to explain the relation between the first and the third blocks. Spatial relation between the subject and other objects can also be presented. The third block captures other objects which may be either a part of background or a target for subject’s action.

Hierarchical Sentence Generation. In this work we define a CFG based presentation for expressing activities by multiple humans. [Ryoo and Aggarwal \[2009\]](#) used CFG for hierarchical presentation of human actions where complex actions were composed of simpler actions. In contrast, we allow a scenario where there is no interaction between humans, *i.e.*, they perform individual actions without a particular relation — imagine a situation whereby three people are sitting around a desk while one person is passing behind them.

The approach is hierarchical in the sense that we start with creating a single human grammar, then build up to express interactions between two or more than two humans as a combination of single human activities. Figure 5.7 (middle part) presents examples involving two subjects. There can be three scenarios; firstly two persons interact with each other to generate some common single activity (*e.g.*, ‘hand shake’ scene). The second scenario involves two related humans performing individual actions but they do not create a single action (*e.g.*, both persons are walking together, sitting or standing). Finally two persons happen to be in the same scene at the same time, but there is no particular relation between them (*e.g.*, one person walks, passing behind the other person sitting on a chair). Figure 5.7 (lower part) shows an example that involves an extension of a single human scenario to more than two subjects. Similarly to two-human scenarios, multiple subjects can create a single action, separate actions, or different actions altogether.

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

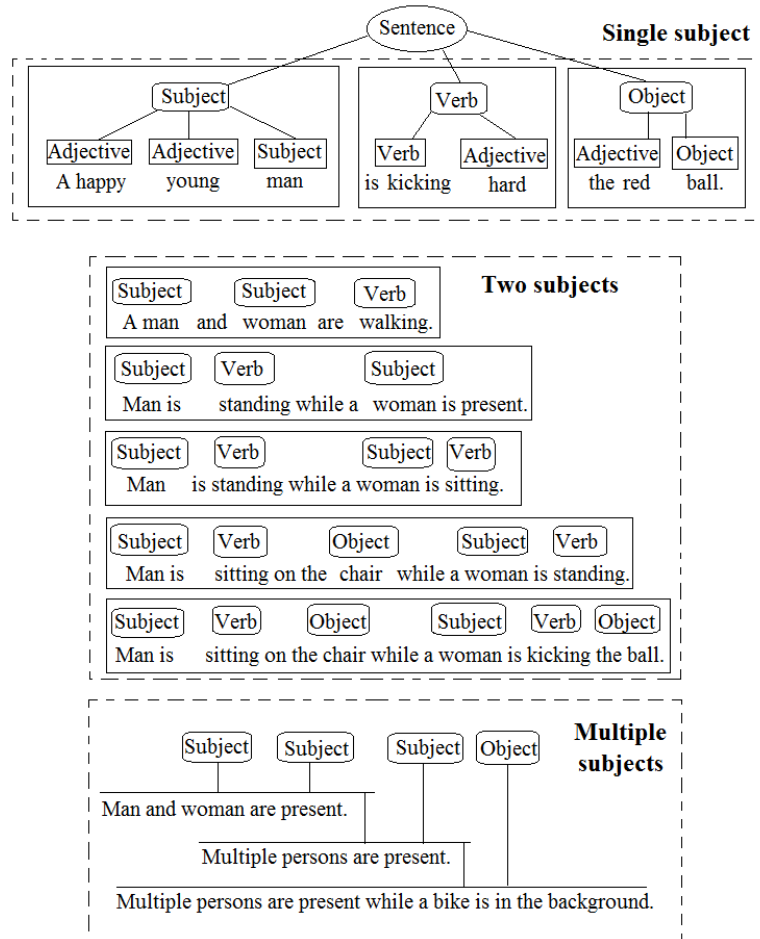


Figure 5.7: Description generation for scenes containing single, two and multiple humans. For elaboration purposes, just one combination of human interaction is shown in two subjects and multiple subjects scenarios, although there can be several scenarios of multiple human interactions. Complete list of templates for multiple humans is presented in the Appendix.

Input: video stream, E (initially empty sentence)
Output: F (populated final sentence)

(1) Find subject of the sentence:
— if one human is present — add one subject to E
— if two humans are present — add two subjects to E
— if more than two humans ... — add multiple subjects to E

(1.1) Find age, gender, emotion (adjective) for subject(s):
— if age is identified — add age to the subject in E
— if gender is identified — add gender to the subject in E
— if emotion is identified — add emotion to the subject in E

(1.2) Find actions (verb) for subject(s):
— if action is identified — add action to the subject in E

(2) Find other HLFs (object) in the video sequence:
— add the object to E
— find the spatial relation between human(s) and HLFs and add keywords to E
— **transfer $E \rightarrow F$ and clear E**

(3) If no human is identified in the video —
— find other HLFs and add these HLF(s) as subject(s) to E
— if HLF is moving — attach ‘*moving*’ (verb) in the E
— if one HLF is moving and the other is static — attach ‘*moving*’ with the moving HLF in the E , and static HLF is considered a part of background
— **transfer $E \rightarrow F$ and clear E**

(4) If no HLF is identified in the video —
— find scene settings (indoor, outdoor)
— if scene settings identified — use the fixed template (*e.g.*, ‘*this is an outdoor scene*’)
— if scene settings are not identified — find any motion in the video and use the fixed template (*e.g.*, ‘*there is movement in the scene*’, or ‘*this is a static scene*’)
— **transfer $E \rightarrow F$ and clear E**

Figure 5.8: Procedure for generating natural language descriptions.

5.3.1 Summary of Generation Procedure

Given the components discussed so far, this section provides a compact procedure of description generation for visual images. Given a video, the following strategy is applied to create a natural language description. First, subject(s) should be identified; there can be one, two or more than two (many) humans present in the video. Determiners and cardinals (*e.g.*, ‘*the*’, ‘*an*’, ‘*a*’, ‘*two*’, ‘*many*’) are selected based on the number of subjects. Age, gender and emotion are selected as an adjective for each subject. Action and pose (verb) is also identified. In the presence of human(s), non-human objects are considered either as objects operated by a human or as a part of background. The most likely preposition (spatial relations) is calculated and inserted between the subject, verb and objects.

Suppose that human is absent in the video, a non-human object may be used as a subject. If they are moving, verb (‘*moving*’) will be attached. If one is moving and the other is static, the verb is attached with the moving object and the static one is considered as a part of background. In case no object is identified, we try to find the scene settings (*i.e.*, indoor or outdoor) and express the scene using a fixed template. Finally, if the scene setting is not identified, we try to detect any motion and express the scene using a fixed template. The above procedure is outlined in Figure 5.8.

Sample Scenario for Description Generation. For elaboration of concept, Figure 5.9 presents a real scenario where two humans are present in an indoor scene. Firstly, using image processing methods, HLFs are extracted from the visual image. A list is generated for each of these HLFs containing HLFs with their corresponding attributes *e.g.*, for human, age, gender, action and emotion is enlisted. Secondly spatial relation between these HLFs are calculated. In this figure, two humans are present, each of which is involved in different action *i.e.*, standing and sitting.

For description generation, first HLFs related to each human are identified. Second, spatial relations between these humans are calculated. Finally, remaining objects in the visual image are considered for complete generation of the natural language description. For this particular examples, extracted HLFs for first subject are ‘*human, male, stand and speak*’, and for the second subject are ‘*human, male, sit and speak*’, spatial relation between humans is ‘*second subject is on the right of first subject*’, and ‘*second subject is sitting on the chair*’. ‘*Four chairs*’ are additionally identified as part of the visual scene. Scene settings are identified as indoor scene. Based on these extracted HLFs, four sentences need to be generated for this visual scene, one for subject 1, one for subject 2, one for chairs and one for the scene settings. For subject 1, 2nd row from table 5.2 is selected for candidate templates. Since there is no information about cardinal or adjective, cell 1 is selected for the 2nd row for the starting phrase. Generated sentence for subject 1 is ‘*A man is standing and speaking.*’ For subject 2, 3rd row is selected since subject has information about ‘*subject, verb and object*’ and again, ‘*determiner +*’ is selected for starting phrase. Generated sentence for subject 2 is ‘*A man is sitting on the chair and talking.*’ Now for ‘*chair*’, since there is no human, objects *i.e.*, chairs are considered as

¹ Non human subject is also allowed in the mechanism.

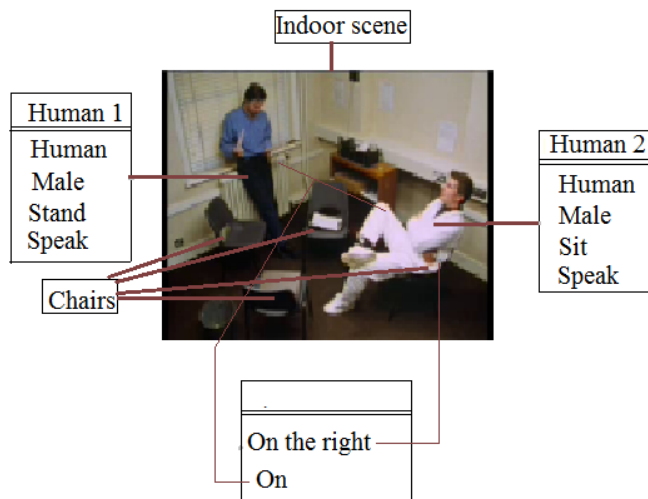


Figure 5.9: Hierarchical Generation: Description for multiple humans consist of individual human's descriptions. Boxes present HLFs and lines show these HLFs inside the real image.

subjects. There is no movement information so, row 1 from table 5.2 is selected. For starting phrase, there is information about cardinal, *i.e.*, four, so 2nd cell of row 1 is selected and generated sentence is 'Four chairs are present.' Finally, scene settings are presented using a fixed template, *i.e.*, 'this is an indoor scene.'

Now, four separate sentence for each of the main HLFs of the visual scene. Next step is to explore interaction between HLFs, such as subject 1 and subject 2 in Figure 5.9 are talking to each other. For this interaction, both the sentences related to each of these subjects need to be joined together to show the relationship between each of these subjects. For this joining process, initially one of the two sentences need to be selected as the 'main sentence' and the other sentence is called 'sub-sentence'. This sub-sentence is merged into the main sentence. Selection of main sentence is as follows:

1. Count the number of high level features in both the sentences. The sentence having higher number of HLFs is selected as the main sentence.
2. If number of HLFs is same in both the sentences then the sentence having subject, verb and object information has the higher priority. If both the sentences have subject, verb and object information, then they are just glued together using function words.

Based on these rules, again using table 5.2, 3rd row of cell 1 is selected as the starting phrase since 'subject 2 related sentence is the main sentence', as it has more information in the form of subject + verb + object, and subject 1 related sentence is considered as sub-sentence. It is worth mentioning that information extends down the columns such as 'subject + verb + object' row can have templates which have 'subject + verb' but not vice versa, *e.g.*, 'a man is sitting on the chair while a woman is standing' can be generated but 'a woman is standing while a man is sitting on the chair' is not possible due to the assumption that main sentence is

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

the one having more HLFs. Generated description for these two sentences combined with the spatial relation is ‘A man is sitting on the chair and speaking to a man standing on his left.’

Implementation. For implementing the templates, *simpleNLG* is used [Gatt and Reiter, 2009]. It also performs some extra processing automatically: (1) the first letter of each sentence is capitalised, (2) ‘-ing’ is added to the end of a verb as the progressive aspect of the verb is desired, (3) all words are put together in a grammatical form, (4) appropriate white spaces are inserted between words, and (5) a full stop is placed at the end of the sentence.

5.4 Experiments

In the experiments, video frames were extracted using *ffmpeg*¹, sampled at 1 fps (frame per second), resulting in 2520 frames in total. Most of HLFs required one frame to evaluate. Human activities were shown in 45 videos and they were sampled at 4 fps, yielding 3600 frames. Upon several trials, we decided to use eight frames (roughly two seconds) for human action recognition. Consequently tags were assigned for each set of eight frames, totaling 450 sets of actions.

5.4.1 Machine Generated Annotation Samples

This section provides real application scenarios for language description framework for seven video categories of Corpus 1 (see Chapter 3). For each of seven categories in the dataset, Figures 5.10 to 5.16 present machine generated annotation and two hand annotations. For figure 5.10, the list of extracted HLFs and spatial relation between them is also presented. Note that, figure 5.10 presents a sample video from action category together with list of extracted HLFs, machine generated natural language description and hand annotations.

Action videos: (Figure 5.10). Main interest was to find humans and their activities. Successful recognition of man, woman and their actions (*e.g.*, ‘sitting’, ‘standing’) led to well phrased description. On the other hand other HLFs such as age, emotion and speaking were not recognized. The bus and the car at the background were also identified. In hand annotations dressing was noted and location was reported as a park. Further, age of woman was presented as young and sitting on the chair was also pointed. Finally, interaction between them such as ‘talking’ was also reported.

¹ Ffmpeg is a command line tool composed of a collection of free software and open source libraries. It can record, convert and stream digital audio and video in numerous formats. The default conversion rate is 25 fps. See <http://www.ffmpeg.org/>



Extracted HLFs:

Subject 1: Human, female, sit. Subject 2: Human male, stand.
 Subject 3: Bus. Subject 4: Car.

Spatial relations: between humans, 'human + female is to the left of human + male', 'car and bus are in background'.

Machine generated original:

A woman is sitting to the left of a standing man. There is a bus in the background. There is a car in the background.

Hand annotation:1

Two persons are talking; One is a man and other is woman; The man is wearing formal clothes; The man is standing and woman is sitting; A bus is travelling behind.

Hand Annotation 2:

Young woman is sitting on a chair in a park and talking to man who is standing next to her.

Figure 5.10: An action scene of two humans as seen in '20041101_160000_CCTV4_DAILY_NEWS_CHN' from the HLF extraction task.

Close-up: (Figure 5.11). Main interest was to find human's age, gender and emotion information. Machine generated description was able to capture human gender and emotion information while age information was not recognized. Humans in the background were also successfully identified, although information related to their age, gender or emotions was not identified. Hand annotations explained the sequence more, *e.g.*, dressing, identity of a person as policeman, hair color and windy outdoor scene settings. Even, human annotators explained dressing and identity of background people too such as policeman and people wearing hats.

News video: (Figure 5.12). Main interest was to find the applicability of the proposed framework in a constrained environment. HLFs such as 'man', 'standing', 'speaking' and 'TV' were successfully described for this sequence, but the scene setting ('news studio') were missed. Hand annotations were able to identify the person in the TV, dressing of news reporter and could further describe this scene as a news report.

Meeting scene: (Figure 5.13). Main interest was to finding interaction between multiple humans with distinct objects that are seen in meeting videos. Presence of multiple humans with varying actions and many objects in the TV studio makes this videos difficult to describe by the approach with the current level of development. There were many items missed by the

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES



Machine generated:

A serious man is speaking. There are humans in the background.

Hand annotation:1

A man is talking to someone; He is wearing a formal suit; A police man is standing behind him; Some people in the background are wearing hats.

Hand Annotation 2:

A man with brown hair is talking to someone; He is standing at some outdoor place; He is wearing formal clothes; He looks serious; It is windy.

Figure 5.11: Closeup of a man talking to someone in the outdoor scene — seen in ‘MS206410’ from the 2007 rushes summarisation task.



Machine annotation:

A man is speaking. A man is standing while a tv in the background. There is a person in the tv. A man is speaking.

Hand annotation:1

A news reporter is standing and presenting news report; He is wearing a formal suit; There is a big TV-monitor; There is a famous person ‘Osama bin Laden’ in the video.

Hand Annotation 2:

A man is speaking; And another man is listening; There is a TV; A man is in the TV-monitor; The man is wearing formal clothes; It seems a news report.

Figure 5.12: A news reporter in the TV news studio — seen in ‘20041030_133100_MSNBC_MSNBCNEWS13_ENG’ from the HLF extraction task.



Machine annotation:

Three persons are present; A man is smiling; An old man is present; Three men are present.

Hand annotation:1

Three men are attending a formal meeting in a news program; They are wearing formal suits; They are talking during the meeting; One man smiles during the meeting.

Hand Annotation 2:

There are several men having a meeting; They are all wearing dark suits; One of them is smiling; Two men are talking while other one is writing.

Figure 5.13: A meeting scene in the TV studio — seen in ‘20041031_200001.LBC.LBCNEWS.ARB02250253’ from the HLF extraction task.

machine generated annotation.

Grouping: (Figure 5.14). Main interest was to find out scenarios related to human interaction where scene settings for meeting scenes are not much obvious. Presence of humans, their emotion and gender helped in describing this scene by the automatic approach. Hand annotations focused on military dressing and were able to describe the theme, *i.e.*, army meeting.

Traffic scene: (Figure 5.15). Main interest was to find out scene settings which are more obvious than foreground objects and sceneries where foreground objects are missing altogether. Humans were absent in most of traffic video. Object detector was able to identify most prominent objects (*e.g.*, car, bus) for description. Hand annotations produced further details such as colours of car and other objects (*e.g.*, flyover, bridge). This sequence was also described as a highway.

Indoor/outdoor scene: (Figure 5.16). Main interest was to find out scene settings which are more obvious than foreground objects and scenarios where foreground objects are missing altogether. Humans were absent and objects movement was also minimal, which generated a description: ‘*this is a static scene*’. Green colour is identified as comprising of the major portion and, ultimately, it was described as an outdoor scene. Hand annotations were able to state specific objects such as pavilion, bushes, trees, fences and park. They also talked about the sky and the day scene.

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES



Machine annotation:

Many persons are present. Some of them are men. Persons are walking.

Hand annotation:1

Many people are standing; This is an indoor scene; Many people are wearing army uniforms; Many people are walking; This seems to be a meeting between higher military officers.

Hand Annotation 2:

A lot people are in a formal meeting.

Figure 5.14: A grouping scene comprising of multiple persons — seen in ‘20041105.110000-CCTV4_NEWS3_CHN’ from the HLF extraction task.



Machine annotation:

Many cars are present. Cars are moving. A bus is present.

Hand annotation:1

There is a red bus, one yellow and many other cars on the highway; This is a scene of daytime traffic; There is a blue road sign on the big tower; There is also a bridge on the road.

Hand Annotation 2:

There are many cars; There is a fly-over; Some buses are running on the fly-over; There is vehicle parapet; This is a traffic scene on a highway.

Figure 5.15: A traffic scene with many vehicles — seen in ‘20041101.110000-CCTV4_NEWS3_CHN’ from the HLF extraction task.

**Machine annotation:**

This is a static scene. There is greenery. This is an outdoor scene.

Hand annotation:1

There is a kiosk; There is a pavilion in the meadow, with the light blue sky as background; There is no movement and no humans.

Hand Annotation 2:

A pavilion is lying on the ground; There are a lot of bushes and trees; Fences are present around the pavilion; It is a day scene and looks like some park.

Figure 5.16: A park scene — seen in ‘CU497924_27062726’ from the HLF extraction task.

5.4.2 Evaluation with ROUGE

Difficulty in evaluating natural language descriptions stems from the fact that it is not a simple task to define the criteria. We adopted Recall-Oriented Understudy for Gisting Evaluation (ROUGE), widely used for evaluating automatic summarisation [Lin, 2004], to calculate the overlap between machine generated and hand annotations. However, there is larger inherent variability in generating sentences from videos than summarizing or translating a sentence from one language to another. This means ROUGE could penalize many correctly generated sentences, and be poorly correlated with human judgment of quality. Nevertheless we report ROUGE scores in the absence of any other automatic evaluation method that serves our needs perfectly.

Figure 5.17 shows the results where higher ROUGE score indicates closer match between them. For elaboration purposes, scores between hand annotations are also shown as the upper bound. In overall scores were not very high, demonstrating the fact that humans have different observations and interests while watching the same video. Descriptions were often subjective, dependent on one’s perception and understanding, that might have been affected by their educational and professional background, personal interests and experiences. Further, description length is different between machine generated and hand annotations and within hand annotations. Finally, there are many repetitive words in annotations which further reduce the ROUGE scores. Nevertheless ROUGE scores were not hopelessly low for machine generated descriptions; Closeup, Action and News videos had higher scores because of presence of humans with well defined actions and emotions. Indoor/Outdoor videos show the poorest results due to the limited capability of image processing techniques.

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

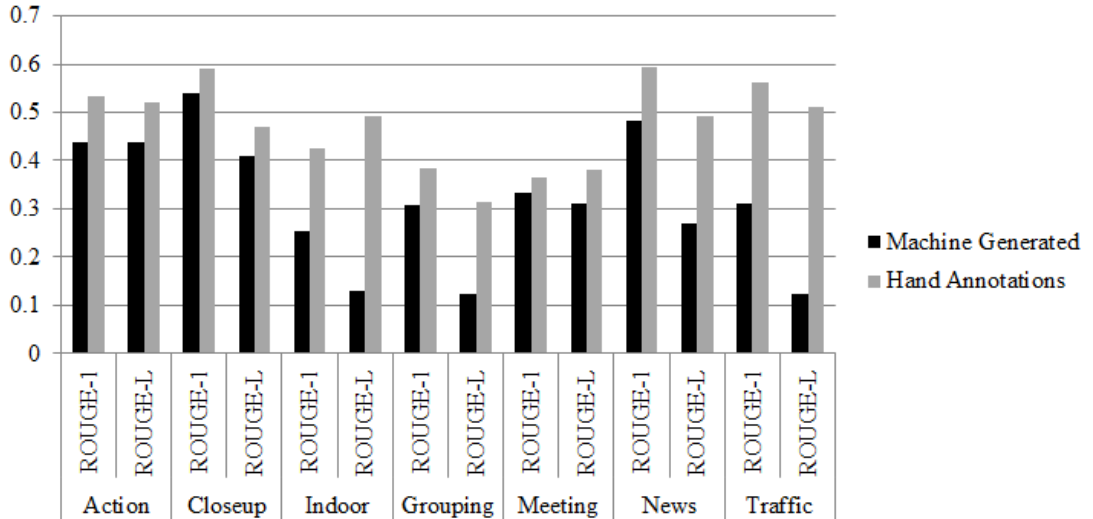


Figure 5.17: Comparison of ROUGE scores between machine generated descriptions and 13 hand annotations for calculating overlap in descriptions. ROUGE 1 shows overlap similarity using uni-grams while ROUGE-L is based on longest common subsequence.

5.4.3 Task Based Evaluation

Similar to human in the loop evaluation [Nwogu et al., 2011], a task based evaluation was performed to make qualitative evaluation of the generated descriptions. Each human subject was instructed to find a video that corresponded to a natural language description. Each subject was provided with one textual description and 20 video segments at one time. The same set of 20 video clips were repeatedly used, where four videos were selected from each of ‘Close-up’, ‘Action’, ‘News’, ‘Meeting’ and ‘Grouping’ category.¹ This resulted in a pool of candidates, consisting of clearly distinctive videos (between categories) and videos with subtle differences (within a single category). Once a choice was made, each subject was provided with the correct video stream and the following questionnaire:

question 1: how well the description explained the actual video, rating from ‘explained completely’, ‘satisfactorily’, ‘fairly’, ‘poorly’, or ‘does not explain’;

question 2: usefulness for including the following visual contents into descriptions, ratings from ‘most useful’, ‘very useful’, ‘useful’, ‘slightly useful’ to ‘not useful’:

- Usefulness of human structured related information: which further corresponds to information such as age, gender, emotion and expressions.
- Usefulness of human action related information: which includes human actions and movements information.

¹Since Traffic and Indoor/outdoor categories usually contains videos without humans, these two categories were not used in the evaluation task.

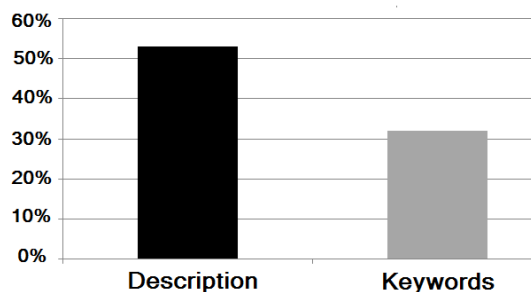


Figure 5.18: Correctly identified videos based on description and keywords alone.

- Usefulness of HLFs excluding humans: objects other than humans and any other motion that may be present in the video.
- Usefulness of background: which consists of scene settings, background, and color information.

Seven human subjects conducted this task searching a corresponding video for 10 descriptions, where two descriptions were selected from each of five categories *i.e.*, They did not involve creation of the dataset, hence they saw these videos for the first time. A baseline performance was measured by replacing a description with keywords. Keywords consisted of a complete set of HLFs that were used for deriving the natural language description. For fairness, subjects were provided with keywords and descriptions for different videos. This arrangement was needed because use of the same video for both keywords and descriptions almost always affected the performance when they saw the same video for the second time.

Figure 5.18 presents results for correctly identified videos for this task based evaluation. Description based retrieval performed better than the keyword baseline by a clear margin (roughly 20% absolute or more)¹, indicating that transforming keywords into more verbose descriptions was a valuable exercise. In the following we summarise the outcomes from the questionnaires which were filled after selection of the correct video.

How well the description explained the actual video. This question measures the scale for using natural language to describe videos. Figure 5.19(a) shows that roughly 50% of subjects responded that natural language descriptions provided satisfactory explanation (or better) of the video. Only 25% of subjects stated that keywords were satisfactory or better, and nearly 40% considered keywords explained the video poorly or worse.

Usefulness of human structure related information. As most videos in the dataset were related to humans, their emotions and activities, this question was very important. Much emphasis was placed on the effect of human structure related information such as age, gender, emotions (facial expressions) and body gestures. The question aimed to find the effect of human related descriptions for correct identification of videos. Figure 5.19(b) shows that more than

¹It is interesting to note that the correct identification rate went up to 70% for three subjects who also conducted creation of the dataset.

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

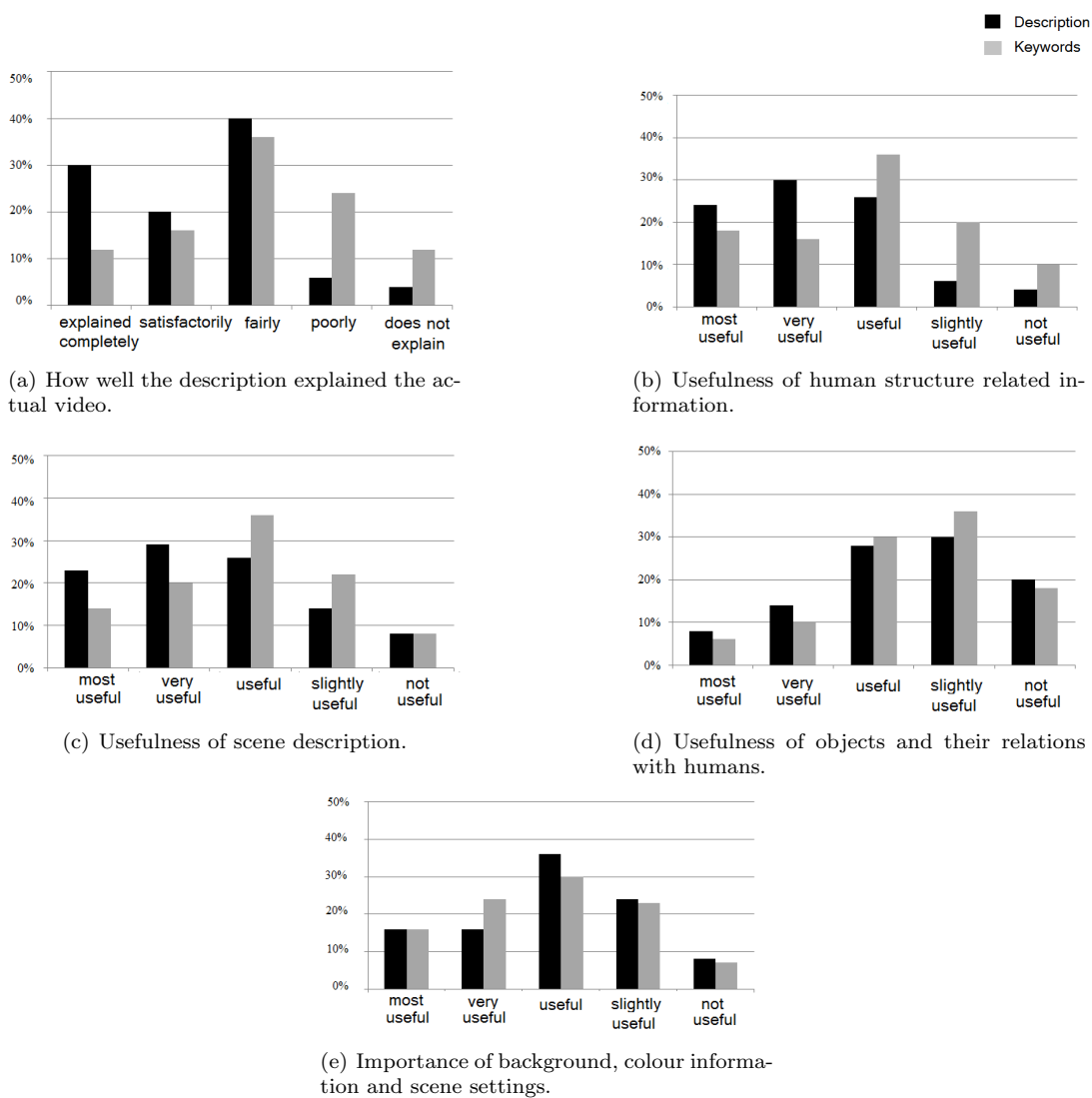


Figure 5.19: Outcomes from the questionnaire collected for the task based evaluation, comparing natural language descriptions and keywords based evaluation.

80% of subjects considered this information was useful, of which more than 50% said it was very useful. Even for keywords based evaluation, roughly 60% of subjects found this information useful or higher.

Usefulness of scene description. This questions finds the effect of scene description presented in a text form. The result was encouraging, as shown in Figure 5.19(c), since roughly 75% of subjects found scene description was useful for a video retrieval task, of which more than 50% found it very useful. On the other hand, roughly 30% subjects found it very useful for keywords alone evaluation. The number dropped to 40% for keyword based identification.

Usefulness of objects and their relations with humans. For better understanding of visual scene and its semantics, non-human objects play a very important role. As presented in Figure 5.19(d), the outcome for this question was not very encouraging; roughly 60% of participants were unable to find well formed explanation of objects and their relations with humans in descriptions generated although descriptions were still considered better than keywords. Overall, it is difficult to argue superiority of description over keywords for this question since both the evaluations are quite comparable for this question.

Importance of background, colour information and scene settings. Often, background and scene settings help in better understanding of the visual scene when humans and other objects are present. However humans and non-human objects are not always present (or failed to be identified all together by the image processing techniques). In such cases, it is particularly useful if we are able to identify visual scene setting and colour information of a video. This question measures the importance of background and scene setting information. Figure 5.19(e) shows that most subjects thought they were useful (or slightly so) but not too much. Similar results were obtained for keywords based evaluation. This might have been the consequence of the dataset biased towards videos with human centred contents.

Based on task based evaluation it can be concluded that ‘for identifying best matching video segments descriptions are much more useful than keywords alone’. Reasons for this usefulness include better presentation of information with clear semantic meaning and context information.

5.5 Summary

Initially this chapter started with the results of HLFs extraction task. Based on the outcome of this task, a list of HLFs was finalized for sentence generation. Then a template based approach using context free grammar was introduced for sentence generation. Hierarchal framework for sentence generation for dealing with cases of multiple humans was presented and discussed with example scenarios. Complete procedure for language description with sample video segments together with machine generated and two sets of hand annotations was provided. Quantitative evaluation using ROUGE scores and qualitative evaluation by task based evaluation was discussed. Although ROUGE scores are not satisfactory, still task based evaluation showed improvement due to description generation over keywords alone.

5. NATURAL LANGUAGE DESCRIPTIONS FOR VISUAL IMAGES

Although a complete description generation framework is presented in this chapter, still there are some issue which need to be addressed for making this framework generic in nature. Next chapter continues the discussion of this framework towards a generic framework which mainly deals with missing and erroneous information and quite stable for other video categories.

Chapter 6

Dealing with Missing and Erroneous Data

6.1 Introduction

Previous chapter presented a ‘framework for generating natural language descriptions for video sequences’. Information about human structure, actions, behavior and their relations with other objects was successfully generated using natural language syntax. The work started with implementation of conventional image processing techniques to extract high level features from individual frames. These features were converted into natural language descriptions using context free grammar. For evaluating that framework, seven video categories were selected containing 140 video segments in total, which were hand annotated by 13 human subjects. Comparing hand annotated and machine generated descriptions using ROUGE score provided quantitative evaluation, while a task based evaluation provided qualitative evaluation of the proposed framework.

This chapter further continues discussion of language description for dealing with noisy data and to accommodate other video categories in addition to already discussed seven categories. To this end, scenarios for dealing with missing and erroneous data to produce coherent descriptions for visual images are presented. This chapter also explains application of framework to different video genres. Tasks discussed in this chapter include:

1. This chapter is not aiming to improve currently available image processing outcomes, but idea is to prepare a framework that can accommodate erroneous / missing HLFs.
2. Previous chapter presented a framework for natural language description of visual images for seven video categories. This chapter further discusses application of that framework to handle different video categories.

Section 6.2 presents application scenarios of this research work which include (i) dependence of description on extracted HLFs (ii) dealing with missing data (iii) describing video sequences,

6. DEALING WITH MISSING AND ERRONEOUS DATA



Figure 6.1: Description depends on extracted high level features.

where human subjects are absent and (iv) discussing scenarios with erroneous extraction of HLFs. Finally Section 6.3 discusses scalability factor of this framework for additional video categories.

6.2 Application Scenarios

This section overviews different scenarios for application of the sentence generation framework. Initially, some discussion is provided to enlighten the effect of quantity of extracted HLFs for description generation. Using natural language processing methods for filling missing information is discussed next. Scenarios showing description generation with erroneous extracted HLFs are provided next. Finally, application of description framework in scenarios where humans are absent is discussed. Although syntactically and semantically correct sentences can be generated in all scenes, immaturity of image processing would cause some errors and missing information.

Description Depends on Extracted Visual Features. It is anticipated that the larger number of features leads to the higher quality in description. However it is not feasible to produce a full list of features within the current technologies. Further the processing time may become an issue when the feature size is very large. A balance between the quality of description and quantity of features needs to be explored. Figure 6.1 shows a video montage showing a woman walking in an outdoor scene. Based on the quantity of extracted HLFs, description generation will vary as follows.

- ‘*This is an outdoor scene.*’ (1 feature)
- ‘*There is a human in an outdoor scene.*’ (2 features)
- ‘*There is a woman in an outdoor scene.*’ (3)
- ‘*A woman is walking in an outdoor scene.*’ (4)
- ‘*A woman is walking while there is a motor bike in the background. This is an outdoor scene.*’ (6)
- ‘*A woman is walking while there are two humans in the background. This is an outdoor scene.*’ (7)
- ‘*A woman is walking while there is a man and a woman in the background. This is an outdoor scene.*’ (9)



Figure 6.2: Description depends on extracted high level features.

First HLFs related to each human are identified. Second, spatial relations between these humans is calculated. Finally, remaining objects in the visual image are considered for complete generation of the natural language description. It is worth mentioning that extraction of HLFs lead to different descriptions for the same visual scene, such as extracting only one HLF provides very basic information such as ‘this is an outdoor scene’, on the other hand extraction of 9 HLFs provides detailed description of the visual scene, such as ‘A woman is walking while there is a man and a woman in the background. This is an outdoor scene.’ Note that woman is considered as combination of two HLFs which are human and female. Similarly, two humans correspond to two HLFs.

Figure 6.2 presents another example where multiple humans are present in a dining scene. For this figure, generated description is as follows,

- ‘This is an indoor scene.’ (1 feature)
- ‘There are humans in an indoor scene.’ (2 features)
- ‘There are four humans in an indoor scene.’ (5)
- ‘There are two men one woman and one human in an indoor scene.’ (8)
- ‘Two men one woman and one human are sitting in an indoor scene.’ (12)
- ‘Two men one woman and one human are sitting and talking in an indoor scene.’ (16)
- ‘Two men one woman and one human are sitting and talking. There is one table and one chair in an indoor scene.’ (18)

Although grammatically correct description is produced for both the example scenarios, still quality of description improves with the addition of more HLFs, *i.e.*, from start to end in provided descriptions. On the other hand, hand annotations may still come up with more HLFs such as scene location as restaurant, dressing and emotions of humans involved and other objects such as portrait hanging on the wall, food plates on the table *etc.*

Missing HLFs. There would be many missing information due to limitedness of current image processing methods. For example, human gender was not detected in Figure 6.3(a).

6. DEALING WITH MISSING AND ERRONEOUS DATA

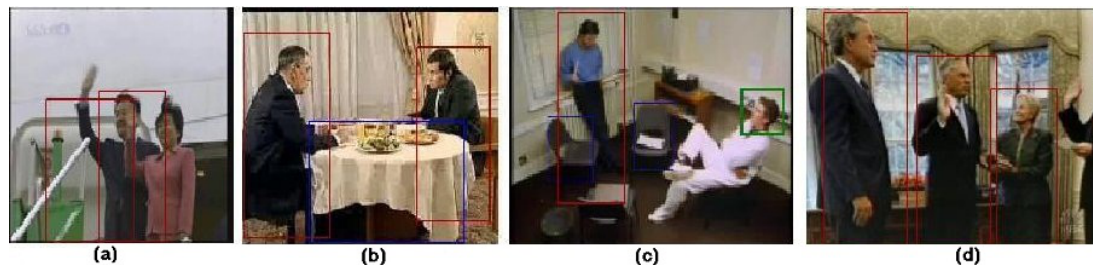


Figure 6.3: Generated description: (a) *subject + subject + verb*: ‘humans are waving hands’; (b) *subject + subject + object*: ‘two persons around the table’; (c) *subject + verb, noun phrase / subject, noun phrase / subject*: ‘a man is standing; a person is present; there are two chairs’; (d) *subject + subject + subject + verb*: ‘multiple persons are present’.

Human action (‘*sitting*’) was not identified in Figure 6.3(b). Further, detection of food on the table might have led to more semantic description of the scene (*e.g.*, ‘*dinning scene*’). In 6.3(c), gender of the second person and his action with reference to chair was missing altogether. In 6.3(d), fourth human and actions by two humans (‘*raising hands*’) were not extracted. In current work, there is no remedy for this shortcoming, *i.e.*, descriptions are generated for the extracted HLFs. One approach to include this missing information can be based on using ‘*natural language processing methods such as language modeling or parsing scores with and without missing words*’. Currently this idea is not fully implemented for this work and need to be addressed in future work.

Non human subjects. Suppose a human is absent, or failed to be extracted, the scene is explained on the basis of objects. They are treated as subjects for which sentences are generated. Firstly, algorithm tries to find out objects in the given scene. Secondly, movement information is extracted from the sequence. There can be three scenarios fulfilling these two conditions. (i) both of them are identified, then description is generated by combining both these features (Figure 6.4(a)). (ii) One of them is detected, such as Figure 6.4(c), description is generated based on identified HLF. (iii) none of them is identified: in such scenarios algorithm tries to find scene settings such as indoor or outdoor. Further information about ‘movement’ is included for better explanation the visual scene, *e.g.*, Figures 6.4(b) and 6.4(d).

Errors in HLF extraction. Several erroneous HLFs are extracted due to shortcomings of image processing methods. In Figure 6.5(c), one person was found correctly but the other was erroneously identified as female. Description generated was ‘*a smiling adult man is present with a woman*’. Detected emotion was ‘*smile*’ in 6.5(d) though real emotion was ‘*serious*’. Description generated was ‘*a man is smiling*’.

6.3 Scalability of Work for other Video Categories

This section provides discusses evaluation of language description framework of NLDV - Corpus 3 (section 3.4, Chapter 3). Natural language descriptions of those video segments were auto-



Figure 6.4: generated description: (a) ‘multiple cars are moving’; (b) ‘This is an outdoor scene. This is a static scene.’; (c) ‘There is a chair in an indoor scene.’; (d) ‘This is an outdoor scene. There is some movement there.’.



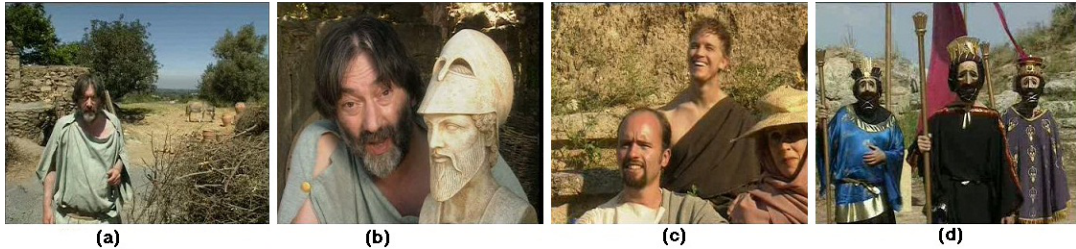
Figure 6.5: Image processing can be erroneous: (a) only three cars are identified although there are many vehicles prominent, (b) five persons (in red rectangles) are detected although four are present; (c) one male is identified correctly, other male is identified as ‘female’; (d) detected emotion is ‘smiling’ though he shows a serious face.

6. DEALING WITH MISSING AND ERRONEOUS DATA

matically generated by our framework. We followed the same steps as discussed in Chapter 5 for description generation. Initially high level features (HLFs) are identified in individual video frames. Then spatial relations between HLFs were calculated. Extracted HLFs and spatial relations between them were then presented by syntactically and semantically correct expressions using context free grammar. Five human subjects manually annotated these video segments with natural language descriptions. Overlap between human annotated and machine generated descriptions was calculated using ROUGE score and compared against the scores of previous seven categories, which presents quantitative evaluation of this description generation task. Similar to task based evaluation of Chapter 4, a task based evaluation was performed by human subjects, providing qualitative evaluation of generated descriptions.

Costume Videos. These are the video sequences which contain costumes, sets and properties in order to capture the ambience of a particular era such as ‘Roman, Egyptian or old Chinese civilizations *etc.*’ Figure 6.6 presents four images together with machine generated description and two sets of hand annotations where each image is taken from one of the videos in the costume category. These video segments were selected ranging from simple to very difficult for high level feature extraction task, *i.e.*, first segment only contains one human where complete body and face was obvious. Second segment contains one human and one human’s face like sculpture, third segment contains three humans and finally fourth segment although contains three humans, but they are wearing masks.

Related to HLF extraction, human detection was successfully performed in all the four video segments except figure 6.6(b), where sculpture was wrongly identified as a human. Human emotion such as ‘serious and smiling’ was also identified in first three segments. Gender identification had poor results, *i.e.*, in figure 6.6(a) and figure 6.6(b), human was detected as a woman although a man was present, while in figure 6.6(c), only one man was identified correctly whereas for other two humans no information related to gender was achieved. Since, humans in these videos segments had special facial appearances such as beard, French cut, wearing hats *etc.*, human’s age identification has very poor results. Although, figure 6.6(d) showed human walking scene, but again this action was not identified due to special appearance and clothing of humans *i.e.*, wearing too broad clothes and holding big sticks in their hands. Humans and their dressing was most important feature in hand annotations. All the annotators paid full attention to human’s dressing (Greek or Roman dress), hats and facial masks. Facial features such as beard and French beard were also annotated. Explaining individual humans and their actions was also done by most of the annotators such as ‘one man with French beard, woman with a hat and a smiling man in the background’ in figure 6.6(c). Almost all the annotators were able to differentiate between real humans and human sculpture. Finally, annotators also noticed other objects that were present in those segments such as ‘sticks, donkey, stones and trees *etc.*’ Although, our framework was able to generate descriptions for each of these video segments, still they were quite different from hand annotations. For such category human dressing becomes a mandatory HLF which need to be added in feature extraction task. Also performance of HLF extraction tasks related to human’s age, gender and emotion are not satisfactory for such videos due to diversity in human’s appearances. Finding scene settings such as ‘village or



- (a) A man in a roman costume:**
(MG:) A serious woman is present in an outdoor scene.
(HA 1:) An old man wearing Greek garments is walking in a village setting. A donkey is eating grass in the background. There are trees, bushes and an old wall in the background.
(HA 2:) A man with a beard in roman clothes is facing camera and speaking. There is a donkey, trees and stones in the scene.
- (b) A man with a sculpture:**
(MG:) A serious woman and a human are present. This is an indoor scene.
HA 1: An old man wearing Greek garments is present with a sculpture.
HA 2: A man with a beard in roman clothes is sitting with a sculpture. he is facing camera and speaking. He is pointing towards the sculpture and explaining something about the sculpture.
- (c) Three humans in roman costumes:**
(MG:) Three humans are present. One man is smiling. Two humans are present while a man is in the background. This is an outdoor scene.
(HA 1:) Two men and one woman wearing Greek garments are show. They look happy and looking at something.
(HA 2:) Two men and one woman are present. One men has French beard and looks worried. Woman is wearing a hat and looks serious. Man standing behind is smiling and looks happy.
- (d) Three humans with masks and roman costumes:**
(MG:) There are three humans. This is an indoor scene.
(HA 1:) Three human wearing Greek garments are walking.
(HA 2:) Three men in roman clothes are wearing masks and holding big and heavy sticks. It looks like old army parade scene.

Figure 6.6: Costume videos: (a) is seen in video ‘MRS157445’ from the 2008 BBC rushes videos. (b) is seen in ‘MRS157446’ from the 2008 BBC rushes videos. (c) is seen in ‘MRS157475’ from the 2008 BBC rushes videos. (d) is seen in ‘MRS157475’ from the 2008 BBC rushes videos.

army parade scene’ is still far away.

Crowd videos. Grouping and meeting videos are quite structured videos containing less number of humans, where there is proper sequence of events and easy to comprehend, on the other hand crowd videos usually contain large number of people and its usually difficult to differentiate between individual activities of humans. Figure 6.7 presents four images together with machine generated description and two sets of hand annotations where each image is taken from one of the videos in the crowd category.

Machine generated descriptions for such videos are based on structure and activities of individual humans. Information related to individual humans was successfully obtained in some of the videos. Although, for all the video segments, ‘presence of multiple humans’ was successfully recognized, still interaction or relationship between different humans was not identified . For some videos human actions were also identified such as ‘standing and waving hands’ in figure

6. DEALING WITH MISSING AND ERRONEOUS DATA



(a) Humans in suits in a street:

(MG:) Many humans are present. This is an outdoor scene. Humans are standing. Humans are waving hands.

(HA 1:) Many men wearing hats are shown.

(HA 2:) One man is wearing a police uniform is standing in front. In background, men wearing suits and hats are standing. All of them are carrying sticks and moving their hands.

(b) Election campaign:

(MG:) Many humans are present. This is an outdoor scene. Humans are waving hands.

(HA 1:) Several happy humans are waving hands, while some are clapping in the background.

(HA 2:) This is a scene of election campaign in US elections. US president Bill Clinton is waving hands to the crowd while other humans are clapping and waving hands.

(c) Humans attending the lecture:

(MG:) Many humans are present. This is an outdoor scene.

(HA 1:) Several women are sitting and listening to a speech.

(HA 2:) This seems to be religious lecture scene. Women wearing veils are sitting on chairs and listening to someone.

(d) Humans in the procession:

(MG:) Many humans are present. This is an outdoor scene.

(HA 1:) Many humans are walking in a street and shouting.

(HA 2:) this is a procession scene, where many humans are present on a road and raising slogans.

Figure 6.7: Costume videos: (a) is seen in ‘MRS025913’ from the 2007 BBC rushes videos. (b) is seen in ‘20041031_133000_MSNBC_MSNBCNEWS13_ENG’ from the 2004 HLF extraction task. (c) is seen in ‘20041101_190001_NTDTV_NTDNEWS19_CHN’ from the 2004 HLF extraction task. (d) is seen in ‘20041101_190001_NTDTV_NTDNEWS19_CHN’ from the 2004 HLF extraction task.

6.7(b) and 6.7(c). Since, humans faces capture very small portion of the complete images and not much clear, HLFs such as human’s age, gender and emotion were not recognized. Based on these HLFs, descriptions are generated where humans are treated as a group such as ‘multiple humans’.

Again, almost all hand annotators paid much attention to humans and their dressing information such as ‘wearing veils and hats *etc.*’ Annotators were able to identify humans gender, emotion and action information such as ‘men, women, happy humans, clapping and walking *etc.*’ Human’s identity also becomes a unique feature for such videos such as ‘police man, Bill Clinton *etc.*’ Presence of other HLFs such as ‘chairs, sticks and road’ was also noticed by annotators. Finally, overall semantic of the scene such as ‘election campaign video, procession scene and religious lecture’ was also noticed by annotators.

For such video categories, our framework was able to generate some useful descriptions which are also meaningful. On the other hand, HLFs extraction methods were not satisfactory for these video segments due to presence of large number of humans. Since our framework is



<p>(a) Lawn tennis scene: (MG:) A woman is waving hands while there are other humans in the background. (HA 1:) A tennis player is doing service during a tennis match. (HA 2:) A long haired man is playing lawn tennis. He is doing service with left hand. he is wearing dark clothes. People are sitting and cheering this game.</p> <p>(b) Golf scene: (MG:) A human is present. This is an outdoor scene. This is a static scene. (HA 1:) A man is playing golf. (HA 2:) This is a scene of golf ground. Two persons are shown then one tries to pocket the ball. He is wearing a cap and old man.</p> <p>(c) Lawn tennis scene containing two persons: (MG:) A human is present. This is an outdoor scene. There is some movement. (HA 1:) Two players are playing lawn tennis. (HA 2:) Two men are playing tennis. One man serves and the other picks. The person in the dark nicker wins in the end. People are clapping.</p> <p>(d) Ice Hockey Scene: (MG:) This is an outdoor scene. There is some movement. (HA 1:) Three players are playing ice hockey. (HA 2:) This is a ball tackling scene of ice hockey. One man wearing light shirt falls down while tackling two other player.</p>
--

Figure 6.8: Sports videos: (a) is seen in ‘20041103_190000_NTDTV_NTDNEWS19.CHN’ from the 2004 HLF extraction task. (b) is seen in ‘20041115_190001_NTDTV_NTDNEWS19.CHN’ from the 2004 HLF extraction task. (c) is seen in ‘20041101_120001_NTDTV_NTDNEWS12.CHN’ from the 2004 HLF extraction task. (d) is seen in ‘MRS308363’ from the 2007 BBC rushes videos of TRECVID.

designed for extracting useful information about single individuals and later combining them for multiple humans but crowd videos usually contain humans which are not easily differentiable. Another direction which needs to be explored is related to semantic scene identification such as election campaign or religious lecture scene which at current times is in its infancy stages.

Sports Videos. Sports videos usually contain very fast human actions and video specific objects such as goal posts, cricket bats, hockey can help in better explaining these videos. Figure 6.8 presents four images together with machine generated description and two sets of hand annotations where each image is taken from one of the videos in the sports category. Two scenes show segments from lawn tennis games and one for golf and one for ice hockey games.

Presence of at least one human was successfully achieved for first three segments shown by figures 6.8(a) to 6.8(c). Since human structure or face is not quite obvious in 6.8(d), human’s presence was not detected in that segment. Although, for 6.8(a), gender was detected as ‘female’ but it was not quite right since a human could easily judge it as a man. Also for 6.8(a) human action *i.e.* ‘waving hands’ was identified which might be similar to ‘service action’ in lawn tennis

6. DEALING WITH MISSING AND ERRONEOUS DATA

game. On the other hand for all other segments, no information related to human's age, gender, emotion or action was obtained. As for previous categories, information about scene settings and movement information was added to further elaborate the visual scene.

Annotators usually picked the scene settings which were specific to games scenes quite easily. All the annotators were able to categorize these segments into their respective games such as 'tennis, golf and ice hockey.' Interestingly, not all the annotators explained the scenes in more detail such as 'service scene of tennis, ball tackling scene in hockey'. One reason for this explanation could be their knowledge for the relevant game, for example, 'ball tackling' might not be a commonly known scenario. Annotators mostly noticed number of humans involved in the sports scene such as 'two players for lawn tennis' and 'three players for ice hockey'. Sometimes, background information was also described such as 'people clapping and cheering' in the background. Again, dressing information was also listed such as 'wearing caps and dark clothes.'

For generating descriptions for such videos, specific information for sports videos need to be extracted. This information can usually be achieved from scene settings such as lawn tennis court, ice hockey ground, golf field or dressing of humans, *i.e.*, different games require their players to wear different clothes and outfits. Although language description framework was able to generate descriptions for such video segments, some extra features need to be extracted for complete semantic annotation of these segments.

Violence Videos. Violence videos are usually based on scenes showing damages, gun shots and army related equipments such as tanks, cannons, troops *etc.* These videos are a common inclusion in news videos these days due to heightened sense of alertness caused by political instability across the globe. Figure 6.9 presents four images together with machine generated description and two sets of hand annotations where each image is taken from one of the videos in the violence category.

Presence of humans was successfully identified in 6.9(a) and 6.9(b). For 6.9(a), human gender was incorrectly detected as 'woman' due to presence of army cap and relatively smaller portion of face. For 6.9(b), human in the foreground and backgrounds were successfully identified. In 6.9(c), human detection step failed to identify any human due to side pose of human and very dark image quality. Interestingly, for the video segment shown in 6.9(d), two cars were detected correctly, which even out performed human judgements, since no human annotator judged that there were two cars. Again for 6.9(c) and 6.9(d), scene settings and movement in the video segment was successfully identified.

For hand annotations, identity of humans was the most interesting feature such as soldier and US soldier. War related equipments such as guns, tanks, trollers were also noticed by the annotators. Specific HLFs such as fire, burning which are related to violence videos were also described. Scene settings such as battlefield are also mentioned. Finally, annotators described such scenes with words which have semantic meaning such as war scene, terrorism scene *etc.*

Although, descriptions were generated for video segments of this category, but they were missing information about the semantics of these sequences. For generating descriptions for such videos, information related to human's dressing such as uniform, hats *etc.*, specific objects



- (a) **Soldier smoking a cigarette:**
 (MG:) A woman is present in an indoor scene.
 (HA 1:) A soldier is smoking cigarette and speaking.
 (HA 2:) A soldier is lighting cigarette and then smoking. There are other soldiers behind him.
- (b) **Soldier carrying a gun while tanks in background:**
 (MG:) A human is standing while there is human in the background. This is an outdoor scene.
 HA 1: A soldier with a gun is standing while a bunker is moving in the back drop.
 HA 2: A US soldier is standing with gun in a battlefield. There are two other soldiers in his background. There is one tanker and one troller in the background.
- (c) **Person with gun shots:**
 (MG:) This is an outdoor scene. There is some movement.
 (HA 1:) A man is firing with a gun.
 (HA 2:) This seems to be a terrorism scene. A man is firing, then another man comes for his help and both start firing.
- (d) **Burning jeep:**
 (MG:) There are two cars. This is an outdoor scene. There is some movement.
 (HA 1:) A jeep is burning.
 (HA 2:) A jeep has caught fire and collided with a tree and burning.

Figure 6.9: Violence related videos: (a) is seen in ‘20041116_220001_CNN_AARONBROWN_ENG’ from the 2004 HLF extraction task. (b) is seen in ‘20041106_200000_LBC_LBCNEWS_ARB’ from the 2004 HLF extraction task. (c) is seen in ‘20041117_133000_MSNBC_MSNBCNEWS13_ENG’ from the 2004 HLF extraction task. (d) is seen in ‘20041116_150000_CNN_LIVEFROM_ENG’ from the 2004 HLF extraction task.

such as tanks, guns *etc.*, and color information such as fire and smoke color need to be extracted.

Animal Videos. Videos showing shots of animals and their activities are quite common. Figure 6.10 presents four images together with machine generated description and two sets of hand annotations where each image is taken from one of the videos in the animal category. Note each of these video segments are quite different in contents such as Figure 6.10(a) shows a ‘tiger on grass’, while Figure 6.10(b) shows a ‘fish in water’.¹

Hand annotations mainly focussed on the main object such as tiger, whale fish, dog and sheep present in the segments. Structure *i.e.*, outlook of objects such as ‘bengal tiger’, activities such as ‘moving, dancing’, emotions such as ‘angry’ related to objects were also given importance by the annotators. Mostly annotators paid much attention to the location of the object such as ‘in the park, water, under the tree and on the road.’ Objects in the background were also

¹With current image processing methodologies, it is still impossible to develop an object recognizer which can recognize all the objects.

6. DEALING WITH MISSING AND ERRONEOUS DATA



(a) Tiger in a fence scene:

(MG:) This is an indoor scene. There is some movement there. (HA 1:) A bengal tiger is shown in a park. It is moving and looks angry. (HA 2:) A tiger is set free in a park, but it is inside the fence. Tiger is moving around.

(b) Whale in water:

(MG:) This is an outdoor scene. There is some movement there. HA 1: A whale is dancing in the water. Some other whales are also dancing. HA 2: A big fish is going up and down in the water. There are many other fishes shown.

(c) Dog under the trees:

(MG:) This is an outdoor scene. There is a static scene. (HA 1:) two men and two women are sitting on the tree. A dog is barking under the tree. (HA 2:) Four persons are afraid of a dog and climbed on a tree, while the dog is waiting for them down the tree.

(d) Cattle scene:

(MG:) This is an outdoor scene. There is some movement there. (HA 1:) Many sheep are wondering on a road. (HA 2:) This is a cattle scene. Some goats and sheep are present on a high mountain, which is very barren.

Figure 6.10: Animal videos: (a) is seen in ‘20041101_120001_NTDTV_NTDNEWS12.CHN’ from the 2004 HLF extraction task of TRECVID. (b) is seen in ‘20041104_220001_CNN_AARONBROWN_ENG’ from the 2004 HLF extraction task of TRECVID. (c) is seen in ‘MRS134704’ from the 2007 BBC rushes videos of TRECVID. (d) is seen in ‘MRS157454’ from the 2007 BBC rushes videos of TRECVID.

mentioned such as ‘high mountains and fence’. While mostly annotations present objects which are more obvious, still it is interesting to note that, in presence of humans, descriptions related to humans become more important, (Figure 6.10(c)), *i.e.*, annotators first described human activities such as ‘climbing on the tree, afraid of dog *etc.*’, and then generated description about other objects and scene settings.

For machine generated descriptions, since humans were absent in most of these videos, other objects such as car, cup, table, chair, bicycle and TV-monitor were looked for. In absence of these objects, scene settings and movement detection was performed. Scene setting was successfully identified as ‘outdoor scene’ in three of the video segments except for Figure 6.10(a), where light quality was not so good and object captures most of the the frame size. On the other hand motion was successfully detected in all the segments.¹

Although, some meaningful descriptions were generated for these video segments, they were far lacking in contents as compared to hand annotations. First, there was no information about specific objects (animals) such as tiger, fish *etc.* Anyhow, this deficiency could be removed by adding more object recognizers. Second, human subject always provided specific scene settings

¹In Figure 6.10(c), dog is barking but there is no movement and it was successfully described as a static scene.

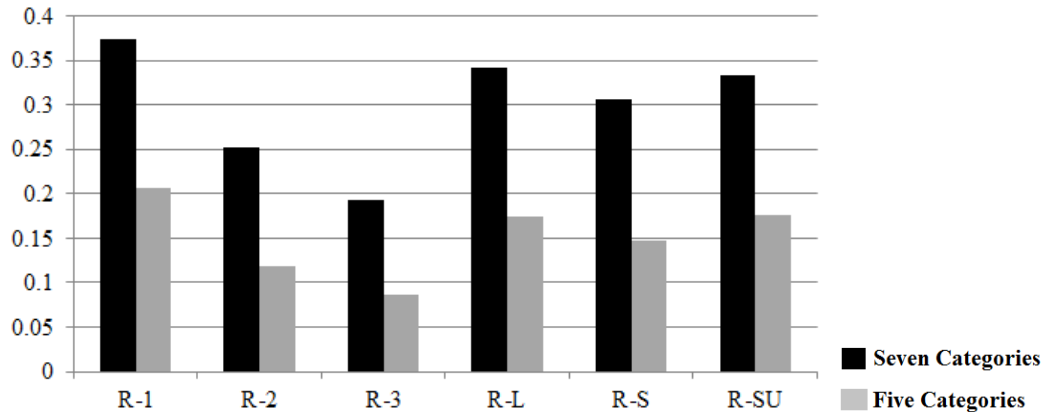


Figure 6.11: Comparison of ROUGE scores for calculating overlap in hand annotations and machine generated descriptions for seven (section 3.2) and five categories (section 3.4) respectively. ROUGE 1-3 shows n -gram overlap. ROUGE-L is based on longest common subsequence. ROUGE-S skips bigram co-occurrence without gap length. ROUGE-SU shows results when skipping bigram co-occurrence with unigrams.

such as ‘in park, in water *etc.*’, but machine generated descriptions could only provide general scene settings which are either indoor or outdoor which are very broad categories. Again, enhancing high level feature extraction task to include these scene locations can come to remedy this shortcoming. On the other hand, both descriptions are able to find movement information in the video segments. Finally, it is worth mentioning that machine generated descriptions are syntactically and semantically correct and algorithm did not fail to generate descriptions, though these descriptions are at very basic level *i.e.*, without much useful contents.

6.4 Evaluation of Generated Descriptions

For evaluating machine generated descriptions for five categories, same evaluation strategy of Chapter 4 was applied. Initially ROUGE scores between machine generated and hand annotations were calculated. Then a task based evaluation was performed to find the matching video given the textual description.

6.4.1 Evaluation with ROUGE

To find the overlap between machine generated and hand annotations¹, ROUGE score was calculated for five video categories. Figure 6.11 shows comparison of ROUGE scores between five categories against the previous seven categories. For making comparison, average of all scores were taken for each of the ROUGE measure, *i.e.*, ROUGE-1 to ROUGE-SU. Roughly, ROUGE scores for these five categories are half of the scores for seven categories.

¹five hand annotations for each of the 20 videos

6.4.2 Task based Evaluation

For qualitative analysis of generated descriptions for these four categories, a task based evaluation similar to Chapter 4 was performed. Each human subject was instructed to find a video that corresponded to a natural language description. Each subject was provided with one textual description and 20 video segments at one time. The same set of 20 video clips were repeatedly used, consisting of clearly distinctive videos (between categories) and videos with subtle differences (within a single category). Once a choice was made, each subject was provided with the correct video stream and the same questionnaire as presented in the Section 1.5.3. Five human subjects¹ conducted this task searching a corresponding video for 10 descriptions, where two machine generated descriptions were selected from each of five categories.

Figure 6.12 presents results for task based evaluation for five categories against the previous seven categories. Almost 25% of times as compared to 53% for seven categories, human subjects were able to find the correct matching video against the given description. Answers for other questionnaires are much lower in comparison to seven categories evaluation. About 43% subjects agreed that descriptions were able to explain the video sequences clearly. Same number of subjects found human structure information useful for the identification of video sequence. Objects were not found useful for this video matching task. Similarly, information related to scene description and background was not found much useful.

6.5 Discussion of Framework Shortcomings

Although our framework was able to generate syntactically and semantically correct descriptions for all the discussed scenarios, still there are several shortcomings. Following are some of the areas which need attention to improve the quality of generated descriptions.

- Human face related features such as age, gender and emotion are mostly view point specific, *i.e.*, they are identifiable in scenarios where complete face with frontal view is quite obvious. There is need of improvement for side views of faces.
- Human body related features such as actions like walking, standing, sitting *etc.*, and gestures are usually dependent on the availability of body structure. For example, walking and running can be differentiated if complete human body is obvious.
- Finally, it is always good to prepare HLFs for particular category *e.g.*, tiger, fish and dog *etc.*, for animal videos, but it is difficult to achieve.
- Scene settings which can help to describe video categories such as tennis court, golf ground for sports videos.
- Human dressing is an interesting feature which covers various video categories and need to be extracted.

¹They did not involve creation of the dataset, hence they saw these videos for the first time.

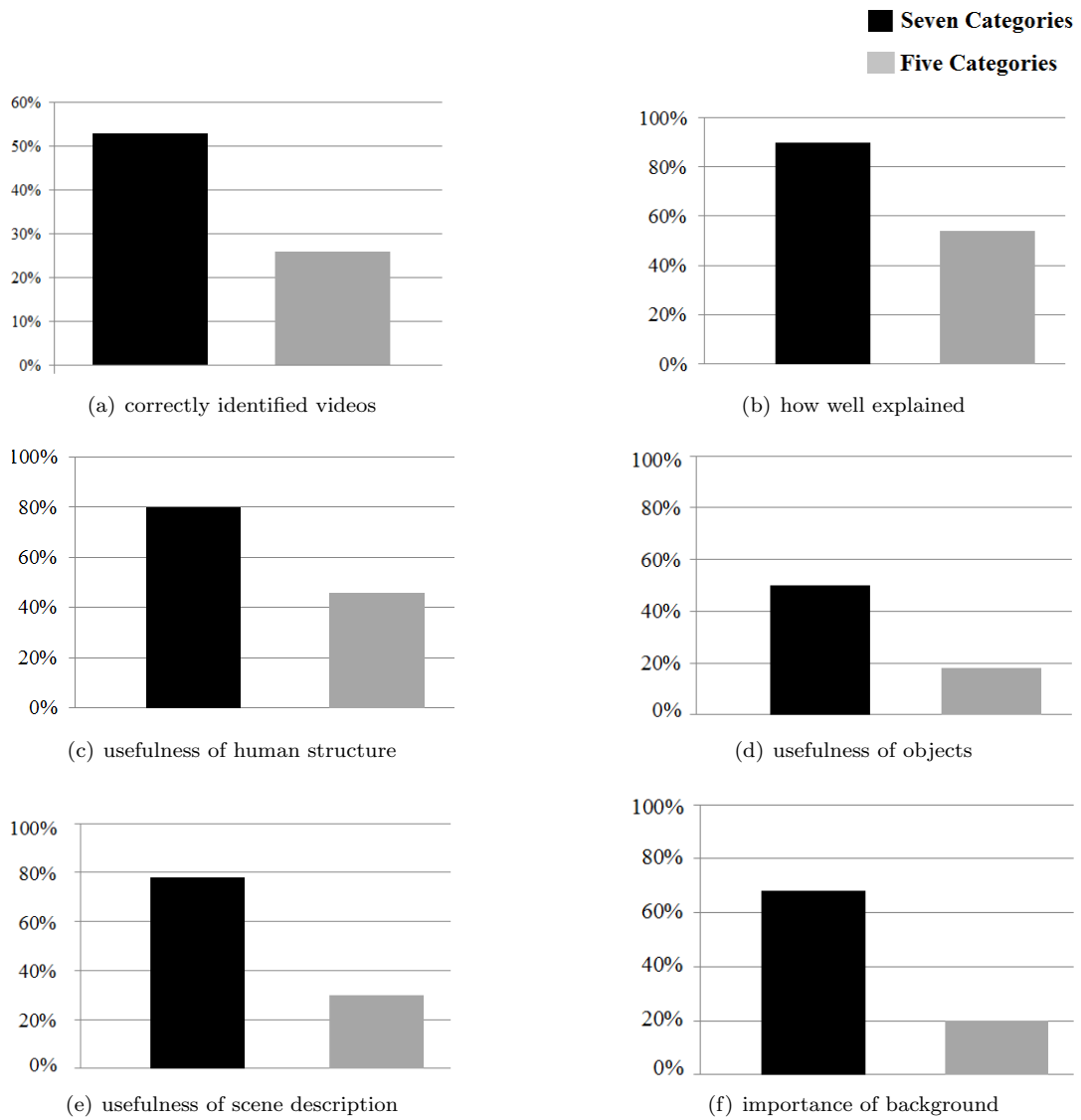


Figure 6.12: Outcomes from the questionnaire collected in Task 1 for the task based evaluation, comparing natural language descriptions and keywords based evaluation.

- There is still no semantic information involved such as ‘election campaign scene, or religious lecture’.

6.6 Summary

This chapter presented discussion about extending the framework of natural language descriptions for visual images. One extension was related to dealing with shortcomings of high level feature extraction task, while other discussed application of framework to unseen video categories. Four application scenarios were discussed which resulted due to limitedness of HLFs extraction task. They include, dependence of generated description on the quantity of HLFs, dealing with missing HLFs, handling errorless HLFs, and describing videos with no human subjects. Other extension explained description generation for five unseen categories, *i.e.*, animal, costume, crowd, sports and violence videos. ROUGE scores between hand annotations of these five categories and machine generated descriptions depicted that our framework was able to generate meaningful descriptions.

Previous chapter and this chapter discussed language descriptions for individual frames. Next chapter continues the discussion for complete video sequences which may consist of several visual frames. The problem is that a frame based generation procedure results in many identical descriptions produced from adjacent frames. Hence simple concatenation of descriptions may lead to redundancy, lacking coherency. Further, there is no temporal information attached with frames which is necessary for full description of video sequences.

Chapter 7

Natural Language Descriptions for Video Streams

7.1 Introduction

Previous two chapters were concerned with creation of sentence-length description for major objects and events in a video frame. This chapter further extends these frame based descriptions for complete video streams. When describing a video sequence, simply joining frame based descriptions will have several shortcomings. Descriptions of individual frames are crude, repeated and in some cases missing useful information due to sparseness of HLFs that can be produced by current technologies. Image processing errors can be accumulated. Lack of temporal information may cause a further problem. In this section we first introduce a ‘unit’, aiming to create smooth and coherent descriptions while alleviating the effects of these shortcomings. By structuring a video sequence based on units, we are able to remove redundancy caused by repeated expressions and to accommodate temporal information into a description. We further explore an approach to paraphrasing unit based descriptions aiming at creating compact and coherent natural language. Temporal information is also incorporated during this process.

7.2 Identifying Units for Description

Definition. We consider a sequence of video frames from which some HLFs (*e.g.*, human, objects or their moves) can be identified. For example in a scene where a man walks out from a room after desk work, the following actions may be identified: ‘*sitting*’ (in front of a desk), ‘*standing up*’ (from a chair), and ‘*exiting*’ (from a room), each of which can span over multiple frames. It may also contain another identifiable features such as facial expressions (*e.g.*, serious in the beginning and smiling later) and some objects in the background. In this work, we refer to a sequence of frames with an identical set of visual HLFs as a unit. Using this definition, the length of individual units may be affected by the availability, as well as the quality, of HLF

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS

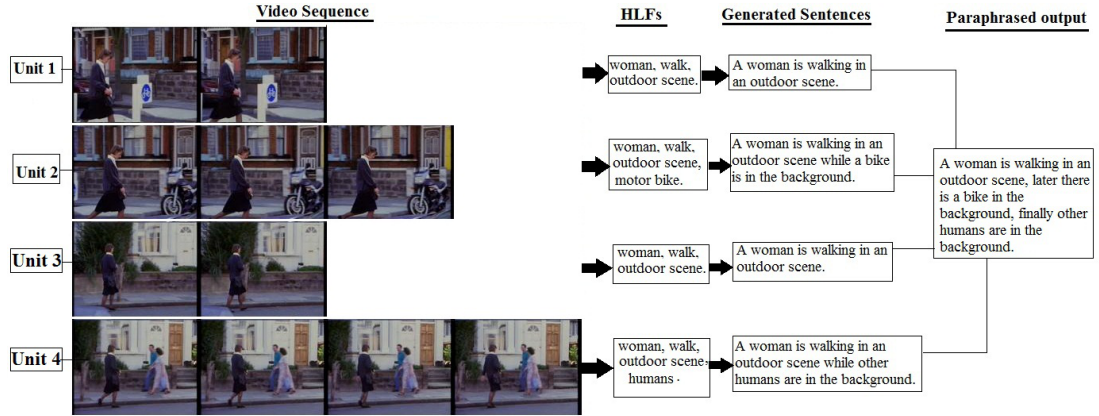


Figure 7.1: Four frame sequences extracted from a scene, ‘a woman walking on a road’ — seen in video ‘MS212890’ from the 2007 TREC Video rushes summarisation task. Each row represents a single unit.

extraction techniques.

Examples. Each row in Figure 7.1 represents a distinct unit, consisting of a frame sequence of variable length, extracted from a single scene where a human is walking on a road. We assume that image processing techniques resulted in an identical set of visual HLFs for frames in each row. A set of HLFs (for each row of Figure 7.1) could lead to each line of the following expressions:

a woman is walking;
a woman is walking while a bike is in the background;
a woman is walking;
a woman is walking while other humans are in the background.

A full description of the scene can be derived from the above four lines. However a simple concatenation of the four is crude because the same statement (*i.e.*, ‘a woman is walking’) is repeated, hence paraphrasing technique may be explored. Further, consideration of temporal information means that, as soon as a woman and her action is identified, this particular expression (‘a woman is walking’) is no longer required in the rest (until some change happens). The better description of the scene in Figure 7.1 may be

a woman is walking; then a bike is in the background; later other humans are in the background.

where ‘then’ and ‘later’ indicate the order of occurrences¹.

Figure 7.1 further elaborates the concept of unit in videos. Each row presents a single feature unit where HLFs remain constant for a fixed number of frames. Natural language description is created for each unit of frame sequence (*i.e.*, each row in the figure). The next step is to derive

¹ One of the hand annotation for this video clip is as follows: ‘A woman appears from left. She is walking while a bike in the background. Later she comes across other humans.’

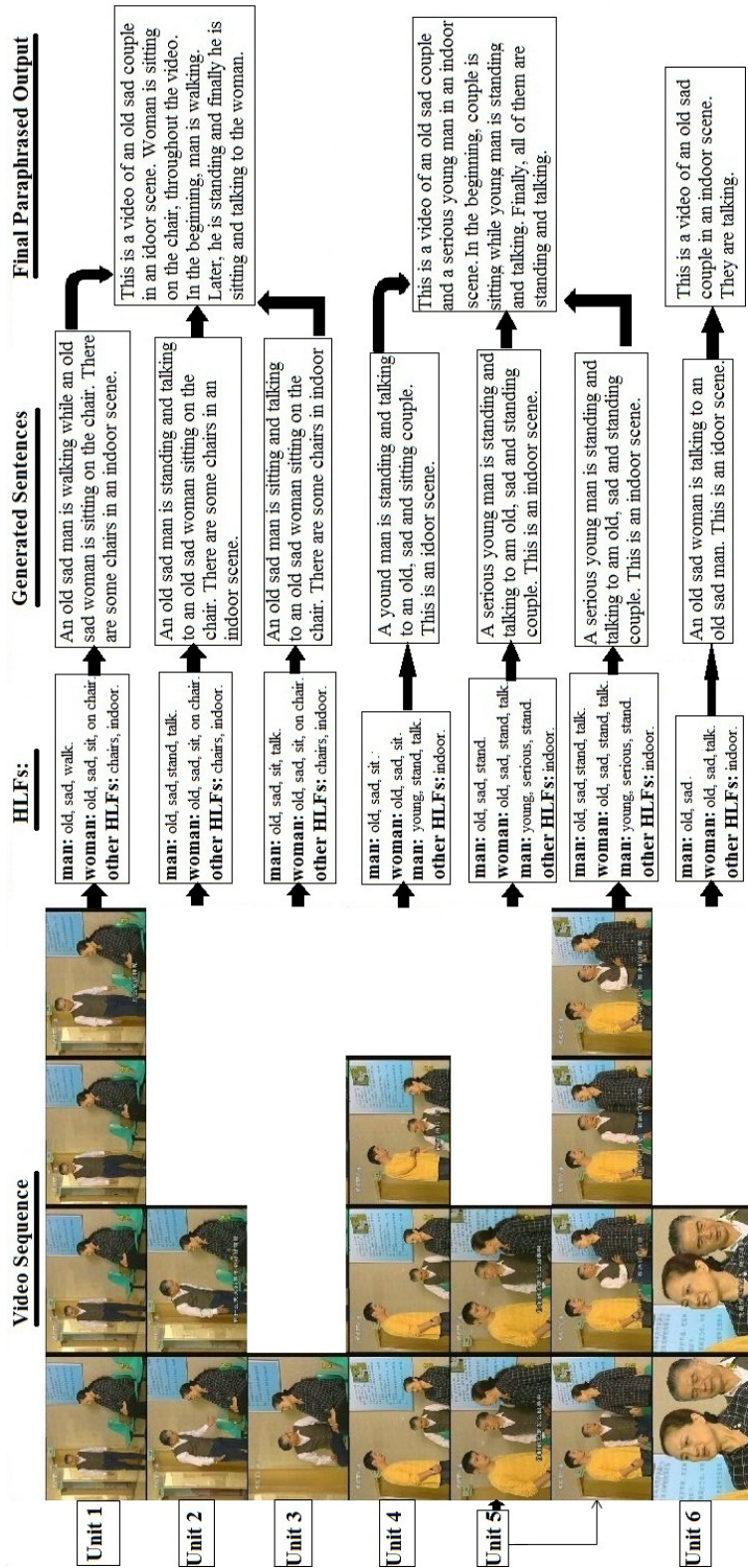


Figure 7.2: Seven frame sequences extracted from a scene, 'indoor scene with a man and a woman' — seen in video '20041101_160000_CCTV4_DAILY_NEWS_CHN' from the 2005 TREC Video search and retrieval task.

a full description of the scene based on individual descriptions. A paraphrasing operation is outlined at this stage.

7.3 Paraphrasing Unit-based Descriptions

By identifying description units we should be able to reduce dramatically, although not fully, redundant and repeated expressions, resulting from simply concatenating frame based descriptions. In order to further improve compactness and coherency, we consider joining multiple unit-based descriptions into a single sentence. We refer to this operation as paraphrasing. The following two problems should be addressed:

- creation of paraphrasing candidates from original descriptions;
- decision of whether to keep the originals, or to replace them with one of paraphrases.

We use a few rules to derive paraphrasing candidates, then calculate statistical language modelling and probabilistic parsing scores to choose the most syntactically appealing expression [Dao and Simpson, 2002]. For example, ‘*a woman is smiling*’ and ‘*a woman is walking*’ can be paraphrased into a sentence ‘*a woman is smiling and walking*’ assuming that the paraphrase has the higher syntactic score.

Creating paraphrasing candidates. Suppose that two original descriptions are given. We aim to create multiple, if possible, paraphrasing candidates by applying a relatively straightforward set of rules:

1. Simply joining both descriptions using conjunction, *i.e.*, ‘*and*’ operator;

(example) ‘*a man is walking*’ + ‘*a woman is happy*’ \Rightarrow ‘*a man is walking and a woman is happy*’;

2. Joining originals using most frequent function words (preposition), *e.g.*, ‘*while*’, ‘*then*’, ‘*after*’, ‘*for*’, ‘*on*’, ‘*with*’, ‘*but*’, ‘*by*’, ‘*because*’, ‘*then*’, ‘*only*’, ‘*between*’, ‘*though*’, and ‘*more*’. The choice is made by calculating the language modelling (LM) score;

(example) ‘*a man is walking*’ + ‘*a woman is happy*’ \Rightarrow ‘*a man is walking while a woman is happy*’;

3. Rephrasing sentences:

- (a) Finding the same phrase between both sentence. If it occurs, keep it in the first sentence and discard from the second;

(example) ‘*a happy old man is walking*’ + ‘*a happy old man is standing*’
 \Rightarrow ‘*a happy old man is walking and standing*’;

- (b) Dealing with adjectives: if an adjective is found, preference is to place it at the beginning of the sentence;

(example) ‘*a man is walking*’ + ‘*a man is happy*’ \Rightarrow ‘*a happy man is walking*’;

(c) Dealing with verbs: verbs are joined by the conjunction operator.

(example) ‘*a man is walking*’ + ‘*a man is talking*’ \Rightarrow ‘*a man is walking and talking*’.

For 3.(a) above, greedy string tiling (GST) algorithm¹ can be applied to find similar phrases between two sentence [Wise, 1993]. Common tiles² between both sentences are rephrased together. Rule 3(a) normally will be used in conjunction with other rules, such as 3(b) or 3(c).

Examples. Here are a few examples for creating paraphrasing candidates. In the first example, rule 3(a) and 3(c) result in the same paraphrase:

originals:

a man is walking;
a man is happy;

candidates:

a man is walking and a man is happy; (rule 1)
a man is walking then a man is happy; (rule 2)
a man is walking and happy; (rule 3(a), 3(c))
a happy man is walking; (rule 3(b))

And here is the second example:

originals:

an old man is sitting on the chair;
a man is smiling;

candidates:

an old man is sitting on the chair and a man is smiling; (rule 1)
an old man is sitting on the chair while a man is smiling; (rule 2)
an old man is sitting on the chair and smiling; (rule 3(a), 3(c))
a smiling old man is sitting on the chair; (rule 3(b))

The last example consists of four sentences:

originals:

¹ The advantages of using GST, in comparison to alternative string similarity algorithms such as a longest common subsequence or an edit distance, is its ability to detect block moves: treating the transposition of a substring of contiguous words as a single move instead of considering each word separately.

² A tile is a consecutive subsequence of the maximal length that occurs as one-to-one pairing between two input sentences.

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS

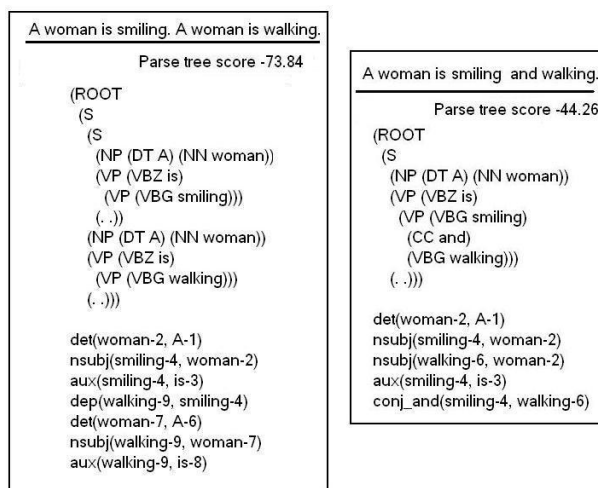


Figure 7.3: The paraphrase had the higher parsing score than its original.

a woman is walking;
a woman is walking while a bike is in the background;
a woman is walking;
a woman is walking while other humans are in the background;

where a phrase ‘a woman is walking’ appears in all sentences, hence it is kept in the first sentence and dropped from the rest. The modified originals and the paraphrasing candidates are the following. Only rules 1 and 2 are able to produce the paraphrase:

modified originals:

a woman is walking;
a bike is in the background;
other humans are in the background;

candidates:

a woman is walking and a bike is in the background and other humans are in the background; (rule 1)
a woman is walking, then a bike is in the background, later other humans are in the background; (rule 2)

Language modelling. A statistical language model assigns a probability to a sequence of words. A language modelling score can indicate which one is more syntactically likely between the original description and its paraphrases. In the experiments (Section 7.5) we derived a trigram language model from the Penn Treebank [Marcus et al., 1993], using *SRILM* toolkit [Stolcke, 2002]. Since LM score is based on multiplication of probabilities, sentences having higher number of words will achieve lesser likelihood scores. To counter this effect, likelihood

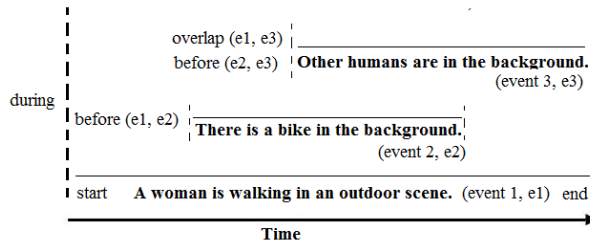


Figure 7.4: TimeML applying temporal relations to machine generated descriptions for Figure 7.1. Three events are identified and relations between events are shown as ‘start’, ‘before’, and ‘overlap’.

scores for candidate sentences were multiplied by the number of words in those sentences which produced normalized likelihood scores. As an example, using this language model, the paraphrase:

a woman is smiling and walking.

has the greater log likelihood score than its original in two phrases:

a woman is smiling; a woman is walking.

As a consequence the paraphrase is chosen.

Probabilistic parsing. The syntactic structure of a natural language description can be parsed using a probabilistic parser. Because a parse tree determines the terminal yield it is sufficient to calculate the probability of the tree. A probabilistic parser is able to score sentences before and after the paraphrasing operation. An example in Figure 7.3 was calculated using a probabilistic parser by [Klein and Manning, 2003].

7.4 Incorporating Temporal Information

The TARSQI toolkit [Verhagen et al., 2005] is used to find these relations between sentences. Figure 7.4 shows temporal relations for the video sequence in Figure 7.1. There are three events identified — ‘walking’, ‘walking while a bike is in the background’ and ‘walking while other humans are in the background’. ‘Duration’ relations is common among these three events. First two events have ‘before’ and ‘after’ relation between them. Last event has two sets of subjects involved; firstly a walking woman and secondly other humans. There is ‘overlap’ relation between woman walking and other people walking. Most frequent relations in video sequences are ‘before’, ‘after’, ‘start’ and ‘finish’ for single humans, while ‘overlap’, ‘during’ and ‘meeting’ are common for multiple humans. Once these relations are identified between sentences, keywords are manually defined to present them. Figure 7.5 shows a list of keywords used for TimeML to represent relations. Finally they are put into templates to produce temporally coherent description of video sequences.

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS

after	then, after, later, next, thereafter
during	while, at the same time, meanwhile, throughout
before	previous, afterwards, prior to, since
start	initiation, at the beginning, at the start

Figure 7.5: Keywords for temporal relations between sentences. These keywords are defined for explaining relations based on Allen algebra.

7.5 Experiments

We start this section by presenting a relatively simple scene with three emotion units identified in a single camera shot (Figure 7.6). Each row represents a unit, indicating changes of the emotional states. It is needless to say that the actual number of frames in one unit was many more than what is shown. The figure also includes hand annotation, the original machine generated description and its paraphrase. More complex example in Figure 7.7 is a scene with a mixture of human emotions, expressions, background and presence of multiple humans. In the following we aim to evaluate the machine generated descriptions using the hand annotation as a groundtruth.

7.5.1 Evaluation with ROUGE

Table 7.1 presents the ROUGE scores for video segments of NLDV- Corpus 1. In seven categories it compares (1) simple concatenation of machine generated original descriptions and (2) their paraphrases with temporal information incorporated. Hand annotations were often subjective, and dependent on one’s perception and understanding, that could be affected by educational and professional background, personal interests and experiences. Still there was reasonable similarity between machine generated descriptions and hand annotations as depicted by ROUGE scores. ‘Action’, ‘Close-up’ and ‘News’ videos had higher scores, probably because of the presence of humans with well defined activities and emotions. ‘Indoor/Outdoor’ videos showed the poorest results, clearly due to the limited capability of image processing techniques. The similarity score was improved in many cases after the paraphrasing operation. In many videos, paraphrases were much shorter than their ground truth hand annotation. This means that, although handy, ROUGE might not have been the most suitable measure for evaluation. To remedy this, a task based evaluation strategy was explored in the section below.

Secondly, ROUGE scores for NLDV- Corpus 2 are presented in the table 7.2. Again, though, there is improvement due to introduction of unit based description, but overall scores are very low. Since, length of hand annotations and machine generated descriptions is quite different, this measure did not seem to be appropriate for this evaluation.

7.5.2 Task Based Evaluation

To shed light on the advantage of description for video retrieval application, two sets of task based evaluations were performed. Firstly, human subjects were provided with a description



Hand annotation: A woman is crying then looks serious. Finally she looks happy and smiling.
Machine generated original: A woman is crying. A woman is serious. A woman is smiling.
Paraphrased: A woman is crying, then she is serious. Later on she is smiling.

Figure 7.6: Three description units in one emotional scene taken from Corpus 1 — seen in video ‘MRS144765’ from the 2007 TREC Video rushes summarisation task. Shown under the montage are one of hand annotations, machine generated original description and its paraphrase.

		-1	-2	-3	-L	-S	-SU
Action	original	0.4369	0.3087	0.2994	0.4369	0.3563	0.3686
	paraphrase	0.4839	0.3327	0.3191	0.4919	0.4123	0.4486
Close-up	original	0.5385	0.3109	0.2106	0.4110	0.4193	0.4413
	paraphrase	0.5787	0.3202	0.2198	0.4622	0.4587	0.4713
News	original	0.4814	0.3627	0.2712	0.3852	0.3618	0.3712
	paraphrase	0.4839	0.3327	0.3191	0.4919	0.4123	0.4486
Meeting	original	0.3330	0.2462	0.2400	0.3330	0.2648	0.2754
	paraphrase	0.3216	0.2154	0.2096	0.3187	0.2543	0.2544
Grouping	original	0.3067	0.2619	0.1229	0.3067	0.2229	0.3067
	paraphrase	0.3213	0.2703	0.1312	0.3315	0.2492	0.3188
Traffic	original	0.3121	0.1268	0.1250	0.3121	0.3236	0.3407
	paraphrase	0.3121	0.1268	0.1250	0.3121	0.3236	0.3407
Indoor/Outdoor	original	0.2544	0.1877	0.1302	0.2544	0.2302	0.2544
	paraphrase	0.2544	0.1877	0.1302	0.2544	0.2302	0.2544

Table 7.1: ROUGE scores between machine generated descriptions and 13 hand annotations. ROUGE 1-3 shows n-gram overlap similarity between reference and model descriptions. ROUGE-L is based on longest common subsequence. ROUGE-S skips bigram co-occurrence without gap length. ROUGE-SU shows results when skipping bigram co-occurrence with unigrams.

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS



Machine generated original:

A serious man is speaking while there are humans in the background. A sad man is speaking while there are humans in the background. It is an outdoor scene. A sad woman is speaking while there is a man in the background. It is an outdoor scene. There are many humans. There are many men present. One of the men is old and two are young. It is an indoor scene. There are many human. There are two women, one child and other humans. A sad woman is silent and looking into camera. A child is looking into camera. A sad woman is silent and looking into camera. A child is looking into camera. A man is present while there is another human in front of him. A serious woman is speaking.

Paraphrased:

A sad man is speaking with a sad woman while there are humans in the background in an outdoor scene. One old man is present with two young men and some other humans in an indoor scene. A sad woman is silent and looking into camera. A child is looking into camera. A man is present while there is another human in front of him. A serious woman is speaking.

Figure 7.7: This video montage taken from Corpus 2, is seen in video ‘MS2063001’ from the 2008 TREC Video rushes summarisation task. Due to space limitation hand annotation is not shown for this example.

ROUGE	-1	-2	-3	-L	-S	-SU
Machine generated original	0.2871	0.1759	0.1169	0.2778	0.2244	0.2531
Paraphrased	0.3214	0.2318	0.1977	0.3364	0.2911	0.2887

Table 7.2: ROUGE scores between machine generated descriptions and hand annotations for NLDV - Corpus 2.

and required to find a matching video. Secondly, human subjects were provided with a video clip and required to find a matching description.

The evaluation strategy for **Task 1** was designed as follows: human subjects were instructed to find a video that corresponded to a natural language description. Each subject was provided with one textual description and 20 video segments at one time. The same set of 20 video clips were repeatedly used, a half of which were selected from the ‘Close-up’ category and the rest were from the ‘Action’ category. This resulted in a pool of candidates, consisting of clearly distinctive videos (between categories) and videos with subtle differences (within a single category). Once a choice was made, each subject was provided with the correct video stream and the following questionnaire:

question 1: how well the video stream were explained, rating from ‘explained completely’, ‘satisfactorily’, ‘fairly’, ‘poorly’, or ‘does not explain’;

question 2: fluency, rating from ‘very fluent’, ‘satisfactory’, ‘fair’, ‘poor’, to ‘does not make sense’;

question 3: usefulness for including the following visual contents into descriptions, ratings from ‘most useful’, ‘very useful’, ‘useful’, ‘slightly useful’ to ‘not useful’:

- scene description in a text form;
- references to humans, age, gender, emotion and expression;
- references to objects and relationship with humans;
- background, colour information or scene settings.

Five human subjects conducted this task searching a corresponding video for each of five descriptions. They did not involve creation of the dataset, hence they saw these videos for the first time. A baseline performance was measured by replacing a description with keywords. Keywords consisted of a complete set of HLFs that were used for deriving the natural language description. For fairness, subjects were provided with keywords and descriptions for different videos. This arrangement was needed because use of the same video for both keywords and descriptions almost always affected the performance when they saw the same video for the second time.

For **Task 2**, subjects were provided with a video segment at each time and instructed to choose the corresponding description out of ten candidates, consisting of five from the ‘Action’ category and another five from the ‘Close-up’ category. The remaining procedure was the same as Task 1: five human subjects conducted this task, searching a corresponding description

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS

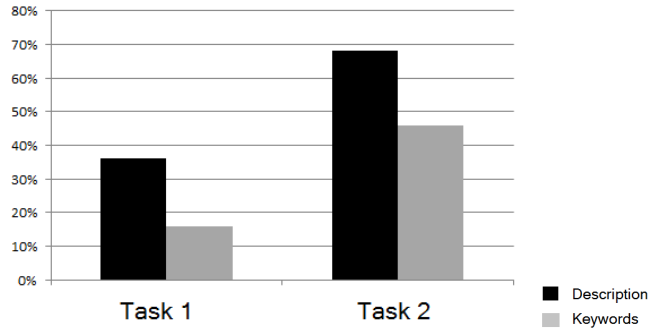


Figure 7.8: Correctly identified videos based on description and keywords alone.

for each of five videos. None of them involved in dataset creation nor Task 1. A baseline performance was measured by replacing a description with keywords.

Figure 7.8 presents results for correctly identified videos for both evaluation tasks. For both Tasks 1 and 2, description based retrieval performed better than the keyword baseline by a clear margin (roughly 20% absolute or more), indicating that transforming keywords into more verbose descriptions was a valuable exercise. Task 2 resulted in higher performance for both the keyword based baseline (46% correct) and description based evaluation (68%). It was probably because Task 2 was inherently the simpler task, not only because the number of candidates were less (ten as opposed to 20 in Task 1) but also comparison of descriptions was more efficient than comparison between video streams. In the following we summarise the outcomes from the questionnaire collected in Task 1. The same set of questionnaire was set for Task 2 however, the outcomes were similar to those for Task 1, hence we do not present them in this paper.

How well the video stream were explained. This question measures the scale for using natural language to describe videos. Figure 7.9(a) shows more than 50% of subjects responded that natural language descriptions provided satisfactory explanation (or better) of the video. Only 20% of subjects stated that keywords were satisfactory or better, and more than 40% considered keywords explained the video poorly or worse.

Fluency. This question sheds light on the fluency factor of generated descriptions¹. Figure 7.9(b) indicates that about 50% of subjects said that fluency in natural language descriptions was at least satisfactory, while 10% of cases they noticed poorly fluent expressions.

Usefulness of scene description. This questions finds the effect of scene description presented in a text form. The result was encouraging, as shown in Figure 7.9(c), since roughly 60% of subjects found scene description was useful for a video retrieval task. The number dropped to 40% for keyword based identification.

Usefulness of human structure related information. As most videos in the dataset were related to humans, their emotions and activities, this question was very important. Much emphasis was placed on the effect of human structure related information such as age, gender,

¹ No comparison is made against keywords since measuring fluency with keywords does not make sense.

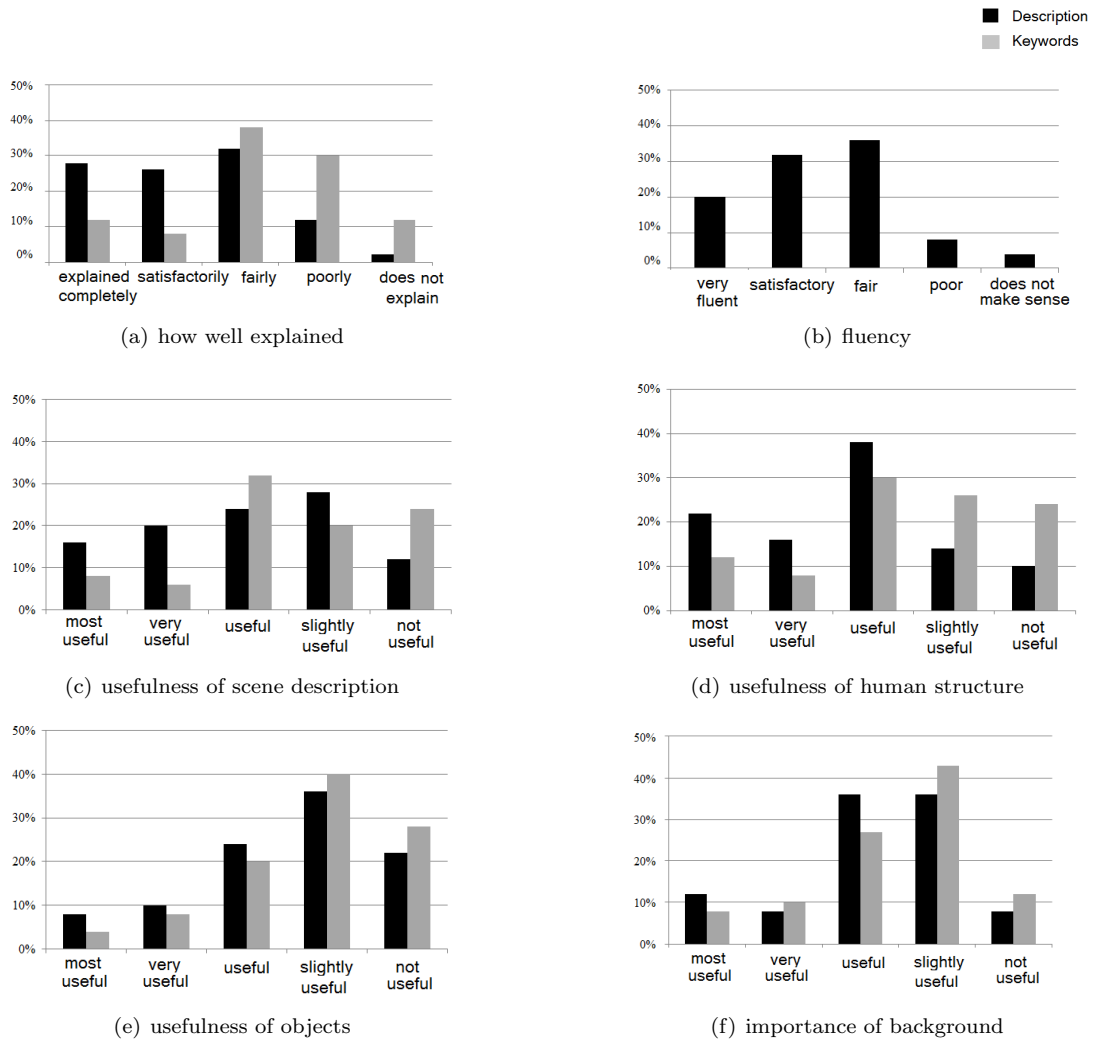


Figure 7.9: Outcomes from the questionnaire collected in Task 1 for the task based evaluation, comparing natural language descriptions and keywords based evaluation.

7. NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO STREAMS

emotions (facial expressions) and body gestures. The question aimed to find the effect of human related descriptions for correct identification of videos. Figure 7.9(d) shows that more than 70% of subjects considered this information was useful, of which nearly 40% said it was very useful. Even for keywords based evaluation, roughly 50% of subjects found this information useful.

Usefulness of objects and their relations with humans. For better understanding of visual scene and its semantics, non-human objects play a very important role. As presented in Figure 7.9(e), the outcome for this question was not very encouraging; roughly 60% of participants were unable to find well formed explanation of objects and their relations with humans in descriptions generated although descriptions were still considered better than keywords.

Importance of background, colour information and scene settings. Often, background and scene settings help in better understanding of the visual scene when humans and other objects are present. However humans and non-human objects are not always present (or failed to be identified all together by the image processing techniques). In such cases, it is particularly useful if we are able to identify visual scene setting and colour information of a video. This question measures the importance of background and scene setting information. Figure 7.9(f) shows that most subjects thought they were useful (or slightly so) but not too much. Similar results were obtained for keywords based evaluation. This might have been the consequence of the dataset biased towards videos with human centred contents.

7.6 Summary

Previous two chapters presented a framework for ‘*language description generation for individual frames*’. To describe a complete video sequence, simply joining frame based descriptions results in several shortcomings. Descriptions of individual frames are crude, redundant and lack temporal information. This chapter introduced the notion of ‘*unit*’ to remedy these shortcomings and generate smooth, coherent and well phrased descriptions of video sequences. Different scenarios for explaining the concept of Unit were presented. Paraphrasing original descriptions using language model and parsing scores was the main step for this Unit generation framework. Once this paraphrasing was done, temporal information was inserted between individual units to generate complete description. Evaluation of the proposed framework was done using two sets of video corpora from Chapter 3. For both corpora, there was clear improvement in ROUGE scores after this Unit generation. Task based evaluation further provided the proof of supremacy of generated descriptions when compared against the keywords alone.

Chapter 8

Use of Natural Language Descriptions for Video Scene Classification

8.1 Introduction

The work presented in this chapter is concerned with classification of video sequences into their proper categories using natural language descriptions. Based on the generated description for a given video sequence, it is assigned a category from one of the seven categories of NLDV - Corpus 1 (section 3.2). For example, a video description such as ‘*A man is walking. He is standing. He is sitting*’ should be assigned ‘*Action*’ category due to the presence of words such walk, stand and sit. This task becomes quite important for categorizing video sequences, *i.e.*, to properly generate video categories containing similar video sequences.

Two paradigms are presented for video scene classification, *i.e.*, (i) classification based solely on hand annotations, (ii) classification based on machine generated descriptions where hand annotations are used as a supplement for machine generated descriptions. The problem for video scene classification based on annotation is that annotation does not always provide sufficiently detailed information about the scene due to small amount of natural language descriptions created for the video stream. We aim to address this problem by incorporating additional information derived from individual scene classes.

The approach incorporates a conventional *tfidf* term-document matrix with scene class specific information derived using the maximum a posteriori (MAP) estimates and the chi-square statistic. Further latent semantic analysis (LSA) is applied to find co-occurrence terms between documents. The experiment adopts the k-nearest neighbour (kNN) and the support vector machine (SVM) classifiers to evaluate the effectiveness of scene class information and co-occurrence terms. Finally for machine generated descriptions, a new technique based on pseudo semantic tree (PST) is proposed, where human annotations complement the limited

8. USE OF NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO SCENE CLASSIFICATION

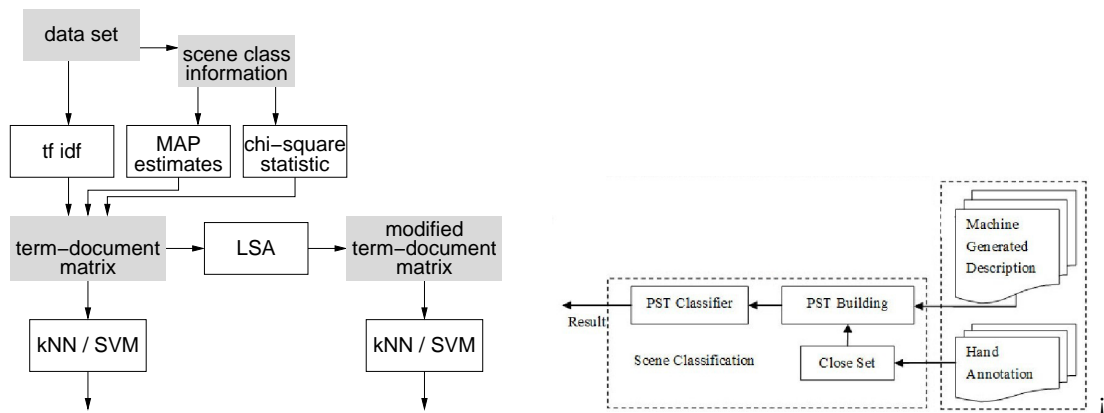


Figure 8.1: (left panel) The flowchart for the approach. It is divided into two stages, feature extraction and classification. (right panel) Pseudo semantic tree for scene classification, where human annotations are applied as effective supplement to machine generated descriptions.

size of machine generated descriptions, and out-of-vocabulary terms can also be handled.

8.2 The Approach

A short video clip of 10 to 30 seconds can be annotated with a few sentences. When the amount of (meta)data is small, it is often difficult to achieve a good performance based only on information derived from the data in a conventional fashion. When annotation is short (which is particularly the case for natural language descriptions we prepared for video data), each document vector is very sparse because terms are drawn from the entire vocabulary. In this situation, we consider incorporating complementarity knowledge extracted from individual video scene classes.

Hand Annotations. The work presented here is concerned with classifying video scenes based on their natural language description such as ‘*a man is standing in front of a desk*’. It is different from typical retrieval and classification approaches in that it does not rely on the large amount of data. Instead, class specific information is extracted for each video scene using the chi-square statistic and the MAP estimates. It complements the insufficient amount of information in the conventional *tf-idf* term-document matrix. Additionally we aim to apply LSA; it may be considered as a space mapping tool that is able to find the term co-occurrences in a term-document matrix. The *tf-idf* term-document matrix will be projected to another matrix, which turns out to be an approximation of the original matrix. The scene classification is finally achieved using the kNN and the SVM classifiers. The approach is illustrated in left panel of Figure 8.1.

Machine Generated Descriptions. Initially same approach is applied as that of hand annotations (shown in left panel of Figure 8.1). Machine generated descriptions are sometimes crude and erroneous expressions. Further the data size, derived from the HLFs, is limited and

frequent terms in <i>hand</i> annotations				
Action	men	women	people	walk
Closeup	men	women	talk	wear
News	men	wear	women	news
Meeting	men	people	talk	sit
Grouping	men	people	wear	women
Traffic	car	people	run	street
In/Outdoor	green	tree	women	walk

frequent terms in <i>machine generated</i> descriptions				
Action	men	walk	women	stand
Closeup	men	serious	speak	women
News	speak	camera	face	men
Meeting	men	person	present	sit
Grouping	men	person	present	back
Traffic	car	move	present	person
In/Outdoor	scene	outdoor	present	static

Table 8.1: Four most frequent words in hand annotations and machine generated descriptions for seven scene categories.

very small. An automatic description produced for a half minute video may consists of a few sentences with several words for each. To address these issues a pseudo semantic tree (PST) is created for each scene category, where human annotations complement the limited size of machine generated descriptions.

The main contributions of the work fall in the following four areas:

- (1) Combining information extraction into a classification framework in order to handle the sparse data problem in classification of video streams.
- (2) Using LSA to find term co-occurrences in the term-document matrix.
- (3) Comparing the kNN and SVM classifiers in the video classification task based on its annotation.
- (4) Video scene classification based on these automatically generated descriptions. Content words in these descriptions are limited. A new pseudo-semantic tree classifier is proposed to combine useful knowledge in human annotation with automatic descriptions.

Analysis of Machine Generated Descriptions and Hand Annotations. Hand annotations contained roughly 2500 unique words after removal of stop words and application of stemming. On the other hand, less than 100 unique words were found in machine generated descriptions. Table 8.1 presents four most frequent words in each scene category. Since many of frequent words (*e.g.*, men, women, person) do not contribute much when finding a particular category, it can be concluded that scene classification cannot be achieved by a frequency alone. The maximum a posteriori (MAP) estimation may be used with *tf-idf* to mine the relationship between words and different scene categories. Furthermore, many of frequent words are pronouns, expressing less information than verb, adjective or adverb. It is also evident that most words are shared between hand annotations and machine generated descriptions; we conclude that hand annotations can be an effective complement for machine generated descriptions.

8. USE OF NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO SCENE CLASSIFICATION

class	four most relevant terms			
action	wave	walk	report	run
closeup	sad	cry	angry	look
grouping	woman	clap	interview	cheer
in/outdoor	tree	green	mountain	hill
meeting	meet	table	group	discuss
news	news	report	present	tv
traffic	car	run	road	man

Table 8.2: Four most relevant terms for each scene class found by the chi-square test in hand annotations.

class	four most relevant terms			
action	wave	walk	news	report
closeup	sad	cry	angry	laugh
grouping	lecture	clap	cheer	speaker
in/outdoor	mountain	tree	hill	green
meeting	meet	table	discuss	journalist
news	news	monitor	report	present
traffic	run	car	traffic	fast

Table 8.3: Four most relevant terms for each scene class found by the MAP estimator in hand annotations.

8.3 Feature Extraction

Table 8.2 presents four most relevant terms for each of seven scene classes found by the chi-square test. Most of terms listed in the table appear reasonable selections, characterising the unique features for each scene class. For example, consider the **closeup** scene; terms such as ‘sad’, ‘cry’, ‘angry’ and ‘look’ represent human emotions well. Each term has its own contribution to a particular class. There are some terms that are relevant to multiple scene classes; *e.g.*, ‘run’ is related to the **action** and the **traffic** scenes.

Table 8.3 presents four most relevant terms for each of seven scene classes found by the MAP estimator. Once again most of terms listed in the table are sensible selections. There are many common selections listed in Tables 8.2 and 8.3. There also exist several substitutions. For the **action** category, ‘news’ probably is not the most suitable keyword. On the other hand, ‘lecture’ in Table 8.3 may represent the **grouping** scene better than ‘woman’ in Table 8.2. Based on these observations, we consider combination of the two in the follow experiments.

By observing Tables 8.2 and 8.3 it is difficult to conclude which one of the chi-square statistics and the MAP estimates is better. Table 8.4 presents the average non-zero entries rate of a document for seven scene classes with different combination of features. Columns with $tfidf + \widehat{\chi^2}$ and $tfidf + \widehat{map}$ imply that $tfidf$ scores were combined with the chi-square statistics and the MAP estimates, respectively — that is, weights were set as $w_t = \frac{1}{2}, w_c = \frac{1}{2}, w_m = 0$ for the former and $w_t = \frac{1}{2}, w_c = 0, w_m = \frac{1}{2}$ for the later. There are not many non-zero entries in a typical $tfidf$ matrix. It reflects the fact that only a few terms are actually used in each

class	$tf\text{-}idf$	$tf\text{-}idf + \widehat{\chi^2}$	$tf\text{-}idf + \widehat{map}$
action	0.0083	0.1954	0.3830
closeup	0.0068	0.1688	0.2954
grouping	0.0075	0.1499	0.2680
in/outdoor	0.0069	0.1591	0.2863
meeting	0.0074	0.1522	0.2828
news	0.0072	0.1469	0.2710
traffic	0.0071	0.1500	0.2744

Table 8.4: The average non-zero entries rate of a document for seven scene classes with different combination of features. The non-zero entries rate is the number of terms, that are not zero, divided by the number of terms in that document.

$tf\text{-}idf$	terms	$\widehat{tf\text{-}idf}$	top 15 terms
0.6069	aircraft	0.3213	room
0.5070	wait	0.3166	go
0.4519	go	0.3122	come
0.3976	come	0.2603	formal
0.3161	formal	0.2403	meet
0.3161	room	0.1794	many
0.2477	meet	0.1633	wait
0.1847	many	0.1557	cloth
0.1596	cloth	0.1292	people
0.1271	people	0.1190	wear
0.1113	wear	0.1103	corner
		0.1068	start
		0.1052	get
		0.0963	aircraft
		0.0660	aeroplane

Table 8.5: Comparison of terms, before and after the approximation by LSA projection, in the descending order of $tf\text{-}idf$ scores. They are extracted from video ‘20041101_150000_CNN.LIVEFROM.ENG’ in the TREC Video 2005 development dataset for the search task.

video clip annotations. It can be seen that by combining a typical $tf\text{-}idf$ matrix with scene class based information, we may be able to solve this to some extent.

8.4 LSA (Latent Semantic Analysis)

Table 8.5 presents comparison of terms, before and after the approximation made by LSA projection, in the descending order of $tf\text{-}idf$ scores. Before the approximation $tf\text{-}idf$ scores were zero for all terms except these 11 terms listed in the table. After LSA projection, most of terms were no longer zero although only 15 terms with the largest $\widehat{tf\text{-}idf}$ scores were shown. Although it may not be very surprising, nevertheless, it is interesting to note that the ordering of terms in the table was very similar before and after the approximation, apart from a few terms such as ‘aircraft’. Further, approximation by LSA projection does increase the number of non-zero terms. In a sense it has an effect of query expansion. Terms such as ‘start’ and ‘get’ appear in

8. USE OF NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO SCENE CLASSIFICATION

the list after the approximation, probably because they co-occur with the terms such as ‘come’ and ‘go’. In a similar fashion, ‘aeroplane’ has the same meaning with the term ‘aircraft’.

8.5 Pseudo Semantic Tree

Use of semantic tree structure in this work is different from the traditional ones (hence it is referred to as a ‘pseudo semantic tree’) in that it represents scene categories hierarchically at a term level and a concept level assigned with different weights. The PST classifier is able to supplement information for machine generated descriptions using knowledge available in hand annotations. During the training procedure it builds a pseudo semantic tree for each scene category. Classification is made with a tree for each scene category by calculating the total score for terms in a test document. The maximum score indicates the class to be assigned.

Closed Set Generation. Hand annotations are used to generate ‘closed sets’; all words having the similar semantic meanings are assigned to the same closed set. Initially a vocabulary is created from hand annotations. Then the semantic similarity of two terms in the vocabulary is calculated using *WordNet* [Miller, 1995]. Lastly a threshold is applied to generate closed sets based on relationship between two terms. All terms in the vocabulary are assigned to one of closed sets using the ‘first fit rule’. *WordNet* is a large lexical database of English language which includes nouns, verbs, adjectives and adverbs. Hyponyms and synonyms can be easily extracted.

The semantic similarity is computed using the synonyms sets (synsets) and the hyponymy relationship. First, synsets for each term in the vocabulary are generated by *WordNet*. The similarity between synsets is calculated based on the approach by [Hirst and St-Onge, 1998]. Two lexicalised concepts are semantically ‘closed’ if their *WordNet* synsets are connected by a path that is not too long and that does not change direction too often. The strength of the relationship is given by

$$R_{HS}(c_1, c_2) = C - \text{path length} - k \times d \quad (8.1)$$

where c_1 and c_2 are synsets of two terms, C and k are constants, and d is the number of direction changes in the path. If no such path exists, $R_{HS}(c_1, c_2)$ is zero and the synsets are deemed unrelated.

PST Construction. Pseudo semantic tree is defined for machine generated descriptions based on the closed set generated by *WordNet* and the term weight in the next section. Machine generated descriptions are divided into the training set and the test set. We adopt the leave-one-out policy for construction of PST. Now PST is defined as follows:

Definition 1

$$PST = (L, F, C, L_w, C_w, Root)$$

where *Root* is the root of the tree, *F* is a function to map a lexical set *L* to a concept set *C*. L_w and C_w are the term weight and the concept weight for the corresponding lexical node and concept node.

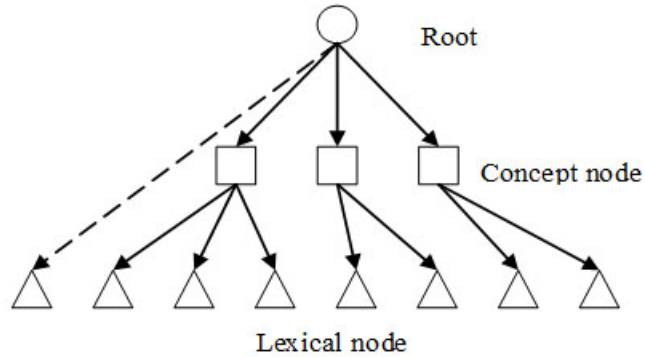


Figure 8.2: Three levels in PST. The dashed line on the left implies that some terms may not have a concept node.

The function F is derived based on the closed set generated from *WordNet*. L_w can be assigned as term weight and C_w is the average term weight in the closed set.

Three kinds of nodes, *i.e.*, root, concept and lexical nodes, are constructed in PST (see Figure 8.2). The terms falling into the same closed set are treated as lexical nodes (triangles in the figure), which can be combined as a concept node (shown by the rectangles). All concept nodes are connected with the root node (circle). In this work the vocabulary is derived from hand annotations; consequently some terms in machine generated descriptions may be out of vocabulary. Such terms can be seen as a lexical node without a concept node (shown by the dashed line on the left in the figure).

8. USE OF NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO SCENE CLASSIFICATION

Scene Classification Based on PST. For each term of a machine generated description, PST matching scores for each scene category are computed using the following steps:

- (i) Find the lexical node in PST. If the corresponding lexical node is found, assign the score as the lexical weight. Otherwise, goto step (ii).
- (ii) Obtain the concept information from the closed set, and find the concept node in PST. If the corresponding concept node is found, assign the score as the concept weight. Otherwise, goto step (iii).
- (iii) It is an out-of-vocabulary term. Find it in the closed set derived from hand annotations, and assign the score as a constant a multiplied with the term weight, then goto (iv). Normally, the constant a is less than 1 to weaken the effect of out-of-vocabulary terms.
- (iv) Traverse all terms in machine generated descriptions and sum the score for each term.

8.6 Experiments

Two sets of experiments are presented to evaluate the approach of video scene classification based on hand annotations. Firstly comparisons were made between

$tf\cdot idf$
 $tf\cdot idf$ + MAP estimates
 $tf\cdot idf$ + chi-square statistic
 $tf\cdot idf$ + MAP estimates + chi-square statistic

without LSA approximation. The purpose was to measure the improvement made by using the class oriented information, such as chi-square statistic and MAP estimates, with the conventional $tf\cdot idf$ based formulation. The combination of the above features were made using Equation (2.10), where equal weights were set for all terms. Secondly the same set of comparisons as the above were made after the LSA approximation. The first and the second sets of comparisons should shed light on the effect of LSA approximation.

The kNN and the linear kernel SVM classifiers were used to evaluate the approach. For the kNN classifier, the number of nearest neighbour ranging between 1 and 20 were tested. Cross validation was adopted; all experimental results below shows the overall average of leave-one-out.

Use of Scene Class Oriented Information. Table 8.6 presents the video scene classification performance based on natural language descriptions as presented by the hand annotations. LSA approximation was not applied for this set of experiments. Note that annotations were initially filtered by a stop list before the term-document matrix was built. The table indicates that the performance could be enhanced by incorporating the class oriented information over the conventional $tf\cdot idf$ based approach. With the $tf\cdot idf$ alone, the kNN classifier with $k = 20$ achieved 67.97%, and 82.98% was measured with the SVM classifier with a linear kernel.

classifier	<i>tf-idf</i>	+MAP	+CHI	+MAP+CHI
kNN 1	62.36	68.98	72.70	76.73
5	63.24	69.42	73.77	79.00
10	65.89	72.76	77.93	81.34
20	67.97	74.46	78.75	81.34
SVM	82.98	91.49	94.89	97.98

Table 8.6: Video scene classification performance (% correct) using hand annotations. The performance is improved by incorporating the class oriented information, such as the MAP estimates (+MAP) and the chi-square statistic (+CHI), over the conventional *tf-idf* based approach. No LSA approximation was made. Evaluations were made with the kNN (with $k = 1, 5, 10, 20$) and the SVM classifiers.

When the scene class information was augmented with the *tf-idf* scores using the MAP estimates, the performance was improved with 6.49% and 8.51% absolute, respectively, for the kNN and the SVM classifiers. Further improvements (10.78% and 11.91%) were recorded when the chi-square statistic was used. In general, it appeared that the chi-square statistic was more effective than the MAP estimates when incorporating the scene class oriented information. Finally, using all of *tf-idf* scores, the MAP estimates and the chi-square statistic together, the performance had reached 81.34% and 97.98%, that accounted for 13.37% and 15% absolute improvement from the conventional *tf-idf*.

Concerning the kNN classifier, it was not too surprising that the performance was improved by increasing the number of nearest neighbours. By taking the larger number of neighbours into consideration, the robustness and the tolerance against outliers (*i.e.*, those samples that were numerically distant from the majority of data) were enhanced. In comparison to the kNN, the SVM made better classification. Even with the *tf-idf* alone, the SVM performance was better than the KNN when the *tf-idf* scores, the MAP estimates and the chi-square statistic all together.

Effect of LSA approximation. Table 8.7 shows the video scene classification performance when LSA approximation was made. Technically, LSA was applied on X' whose matrix form was given by Equation (2.11), *i.e.*, after scene class information was incorporated with the *tf-idf* scores. The table indicates the clear improvement for most of cases. In particular, the kNN and the SVM classifiers had achieved 83.86% and 98.11%, respectively, when all three sources of information (*i.e.*, *tf-idf*, the MAP estimates and the chi-square statistic) were combined. They accounted for 2.52% and 0.13% absolute improvement when compared against the earlier experiments without LSA. This outcome seems to indicate that LSA approximation does have a great effect on the classification, because it has an ability for manipulating the co-occurrence of terms.

Finally, Figure 8.3 presents comparison of LSA approximation applied before or after the features combination (Equation 2.10). Note that, when LSA was applied before the combination, only the *tf-idf* term was approximated, then augmented with the LSA estimates and the chi-square statistic. The figure shows the better classification results, although with a very small margin, when LSA was used after the features combination.

8. USE OF NATURAL LANGUAGE DESCRIPTIONS FOR VIDEO SCENE CLASSIFICATION

classifier	<i>tf-idf</i>	+MAP	+CHI	+MAP+CHI
kNN 1	62.80	67.02	70.18	74.72
5	64.19	69.86	74.84	80.83
10	67.09	74.21	77.87	82.41
20	68.54	75.66	79.38	83.86
SVM	82.80	92.50	95.15	98.11

Table 8.7: Video scene classification performance (% correct) using hand annotations. The evaluation setup was the same as the previous results shown in Table 8.6, except that LSA approximation was applied before classification.

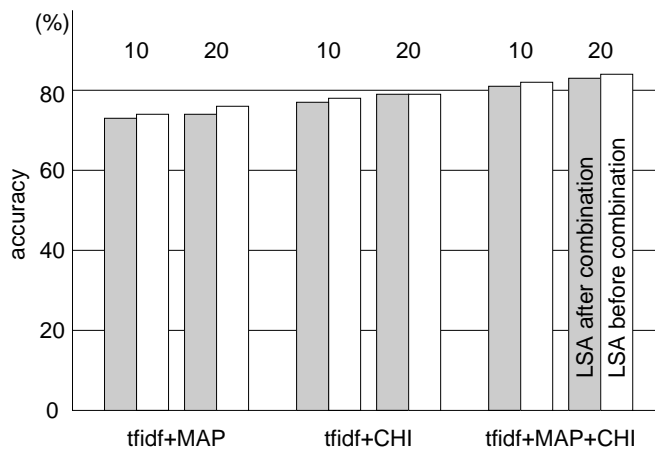


Figure 8.3: Comparison of two approaches to LSA approximation. It can be achieved before or after features combination is made by Equation (2.10). This evaluation has been made using the kNN with $k = 10, 20$.

1-NN	5-NN	10-NN	20-NN	SVM
58.09	68.38	74.26	72.06	79.56

Table 8.8: Video scene classification performance (% correct) using the small size data of machine generated annotations. The k NN (with $k = 1, 5, 10, 20$) and the SVM classifiers measured the LSA extended $tfidf$ weights.

t_1	t_2	t_3	$w_f t_1 + w_t t_2 + w_m t_3$	$w_t t_2 + w_m t_3$
74.45	80.29	81.75	83.21	84.67

Table 8.9: Video scene classification performance (% correct) using the PST classifier. Three term weights, t_1 (frequency), t_2 ($tfidf$), t_3 (MAP), and their combinations were tested. w_* is the weighting factor for combination, where equal weights were set in the experiments.

Use of Machine Generated Descriptions for Video Scene Classification. Table 8.8 presents the classification results using machine generated descriptions alone. The k NN with $k = 1, 5, 10, 20$ and the linear kernel SVM classifiers measured the effect of LSA extended $tfidf$ weights. It is not surprising that the k NN classifier achieved better when the neighbour size (k) was larger. A larger k improved the robustness and the tolerance against outliers (*i.e.*, samples that were numerically distant from the majority). The k NN classifier achieved 74.26% correct when $k = 10$, then the performance declined with even larger k . It was probably caused by the seriously limited sample size in each scene category. The SVM classifier with a linear kernel achieved 79.56%, outperforming the k NN by more than 5%.¹

PST Classifier. Table 8.9 shows the classification results based on pseudo semantic tree. The PST classifier was evaluated using 10-fold cross validation; the overall average was measured using the leave-one-out strategy. A weighting scheme based on the term frequency (t_1 in the table) was tested in order to set a baseline for the PST classifier. The $tfidf$ weights (t_2) were calculated for combination of a concept (*i.e.*, closed set) and a scene category. The outcome was better than the frequency term weighting scheme because a broader range of information was taken into consideration by the inverse document frequency factor. The term weighting scheme using MAP estimates (t_3) achieved the best performance with 81.75%. The experiment clearly indicated that the PST classifier outperformed approaches using the LSA extended $tfidf$ weight.

Finally combinations of different weighting schemes were considered. In Table 8.9, w_* was the weighting factor for combination. The performance was enhanced to 83.21% when all three term weights (*i.e.*, frequency, $tfidf$, and MAP) were combined. The best performance was 84.67% when the $tfidf$ and the MAP term weights were used together.

¹Although these results are shown including LSA approximation, still they are worse than hand annotations results shown in Table 8.7.

8.7 Summary

In this chapter we presented the approach for scene classification of video clips based on their annotations. In particular, we addressed a problem when the amount of information was small. To make a sound classification, incorporation of scene class based information (the MAP estimates and the chi-square statistic) was effective. LSA approximation was also tested, resulting in further improvement in classification. We argued that this improvement was caused by the increased number of co-occurrence terms in the term-document matrix. Finally, the kNN and the SVM classifiers were also compared for the task, resulting in superior performance by the latter. PST based classification was also proposed, which incorporates information from hand annotations and handles out-of-vocabulary terms in machine descriptions.

Chapter 9

Conclusions

Natural language description is an effective and pragmatic solution to the information abundance problem in video streams where the amount of contents is growing at a fast speed. Video is a much richer medium than text and is also less easy to browse, search and index. Automatic processing is necessary as manual processing would be impractical. Accurate processing will enable easier browsing, searching and indexing and will transcend into other fields. To make videos fully understandable, descriptions are more useful than the keywords alone. Descriptions capture relationship between participating keywords and make the context information clear, which further helps in complete understanding of the video sequences. Descriptions can further help in differentiating video sequences from each other.

This thesis is concerned with the automatic generation of natural language descriptions that can be used for video indexing, retrieval and summarization applications. Initially, hand annotations were generated for two video corpora which consisted of manually segmented video clips from TREC Video dataset (Chapter 3). Analysis of this data presented insights into humans interests on video contents. For machine generated descriptions, conventional image processing techniques are applied to extract high level features (HLFs) from individual video frames (Chapter 4). Natural language description is then produced based on these HLFs (Chapter 5). Although feature extraction processes are erroneous at various levels, approaches are explored to put them together for producing coherent descriptions. For scalability purpose, application of framework to several different video genres is also discussed (Chapter 6). For complete video sequences, a scheme to generate coherent and compact descriptions for video streams is presented which makes use of spatial and temporal relations between HLFs and individual frames respectively (Chapter 7). Calculating overlap between machine generated and human annotated descriptions concludes that machine generated descriptions capture context information and are in accordance with humans watching capabilities. Further, a task based evaluation shows improvement in video identification task as compared to keywords alone. Finally, application of generated natural language descriptions, for video scene classification is discussed (Chapter 8).

9.1 Original Contributions

1. **Corpus Generation and Analysis.** Two video corpora consisting of manually crafted videos from TRECVID provided videos, together with metadata in the form of natural language descriptions were introduced in Chapter 3. First corpus consisted of short video clips ranging from 10 to 30 seconds in length. It consisted of 140 segments of videos — 20 segments for each of the following seven categories: human actions, human close up, news, meeting, grouping, traffic, and indoor / outdoor videos. 13 human participants manually annotated this dataset in three flavors, identification of important concepts (keywords), a very short summary (title) comprising of one phrase or sentence (presenting theme or main idea) and a complete natural language description comprising of several sentences (detailed description of the visual scene). Analysis of this corpus presented insights of human behavior and interest while watching videos. This corpus was used as training and testing dataset in Chapter 5. Second corpus consisted of 10 video segments — each segment contained multiple camera shots, combined together to present a coherent and complete video story, where length of complete video was spanning between 3 and 5 minutes. This corpus was used for evaluation of approach discussed in Chapter 7.
2. **Framework for Natural Language Description of Images.** A framework for generation of natural language descriptions for human actions, behavior and their relations with other objects was presented in Chapter 4. Initially, conventional image processing techniques were used to extract high level features¹ from visual frames. These features were converted into natural language descriptions using context free grammar and a template based approach. Hierarchical sentence generation which covered scenarios related to multiple humans was presented. Quantitative evaluation based on the overlap between machine generated descriptions and hand annotations depicted the fact that these machine generated descriptions were able to capture the important visual contents together with relationships between these contents. Further, a task based evaluation presented qualitative evaluation and concluded that human subjects were able to identify video streams based on given textual descriptions.
3. **Dealing with missing and erroneous data.** Of all the various sources available, this work focussed on the processing of broadcast news and Rushes videos. Different scenarios taken from Rushes and news videos were discussed and evaluated. For dealing with noisy information caused by scarcity of HLFs, scenarios were presented to overcome these shortcomings and to generate coherent and smooth descriptions. These scenarios included dealing with missing information, absence of human subjects and dealing with erroneous HLF extraction outputs (Chapter 6). Further evaluation of the framework was performed for five more categories in addition to seven video categories of Chapter 3. Evaluation results for overlap similarity and task based evaluation concluded that framework was able to generate well phrased descriptions even with the limited HLFs.

¹Important HLFs were decided on the analysis of Corpus 1 meta-data, which was introduced in Chapter 3.

-
4. **Natural Language Descriptions for complete video sequences.** For generating descriptions of complete video sequences, notion of ‘unit’ was introduced (Chapter 7). Simple concatenation of individual frames descriptions resulted in redundant and ill formed descriptions. Further, temporal information was lost in this concatenation. ‘Unit’ based descriptions lead to well phrased, smooth and coherent descriptions for lengthy video descriptions by making use of spatial and temporal information across frames. Paraphrasing descriptions of individual frames using language modeling and parsing scores further generated a compact and coherent description.
 5. **Use of Natural Language Descriptions for Video Scene Classification.** A novel approach for video scene classification based on natural language descriptions was introduced in Chapter 8. Incorporating complementarity knowledge extracted from individual video scene classes came to rescue for classification task. Finding co-occurrence terms between documents further improved classification problem. For classification of visual scenes based solely on machine generated descriptions, human annotations were used as a complement to the limited size of machine generated descriptions, and out of vocabulary terms were also handled.

9.2 Future Work

Although fully functional framework for natural language descriptions of video streams was presented in this work, still there is room of improvement. Following directions will be explored for future work.

1. **Improving HLFs Extraction:** Current image processing methods are limited to specific objects and HLFs (as discussed in Chapter 4). Some of the further improvements include: detection of groups, extension of behavioral models, more complex interactions among humans and other objects.
2. **Probabilistic Methods for HLFs Extraction:** Currently, HLFs extraction methods are rule based which generate fixed labels to each of the identified HLFs. An interesting idea will be to make HLFs extraction methods probabilistic, *i.e.*, instead of providing definite labels, probabilities are generated for each of the HLFs. This probabilistic framework will help in reducing the chances of lost information due to rule based selection.
3. **Use of NLP¹ Techniques for Finding Missing Words:** Language modeling and parsing scores can be used to find out missing words based on the given words. The premise behind using language models for predicting missing words is the idea that a word is primarily dependent on the previous few words. To do this, an ngram language model is built from some training data by recording how often each word follows a sequence of words. The number of words in the sequence of prior words determines the order of

¹natural language processing

9. CONCLUSIONS

the ngram model. If only the previous two words are considered, then it is a 2nd-order Markov model and is called a trigram model. Similarly, a 1st-order model is called a bigram model and a model that ignores the previous words is called either a unigram model or frequency model. As a simple example, suppose if a decision is to be made to insert a missing word x between a pair of words α and β in the meaningful representation. Then, comparing probability of the sentence with missing word *i.e.*, $p(\alpha x \beta)$ and without missing word *i.e.*, $p(\alpha \beta)$ leads including or exclusion of missing word. If $p(\alpha x \beta) > p(\alpha \beta)$ then a new missing word x is inserted between α and β . For example, if extracted HLFs are ‘bird’ and ‘sky’, then missing word can be easily identified as ‘fly’ which generate a meaningful sentence such as ‘A bird is flying in the sky.’

4. **Use of Synonyms and Hyponyms for Better Description Generation:** Synonyms can be used for smooth description generation. List of synonyms can be generated from WordNet and the the best synset can be used for description based on LM score. Hyponyms can be applied for the analysis of errors/shortcomings of the system. Based on hyponyms, it is possible to explain some of the errors such as ‘run is a kind of rapid walk’ and ‘clap is a kind of hand waving’ *etc.*
5. **Use of Audio Data for Dealing with Missing and Erroneous Data.** Audio tracks are available for most of the video sequences. These audio tracks can be converted into text using speech recognition systems, which are quite common these days. Similar to the concept of language modeling scores which are based on very large corpora, these extracted tracks from audio data can be helpful for finding missing words and removing erroneous outputs from image processing task.
6. **Video Retrieval System using Language Descriptions:** Recently, some video retrieval systems have been proposed which are based on keywords based indexing and retrieval. Still there is no such system available which can retrieve videos based on complete textual descriptions such as ‘show me a video scene, where Obama is shaking hand with Pervez Musharaf’. Such queries require relationship between individual keywords.
7. **Video Summarization:** We are working on automatic text summarization framework molded for video sequences. Use of Rhetorical Structure Theory (RST) is being investigated for this summarization task. Idea is to extract the most salient sentences from the complete video description. These salient sentences are joined together to present the extractive summary(*i.e.*, summary contains the most important parts of the original document).

9.2.1 RST Based Video Summarization Framework.

1

¹This section is based on our current work for the paper [Khan and Gotoh, 2012c], entitled, ‘Summarizing Video Contents using Natural Language Descriptions’

According to RST, a rhetorical relation typically holds between two contiguous spans, of which one span (the nucleus) is more central to the writer's intention than the other (the satellite), whose sole purpose is to increase the reader's understanding or belief of what is said in the nucleus. Sometimes, two related spans are of equal importance, in which case there is a multinuclear relation between them. The smallest units of discourse are elementary discourse units (DUs) which are usually sentences although they can be clauses. All the rhetorical relations that can possibly occur in a text can be categorized into a finite set of relation types as defined by Mann and Thompson [1988] for English texts. Once the most important sentence has been identified, other sentences which are most closely related to the salient sentence are extracted and the complete summary is generated. RST is used to find this relation. In short, the following method is proposed.

Firstly RST tree is generated for complete description. The most salient sentence is selected as a nucleus. Other sentences are taken as satellites of this nucleus sentence and a tree is generated. This RST tree is transformed into a weighted graph [Bosma, 2005], in which each vertex represents a sentence. The weight of an edge represents the distance between the two sentences. Now we try to find out those sentences which are most relevant to the nucleus sentence. Given an appropriate assignment of weights in the graph, such a graph can be used to determine which sentences are the most relevant to the nucleus. Finally a graph consisting of all the most relevant sentences is maintained and presents a summary of the original document.

Constructing RST Tree. RST analysis of text requires the division of text into DUs (segments) that are the primitive elements for analysis. As video is combination of objects and activities (HLFs), a strategy might be to identify units with the semantics of the video stream, such as meaningful objects and/or events that are represented within the video. For this work we choose sentences as DU as they capture relations between objects and activities.

Identification of Nucleus. Nucleus is the most important part in any text and is the central factor for presenting text. As HLFs are the most important factors in videos, one idea of selecting the nucleus is to select the sentence having most number of HLFs.

Defining Relations Among Sentences. Once nucleus is identified, other sentences are related to nucleus using rhetorical relations. RSTTool¹ is used to draw RST tree.²

Conversion to Graph. Converting RST tree to a graph is straight forward. RST tree can be converted to a discourse graph by means of the following steps. For each elementary discourse unit in the RST tree, create a vertex associated with it. For each directed relation, create an edge from the nucleus to the satellite of the relation.

The result of the transformation is an a-cyclic directed graph of which the vertices correspond to elementary discourse units, and the edges define relations between them. Figure 9.1(a) shows an example of a rhetorical structure and a discourse graph that was created as described above. If in RST one sentence was related to a text span of two sentences, it is related to the nucleus of the two sentences in the discourse graph. If a multinuclear relation is involved, as in Figure

¹<http://www.wagsoft.com/RSTTool/>

²Note: Selection of relations among sentences is very much subjective and depends on authors choice.

9. CONCLUSIONS

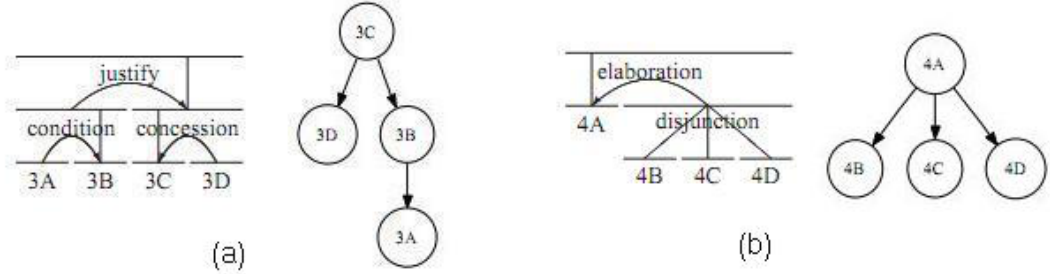


Figure 9.1: (a) Rhetorical structure example and a discourse graph created for this rhetorical structure. (b) Rhetorical structure containing a multinuclear relation and the corresponding discourse graph. *Bosma [2005]*

9.1(b), each of the sentences participating in the multinuclear relation (in the example: sentences 4B, 4C and 4D) is connected with the nucleus of the multinuclear span.

Determining Weights. Weights are assigned to edges and vertices based on the following features.

1. Each edge has a basic weight, which is the same for all edges in the graph. This makes the distinction between directly and indirectly related sentences explicit. Two sentences are less closely related if the path that connects them consists of more edges.
2. For each edge, in case of repetition of sentence, one sentence is eliminated and the other is assigned double weights to present both the sentences.
3. For each edge, number of HLFs in the satellite are counted. Sentence having higher number of HLFs has more weighting factor.
4. For each vertex, a weight is added depending on the number of words in the sentence.

The weights of edges and vertices are calculated as follows.

$$weight(e) = a + b + c \cdot \frac{1}{HLFs(sat(r))} \quad (9.1)$$

where e is the edge that was created for the relation r , where $sat(r)$ is the satellite of r , a is the basic weight, b is a weight for handling repetitions, c is constant factor of the 'number of HLFs'.

$$weight(v) = d \cdot \frac{1}{words(s)} \quad (9.2)$$

where d is the number of words in the sentence. The constants a , b , c , d are used to balance the four factors of the distance between two sentences.

Example. This example shows extraction of useful sentences from a text, based on its RST analysis. First sentence is taken as the entry point(nucleus). Based on the weights of other

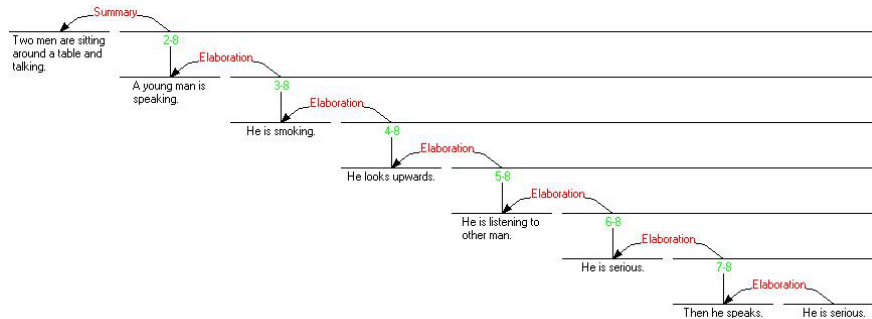


Figure 9.2: Rhetorical structure tree for the video sequence MRS042546 taken from 2007 BBC Rushes summarization task of TRECVID.

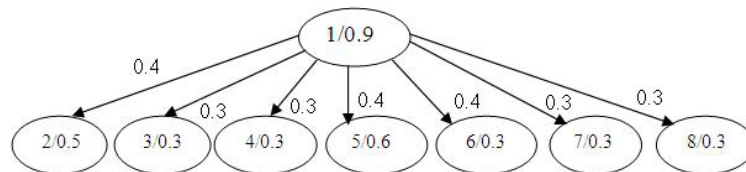


Figure 9.3: Graph for the RST tree shown in Figure 9.2

sentences in the graph best sentences are chosen to present in the summary. RST analysis of the following (segmented) text is shown in Figure 9.2.

[Two men are sitting around a table and talking.]¹ [A young man is speaking.]² [He is smoking.]³ [He looks upwards.]⁴ [He is listening to other man.]⁵ [He is serious.]⁶ [Then he speaks.]⁷ [He is serious.]⁸

It should be noted that this is a simple scenario in the sense that there is only one nucleus and all other sentences are elaboration of the concept presented in the nucleus. Figure 9.3 presents equivalent graph for this RST tree. Note that all the nucleus occupy same level of importance and are candidates for the summary.

Now using the weights attached to each edge and vertex, final candidates for summarization are selected. From the graph shown in Figure 9.3, it is clear that most important sentences are 2, 5 and 6 as they have higher weights as compared to other edges.

[A young man is speaking.][He is listening to other man.][He is serious.]

Now we use our paraphrasing framework from Section 7.3 of Chapter 6 to generate a smooth and coherent summary of the complete video sequence.¹ For paraphrasing, first there is need to merge similar sentences, such as sentence 1 and sentence 3 can be merged into ‘A serious young man is speaking.’ Now these three sentences are converted into two sentences which are (1) ‘A serious young man is speaking.’ and (2) ‘He is listening to other man.’ There is need to

¹We are currently exploring approaches to extend this two framework for lengthy descriptions.

9. CONCLUSIONS

replace words such as listening and speaking with some semantic word which can replace both words such as *'talking to'*, or *'in conversation with'*. Based on language modeling score, this words can be finalized. Final generated summary of the complete video sequence would be *'A serious young man is in conversation with an other man.'* or *'A serious young man is talking to an other man.'*

References

- H. Abdi, D. Valentin, B. Edelman, and A.J. O’Toole. More about the difference between men and women: Evidence from linear neural network and the principal-component approach. *PERCEPTION-LONDON*-, 24:539–539, 1995. [19](#)
- A. Abella, J.R. Kender, and J. Starren. Description generation of abnormal densities found in radiographs. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 542. American Medical Informatics Association, 1995. [31](#)
- A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1250–1258. Association for Computational Linguistics, 2010. [32](#)
- J.F. Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984. [73](#)
- J.F. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994. [46](#)
- A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M.R. Naphade, A. Natsev, J.R. Smith, et al. IBM research TRECVID-2005 video retrieval system. In *TRECVID Workshop, Washington DC*. Citeseer, 2005. [14](#), [15](#)
- SK Antani, M. Natarajan, JL Long, L.R. Long, and G.R. Thoma. Developing a comprehensive system for contentbased image retrieval of image and text from a national survey. In *Proc. of SPIE Medical Imaging3: PACS and Integrated Medical Information Systems*, pages 152–161. Citeseer, 2005. [14](#)
- P. Baiget, C. Fernández, X. Roca, and J. González. Automatic learning of conceptual knowledge in image sequences for human behavior interpretation. *Pattern Recognition and Image Analysis*, pages 507–514, 2007. [31](#)
- M.S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 592–597. IEEE, 2004. [20](#)

REFERENCES

- M.A. Berbar, H.M. Kelash, and A.A. Kandeel. Faces and facial features detection in color images. In *Geometric Modeling and Imaging—New Trends, 2006*, pages 209–214. IEEE. [16](#)
- D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. [33](#)
- A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001. [22](#)
- RM Bolle, B.L. Yeo, and MM Yeung. Video query: Research directions. *IBM Journal of Research and Development*, 42(2):233–252, 2010. ISSN 0018-8646. [30](#)
- A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. *Computer Vision—ECCV 2006*, pages 517–530, 2006. [33](#)
- W. Bosma. Query-based summarization using rhetorical structure theory. In *15th Meeting of CLIN, LOT, Leiden*, pages 29–44. Citeseer, 2005. [145](#), [146](#)
- G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O’Reilly Media, 2008. [59](#)
- C. Bregler and S.M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 494–499. IEEE, 1995. [28](#)
- R. Brunelli and T. Poggio. Hyperbf networks for gender classification. In *Proceedings of the DARPA Image Understanding Workshop*, volume 314. San Diego: CA, 1992. [19](#)
- P. Bull and G. Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187, 1985. [27](#)
- C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998. [15](#), [23](#)
- R.H. Chan, C.W. Ho, and M. Nikolova. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *Image Processing, IEEE Transactions on*, 14(10):1479–1485, 2005. [64](#)
- S.F. Chang, W.Y. Ma, and A. Smeulders. Recent advances and challenges of semantic image/video search. In *Proc. of ICASSP, 2007*. [2](#)
- S.F. Chang, J. He, Y.G. Jiang, E. El Khoury, C.W. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search. In *TRECVID workshop*, volume 1, 2008. [23](#)
- B. Chen. Exploring the use of latent topical information for statistical chinese spoken document retrieval. *Pattern Recognition Letters*, 27(1):9–18, 2006. [33](#)

- H.S. Chen, H.T. Chen, Y.W. Chen, and S.Y. Lee. Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 171–178. ACM, 2006. [22](#), [64](#)
- G.C.H. Chuang and C.C.J. Kuo. Wavelet descriptor of planar curves: Theory and applications. *IEEE Transactions on Image Processing*, 5(1):56–70, 1996. [14](#)
- I. Cohen, A. Garg, and T.S. Huang. Emotion recognition from facial expressions using multilevel hmm. *Science And Technology*, 2000. [20](#)
- T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, et al. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. [27](#)
- D. Cunado, M.S. Nixon, and J.N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003. [65](#)
- J.E. Cutting and L.T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9(5):353–356, 1977. [22](#)
- J.Z. Da Deng. Combining multiple precision-boosted classifiers for indoor-outdoor scene classification. 2005. [26](#)
- L.G. Da Silveira, J. Facon, and D.L. Borges. Visual speech recognition: a solution from feature extraction to words classification. In *Computer Graphics and Image Processing, 2003. SIBGRAPI 2003. XVI Brazilian Symposium on*, pages 399–405. IEEE, 2003. [70](#)
- H. Dalianis. Aggregation as a subtask of text and sentence planning. In *Proceedings of the 9th Florida Artificial Intelligence Research Symposium*, pages 1–5, 1996. [28](#)
- T.N. Dao and T. Simpson. Measuring similarity between sentences. In *Proceedings of the Document Understanding Conference*, 2002. [118](#)
- E. de Lera and M. Garreta-Domingo. Ten emotion heuristics: guidelines for assessing the user’s affective dimension easily and cost-effectively. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 2*, pages 163–166. British Computer Society, 2007. [20](#)
- S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990. [35](#)
- A. Del Bimbo. *Visual information retrieval*. Morgan Kaufmann, 1999. [14](#)
- W.T. Dempster and G.R.L. Gaughran. Properties of body segments based on size and weight. *American Journal of Anatomy*, 120(1):33–54, 1967. [65](#)
- P. Ekman and W.V. Friesen. Facial action coding system. 1977. [20](#)

REFERENCES

- A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002. [17](#)
- M. Elhadad. *Using argumentation to control lexical choice: a functional unification implementation*. PhD thesis, Citeseer, 1992. [29](#)
- B.D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101, 2004. [47](#)
- M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. [67](#)
- A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009. [32](#)
- A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. *Computer Vision–ECCV 2010*, pages 15–29, 2010. [12](#)
- Y. Feng and M. Lapata. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249. Association for Computational Linguistics, 2010. [32](#)
- R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from googles image search. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1816–1823. Citeseer, 2005. [25](#)
- S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Proceedings Conference on Computer Vision and Pattern Recognition*, pages 1–8. Citeseer, 2007. [25](#)
- R. Fisher, J. Santos-Victor, and J. Crowley. Caviar: Context aware vision using image-based active recognition, 2005. [12](#)
- Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995. [59](#)
- Y. Fu and T.S. Huang. Human age estimation with regression on discriminative aging manifold. *Multimedia, IEEE Transactions on*, 10(4):578–584, 2008. [18](#)
- Y. Fu, G. Guo, and T.S. Huang. Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1955–1976, 2010. [18](#)

-
- H. Fujiyoshi and A.J. Lipton. Real-time human motion analysis by image skeletonization. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pages 15–21. IEEE, 1998. [22](#)
- H. Fujiyoshi, A.J. Lipton, and T. Kanade. Real-time human motion analysis by image skeletonization. *IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E SERIES D*, 87(1):113–120, 2004. [66](#)
- R.M. Gagne, W.W. Wager, K.C. Golas, J.M. Keller, and J.D. Russell. Principles of instructional design. *Performance Improvement*, 44(2):44–46, 2005. [28](#)
- C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010. ISSN 1077-3142. [30](#)
- A. Gatt and E. Reiter. 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, 2009. [30](#), [88](#)
- X. Geng, Z.H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2234–2240, 2007. [18](#)
- B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski. Sexnet: A neural network identifies sex from human faces. *Advances in neural information processing systems*, 3:572–577, 1991. [19](#)
- K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 19–25. Citeseer, 2006. [25](#)
- S. Gutta and H. Wechsler. Gender and ethnic classification of human faces using hybrid classifiers. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 6, pages 4084–4089. IEEE, 1999. [19](#)
- I. Haritaoglu, D. Harwood, and L.S. Davis. Ghost: A human body part labeling system using silhouettes. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 77–82. IEEE, 1998. [64](#)
- A. Hauptmann, MY Chen, M. Christel, C. Huang, W.H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, et al. Confounded expectations: Informedia at trecvid 2004. In *Proc. of TRECVID*. Citeseer, 2004. [15](#)
- T.J. Hazen, F. Richardson, and A. Margolis. Topic identification from audio recordings using word and phone recognition lattices. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 659–664. IEEE, 2007. [33](#), [34](#)

REFERENCES

- P. Héde, P.A. Moëllic, J. Bourgeois, M. Joint, and C. Thomas. Automatic generation of natural language descriptions for images. *Proceedings of Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications Ordinateur)(RIAO)*. Avignon, France, 2004. 30, 31
- G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:332, 1998. 134
- E. Hjelmås and B.K. Low. Face detection: A survey. *Computer vision and image understanding*, 83(3):236–274, 2001. 16
- T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, page 21. Citeseer, 1999. 33
- W.B. Horng, C.P. Lee, and C.W. Chen. Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering*, 4(3):183–192, 2001. 60, 61
- S. Ioannou, G. Caridakis, K. Karpouzis, and S. Kollias. Robust feature detection for facial expression recognition. *Journal on Image and Video Processing*, 2007(2):5–5, 2007. 62
- K. Iwano, S. Tamura, and S. Furui. Bimodal speech recognition using lip movement measured by optical-flow analysis. In *International Workshop on Hands-Free Speech Communication*, 2001. 27, 70
- A. Jain, J. Huang, and S. Fang. Gender identification using frontal facial images. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 4–pp. IEEE, 2005. 19
- C. Jaynes, A. Kale, N. Sanders, and E. Grossmann. The terrascope dataset: A scripted multi-camera indoor video surveillance dataset with ground-truth. In *Proceedings of the IEEE Workshop on VS PETS*, volume 4. Citeseer, 2005. 12
- R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. In *Computer Vision, 1998. Sixth International Conference on*, pages 370–375, 1998. 27, 28
- M.U.G. Khan and Y. Gotoh. Describing video contents in natural language. In *EACL 2012, Workshop on Innovative hybrid approaches to the processing of textual data*, 2012a. 7
- M.U.G. Khan and Y. Gotoh. Generating natural language tags for video information management. In *Submission to the International Journal of Information Processing and Management, 2012*, 2012b. 7
- M.U.G. Khan and Y. Gotoh. Summarizing video contents using natural language descriptions. In *Preparation for International Conference on Computational Linguistics, COLING.*, 2012c. 8, 144

-
- M.U.G. Khan, L. Zhang, and Y. Gotoh. Human focused video description. In *The 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications*, 2011a. 7
- M.U.G. Khan, L. Zhang, and Y. Gotoh. Towards coherent natural language description of video streams. In *2nd IEEE International Workshop on Stochastic Image Grammars*, pages 664–671, 2011b. 7
- M.U.G. Khan, R.N.A. Nawab, and Y. Gotoh. Natural language descriptions of visual scenes - corpus generation and analysis. In *EACL 2012 workshop, Joint workshop of ESIRMT and HYTRA*, 2012a. 7
- M.U.G. Khan, L. Zhang, and Y. Gotoh. Natural language description of video streams. In *ICIP 2012, International conference on Image processing*, 2012b. 7
- W. Kim, J. Park, and C. Kim. A novel method for efficient indoor–outdoor image classification. *Journal of Signal Processing Systems*, 61(3):251–258, 2010a. 27
- W. Kim, J. Park, and C. Kim. A novel method for efficient indoor–outdoor image classification. *Journal of Signal Processing Systems*, pages 1–8, 2010b. ISSN 1939-8018. 67, 68
- D. Klein and C.D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics- Volume 1*, pages 423–430. Association for Computational Linguistics, 2003. 121
- A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002. 30, 31
- A. Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga. Recognition and textual description of human activities by mobile robot. In *Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on*, pages 53–53. IEEE, 2008. 31
- Y.H. Kwon and N.D.V. Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999. 18
- A. Lanitis, C.J. Taylor, and T.F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 442–455, 2002. 18
- I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2): 107–123, 2005. 21
- I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. *Spatial Coherence for Visual Motion Analysis*, pages 91–103, 2006. 21

REFERENCES

- S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. Ieee, 2006. [33](#)
- B. Lee, J. Chun, and P. Park. Classification of facial expression using svm for emotion care service system. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on*, pages 8–12. IEEE, 2008a. [20](#)
- M.W. Lee, A. Hakeem, N. Haering, and S.C. Zhu. SAVE: A framework for semantic annotation of visual events. In *CVPR Workshop*, 2008b. [31](#)
- M.K. Leung and Y.H. Yang. First sight: A human body outline labeling system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(4):359–377, 1995. [22](#)
- S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, and Y. Choi. 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, 2011. [32](#)
- Y. Li, J. Chen, W. Gao, and B. Yin. Face detection: a survey. In *The 16th National Conference on Computer Science and Technology Application*, 2004. [16](#)
- R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Proc. SPIE*, volume 3656, pages 290–301. Citeseer, 1999. [49](#)
- R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP*, volume 1, pages 900–903. Citeseer, 2002. [25](#), [67](#)
- C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *WAS*, 2004. [93](#)
- Y.C. Lin and S.C. Tai. A fast linde-buzo-gray algorithm in image vector quantization. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 45(3):432–435, 1998. [60](#)
- K.A. Lindquist and L.F. Barrett. Constructing emotion. *Psychological science*, 19(9):898, 2008. [20](#)
- D. Liu and T. Chen. A topic-motion model for unsupervised video object discovery. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. [33](#)
- J. Luo and A. Savakis. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pages 745–748. IEEE, 2001. [26](#)
- I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos. Face detection and recognition of natural human emotion using markov random fields. *Personal and Ubiquitous Computing*, 13(1):95–101, 2009. ISSN 1617-4909. [60](#), [63](#)

- J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001. 15
- W.C. Mann and S.A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988. 145
- C.D. Manning, P. Raghavan, H. Schütze, and Ebooks Corporation. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, UK, 2008. 34
- Christopher Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. 47
- M. Manoria, A. Hemlata, PK Chande, and S. Jain. Video mining for face retrieval using skin color model. In *The International Congress for global Science and Technology*, page 17, 2007. 17
- M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993. 120
- M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. 2007. 24, 25
- M. Marszalek, I. Laptev, and C. Schmid. Actions in context. 2009. 12
- B.M. Mehtre, M.S. Kankanhalli, and W.F. Lee. Shape measures for content based image retrieval: a comparison. *Information processing and Management*, 33(3):319–337, 1997. 14
- K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *Proc. CVPR*, volume 1, pages 26–36. Citeseer, 2006. 25
- G.A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 134
- B. Moghaddam and M.H. Yang. Gender classification with support vector machines. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 306–311. IEEE, 2000. 19
- A.S.S. Mohamed, Y. Weng, S.S. Ipson, and J. Jiang. Face detection based on skin color in image by neural networks. In *Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on*, pages 779–783. IEEE, 2007. 16, 17
- T. Mondal, A. Nath, A. Das, and M. Banerjee. An approach of face detection using geometrical definition of human face. 17
- P. Muller and A. Reymonet. Using inference for evaluating models of temporal discourse. 2005. 73
- J. Mutch and D.G. Lowe. Multiclass object recognition with sparse, localized features. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11–18. Citeseer, 2006. 25

REFERENCES

- H.H. Nagel. Steps toward a cognitive vision system. *AI Magazine*, 25(2):31, 2004. [31](#)
- M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151, 2001. [2](#)
- M.R. Naphade and J.R. Smith. On the detection of semantic concepts at TRECVID. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667. ACM New York, NY, USA, 2004. [15](#), [23](#)
- R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 4, pages 39–39. IEEE, 2003. [73](#)
- C.W. Ngo, T.C. Pong, H.J. Zhang, et al. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, 4(4):446–458, 2002. [14](#)
- J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008. [33](#)
- I. Nwogu, Y. Zhou, and C. Brown. Disco: Describing images using scene contexts and objects. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011. [94](#)
- J. Oh, Q. Wen, S. Hwang, and J. Lee. Video abstraction. *Video data management and information retrieval*, pages 321–346, 2005. [49](#)
- P.P. Ohanian and R.C. Dubes. Performance evaluation for four classes of textural features. *Pattern Recognition*, 25(8):819–833, 1992. [14](#)
- Y.I. Ohta, T. Kanade, and T. Sakai. Color information for region segmentation. *Computer graphics and image processing*, 13(3):222–241, 1980. [26](#)
- AJ OToole, H. Abdi, K.A. Deffenbacher, and D. Valentin. Low-dimensional representation of faces in higher dimensions of the face space. *JOSA A*, 10(3):405–411, 1993. [19](#)
- P. Over, A.F. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the international workshop on TRECVID video summarization*, page 15. ACM, 2007. [13](#)
- P. Pal, A.N. Iyer, and R.E. Yantorno. Emotion detection from infant facial expressions and cries. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 2006. [21](#)
- C.H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998. [33](#)
- C.H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000. [33](#)

-
- A. Payne and S. Singh. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition*, 38(10):1533–1545, 2005. 26
- S. Phimoltares, C. Lursinsap, and K. Chamnongthai. Face detection and facial feature localization without considering the appearance of image context. *Image and Vision Computing*, 25(5):741–753, 2007. 16
- R. Polana and R. Nelson. Low level recognition of human motion. In *Proc. IEEE Workshop on Nonrigid and Articulate Motion*, pages 77–82. Citeseer, 1994. 23
- M.F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1993. 47
- J. Pustejovsky, J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. Timeml: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*, 2003. 72
- J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. The specification language timeml. *The Language of Time: A Reader*. Oxford University Press, Oxford, 2004. 73
- R.J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *Image Processing, IEEE Transactions on*, 14(3):294–307, 2005. 17, 18
- N. Ramanathan and R. Chellappa. Face verification across age progression. *Image Processing, IEEE Transactions on*, 15(11):3349–3361, 2006. 18
- N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009. 18
- Z. Rasheed and M. Shah. Scene detection in Hollywood movies and TV shows. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE Computer Society; 1999, 2003. 25
- E. Reiter and R. Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(01):57–87, 1997. 29, 78, 79
- E. Reiter and R. Dale. *Building natural language generation systems*. Cambridge Univ Pr, 2000. 28, 29
- M. Rizon, M. Karthigayan, S. Yaacob, and R. Nagarajan. Japanese face emotions classification using lip features. In *Geometric Modeling and Imaging, 2007. GMAI'07*, pages 140–144. IEEE, 2007. 20
- D.D. Roberts. The existential graphs of CS Peirce. *Mouton, The Hague*, 1973. 24

REFERENCES

- D. Rose and T.J. Clarke. Look who's talking: Visual detection of speech from whole-body biological motion cues during emotive interpersonal conversation. *Perception*, 38(1):153, 2009. [27](#)
- Y. Rui, A.C. She, and T.S. Huang. Modified Fourier descriptors for shape representation—a practical approach. In *In Proc of First International Workshop on Image Databases and Multi Media Search*, 1996. [14](#)
- M.S. Ryoo and J.K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. 2006. [73](#)
- M.S. Ryoo and J.K. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *International journal of computer vision*, 82(1):1–24, 2009. [83](#)
- G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. [33](#)
- J.R.J. Schirra, G. Bosch, CK Sung, and G. Zimmermann. From image sequences to natural language: a first step toward automatic perception and description of motions. *Applied Artificial Intelligence an International Journal*, 1(4):287–305, 1987. [72](#)
- C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE. [12](#), [21](#)
- N. Serrano, A. Savakis, and J. Luo. A computationally efficient approach to indoor/outdoor scene classification. *Pattern Recognition*, 4:40146, 2002. [26](#)
- J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. *Lecture Notes in Computer Science*, 4170:127, 2006. [24](#)
- J. Sivic, B. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, volume 1. Citeseer, 2005. [25](#)
- A.F. Smeaton, P. Over, and W. Kraaij. High-level feature detection from video in trecvid: a 5-year retrospective of achievements. *Multimedia Content Analysis*, pages 1–24, 2009. [3](#), [13](#)
- A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000. [15](#)
- J.R. Smith and S.F. Chang. Automated binary texture feature sets for image retrieval. In *Proc ICASSP-96*, 1996a. [14](#)
- J.R. Smith and S.F. Chang. Tools and techniques for color image retrieval. *Storage & Retrieval for Image and Video Databases IV*, 2670:426–437, 1996b. [14](#)

-
- J.R. Smith and S.F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia*, page 98. ACM, 1997. [14](#)
- C.G.M. Snoek, M. Worring, and A.W.M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM International Conference on Multimedia*, page 402. ACM, 2005. [24](#)
- R.C. Staunton. An analysis of hexagonal thinning algorithms and skeletal shape representation. *Pattern Recognition*, 29(7):1131–1146, 1996. [65](#)
- A. Stolcke. SRILM—an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904. Citeseer, 2002. [120](#)
- E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77(1-3):291–330, 2008. [24](#), [25](#)
- Z. Sun, G. Bebis, X. Yuan, and S.J. Louis. Genetic feature subset selection for gender classification: A comparison study. 2002. [19](#)
- J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao. Design sparse features for age estimation using hierarchical face model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008. [18](#)
- M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pages 42–51. IEEE, 1998. [26](#)
- S. Tamura, H. Kawai, and H. Mitsumoto. Male/female identification from 8 × 6 very low resolution face images by neural network. *Pattern Recognition*, 29(2):331–335, 1996. [19](#)
- P.N. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006. ISBN 0321321367. [47](#)
- Y.I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001. [20](#)
- Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1683–1699, 2007. [20](#)
- A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *Ninth IEEE International Conference on Computer Vision*, pages 273–280. IEEE, 2008. ISBN 0769519504. [30](#)
- M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. [17](#)

REFERENCES

- M. Verhagen, I. Mani, R. Sauri, R. Knippen, S.B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with tarsqi. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 81–84. Association for Computational Linguistics, 2005. [74](#), [121](#)
- P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. 2001. [25](#), [58](#)
- L. Walawalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma. Support vector learning for gender classification using audio and visual cues: A comparison. *Pattern Recognition with Support Vector Machines*, pages 35–43, 2002. [19](#)
- P. Wilkins, P. Kelly, Ó. Conaire, T. Foures, A.F. Smeaton, and N.E. O’Connor. Dublin City University at TRECVID 2008. 2008. [15](#)
- J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. Citeseer, 2005. [25](#)
- M. J. Wise. String similarity via greedy string tiling and running karp-rabin matching. *Online Preprint, Dec*, 1993. [119](#)
- P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006. Citeseer, 2006. [13](#)
- C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. Pfindex: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7): 780–785, 1997. [22](#)
- C. Xu, Y.F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *Multimedia, IEEE Transactions on*, 10(7):1342–1355, 2008. [33](#)
- J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR’92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992. [22](#)
- G. Yang and T.S. Huang. Human face detection in a complex background. *Pattern recognition*, 27(1):53–63, 1994. [16](#)
- Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 412–420. Citeseer, 1997. [33](#)
- Y. Yang, C.L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. [32](#)

-
- Y. Yang, C.L. Teo, H. Daumé III, C. Fermüller, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 32
- B.Z. Yao, X. Yang, L. Lin, M.W. Lee, and S.C. Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. 30, 31, 32
- J.H. Yoo, D. Hwang, and M.S. Nixon. Gender classification in human gait using support vector machine. In *Advanced concepts for intelligent vision systems*, pages 138–145. Springer, 2005. 65
- D.P. Young and J.M. Ferryman. Pets metrics: On-line performance evaluation service. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 317–324, 2005. 12
- A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111, 1992. 28
- Z. Zeng, W. Liang, H. Li, and S. Zhang. A novel video classification method based on hybrid generative/discriminative models. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 705–713, 2010. 33
- L. Zhang, M.U.G. Khan, and Y. Gotoh. Video scene classification based on natural language description. In *2nd IEEE Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, pages 942–949, 2011. 8
- L. Zhang, M.U.G. Khan, and Y. Gotoh. Video scene classification based on limited dataset of natural language descriptions. In *Submission to the Journal of Zhejiang University*, 2012. 8
- W. Zheng, H. Tang, Z. Lin, and T. Huang. Emotion recognition from arbitrary view facial images. *Computer Vision–ECCV 2010*, pages 490–503, 2010. 30
- X. Zhou, X. Huang, B. Xu, and Y. Wang. Real-time facial expression recognition based on boosted embedded hidden markov model. In *Image and Graphics, 2004. Proceedings. Third International Conference on*, pages 290–293. IEEE, 2004. 20

REFERENCES

Appendix A

Video Annotation Tool

Figure A.1 shows a screen shot of the video annotation tool developed for this research, which is referred to as *Video Description Tool* (VDT). VDT is simple to operate and assist annotators in creating quality annotations. There are three main items to be annotated. An annotator is shown one video segment at one time. Firstly a restricted list of HLFs is provided for each segment and an annotator is required to select all HLFs occurring in the segment. Second, a title should be typed in. A title may be a theme of the video, typically a phrase or a sentence with several words. Lastly, a full description of video contents is created, consisting of several phrases and sentences. During the annotation, it is possible to stop, forward, reverse or play again the same video if required. Links are provided for navigation to the next and the previous videos. An annotator can delete or update earlier annotations if required.

Figure A.2 presents the screen shot of instructions page provided to the participants. First part gives introduction of dataset and explains different categories to help participants understand the nature of the dataset. The annotations are made with open vocabulary — that is, they can use any English words as long as they contain only standard (ASCII) characters. They should avoid using any symbols or computer codes. Annotators were further guided not to use proper nouns (*e.g.*, do not state the person name) and information obtained from audio. They were also instructed to select all HLFs appeared in the video.

Finally, they were presented with a sample video segment. Some of the possible textual annotations were provided with a title and complete description about the video stream. Selection of the HLFs depicted in the video segment was also shown to help the participants understand annotation generation procedure.

A. VIDEO ANNOTATION TOOL

The screenshot shows the Video Annotations Software interface. At the top, there are logos for The University of Sheffield and SPandH, and the title "Natural Language Description of Visual Images". The main area features a video player showing a man in a forest. Below the player are input fields for "Title Sentence" and "Description", a "Save Annotation" button, and a "Select list of HLFs" section. At the bottom, there is a "Previous Annotations" table and an "Edit" button.

Annotations:

- Instructions:** takes to Instructions.html page
- Previous:** Navigate to previous video
- Next:** Navigate to next video
- Title Sentence:** Enter title for the complete video
- Description:** Describe video based on objects, actions and their interations. Human gender, emotion, acitvites are required.
- Save Annotation:** Saves title and annotation in database
- Previous Annotations:** Display your already done annotations for the same video
- Edit:** You can update your annotations here


Select list of HLFs

Gender	Age	Action	Emotion	Objects	Scene Settings
<input checked="" type="checkbox"/> Male	<input type="checkbox"/> Baby	<input type="checkbox"/> Walk	<input checked="" type="checkbox"/> Happy	<input type="checkbox"/> Car	<input type="checkbox"/> Indoor
<input type="checkbox"/> Female	<input type="checkbox"/> Young	<input checked="" type="checkbox"/> Run	<input type="checkbox"/> Sad	<input type="checkbox"/> Cycle	<input checked="" type="checkbox"/> Outdoor
	<input checked="" type="checkbox"/> Adult	<input type="checkbox"/> Jog	<input type="checkbox"/> Serious	<input type="checkbox"/> Table	
	<input type="checkbox"/> Old	<input type="checkbox"/> Sit	<input type="checkbox"/> Surprise	<input type="checkbox"/> Aero-Plane	
		<input type="checkbox"/> Stand	<input type="checkbox"/> Angry	<input type="checkbox"/> TV-monitor	
		<input type="checkbox"/> Hand Wave	<input type="checkbox"/> Frightened	<input type="checkbox"/> Chair	
		<input type="checkbox"/> Clapping	<input type="checkbox"/> Disgust	<input type="checkbox"/> Bus	
		<input type="checkbox"/> Fight		<input type="checkbox"/> Bottle	
				<input type="checkbox"/> Sofa	

Previous Annotations:


Edit	Title	Description
Edit	man is running	man is running in the jungle
Edit	man is running	man is wearing heavy clothes

Figure A.1: Video Description Tool (VDT). An annotator watches one video at one time, selects all HLFs present in the video, describes a theme of the video as a title and creates a full description for important contents in the video.




The University of Sheffield.

Natural Language Description of Visual Images



Instructions:

1. There are about 140 video clips of 10 to 30 seconds;
2. There are seven categories of videos:
 - i) Action Videos: Human posture is visible and performing some action like sit, stand, walk, run.
 - ii) Close Ups: Human face is visible.
 - iii) News: Presence of anchor or reporter and scene settings like presence of weather boards at the background.
 - iv) Meeting: Multiple humans sitting and interacting and objects like chairs, table.
 - v) Traffic: Presence of vehicles like cars, buses, trucks. Traffic signals scenes.
 - vi) Grouping: Multiple human interaction scenes other than meeting scenarios. Table, chairs are not present.
 - vii) Indoor/Outdoor: Scene settings are much obvious than human activities. Park scenes, office scenes(computers and files are visible)
3. The task is to provide each video with
 - (a) a title -- one sentence indicating the main theme of the video.
 - (b) a description with multiple (say, 4 to 6) sentences.
 - (c) Certain high level feature lists are provided; select the ones which are shown in the video
 - i) Gender: Male, Female (If multiple humans appear and one is male other female then select both)
 - ii) Age: Baby, Young, Adult and Old (again there can be multiple persons and you have to select multiples)
 - iii) Action: Select all the actions as shown in the video.
 - iv) Emotions: Multiple humans can have multiple emotions and actions select all emotions and actions.
 - v) Objects: Select all the objects which are shown in the video.
 - vi) Indoor/outdoor: Every video can be either indoor or outdoor. Our interest is in those videos where there is no human present.
4. Each sentence should not be no longer than 5, 6 words;
5. Open Vocabulary; you can describe the visual image using any English word;
6. Title/description can include
 - (a) emotions, activities, actions, location, age, gender, dressing, ...
(eg. two persons are sitting on chairs; they are wearing dark clothes);
 - (b) non-human objects such as scene settings and background
(eg. there is food on the table, this is an indoor scene);
7. Do not use proper nouns (eg. do not state the personal name);
8. Do not use information obtained from audio.



- This video can be described by sentences like:

(title sentence)
Meeting between two persons.

(description with multiple sentences)
Two persons are sitting on chairs;
They are wearing formal clothes;
It seems some interview type meeting;
They are going to start food;
One person is speaking while other is silent.
- Checkboxes such as Male, Sit, Talk, Table, Chair can be clicked.

Gender	Age	Action	Emotion	Objects	Scene Settings
<input checked="" type="checkbox"/> Male	<input type="checkbox"/> Baby	<input type="checkbox"/> Walk	<input type="checkbox"/> Happy	<input type="checkbox"/> Car	<input checked="" type="checkbox"/> Indoor
<input type="checkbox"/> Female	<input type="checkbox"/> Young	<input type="checkbox"/> Run	<input type="checkbox"/> Sad	<input type="checkbox"/> Cycle	<input type="checkbox"/> Outdoor
	<input type="checkbox"/> Adult	<input type="checkbox"/> Jog	<input checked="" type="checkbox"/> Serious	<input checked="" type="checkbox"/> Table	
	<input checked="" type="checkbox"/> Old	<input checked="" type="checkbox"/> Sit	<input type="checkbox"/> Surprise	<input type="checkbox"/> Aero-Plane	
		<input type="checkbox"/> Stand	<input type="checkbox"/> Angry	<input type="checkbox"/> TV-monitor	
		<input type="checkbox"/> Hand Wave	<input type="checkbox"/> Frightened	<input checked="" type="checkbox"/> Chair	
		<input type="checkbox"/> Clapping	<input type="checkbox"/> Disgust	<input type="checkbox"/> Bus	
		<input type="checkbox"/> Fight		<input type="checkbox"/> Bottle	
				<input type="checkbox"/> Sofa	

Figure A.2: Instructions for annotation generation.

A. VIDEO ANNOTATION TOOL

Appendix B

Templates for Sentence Generation

B. TEMPLATES FOR SENTENCE GENERATION

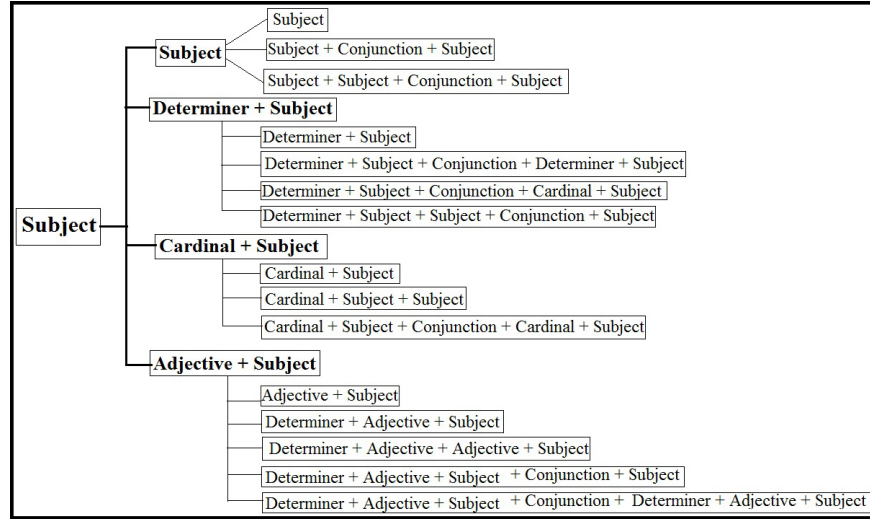


Figure B.1: Template selection where subject information is only available.

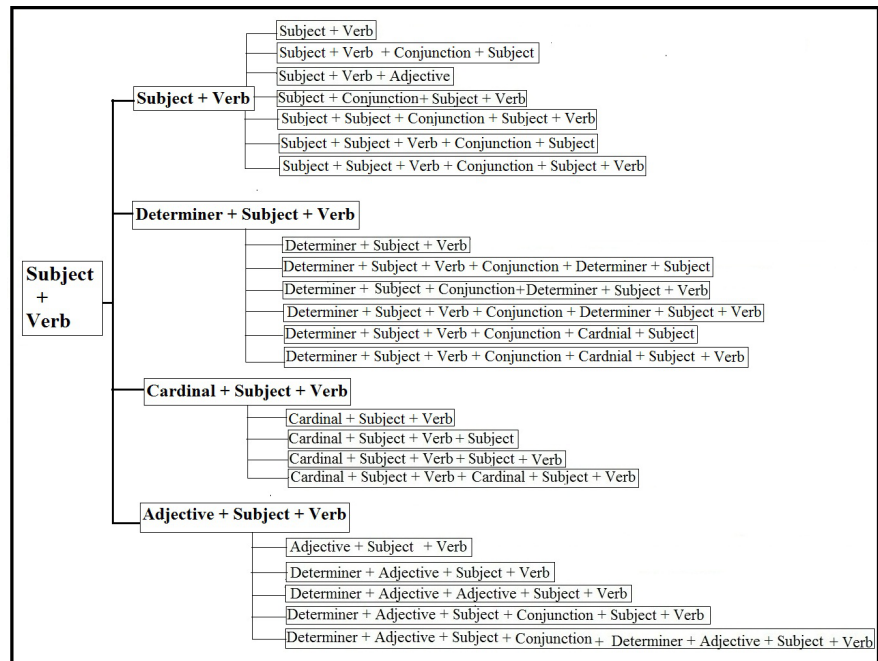


Figure B.2: Template selection where subject information is only available.

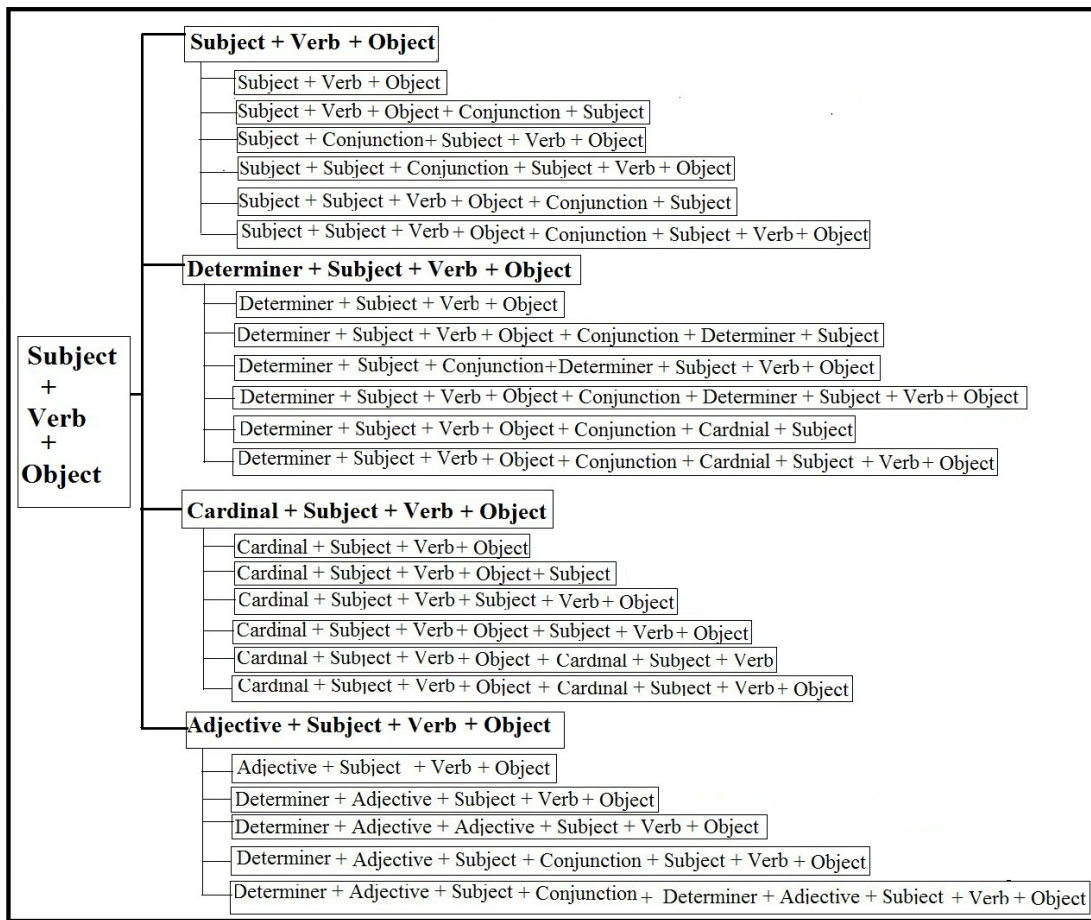


Figure B.3: Template selection where subject information is only available.