Feature Extraction and Representation for Human Action Recognition

by

Xiantong Zhen

Submitted to the Department of Electronic and Electrical Engineering
in partial fulfillment of the requirements for

Doctor of Philosophy

at

The University of Sheffield

September 2013

Feature Extraction and Representation for Human Action Recognition

by

Xiantong Zhen

Submitted to the Department of Electronic and Electrical Engineering
on September, 2013, in partial fulfillment of the requirements for
Doctor of Philosophy

Abstract

Human action recognition, as one of the most important topics in computer vision, has been extensively researched during the last decades; however, it is still regarded as a challenging task especially in realistic scenarios. The difficulties mainly result from the huge intra-class variation, background clutter, occlusions, illumination changes and noise. In this thesis, we aim to enhance human action recognition by feature extraction and representation using both holistic and local methods.

Specifically, we have first proposed three approaches for the holistic representation of actions. In the first approach, we explicitly extract the motion and structure features from video sequences by converting the video representation into a 2D image representation problem; In the second and third approaches, we treat the video sequences as 3D volumes and propose to use spatio-temporal pyramid structures to extract multi-scale global features. Gabor filters and steerable filters are extended to the video domain for holistic representations, which have been demonstrated to be successful for action recognition. With regards to local representations, we have firstly done a comprehensive evaluation on the local methods including the bag-of-words (BoW) model, sparse coding, match kernels and classifiers based on image-to-class (I2C) distances. Motivated by the findings from the evaluation, we have proposed two distinctive algorithms for discriminative dimensionality reduction of local spatio-temporal descriptors. The first algorithm is based on the image-to-class distances, while the second explores the local Gaussians.

We have evaluated the proposed methods by conducting extensive experiments on widely-used human action datasets including the KTH, the IXMAS, the UCF Sports, the UCF YouTube and the HMDB51 datasets. Experimental results show the effectiveness of our methods for action recognition.

To my parents and sisters

Acknowledgments

First of all, we would like to thank my supervisor Dr. Ling Shao for his excellent supervision, without whom I could not have had the chance to do my PhD in Sheffield. It has been my fantastic experience to be working with him. On the one hand, he has given me maximum independence in pursuing my ideas and lots of encouragement. On the other hand he has provided thorough advice on research innovation and drafting papers as well.

I would like to thank my fellow colleagues: Simon Jones, Ruomei Yan, Di Wu, Muhammad Mubashir (MPhil), Fan Zhu, Li Liu, Bo Dong and Mengyang Yu, who have been so helpful during my PhD study. Fan always tries to help me and encourage me when I have difficulties. Simon helped me a lot when I just came and joined the group.

I would also like to thank other members in B28 including Zia Khan, Ji Ni and Russ Drieberg for their help and kindness. My fellow Zia helped me a lot with my study and life in Sheffield.

I would like to thank Feng Zheng (Civil and Structural Engineering). I have been so lucky to be working with him. Feng is a wonderful fellow to collaborate with and I can always find constructive suggestions from him when discussing with him. I have learned a lot from him.

I would like to thank Prof. Lei Zhang. It was my valuable experience to be working with her. I thank Dr. Jun Tang who has been encouraging me and helping me a lot. I have learned very much from him.

I thank my friends, Dr. Jing Li and Dr. Wenting Duan, for their consistent encouragement. I thank my friends, Dr. Ning Ma and Ms. Chunjie Bi, a wonderful couple, for their help in my life in Sheffield. I also thank all my friends, near and far, for their support and care.

I specially thank Postgraduate Administrator, Ms. Hilary J Levesley, who has always been so nice and helpful. Thanks also go to Technical Manager Mr. James Screaton and Safety Officer Ms. Dianne Webster for their support.

I would like to thank my examiners Dr. Krystian Mikolajczyk and Dr. Peter Rockett for their tough work of reviewing thesis. I am particularly grateful for the valuable comments and suggestions from Dr. Peter Rockett.

Thanks to the China Scholarship Council (CSC) for the financial support of my PhD study in Sheffield.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human action recognition

Human action recognition and analysis [41], one of the most active topics in computer vision, has drawn increasing attention and its applications can be found in video surveillance, video annotation and retrieval, and human-computer interaction, etc. The goal of action recognition is to automatically analyse ongoing activities from an unknown video [2]. In the last a few decades, action recognition has been extensively researched while there is still a long way to go for real applications. The challenges of human action recognition come from difficulties such as great intra-class variance, scaling, occlusion and clutter. Human action recognition has been extensively researched through methods based on local and holistic representations.

1.1.1 Local representations

Methods based local representation, also known as local methods, encode a video sequence as a collection of local spatio-temporal features (local descriptors). These local descriptors are extracted from spatio-temporal interest points (STIPs) which can be sparsely detected from video sequences by detectors [59, 25, 85]. In contrast to holistic representations of human actions, local methods enjoy many advantages.

- Avoidance of some preliminary steps,*e.g.*, background subtraction and target

tracking required in holistic methods.

- Resistance to background variation and occlusions. However, local representations also have deficiencies, of which a key limitation is that it can be too local, as it is not possible to capture adequate spatial and temporal information.

Local descriptors can also be obtained from trajectories. The early work by Johnsson [51] indicates that it is sufficient to distinguish human actions by the tracking of joint positions. One of the advantages of using trajectories is being discriminative [51]. Nevertheless, the performance of trajectory-based methods depends on the quality of these trajectories, and in practice extracting trajectories from video sequences would be computationally expensive.

To obtain the final representation of an action, the bag-of-words (BoW) model has been widely used and has achieved good results in human action recognition tasks. The BoW model is actually based on mapping local features of each video sequence onto a pre-learned dictionary, which unavoidably introduces quantisation errors during its creation. The errors would be propagated to the final representation and harm the recognition performance. Additionally, the size of this dictionary needs to be empirically determined, and codewords, *i.e.*, the cluster centres, obtained by k-means, gather around dense regions of local feature space, resulting in less effective codewords of action primitives. Sparse representation has recently been introduced for action representation based on local features [35].

1.1.2 Holistic representations

Methods based on holistic representation, called global methods, treat a video sequence as a whole rather than applying sparse sampling using STIP detectors or extracting trajectories. In holistic representations, spatio-temporal features are directly learned from raw frames in video sequences. Holistic representations have recently drawn increasing attention [47, 111, 71], because they are able to encode more visual information by preserving spatial and temporal structures of actions occurring in a video sequence.

However, holistic representations are highly sensitive to partial occlusions and

background variations. Additionally, they often require pre-processing steps, such as background subtraction, segmentation and tracking, which makes it computationally expensive and even intractable in some realistic scenarios.

1.2 Literature review

In this section, we review related works in two aspects: spatio-temporal features and human action representations.

1.2.1 Spatio-temporal features

Low-level features play a fundamental role in both local and holistic representations of human actions. In the last decades, many spatio-temporal descriptors have been proposed and shown to be effective for action recognition.

Histograms of optical flow (HOF) and histograms of oriented gradients (HOG) are combined by Laptev *et al.* [61] as a descriptor, HOGHOF, which is demonstrated to be better than either of HOG or HOF as a single descriptor.

The 3D gradient is directly extended from its counterpart in the 2D domain, and a histogram of oriented 3D gradients (HOG3D) [54] as a descriptor has been applied to many action recognition tasks. Similarly, the idea of the scale invariant feature transform (SIFT) was extended to spatio-temporal video sequences as a 3D-SIFT descriptor [99].

Inspired by the local binary patterns (LBP), Yeffet and Wolf [139] proposed a global descriptor named local trinary patterns (LTP), which is successfully used for action recognition.

A biologically inspired features based on Gabor filters were first exploited for human action recognition by Jhuang *et al.* [46]. They introduced a biological model of motion processing based on a hierarchical feed forward architecture [91]. The model extends a neurobiological process of motion processing in the visual cortex and considers space-time gradients based and optical flow based S1 units [46]. Their work demonstrates the potential of biologically inspired features for human action

recognition. Based on the spatio-temporal features, a great number of local and holistic approaches have been proposed for human action representations in the past decades.

1.2.2 Local representations

In this section, we review state-of-the-art local methods for action recognition based on spatio-temporal interest points and trajectories. In order to compensate for the loss of structure in local representations, a lot of methods try to improve local representations by exploring spatio-temporal structural information [30], including context information of each interest point [109, 123], relationships between/among spatio-temporal interest points [32, 76, 123, 141] and neighbourhood-based features [57]. The relationship/co-occurrence among visual words in the BoW model and their semantic meaning have also be explored to encode higher-level features [67, 145, 72, 128]. New local descriptors have also be developed [80, 62, 43] to improve the performance of local methods. In addition, aiming to alleviate the quantisation errors in the BoW model, sparse coding has also been introduced into action recognition to learn more compact and richer representations of human actions [35]. In the following, we will give a more detailed description of these methods.

Sun *et al.* [109] proposed to model the spatio-temporal context information in a hierarchical way by exploiting three levels of context, *i.e.*, point-level, intra-trajectory and inter-trajectory context. In their work, trajectories are first extracted using a Scale Invariant Feature Transform (SIFT). The point-level context is the average of SIFT descriptors extracted at the salient points on the trajectory. Intra-trajectory and inter-trajectory context is modelled by the transition matrix of a Markov process and encoded as the trajectory transition and trajectory proximity descriptors.

In order to capture the most informative spatio-temporal relationship between local descriptors, Kovashka and Grauman [57] proposed to learn a hierarchy of spatio-temporal neighbourhood features. The main idea is to construct a higher-level vocabulary from new features that consider the hierarchical neighbouring information around each interest point.

Matikainen *et al.* [76] proposed expressing pair-wise relationships between quantised features by combining the power of discriminative representations with key aspects of Naive Bayes. The relationship between local features is modelled as the distribution of quantised location differences between each pair of interest points. Two basic features namely STIP-HOG and quantised trajectories are considered.

Gaur *et al.* [30] modelled the activity in a video as a "string of feature graphs" (SFGs) by treating a video as a spatio-temporal collection of primitive features (*e.g.*, STIP features). They divide the features into small temporal bins and represent the video as a temporally ordered collection of such feature-bins, each bin consisting of a graphical structure representing the spatial arrangement of the low-level features. A video then becomes a string of such graphs and comparing two videos is to match two strings of graphs.

Claiming that the higher-order semantic correlation between mid-level features (*e.g.*, from the BoW representation) is useful to fill the semantic gap, Lu *et al.* [72] proposed novel spectral methods to learn latent semantics from abundant mid-level features by spectral embedding with nonparametric graphs and hypergraphs. A new semantics-aware representation (*i.e.*, histogram of high-level features) is derived for each video from the original BoW representation, and actions are classified by an SVM with a histogram intersection kernel based on the new representation.

Wang *et al.* [123] presented a novel local representation by augmenting local features with contextual features which capture the interactions between interest points. Multi-scale channels of contextual features are computed and, for each channel, a regular grid is used to encode spatio-temporal information in the local neighbourhood of an interest point. Multiple kernel learning is employed to integrate the contextual features from different channels.

Aiming to encode rich temporal ordering and spatial geometry information of local visual words, Zhang *et al.* [145] proposed modelling the mutual relationships among visual words by a novel concept named the spatio-temporal phrase (ST phrase). A ST phrase is defined as a combination of k words in a certain spatial and temporal structure including their order and relative positions. A video is represented as a bag

of ST phrases which is shown to be more informative than the BoW model.

In order to capture the geometrical distribution of interest points, Yuan *et al.* [141] applied the 3D R transform on the interest points based on their 3D locations. The 3D R transform is invariant to geometrical transformation and robust to noise. $(2D)^2$ PCA is then employed to reduce the dimensionality of the 2D feature matrix from the 3D R transform, obtaining the so-called R features. To encode the appearance features, they combined the R features with the BoW representation. Finally, they proposed a context-aware fusion method to efficiently fuse these two features. Specifically, one feature is used to compute the context of each video and the other to calculate the context-aware kernel for action recognition.

In the BoW model, mid-level features are obtained by k-means clustering which however is unable to capture the semantic relation between low-level features due to only appearance similarity being used. Liu *et al.* [67] proposed using diffusion maps to automatically learn a semantic visual vocabulary from abundant quantised mid-level features. Each mid-level feature is represented by the vector of point-wise mutual information (PMI). Diffusion maps can capture the local intrinsic geometric relations between the mid-level feature points on the manifold.

With the argument that visual words from video sequences belonging to the same class in the BoW model are correlated and jointly reflect a specific action type, Wang *et al.*, [128], by assuming that visual words share a common structure in a low-level space, presented a framework named semi-supervised feature correlation mining (SFCM) to exploit the shared structure. A discriminative and robust classifier for action annotation is trained by taking into account the global and local structural consistency.

Shapovalova *et al.* [103] proposed modelling a video using a global bag-of-words histogram based on local features, combined with a bag-of-words histogram focused latent regions-of-interest. The latent regions of interest are spatio-temporal sub-regions of a video. The model parameters are learned by a similarity constrained latent SVM in which the constraint is to enforce that the latent regions chosen across all videos of a class are coherent.

Instead of using hand-crafted features such as HOGHOF, HOG3D and MBH [123], Le *et al.* [62] introduced an unsupervised deep learning algorithm, named Independent Subspace Analysis (ISA), which learns spatiotemporal features of interest points from unlabelled videos. Convolution and stacking are adopted in the deep learning model to scale the algorithm to large images and learn hierarchical representations.

As indicated by Wang *et al.* [125], dense sampling tends to produce better results than sparsely detected spatio-temporal interest points. Wang *et al.* [123] presented an approach by dense trajectories. Points are densely sampled from each frame and tracked based on displacement information from a dense optical flow field. A novel descriptor based on motion boundary histograms was introduced in their work to encode the trajectory information. The remarkable performance of dense trajectories is largely due to the rich description of scene and contextual information of dense sampling, and the robust extraction of motion information of trajectories.

Also based on dense trajectories, Jiang *et al.* [50] presented a new video representation that integrates trajectory descriptors with the pair-wise trajectory locations as well as motion patterns. Global and local reference points are adopted to characterise motion information with the aim to be robust to camera movements.

Motion is regarded as the most reliable source of information for human action recognition, as it is related to the regions of interest. Jain *et al.* [43] introduced the Divergence-Curl-Shear (DCS) descriptor to encode scalar first-order motion features. These contain the motion divergence, curl and shear, which capture physical properties of the flow pattern. To handle the noisy motion from background and the unstable camera, an affine model is employed for motion compensation to improve the quality of descriptors. Dense trajectories are also used and the Vector of Locally Aggregated Descriptors (VLAD) is used for the final encoding of local features which is shown to be better than a standard BoW model.

Although dense sampling shows increasing performance with the decrease of the sampling step size, it does not scale well with a large number of local patches and becomes computationally intractable for large-scale video datasets. Vig *et al.* [121] proposed selecting informative regions and descriptors by saliency-mapping algorithms.

These regions are either used exclusively or given greater representational weights. By using saliency-based pruning, up to 70% of descriptors can be discarded while maintaining high performance on the Hollywood2 dataset.

Rather than using the BoW model, Guha and Kreidieh [35] introduced sparse representations into the context of human action recognition in video. Over-complete dictionaries are learned from a set of local spatio-temporal descriptors in the training set. It is claimed that the obtained representation based on the dictionaries learned by sparse coding is more compact compared with the BoW model involving clustering and vector quantisation. Three options for dictionaries, *i.e.*, shared, class-specific and concatenated, were investigated.

1.2.3 Holistic representations

Holistic representations play an important role in human action recognition because of their good ability to preserve the structural information of actions.

Bobick *et al.* [8] presented temporal templates through projecting frames onto a single image, namely motion history image (MHI) and motion energy image (MEI). MHI indicates how motion happens and MEI records where it is. This representation gives satisfactory performance under the circumstance where the background is relatively static.

Efros [26] introduce a new motion descriptor based on smoothed and aggregated optical flow measurements over a spatio-temporal volume centred on a moving figure, and an associated similarity measure to be used in a nearest-neighbour framework. To classify the action being performed by a human figure in a query sequence, they retrieved nearest neighbour(s) from a database of stored, annotated video sequences.

Yilmaz and Shah [140] proposed modelling an action based on both the shape and the motion of the performing object. A spatio-temporal volume (STV) is generated from a sequence of 2D contours with respect to time. STV is analysed by the differential geometric surface properties to identify action descriptors capturing both spatial and temporal information. Finally, they formulated the action recognition as graph theoretical problem.

Regarding human actions in video sequences as three-dimensional shapes induced by the silhouettes in space-time volumes, Gorelick *et al.* [33] extracted spatio-temporal features, *e.g.*, local space-time saliency, action dynamics, shape structure and orientation, based on the properties of Possion equation solutions.

Rodriguez *et al.* [92] introduced a template-based method in which a maximum average correlation height (MACH) filter is generalised to analyse videos as 3D spatio-temporal volumes in the frequency domain. MACH is capable of capturing intra-class variability by synthesizing a single action MACH filter for a given action class.

Ali and Shah [3] proposed a set of kinematic features derived from optical flow for holistic representation of actions. The set of kinematic features includes divergence, vorticist, symmetric and antisymmetric flow fields, second and third principal invariants of flow gradient and rate of strain tensor, and third principal invariant of rate of rotation tensor. Each kinematic feature, when computed from the optical flow of a sequence of images, gives rise to a spatiotemporal pattern.

Neural networks and deep learning algorithms are also exploited for learning spatio-temporal global features for holistic representations. By extending restricted Boltzmann machines (RBMs) [39] to the spatio-temporal domain, Taylor *et al.* [111] proposed a novel convolutional gated restricted Boatman machine (GRAM) for learning spatio-temporal features. A probabilistic max pooling technique was adopted into their model.

Similarly, Jr *et al.* [47] developed a 3D convolutional neural network (CNN) based on a two-dimensional model for feature extraction. In a 3D CNN, motion information in multiple adjacent frames is captured through performing convolutions over spatial and temporal dimensions. However, similar to [111, 62], in this model the number of parameters to be adjusted is very large, sometimes too large relative to the available number of training samples, which unfortunately restricts its applicability.

Recently, Saraband and Cors [95] presented a high-level representation, *i.e.*, action bank, in which oriented energy features are used to generate action templates for bank detectors and a spatiotemporal orientation decomposition is realised using broadly tuned 3D Gaussian third derivative filters.

1.3 Motivations

In this thesis, we aim to address action recognition by both holistic and local methods. Most of the holistic methods borrow the ideas from the image domain which have demonstrated to be effective in the video domain for action recognition. However, the multi-scale analysis which plays a key role in image processing and analysis has not been investigated in the video domain. To fill the gap, we adopt the multi-scale analysis, *i.e.*, the Gaussian/Laplacian pyramids into the video domain. Combining with the orientation analysis, *e.g.*, Gabor filters and steerable filters, we propose two global descriptors for holistic action representations. The proposed holistic methods are presented in Chapter 2, 3 and 4.

With regards to local methods, most of state-of-the-art methods remain in the frameworks of the bag-of-words (BoW) model and the sparse coding (SC) algorithm, without considering to improve local descriptors, *e.g.*, histograms of three-dimensional gradients (HOG3D) [54], the three-dimensional SIFT (SIFT3D) [99]. Based on the findings in our evaluation work in Chapter 5, we go beyond the BoW and SC frameworks and propose two dimensionality-reduction methods based on the image-to-class (I2C) distances [9] and local Gaussians to improve the effectiveness of local descriptors. The proposed local methods are presented in Chapter 6 and 7.

1.4 Datasets

In the past decade, many action datasets, including the Weighman, KTH, IXMAS, UCF Sports, UCF-Youthes, Hollywood and HMDB51 dataset, have been released for public use. We give detailed description of the datasets mainly used in this thesis. The sample frames from these datasets are illustrated in Fig. 1-1.

1.4.1 The KTH dataset

KTH [98] is a commonly used benchmark action dataset with 599 video clips. Six human action classes, including walking, jogging, running, boxing, hand waving and

handicapping, are performed by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors with lighting variation (s4). In our holistic methods (Chapter 2, 3, and 4), we use one cycle of the actions in each video clip with the bounding boxes obtained by the work in [138]. The evaluation is based on the leave-one-sample (actor) validation framework. In our local methods, we use the raw video sequences with four cycles. We follow the original experimental setup in [98], *i.e.*, divide the samples into a test set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and a training set (the remaining 16 subjects).

1.4.2 The IXMAS dataset

IXMAS [130] contains 11 action classes. Each action is repeatedly executed three times by ten actors and recorded by five cameras simultaneously. These actions contain checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking and picking up. We treat five cameras separately, namely, training and testing are performed on single view. We use the silhouettes released with the dataset and the same leave-one-out evaluation scheme is adopted on this dataset [130].

1.4.3 The UCF Sports dataset

UCF Sports [92] is a collection of 150 broadcast sports videos of ten different types of actions, including swinging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging at the high bar, golf swinging and walking. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. Due to the unequal numbers of sequences in each action category, we follow the original setting in [92] and adopt five-fold cross-validation with one-fifth of the total number of sequences in each category for testing.

1.4.4 The UCF YouTube dataset

UCF YouTube [67] contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions. The dataset contains a total of 1168 sequences. We follow the original setup [67] using leave-one-out cross validation for a pre-defined set of 25 folds.

1.4.5 The HMDB51 dataset

The Human Motion Database (HMDB51) [58] contains 51 distinct categories with at least 101 clips in each for a total of 6766 video clips extracted from a wide range of sources. The action categories can be grouped in five types: 1) General facial actions: smile, laugh, chew, talk; 2) Facial actions with object manipulation: smoke, eat, drink; 3) General body movements: cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave; 4) Body movements with object interaction: brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw; 5) Body movements for human interaction: fencing, hug, kick someone, kiss, punch, shake hands, sword fight. We follow the experimental settings using three training/test splits and report the average.

1.4.6 Thesis road map

The remainder of this thesis is organised as follows. Chapters 2, 3 and 4 present three holistic methods. In Chapter 2, we describe the method of embedding motion and structure features. In Chapter 3, the spatio-temporal Laplacian pyramid coding (STLPC) is presented which extends the techniques including the multi-scale analysis (the Laplacian pyramid) and orientational analysis (Gabor filters) in the image

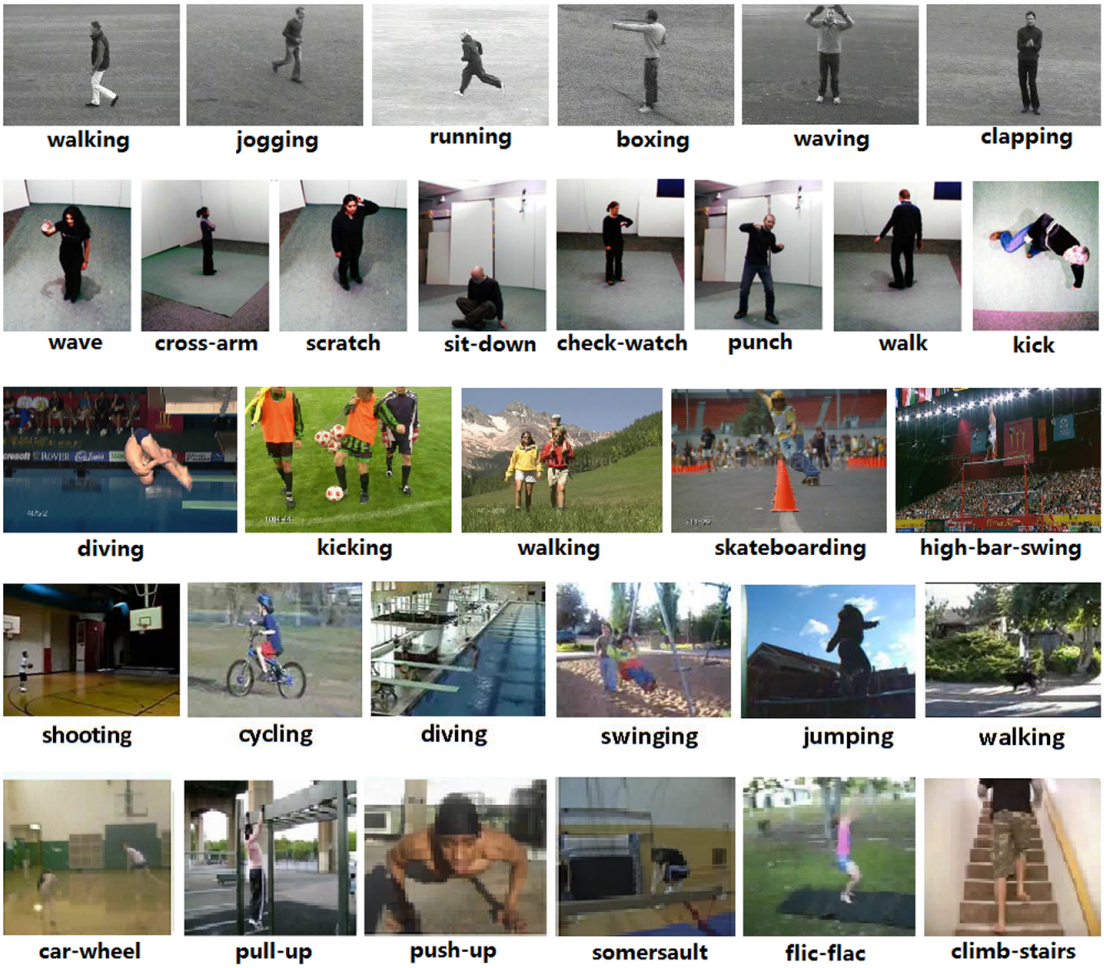

Figure 1-1: Each row illustrates the sample frames respectively from the KTH, IXMAS, UCF Sports, UCF YouTube and HMDB51 datasets.

domain. In Chapter 4, we present a representation of spatio-temporal oriented energies, which is based on the 3D steerable filters. Chapters 5, 6, and 7 deal with local methods. In Chapter 5, we have done a performance evaluation of local methods. Based on the findings in Chapter 5, in Chapters 6 and 7, we propose two dimensionality reduction techniques based on image-to-class distances and local Gaussians, respectively. In Chapter 8, we summarise our work in this thesis and outline the most important directions for future work.

Chapter 2

Motion and Structure Feature Embedding

2.1 Introduction

Actions as spatio-temporal patterns in video sequences contain mainly motion and structure information. Motion refers to the movement of actions while structures, the poses of human body, and the relative positions among them. In this chapter, we propose a unified framework to explicitly encode features of motion and structure of an action, and embed them as a holistic representation.

2.1.1 Motivations

Representations based on detected spatio-temporal interest points are drawing much attention [59], [25], [85], [133], [102]. Human action recognition systems based on the bag of words (BoW) model have achieved good results in many tasks.

Nevertheless, this model also suffers some limitations, one of which is the inability to capture adequate spatial and temporal structure information. Since the BoW model is actually based on mapping local features of each video sequence onto a pre-learned dictionary, it inevitably introduces information loss and errors during quantisation of continuous distributions into bins; the errors would be propagated to the

final representation and compromise the recognition performance. The effectiveness of the codebook would be dependent on the clustering algorithm [36]. Additionally, the size of the codebook needs to be empirically determined which is less flexible for different tasks.

To alleviate the above-mentioned shortcomings, we propose a unified framework for human action representation. Based on the fact that human actions mainly comprise motion and structure information [33, 8], we explicitly extract those features from video sequences and integrate them into a holistic representation. Motion history images (MHI) are used to extract motion information because of their effectiveness of capturing action and computational efficiency [8]. The motivations for choosing five planes to encode structure information lie in two aspects: (1) the three slices of the three orthogonal planes (TOP) record the spatial and temporal structure information simultaneously; (2) the starting and ending slices combined with the middle slice in TOP could provide dynamic structure information about the action.

2.1.2 Overview

Our framework takes the following steps. Firstly, inspired by the work in [13], we apply a preprocessing step, differencing adjacent frames, to the raw video sequence and a 3D volume with difference of frames (DoF) is obtained. Consequently, the action-related information is well preserved and background is largely suppressed.

Secondly, motion and structure information (Section 2.2) are separately extracted in two feature channels. In the motion feature channel, one feature map, *i.e.*, the motion history image (MHI) is obtained which encodes the motion information. In the structure feature channel, five feature maps are extracted from the DoF volume. These five feature maps are the three orthogonal planes with the intersection point falling on the centre of the volume and the starting and ending slices of the volume.

Thirdly, given the feature maps (5 structure planes + 1 motion template = 6 feature maps), each of which is actually an 2D image, a multi-scale analysis technique, *i.e.*, the Gaussian pyramid, is employed due to its success in image processing and analysis [14]. A Gaussian pyramid (Section 2.3) is applied to each of the obtained

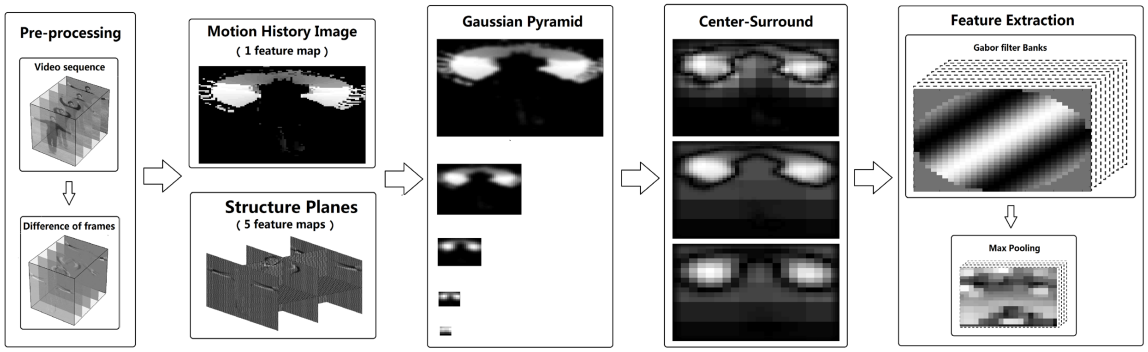

Figure 2-1: Schematic overview of the feature extraction from a raw video sequence.

six feature maps, following which the centre-surround operation [107] is performed on each level of the Gaussian pyramid resulting in a series of sub-band maps. Features with different scales are segregated into different bands.

Subsequently, a two-stage feature extraction step [45] (Section 2.4), namely Gabor filtering and max pooling, is used to select invariant features. Gabor filters are widely used, a common choice of filter bank for the first stage in feature extraction [100], and can capture edge and orientation information [46, 97, 107]. Feature pooling techniques, *e.g.*, the max pooling, have drawn more attention in low-level feature extraction algorithms due to their invariance properties [46, 10, 47, 137, 63].

The obtained features are biologically inspired in that both Gabor filtering and max pooling have biological mechanisms in common with the human visual system [46], [107].

Finally, a dimensionality-reduction technique named discriminative locality alignment (DLA) [144] (Section 2.5) is used to embed the motion and structure features into a low dimensionality space which leads to a more compact and discriminative representation. We will show experimentally that DLA outperforms the state-of-the-art reduction techniques.

Our feature extraction procedure is illustrated in Fig 2-1, in which the last three blocks, *i.e.*, Gaussian pyramid, centre-surround and feature extraction, depict the processes on one feature map. The other five feature maps are identically processed.

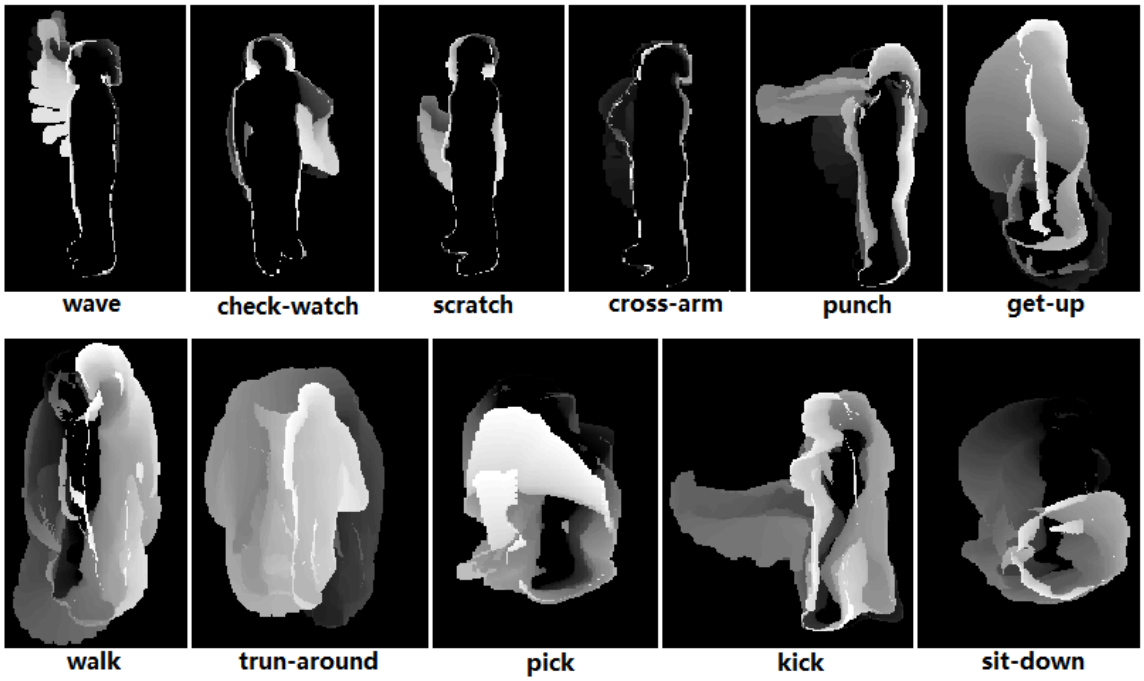

Figure 2-2: Examples of motion history images from the IXAMS dataset

2.1.3 Contributions

The main contributions of the proposed work fall in the following three aspects.

- A unified framework is proposed to integrate motion and structure information for human action representation. The essential cues of actions are effectively captured.
- Multi-scale analysis techniques, *i.e.*, the Gaussian pyramid and centre-surround operation, are introduced for feature extraction in action representation. Effective biologically-inspired features are extracted by Gabor filters and max pooling, obtaining a more informative and discriminative representation.

2.2 Feature maps

In order to explicitly extract the motion and structure features from video sequences, we take the advantages of the motion templates model, *i.e.*, motion histogram image (MHI) and the three orthogonal planes (TOP) used in dynamic texture analysis [149]. We encode the motion and structure information in a set of feature maps.

2.2.1 Motion templates

Motion history images (MHI) proposed by Bobby *et al.* [8] are used to represent the motions of an object in a video. All frames in the video sequence are projected onto one image across the temporal axis and recent motion is emphasised more than that happened in the past. Assume $I(x, y, t)$ is an image sequence and let $D(x, y, t)$ be a binary image sequence indicating regions of motion, which can be obtained from image differencing. The motion history image (MHI), $H_\tau(x, y, t)$, is used to represent how the motion image is moving, and is obtained with a simple replacement and decay operator:

$$H_\tau(x, y, t) = \{_{max(0, H_\tau(x,y,t-1)-1)}^{\tau} \quad {}^{if D(x,y,t)=1}_{otherwise} \tag{2.1}$$

where τ is the duration for defining the range of the motion. An example of MHI from the IXMAS dataset is shown in Fig 2-2.

2.2.2 Structure planes

Three orthogonal planes (TOP), namely XI, XT and YET planes, are orthogonal slices, of which the point of intersection falls on the centre of a DoF volume. Combining the three orthogonal planes and the starting and ending slices of the DoF volume, we obtain five structure planes. These planes contain both the spatial and temporal structures of an action. The three X-Y planes give the dynamic structure (three body poses) of the action, while X-T and Y-T planes record the temporal structures. Therefore, these five planes contain structure information complementary to each other. Fig 2-3 illustrates an example of structure planes.

2.3 Gaussian pyramid

The image pyramid is a data structure designed to support efficient scaled convolution through a reduced image representation. It consists of a sequence of copies of an original image in which both sample density and resolution are decreased in regular steps. A pyramid is a multi-scale representation formed by a recursive method.

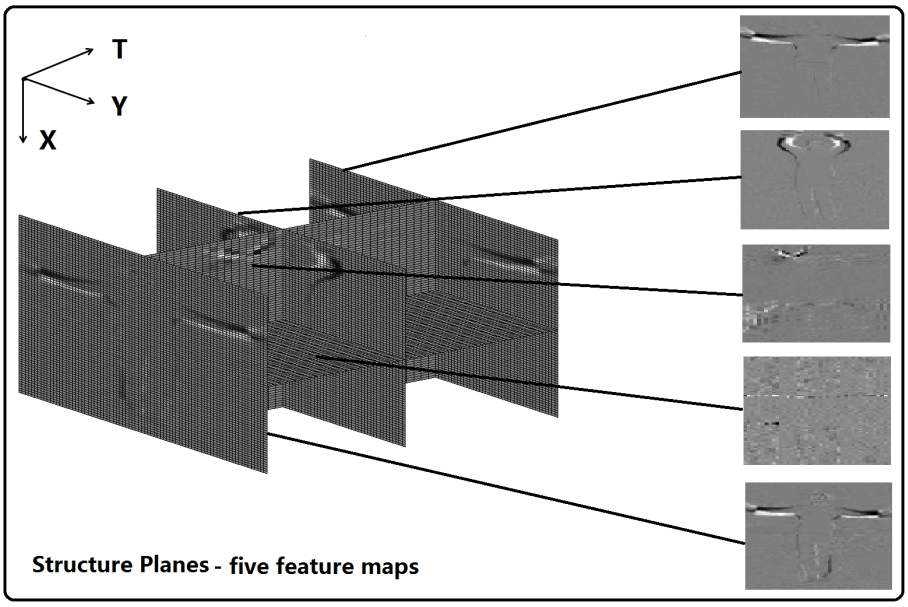

Figure 2-3: An example of structure planes extracted from the volume of difference of frames

Images are composed of features of many different sizes. Therefore, to encode the motion and structure information in the feature maps, a multi-scale analysis technique needs to be used. The Gaussian pyramid is a widely used multi-scale representation of images. By using the dyadic Gaussian pyramid convolved with each of the input feature maps, a series of low-passed images are obtained. A main advantage with the pyramid operation is that the image size decreases exponentially with the scale level and hence also the amount of computations required to process the data. To be precise, the levels of the pyramid are obtained recursively as follows:

$$G_l(i,j) = \sum_m \sum_n w(m,n) G_{l-1}(2i+m, 2j+n) \tag{2.2}$$

where l indexes the level of the pyramid and $w(m,n)$ is the Gaussian weighted function.

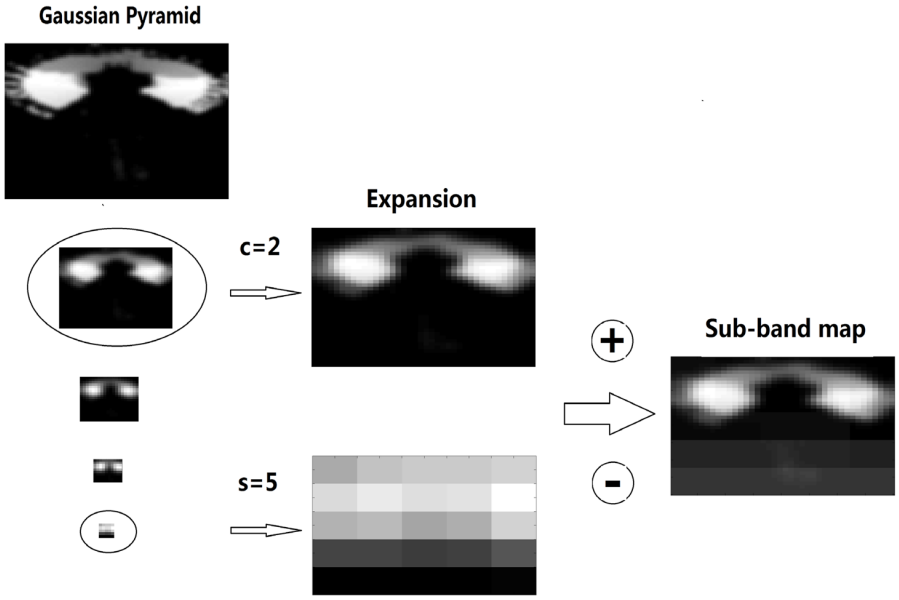

Figure 2-4: Illustration of the centre-Surround operation between Level 2 (centre) and Level 5 (surround) of a pyramid.

2.3.1 Centre-surround mechanism

Centre-surround (CS) fields have long been identified in the human visual system as having properties of edge enhancement that facilitate the detection, location, and tracking of small objects. After the centre-surround operation, features with different scales, such as edges and boundaries, are enhanced and segregated into a series of sub-band images. Here, the centre-surround mechanism is performed between centre levels (c = 2, 3, 4) and surround levels (*i.e.*, s = c + d, with d = 3, 4) of the obtained Gaussian pyramid. Thus six sub-band images are computed at levels of 2-5, 2-6, 3-6, 3-7, 4-7, and 4-8. Because scales are different between centre levels and surround levels, images of surround levels are interpolated to the same size as the corresponding centre levels, and then they are subtracted point-by-point from the corresponding centre levels to generate the relevant sub-band images. Fig 2-4 gives an example of the centre-surround operation.

2.4 Feature extraction

As highlighted in [45], feature extraction using an architecture with two stages, namely a filter bank and a feature pooling technique, performs better than that with a single stage. In the light of [107, 46], we employ a two-stage approach for feature extraction: 1) applying a bank of 2D Gabor filters to each sub-band of the feature maps to intensify the edge information at multiple orientations; and 2) performing a nonlinear max pooling within each band of Gabor filters and over local neighbourhoods to generate invariant features. Therefore, the extracted features are resistant to spatial shifts and insensitive to noise.

Since Gabor filtering and max pooling both share biological mechanisms with the human visual system, features extracted by the two-stage feature extraction module are biologically inspired and share common mechanism with the human visual system and therefore are more useful for recognition. Compared with the hierarchical model in [46], which is expensively computed, our features are more efficient with a low computational cost.

2.4.1 Gabor filtering

Gabor filters are widely used in visual recognition systems [46, 107, 100, 104], and provide a useful and reasonably accurate description of most spatial aspects of simple receptive fields. Due to their properties in common with mammalian cortical cells, such as spatial localisation, orientation selectivity and spatial frequency characterisation, Gabor filters are employed to extract orientation information. The 2D Gabor mother function is defined as:

$$F(x,y) = e^{-\frac{x_0^2+\gamma y_0^2}{2\sigma^2}} cos\frac{2\pi x_0}{\lambda} \tag{2.3}$$

where $x_0 = xcos\theta + ysin\theta$, $y_0 = -xsin\theta + ycos\theta$, the range decides the scales of Gabor filters and θ determines orientations. Gabor filters with eight scales in a range from 7×7 to 21×21 pixels and four orientations: degrees of 0, 45, 90, and 135 are used.

32 feature maps are obtained by convolving the initial input image with these Gabor filters which contain the features with multi-orientation information.

2.4.2 Max pooling

Feature pooling is employed in many modern visual recognition algorithms from pooling over image pixels [82, 107] to pooling across activations of local features on dictionary in sparse coding [137]. It preserves task-related information while removing irrelevant details. Pooling is used to achieve insensitivity to image transforms, more compact representations, and better robustness to noise and clutter [11].

The MAX mechanism was exploited by Disencumber and Pogges [90] in a hierarchical model of object recognition. This max-like feature selection operation is a key mechanism for object recognition in the cortex and provides a more robust response in the case of recognition in clutter or with multiple stimuli in the receptive field [90]. It successfully achieves invariance to image-plane transforms such as translation and scale. A max pooling operation is incorporated in the second stage of feature extraction. Pooling between scales of responses from each band of Gabor filters results in invariance to a range of scales; pooling over local neighbours leads to local robustness to position shifts and to possible localisation errors. The max pooling function can be defined as:

$$h(x, y) = max_{(x,y)\in G(x,y)}[g(i, j)] \tag{2.4}$$

where $g(i, j)$ is the response of the Gabor filters and $G(x, y)$ denotes the neighbourhood (receptive field) of the pixel (x, y). The neighbourhood window of max pooling is the average of the adjacent scales of Gabor filters. For instance, if two adjacent scales are 7 and 9 respectively, the neighbourhood window is then 8×8 .

2.5 Dimensionality reduction

Dimensionality reduction and feature selection have been an active research area in pattern recognition such as face and human gait recognition [65], [37]. The extracted

biologically-inspired features are high dimensional feature vectors so dimensionality reduction is needed to find the intrinsic low-dimensional subspace.

2.5.1 Discriminative locality alignment

Discriminative locality alignment (DLA) proposed recently by Zhang *et al.* [144] is chosen for the dimensionality reduction in this work because its advantages. It can (1) deal with the nonlinearity of the measurement distribution by taking into account the locality of measurements, (2) preserve the discriminative ability by considering different classes in the neighbour measurements and (3) avoid the small sample-size problem because it obviates the need to compute the inverse of a matrix. To be self-contained, we give a brief introduction to discriminative locality alignment. More details can be referred to [144].

Based on a framework named patch (A patch here refers a neighbourhood associated with a sample in the feature space) alignment, discriminative locality alignment unifies spectral analysis-based dimensionality reduction algorithms. This framework consists of two stages: part optimisation and whole alignment. For part optimisation, different algorithms have different optimisation criteria over patches, each of which is built by one measurement associated with its related ones. For whole alignment, all part optimisation are integrated to form the final global coordinate for all independent patches based on the alignment trick [148]. Global patches are usually built for conventional linear algorithms, *e.g.*, PCA and LDA, while local patches are usually formed in manifold learning-based ones, *e.g.*, LLE and LE. Two cases are given in Fig 2-5. As shown in Fig 2-5 (a), global patches should be built based on each measurement and all the others because measurements in this case are Gaussian distributed. In Fig 2-5 (b), measurements are sampled at random from the S-curve manifold embedded in a three-dimensional space. In this case, local patches should be built based on a given measurement and its nearest neighbours to capture the local geometry (locality).

More specifically, given a set of m-dimensional samples $X = [x_1, x_2, ..., x_N]$, each of the samples x_i belongs to one of C classes. What we want to do is finding a mapping

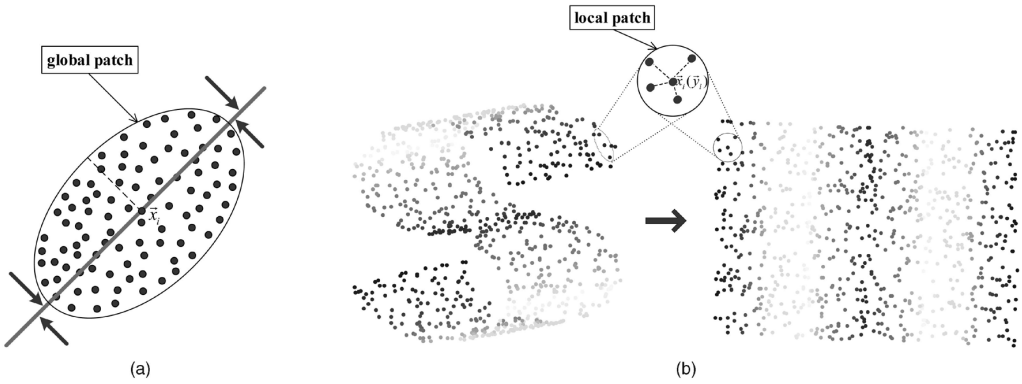

Figure 2-5: The framework of discriminative locality alignment [144].

matrix U to project X to $Y = [y_1, y_2, ..., y_N]$ in lower dimensional space. For each sample x_i, the samples are divided into the same class with the same label information as x_i and different classes with different label information from x_i. Then patches for each sample x_i are built according to the same class and different classes, denoted as $X_i = [x_i, x_{i^1}, ..., x_{i^{k_1}}, x_i, x_{i_1}, ..., x_{i_{k_2}}]$. Each patch contains $k_1 + k_2 + 1$ samples, namely neighbours from the same class and from different classes. Optimisation can be imposed on those patches based on an objective function that minimises the distance between samples in the same class and maximises the distance between samples from different classes. The mapping matrix U is obtained based on the objective function. We can get $Y = [y_1, y_2, ..., y_N]$ in the lower dimensional space by projecting X on the mapping matrix U.

2.6 Experiments and results

The proposed method is evaluated on the baseline KTH dataset, the multi-camera IXMAS dataset, and the realistic UCF sports dataset. To comprehensively evaluate the proposed framework, we will provide experimental results of the comparison with the state-of-the-art methods and an analysis of the proposed method in the following two subsections. All the results of other methods are obtained from the original papers.

2.6.1 Experimental settings

For the KTH dataset, we obtained the bounding boxes according to [138], to capture the main motion area of each action. We adopt the leave-one-out cross validation, *i.e.* videos of 24 subjects for training data and videos of the remaining one subject for testing. For the IXMAS dataset, we used the silhouettes available for each video sequence, we follow the validation settings, *i.e.*, leave one person out, in the original work [130]. For the UCF Sports dataset, we used the bounding boxes associated with each frame supplied with the dataset.

A linear support vector machine (SVM) is employed for action classification. We use the SVM implementation in the publicly available machine learning library LIMBS [18]. The parameters of the linear SVM are kept as the default values ($\sigma = \frac{1}{num.offeatures}$ and $C = 1$) throughout the thesis. We conduct two types of experiments, *i.e.*, *comparison* and *analysis*, to verify the advantages of the proposed framework.

2.6.2 Comparison with the state of the art

Each action in the KTH dataset is executed in four different scenarios; we perform our method on four scenarios separately and also give the results of taking all scenarios at once. Results of the proposed method on the KTH dataset and the comparison with other descriptors are shown in Table 2.1 which indicates that the proposed method achieves the best recognition rates among all the listed methods.

Specifically, our method achieves almost perfect accuracy in scenarios S1 and S4 and a relatively satisfactory result in scenario S2, which contains camera zooming. Although in S3 actors are dressed in quite different clothes, it is still able to achieve a high recognition rate. This suggests that our method is robust to scale variance (S2) and insensitive to clothing variance of human subjects (S3). Note that in our method the average accuracy of four scenarios is slightly higher than that of all scenarios in one, which is theoretically reasonable because actions in all scenarios have greater intra-class variations than those in each single scenario. The comparison

Methods	S1	S 2	S 3	S 4	Average	All-in-one
Our method	**98.7**	**88.1**	**94**	**94**	**93.5**	**93.3**
HMAX [46]	96.0	86.1	89.8	94.8	91.7	-
Schindler *et al.* [97]	93.0	81.1	92.1	96.7	90.7	90.9
Yeffet *et al.* [139]	-	-	-	-	-	90.1
Taylor *et al.* [111]	-	-	-	-	-	90.0
Jr *et al.* [47]	-	-	-	-	-	90.2

Table 2.1: Performance (recognition rate in percentage) comparison of different descriptors on the KTH dataset. Scenaros 1, 2, 3 and 4 are four scenarios in the KTH dataset. 'All in one' is the accuracy of taking four scenarios in one. '-' means not available (recognition rates in %).

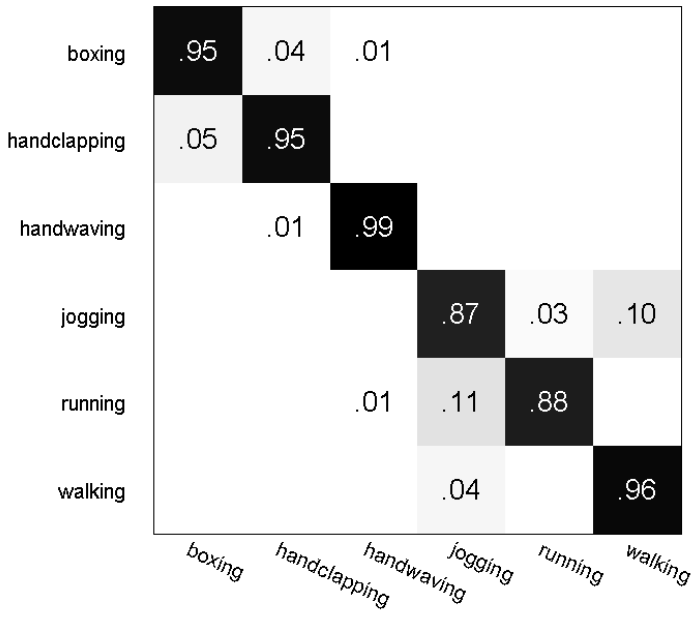

Figure 2-6: The confusion matrix of the proposed method on the KTH dataset.

Methods	Camera 1	Camera 2	Camera 3	Camera 4	Camera 5
Our method	**84.9**	**87.9**	**90.1**	86.9	**78.9**
GMKL [135]	76.4	74.5	73.6	71.8	60.4
AFMKL [135]	81.9	80.1	77.1	77.6	73.4
Weinland *et al.* [131]	84.7	85.8	87.9	88.5	72.6
Liu *et al.* [68]	76.7	73.3	72.1	73.1	-
Yan *et al.* [136]	72.0	53.0	68.1	63.0	-
Weinland *et al.* [130]	65.4	70.0	54.5	66.0	33.6
Junejo et al [52]	76.4	77.6	73.6	68.8	66.1

Table 2.2: Comparison of performance on the IXMAS dataset.Camera 1, 2, 3, 4 and 5 are five cameras in the dataset. '-' means not available (recognition rates in %).

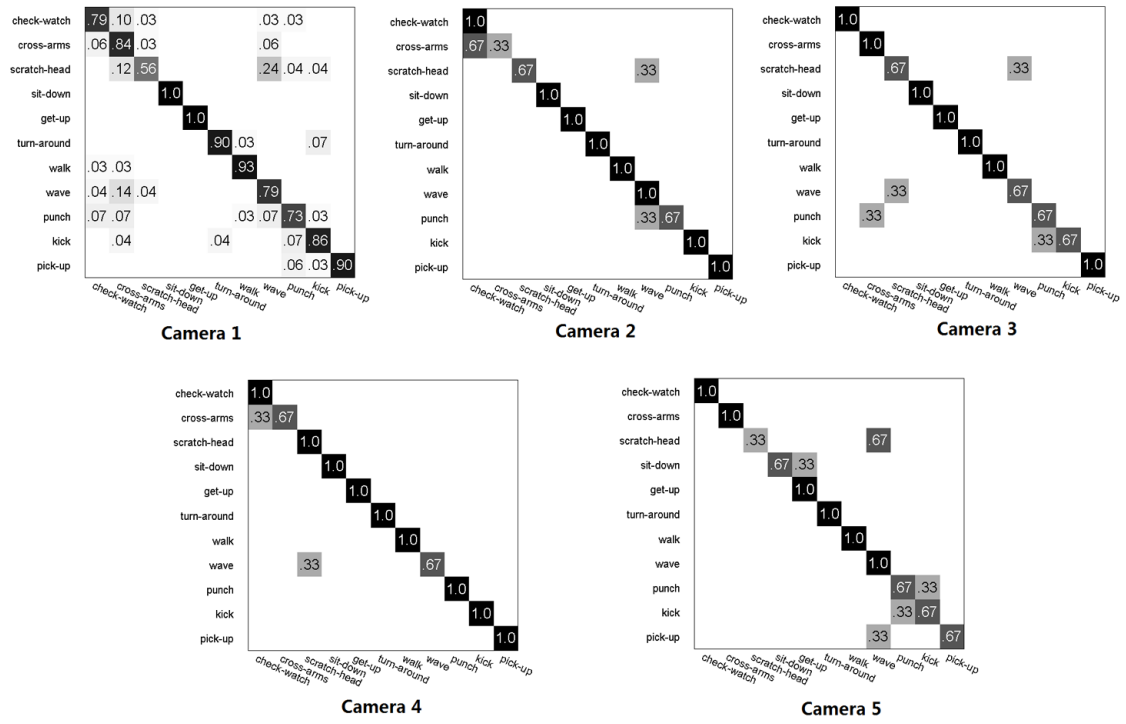

Figure 2-7: The confusion matrices of the proposed method on the IXMAS dataset.

Method	Accuracy
Our method	**93.9**
Yeffet *et al.* [139]	79.3
GMKL [135]	85.2
AFMKL [135]	91.3
Wang *et al.* [123]	88.2
Le *et al.* [62]	86.5
Weinland *et al.* [131]	90.1
Kovashka *et al.* [57]	87.3
Wang *et al.* [124]	85.6
Rodriguez *et al.* [92]	69.2

Table 2.3: Performance comparison of different methods on the UCF sports dataset (recognition rates in %).

with previously proposed methods demonstrates that the proposed method produces comparable results with state-of-the-art methods.

Fig 2-6 shows the confusion matrix of recognition results on the KTH dataset. Interestingly, three actions including *jogging*, *running* and *walking* are confused with each other. *running* is easy to be misclassified as *jogging*. This is reasonable because these three actions share the same motion patterns in the KTH dataset, especially between *running* and *jogging*.

On the multi-camera IXMAS dataset, the proposed method greatly outperforms state-of-the-art methods in all five views, as shown in Table 2.2. Although silhouettes are available for each action, some of them are not well extracted because of noise, missing body parts and self occlusion, especially in camera 5. However, our method achieves a quite good recognition rate in camera 5 in which actions are significantly occluded and are difficult for most of the current methods.

The confusion matrices of recognition rates are illustrated in Fig 2-7. In all the five cameras, *wave* and *scratch* are confused with each other, which is reasonable because *wave* and *scratch* are actions with a lot of similar motion patterns and body poses. Especially in camera 5 which has significant occlusions, the proposed method can still successfully recognise five categories (11 categories in total) of actions.

The evaluation on the realistic UCF sports dataset is presented in Table 2.3. The

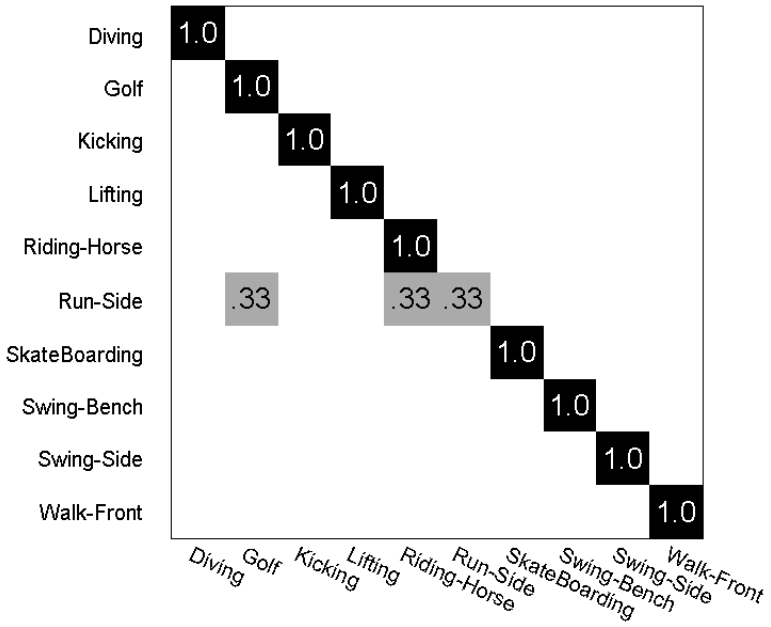

Figure 2-8: The confusion matrix of the proposed method on the UCF sports dataset.

UCF sports dataset is regarded as one of the most challenging datasets for action recognition. Actions in this dataset are all realistic and are performed in different ways with large intra-class variability, which makes it hard for recognition. However, the proposed method still outperforms the best published result - 91.3% - by over 2.6%.

Similarly, we plot the confusion matrix of the recognition rate on the UCF sports dataset in Fig 2-8, in which we can see the proposed method can successfully recognise most action categories except for *Run-side*. A possible explanation is that it has some similar spatial-temporal appearance and motion patterns to *Golf* and *Riding-Horse*.

2.6.3 Analysis

As our framework integrates motion (MHI) and structure features (five structure planes) as a holistic descriptor, we conduct experiments to evaluate their individual contributions to the representation of actions. In addition, to validate the contribu-

	Features	PCA	LDA	LPP	NPE	Isomap	DLA
Scenarios 1	SP	96.7	93.3	88.7	94.0	96.7	97.3
	MHI	77.8	68.7	78.7	69.3	72.7	79.3
	SP+MHI	98.7	98.0	92.7	91.2	**99.3**	**98.7**
Scenarios 2	SP	80.0	79.3	80.7	78.7	81.3	81.3
	MHI	61.3	56.0	62.7	54.7	61.3	62.0
	SP+MHI	84.7	81.3	83.3	80.7	86.0	**88.1**
Scenarios 3	SP	92.5	91.2	88.4	91.2	91.9	94.0
	MHI	76.0	64.9	74.4	64.9	76.0	76.0
	SP+MHI	91.9	91.2	89.8	91.2	92.5	**94.0**
Scenarios 4	SP	94.0	90.0	91.3	90.0	94.0	94.0
	MHI	83.0	72.0	83.2	66.7	83.0	83.0
	SP+MHI	92.7	92.7	85.6	92.7	94.0	**94.0**
All in one	SP	90.6	88.9	86.1	89.1	92.8	91.6
	MHI	76.1	69.4	72.6	64.6	62.4	76.2
	SP+MHI	91.1	91.1	86.6	89.6	92.8	**93.3**

Table 2.4: The performance of the proposed framework with different features on the KTH dataset, and the comparison of DLA with the state-of-the-art dimensionality reduction techniques.

		PCA	LDA	LPP	NPE	Isomap	DLA
Camera 1	SP	79.8	80.0	68.6	80.9	81.0	81.0
	MHI	77.0	69.1	68.8	69.4	77.3	77.8
	SP+MHI	81.0	82.7	71.6	82.7	81.8	**84.9**
Camera 2	SP	84.4	83.6	73.3	83.6	85.0	85.3
	MHI	79.8	78.0	71.3	77.4	78.9	78.9
	SP+MHI	84.1	87.2	77.4	87.2	86.9	**87.9**
Camera 3	SP	81.4	85.1	78.3	84.2	85.6	85.6
	MHI	84.0	82.9	71.9	82.3	84.5	84.9
	SP+MHI	86.5	90.1	79.2	90.0	89.6	**90.1**
Camera 4	SP	79.1	81.4	71.4	78.6	81.1	81.4
	MHI	80.2	76.6	73.2	78.6	81.8	81.8
	SP+MHI	85.5	84.5	72.5	84.5	86.5	**86.9**
Camera 5	SP	76.8	71.8	66.5	71.2	76.1	75.8
	MHI	68.3	60.9	61.5	60.8	69.7	69.5
	SP+MHI	76.3	77.9	65.8	77.9	77.9	**78.9**

Table 2.5: The performance of the proposed framework with different features on the IXMAS dataset, and the comparison of DLA with the state-of-the-art dimensionality reduction techniques.

tion of DLA to the overall performance of the proposed framework and its advantage over other dimensionality reduction techniques, we compare it with many widely used reduction methods including principal component analysis (PCA), linear discriminative analysis (LDA), Locality Preserving Projections (LPP), Neighbourhood Preserving Embedding (NPE) and Isomap. Table 2.4, Table 2.5 and Table 2.6 show the comprehensive analysis of the proposed recognition framework and the comparisons with different dimensionality reduction techniques on the three datasets, *i.e.*, KTH, IXMAS and UCF Sports ('SP' denotes the structure planes and accuracies are in percentages). PCA is employed as the first step in the DLA method to filter the noise. To make fair comparisons, we keep the dimensions to be 100 (KTH), 200 (IXMAS) and 200 (UCF Sports) for all the reduction techniques except for LDA.

Features from structure planes can achieve satisfactory recognition rates on all the three datasets. Features from the MHI exhibit comparable performance to those of the structure planes on the IXMAS dataset, while not achieving high accuracies on the UCF Sports dataset. This is reasonable, because in well-constrained circumstances, *e.g.*, in the IXMAS dataset, MHI is able to encode more accurate motion information, and therefore can give better performance. In addition, the MHI contains only one feature map while structure planes have five feature maps, which would encode more information about an action. However, the combination of features on the structure planes and the MHI can improve the overall performance and is better than either of them, which proves that structure planes and the MHI provide complementary information.

Additionally, as the obtained feature vectors are of high dimensionality, we perform DLA for more compact and discriminative representation. From the comparison results in Table 2.4, Table 2.5 and Table 2.6, we can see DLA performs consistently better than other dimensionality reduction techniques on all the three datasets, which demonstrates that the use of DLA does improve the performance of the framework. Furthermore, with the DLA reduction, we can see from the tables that the combination of structure planes and motion history image consistently outperforms either of them. This again validates that structure planes and motion history image are com-

	PCA	LDA	LPP	NPE	Isomap	DLA
SP	91.8	91.2	88.3	91.9	91.8	92.4
MHI	49.6	44.0	52.9	42.6	54.5	53.2
SP+MHI	93.1	90.5	89.8	91.2	**93.9**	**93.9**

Table 2.6: The performance of the proposed framework with different features on the UCF sports dataset, and the comparison of DLA with the state-of-the-art dimensionality reduction techniques.

plementary features and at the same time manifests that DLA can effectively embed them into a unified and meaningful representation of human actions. Note that even with PCA for dimensionality reduction, our method can still achieve competitive results compared with the state-of-the-art methods. From the experimental results, we can safely draw the following conclusions:

- The structure planes and the motion history image provide complementary information, and therefore the combination of them gives an informative and effective representation of actions.

- The employed dimensionality reduction method, *i.e.*, discriminative locality alignment (DLA), is able to effectively embed the structure and motion features into meaningful representations, and outperforms many widely used dimensionality reduction techniques.

The reasons why our method can achieve better results lie in the following aspects:

- We explicitly model the motion and structure features by the motion templates and structure planes, which encapsulate the main information of the action in a video sequence.

- On top of the extracted feature maps, we use the biologically-inspired features to better represent the information of motion and structure, which provides an informative and invariant representation of actions.

- Last but not least, the powerful DLA reduction technique improves the performance of our method compared with other dimensionality reduction approaches.

2.7 Conclusion

In this chapter, we have presented a framework unifying motion and structure features for human action recognition. By applying the motion template to the volume with difference of frames (DoF), we encode the motion information into the motion feature map, *i.e.*, the motion history image (MHI), and structure feature maps are obtained from the structure planes extracted form the DoF volume. Two dimensional Gaussian pyramid and centre-surround operations are performed on each feature map, and the feature maps are decomposed into sub-band images localised on multiple centre spatial frequencies. Efficient, biologically-inspired features are then extracted through a two-stage feature extraction step, namely Gabor filtering and max pooling.

Finally, the discriminative locality alignment (DLA) technique embeds the high-dimensional features onto a low-dimensional manifold space which leads to a more discriminative and compact representation of actions. Evaluations on three increasingly difficult datasets, KTH, IXMAS and UCF Sports, demonstrate that the proposed framework is a very promising global representation for human action recognition.

Chapter 3

Spatio-Temporal Laplacian Pyramid Coding

3.1 Introduction

In the video domain, many of algorithms actually borrow the ideas from the 2D image domain on image/scene representation and classification. For instance, the three-dimensional histogram of oriented gradients (HOG3D) and the three-dimensional SIFT (SIFT3D) extended from their 2D counterparts have shown the effectiveness for action recognition. Inspired by the success of multi-resolution analysis and the biologically inspired features for action recognition in Chapter 2 and the work in [46], in this chapter we propose a global descriptor, named spatio-temporal Laplacian pyramid coding (STLPC), by extending the multi-scale analysis and biologically-inspired features in the image domain.

3.1.1 Motivations

According to the scale-space theory [66], objects in the world, as meaningful entities, only exist over certain ranges of scales. The multi-scale representation is of crucial importance for describing unknown real-world signals and holds a basic and important role in early vision. Human visual perception treats images on several levels

of resolutions simultaneously. Spatio-temporal features in video analysis share some important properties with those in the image domain. In this chapter, we aim to describe video sequences by combing the multi-scale and orientation analysis, *i.e.*, the spatio-temporal Laplacian pyramid and 3D Gabor filters, respectively.

The Laplacian pyramid provides a multi-resolution analysis, a scheme employed by the human visual system [134] which has shown its effectiveness in image representation [14]. Through introducing the multi-resolution technique to video analysis and action representation, the proposed STLPC algorithm can maximally extract structural and motion features with different scales, and therefore provides an informative holistic representation which can overcome the aforementioned limitations of sparse representations.

The Laplacian pyramid does not take account of orientation which carries important information in video sequences with actions. Gabor filters are widely used for feature extraction and can capture edge and orientation information in the image domain. Similarly, 3D Gabor filters are able to extract spatio-temporal edge and orientation features related to motion occurring in a video sequence.

Direct use of the outputs from the filtering as features is redundant for representation, so feature pooling techniques are needed. Max pooling has drawn most attention in low-level feature extraction algorithms. We introduce max pooling into the spatio-temporal domain which can, to a large extent, cope with shortcomings, such as spatio-temporal misalignment and inaccurate motion localisations existing in traditional holistic representation methods. In addition, after max pooling, we can obtain a compact representation.

3.1.2 Overview

A video sequence is considered as a spatio-temporal intensity volume from which motion cues of human actions are firstly extracted through differencing adjacent frames. Backgrounds are simultaneously suppressed without suffering from expensive computations resulting from tracking or background subtraction.

We then construct a spatio-temporal Laplacian pyramid (STLP) (Section 3.2)

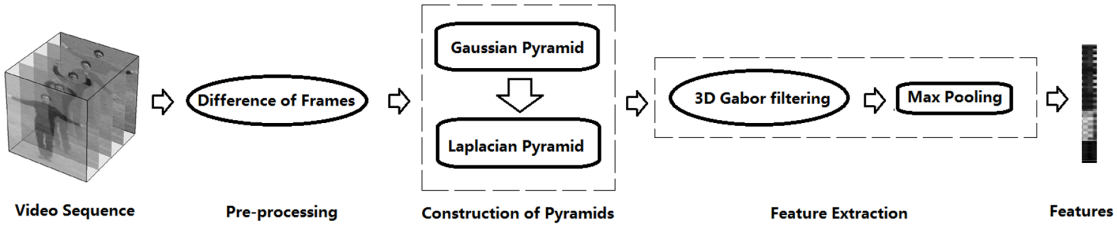

Figure 3-1: Construction of spatio-temporal Gaussian pyramid and Laplacian pyramid.

as follows. The obtained volumes with DOF are repeatedly filtered with Gaussian weighting functions and subsampled to generate volumes with regularly reduced resolutions. These comprise a series of low-pass filtered copies of original volumes, namely a spatio-temporal Gaussian pyramid, in which the bandwidth decreases at one-octave per step. To directly represent the volumes in terms of voxel intensity values, however, is inefficient due to the high correlation among these voxels. Therefore, the smoothed 3D volumes are decomposed into a set of spatio-temporal band-pass filtered volumes called a spatio-temporal Laplacian pyramid by differencing adjacent levels of the Gaussian pyramid. Features with different sizes are appropriately localised at each level of the pyramid, as the band-pass filtered volume represents a particular fineness of detail at each scale.

Subsequently, we apply a feature extraction step (Section 3.3). A bank of 3D Gabor filters is then applied to the original volume and each level of the Laplacian pyramid to enhance edge and orientation information. To extract invariant and discriminative features, a nonlinear max pooling technique is performed within bands of Gabor filters and over spatio-temporal neighbourhoods, resulting in robustness to spatial and temporal shifts, partial occlusions and noise. Our feature extraction process from a raw video sequence is illustrated in Fig. 3-1.

3.1.3 Contributions

We summarise the contributions of the work in this chapter as follows:

- We introduce the multi-resolution analysis techniques, *i.e.*, the spatio-temporal

Gaussian and Laplacian pyramids, for human action recognition and video analysis.

- Based on the spatio-temporal Laplacian pyramid, a novel descriptor is proposed for the holistic representation of human actions.

- Gabor filters and max pooling are extended to the 3D video domain and are successfully applied to spatio-temporal feature extraction in the holistic representation.

In contrast to traditional holistic methods which are heavily dependent on tracking and spatial/temporal alignment algorithms, our method can, to a large extent, handle misalignments and background variations. Moreover our method needs only coarse rather than accurate bounding boxes, which are usually essential in common holistic representation methods. These benefits result from our multi-scale representation, *i.e.*, the spatio-temporal Laplacian pyramid and the max pooling over spatio-temporal neighbours.

3.2 Spatio-temporal Laplacian pyramid coding

A video sequence is viewed as a spatio-temporal intensity volume that contains all structural and motion information of the action, including poses of human figures at any time as well as the dynamic motion information.

The Laplacian pyramid, a multi-resolution analysis technique, decomposes spatio-temporal volumes into different levels with a certain band of frequencies. Salient features residing at different scales are segregated in each level of the pyramid and extracted separately in the following feature extraction step. In contrast to three-dimensional scale invariant feature transform (3D SIFT), in which convolutions are applied spatially [99], our model performs convolutions spatio-temporally with 3D Gaussian kernels. Each layer of the Laplacian pyramid is approximated by the difference of Gaussians.

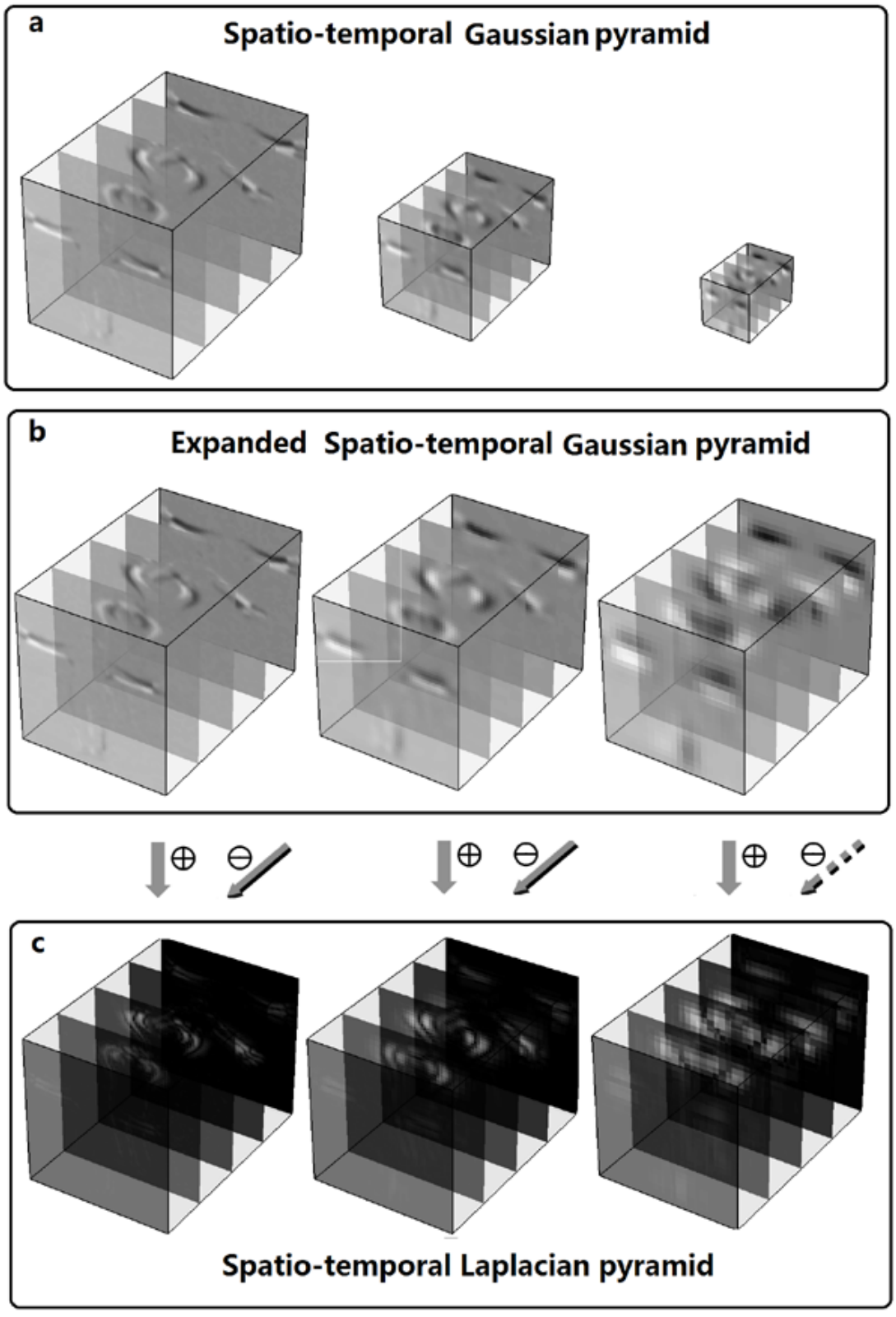

Figure 3-2: Construction of spatio-temporal Gaussian pyramid and Laplacian pyramid.

3.2.1 Spatio-temporal Gaussian pyramid

The first step in creating a Laplacian pyramid is to construct a series of low-pass filtered 3D volumes - a spatio-temporal Gaussian pyramid. The filtering is performed along spatial and temporal dimensions by a procedure equivalent to convolution with local symmetric weighting functions, *e.g.*, 3D Gaussian function, which is given as follows:

$$w(x, y, t) = \frac{1}{(\sqrt{2\pi}\sigma)^3} e^{-\frac{x^2+y^2+t^2}{2\sigma^2}} \tag{3.1}$$

The operation is a kind of multi-scale filtering. Gaussian is chosen as the smoothing kernel because it has been proven to be the only kernel for which local maxima of a signal increase and local minima decrease as the filter bandwidth increases [56, 66]. Given a 3D volume, it is viewed as the bottom or zero level of a Gaussian pyramid. Higher levels of the Gaussian pyramid can be generated by convolving a 3D Gaussian function with several copies of the original 3D volume with reduced resolutions obtained by subsampling. Precisely, levels of a Gaussian pyramid are iteratively obtained as follows:

$$g_l(i, j, k) = \sum_x \sum_y \sum_t w(x, y, t) g_{l-1}(2i + x, 2j + y, 2k + t) \tag{3.2}$$

where l indexes the levels of a Gaussian pyramid and (i, j, k) is the position of a voxel in a 3D volume. The construction of a Gaussian pyramid is computationally efficient because, with increasing levels, the size of the video volume decreases exponentially and the number of required computations reduces as well. The constructed 4-level Gaussian pyramid is shown in Fig 3-2 (b).

3.2.2 Spatio-temporal Laplacian pyramid

As different spatio-temporal features would be salient at different scales (resolutions) in the spatio-temporal space, we aim to extract them separately. The Laplacian

pyramid is a technique for multi-resolution analysis. The construction of the Laplacian pyramid is actually performing Laplacian operators on Gaussians with many scales. As studied in [66], the scale-normalised Laplacian of Gaussian, $\sigma^2\nabla^2 G$, can be approximated by the difference of Gaussians.

To be self-contained, we provide the relationship between difference of Gaussians, D and $\sigma^2\nabla^2 G$ by the heat diffusion equation according to [70]:

$$\frac{\partial G}{\partial \sigma} = \sigma\nabla^2 G, \tag{3.3}$$

where $\partial G/\partial\sigma$ can be approximated by the difference of nearby scales at $k\sigma$ and σ. Note that σ is equivalent to t in the heat kernel equation.

and, we have

$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, \sigma)}{k\sigma - \sigma}. \tag{3.4}$$

Therefore, we can see that Laplacian of Gaussian can be approximated by the difference of Gaussians:

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2\nabla^2 G. \tag{3.5}$$

Having obtained the Gaussian pyramid of the original 3D volume, we expand each level of the Gaussian pyramid into the same size as the bottom level, and represent it as G_l. Now we can generate the Laplacian pyramid. The bottom level of the Laplacian pyramid is obtained by subtracting the first level of the Gaussian pyramid from the expanded version of the previous (bottom) level of the Gaussian pyramid. The higher levels of the Laplacian pyramid are generated with a similar operation, as follows:

$$L_l = G_l - G_{l+1} \tag{3.6}$$

Similarly, l indexes the levels of a Laplacian pyramid. G_l and G_{l+1} are the expanded versions of g_l and g_{l+1}. Fig 3-2 (c) shows an example of a three-level Laplacian

pyramid. Obviously, features such as edges and corners are enhanced at each level of the pyramid, and these features correspond to the areas of motion in the original video sequence. In addition, these enhanced features are separately extracted at each level of the pyramid with different resolutions. The example in Fig 3-2 (c) demonstrates that the Laplacian pyramid is a particularly effective way to represent video sequences spatially and temporally. Salient spatio-temporal features are enhanced for analysis and representations based on the pyramid would be both compact and robust.

3.3 Feature extraction

Although the Laplacian pyramid provides an efficient multi-scale analysis, orientation information has not been taken into account. While actions can be regarded as spatio-temporal patterns in different orientations, we propose employing a bank of 3D Gabor filters to extract the orientation information. We employ a two-stage approach for spatio-temporal feature extraction.

- Spatio-temporally applying a bank of 3D Gabor filters to intensify the edge information at multiple orientations.
- Performing a nonlinear max pooling within each band of 3D Gabor filters and over spatio-temporal neighbourhoods to generate invariant features. Therefore, the extracted features are resistant to spatial and temporal shifts and insensitive to noise. More importantly, motion information encoded in multiple contiguous frames will be exploited.

Our method differs from the C1 model by Jhuang *et al.* [46] in that filtering and pooling are applied to both spatial and temporal dimensions.

3.3.1 3D Gabor filters

Gabor filters are widely used in visual recognition systems [46, 107], and provide a useful and reasonably accurate description of most spatial aspects of simple receptive fields. The Laplacian pyramid representation does not introduce any spatial orientation selectivity into the decomposition process. Gabor filters are employed to extract

Bank	1	2
Filter Size	7 & 9	11 & 13
σ	2.8 & 3.6	4.5 & 5.4
λ	3.5 & 4.6	5.6 & 6.7
θ	$-\pi/4$, 0, $\pi/4$	
ω	$-\pi/4$, 0, $\pi/4$	

Table 3.1: Summary of parameters for Gabor filters used in our implementation.

orientational information due to their properties in common with mammalian cortical cells, such as spatial localisation, orientation selectivity and spatial frequency characterisation.

Inspired by [107, 84], our method uses three-dimensional Gabor filters to localise salient features in spatio-temporal dimensions. In the 3D Gabor filter bank, 4 scales in two bank with a total of 9 orientations are used. In a 3D space, the Gabor filter is defined as:

$$G(x, y, t) = exp[-(\frac{X^2}{2\sigma_x} + \frac{Y^2}{2\sigma_y} + \frac{T^2}{2\sigma_t})] \times \cos(\frac{2\pi}{\lambda_x}X)\cos(\frac{2\pi}{\lambda_y}Y) \tag{3.7}$$

where

$$\begin{pmatrix} X \\ Y \\ T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \times \begin{pmatrix} \cos(\omega) & 0 & \sin(\omega) \\ 0 & 1 & 0 \\ -\sin(\omega) & 0 & \cos(\omega) \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

The parameters used in 3D Gabor filters are listed in Table 3.1. σ and λ are the spatial and temporal scales, respectively, and similarly θ and ω refer to the spatial and temporal orientations, respectively.

3.3.2 Spatio-temporal max pooling

Similar to the feature extraction in Chapter 2, max pooling is extended into the spatio-temporal domain and incorporated in the second stage of spatio-temporal feature selection. More specifically, for the volumes from one bank of Gabor filters with 2 scales and 9 orientations, we first perform max pooling between the two volumes with different scales and at the same orientation. After this first step of max pooling, we have one volume at each orientation. We then pool the volumes over a local neighbourhood, which is equivalent to applying a 3D max filter to the volumes. Fig. 3-3 demonstrates the mechanism of max pooling in our method. On the left are the two volumes of the outputs of Gabor filters in two adjacent scales and at the same orientation. On the right, the first volume is the one pooled between the two scales and the second volume is the one pooled over a local neighbourhood. Pooling between scales of responses from each band of Gabor filters results in invariance to a range of scales; pooling over spatio-temporal neighbours leads to local robustness to position shifts and to possible localisation errors.

After max pooling, we need to flatten the volumes into final feature representations. To make a compact and invariant representation, similar to the Gist feature extraction in scene recognition [42], averaging operations are applied in a fixed $4\times4\times4$ grid of spatio-temporal sub-regions of the volumes from max pooling. The averaging operation is commonly used for feature extraction [42, 104, 107]. The dimensionality of the final feature vector before dimensionality reduction is $(4\times4\times4 = 64)\times N\times L\times O$, where N is the number filter banks, L is the number of levels of the Laplacian pyramid and O is the number of orientations. If 2 banks of Gabor filters, 5 levels of the Laplacian pyramid and 9 orientations are used, the dimensionality is $64 \times 2 \times 5 \times 9 = 5760$.

3.3.3 Discriminative locality alignment

The features are still in a high-dimensional space. Based on the work [151] in Chapter 2, which has presented a comprehensive comparison of dimensionality reduction techniques for action recognition, we also employ discriminative locality alignment

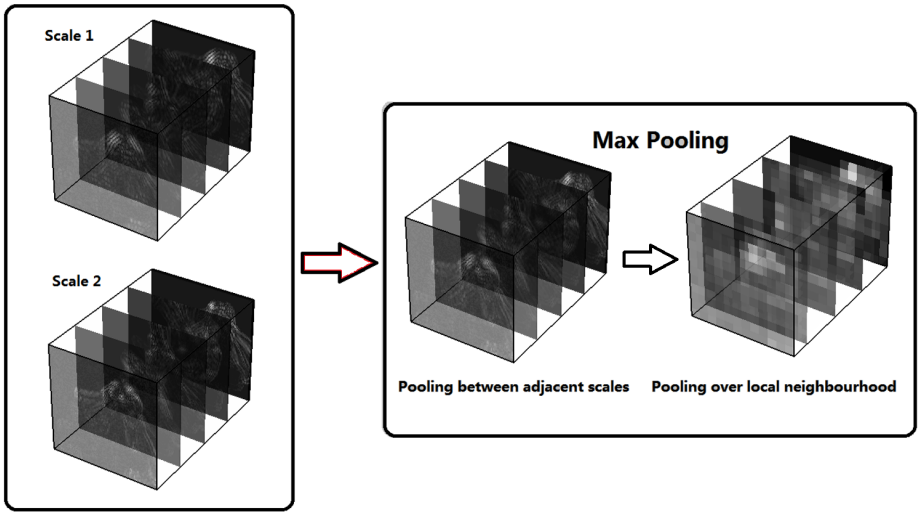

Figure 3-3: On the left are the two volumes of outputs from Gabor filters with adjacent scales at the same orientation. The first volume on the right is the volume pooled between two scales and the second is the volume after pooling over local neighbourhoods.

(DLA) [144] for dimensionality reduction to obtain compact and discriminative representations.

3.4 Experiments and results

The proposed spatio-temporal Laplacian pyramid coding (STLPC) is evaluated on the baseline KTH dataset, the multi-camera IXMAS dataset, the realistic UCF sports dataset and the newly released HMDB51 dataset.

3.4.1 Experimental settings

To demonstrate its effectiveness and efficiency as a holistic descriptor, we compare it with popular descriptors such as the 3D histogram of oriented gradients (HOG3D) [54] and the 3D scale invariant feature transform (SIFT3D) [99]. To make the comparison fair, we replace only our STLPC descriptor with HOG3D and SIFT3D with the rest of the settings the same. All descriptors are used as holistic representations of human actions. For both HOG3D and SIFT3D, the spatio-temporal volume containing the

action is divided into equally-sized small cubes and the final descriptor vector is a concatenation of descriptors calculated from all cubes. For HOG3D, we use 6 bins for orientation quantisation. For SIFT3D, we follow the parameter settings in the original paper [99] using $2 \times 2 \times 2$ and $4 \times 4 \times 4$ configurations of sub-histograms, and 8×4 histograms to represent θ and ϕ. We follows the validation settings in Chapter 2. A linear support vector machine (SVM) is employed for action classification [18].

3.4.2 Comparison with the state of the art

For the KTH dataset in which the same action is executed in four different scenarios, we perform our method on four scenarios separately and also give the results of taking all scenarios in one.

Results of STLPC on the KTH dataset and the comparison with other descriptors are shown in Table 3.2 and Table 3.3. The proposed STLPC algorithm achieves the best recognition rates among all the listed methods. Our method achieves almost perfect accuracy in Scenarios 1 and Scenario 4 and a relatively satisfactory result in Scenario 2, which contains camera zooming. Although in Scenario 3 actors are dressed in quite different clothes, STLPC is still able to achieve a high recognition rate. This shows that STLPC is robust to scale variation (Scenario 2) and insensitive to clothing variance of human subjects (Scenario 3). Note that in our method the average accuracy of four scenarios is slightly higher than that of all scenarios in one, which is theoretically reasonable because actions in all scenarios have greater intra-class variations than those in each single scenario. In addition, the proposed STLPC greatly outperforms the popular descriptors: HOG3D and SIFT3D. Table 3.2 reports a longitudinal comparison with previously proposed methods, which demonstrates that the proposed STLPC outperforms state-of-the-art methods.

On the multi-camera IXMAS dataset, STLPC greatly outperforms state-of-the-art methods in all five views, as shown in Table 3.4. Although silhouettes are available for each action, some of them are not well extracted because of noise, missing body parts and self occlusions, especially in Camera 5. However, our method achieves a quite good recognition rate in Camera 5 in which actions are significantly occluded and are

Methods	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Average	All in one
STLPC	**98.7**	88.0	**96.7**	**98.7**	**95.5**	**95.0**
AFMKL [135]	96.7	**91.3**	93.3	96.7	94.5	-
GKML [119]	96.0	86.0	90.7	94.0	91.7	-
HMAX [46]	96.0	86.1	89.8	94.8	91.7	-
HOG3D	97.3	80.7	93.3	95.9	91.8	91.5
SIFT3D	96.0	74.7	90.7	96.5	89.5	90.5

Table 3.2: Performance comparison of different descriptors on the KTH dataset. 'Average' is the average accuracy of the four scenarios, and 'All in one' is the accuracy of taking four scenarios in one. '-' means not available (in percentages).

Methods	Accuracy (%)
Dollár *et al.* [25]	81.2
Savarese *et al.* [96]	86.8
Niebels *et al.* [83]	81.5
Liu *et al.* [68]	94.2
Zhang *et al.* [146]	92.9
Liu *et al.* [67]	93.8
Wang *et al.* [123]	94.2
Zhang *et al.* [147]	93.5
STLPC	**95.0**

Table 3.3: A longitudinal performance comparison of different methods on the KTH dataset. All the methods compared in the table used leave-one-out cross validation (in percentages).

difficult for most of the current methods. We also apply either HOG3D or SIFT3D as the holistic descriptor for comparison. The comparison with different methods is reported in Table 3.4.

The evaluation on the realistic UCF sports dataset is presented in Table 3.5. Actions in this dataset are all realistic and are performed in different ways with large intra-class variability, which makes recognition hard. However, the proposed method still produces an excellent result and outperforms the best published result - 91.3% - by over 2%. In addition, our SPLPC descriptor consistently performs better than both HOG3D and SIFT3D in this dataset. This result demonstrates that STLPC is effective for recognizing realistic human actions.

Finally, the outcome on general body movements of the HMDB51 dataset is re-

Methods	Camera 1	Camera 2	Camera 3	Camera 4	Camera 5
STLPC	**91.8**	**88.8**	**91.7**	87.4	**81.8**
HOG3D	85.8	85.9	88.2	80.1	78.4
SIFT3D	81.9	81.9	84.2	82.0	70.4
GMKL [135]	76.4	74.5	73.6	71.8	60.4
AFMKL [135]	81.9	80.1	77.1	77.6	73.4
Weinland et al [131]	84.7	80.8	87.9	**88.5**	72.6
Liu *et al.* [68]	76.7	73.3	72.1	73.1	-
Yan *et al.* [136]	72.0	53.0	68.1	63.0	-
Weinland *et al.* [130]	65.4	70.0	54.5	66.0	33.6
Junejo *et al.* [52]	76.4	77.6	73.6	68.8	66.1

Table 3.4: Performance comparison of different methods in five cameras on the IX-MAS dataset. '-' means not available.

ported in Table 3.8. The proposed STLPC algorithm achieves an average accuracy of 37.3% using the three distinct training and testing splits, which demonstrates the potential of STLPC for large scale realistic human action recognition. Moreover, the bounding boxes for the HMDB51 dataset are quite coarse, and some are the same sizes as the original video sequences. With the same experimental setting, STLPC significantly outperforms the result -25.6%- reported in [150], where 3D steerable filters are used for holistic action representation.

3.4.3 Laplacian pyramid

Our final descriptor is based on the combination of the original video sequence (called the bottom level of the Laplacian pyramid) and higher levels of the Laplacian pyramid of the video sequence. We evaluate the performance of the final descriptors with different numbers of pyramid levels. The results are illustrated in Table 3.6, Table 3.7 and Table 3.8. '#level' denotes the number of the Laplacian pyramid levels. '0' means the bottom level, namely the original video sequences.

From the Tables, we can see that combining the original video sequences with higher levels of the Laplacian pyramid does make the descriptor more informative and discriminative and therefore increases the performance of the recognition system. The performance generally improves with the increase of the number of levels, the best

Methods	Accuracy (%)
STLPC	**93.4**
HOG3D	84.4
SIFT3D	77.3
Raptis *et al.* [123]	79.4
Wang *et al.* [123]	88.2
Yeffet *et al.* [139]	79.3
Raptis *et al.* [88]	79.4
GMKL [135]	85.2
AFMKL [135]	91.3
Le *et al.* [62]	86.5
Kovashka *et al.* [57]	87.3
Weinland *et al.* [131]	90.1
Wang *et al.* [124]	85.6
Rodriguez *et al.* [92]	69.2

Table 3.5: Performance comparison of different methods on the UCF Sports dataset.

recognition rates being achieved by the use of three levels of the Laplacian pyramid across all datasets. It is demonstrated in this experiment that the Laplacian pyramid can capture salient structural and motion information with multiple scales residing in the raw video sequences. Therefore it provides an effective representation of human action.

3.4.4 3D Gabor filtering

To validate the use of 3D Gabor filtering on the Laplacian pyramid, we have also conducted experiments to evaluate the performance of the Laplacian pyramid and 3D Gabor filtering, independently. For 3D Laplacian pyramid, the max pooling operation is also applied on each level of the pyramid to obtain the descriptors. The comparison results are illustrated in Table 3.9. Note that both the 3D Laplacian and 3D Laplacian + 3D Gabor use a three-level Laplacian pyramid. On the four datasets, the Laplacian pyramid yields the worst recognition rates. The 3D Gabor filters perform much better than the Laplacian pyramid. As expected, the combination of the Laplacian pyramid with 3D Gabor filters (3D Laplacian + 3D Gabor) improves the performances of both the 3D Laplacian and the 3D Gabor filters. The results have shown the effectiveness

	#level	0	1	2	3	4
KTH	DLA	93.3	94.2	94.3	**95.0**	94.8
	PCA	92.3	93.8	94.2	94.3	93.7
UCF	DLA	92.5	93.4	**93.9**	93.4	**93.9**
	PCA	89.7	92.0	91.9	91.3	91.3

Table 3.6: Performance of STLPC with different levels of the Laplacian pyramid and different dimensionality reduction techniques on the KTH and UCF sports datasets.

	#level	0	1	2	3	4
Cam1	DLA	89.9	90.6	89.0	**91.8**	90.5
	PCA	83.4	85.1	85.6	87.1	88.1
Cam2	DLA	88.1	87.6	**89.4**	88.8	88.6
	PCA	83.0	84.9	84.3	85.5	85.2
Cam3	DLA	90.1	91.7	91.5	**91.7**	91.1
	PCA	87.7	89.0	89.4	89.6	89.1
Cam4	DLA	86.9	87.1	87.1	**87.4**	86.7
	PCA	80.6	81.2	83.7	82.7	83.0
Cam5	DLA	80.4	80.6	80.6	**81.8**	80.9
	PCA	73.5	77.8	78.8	79.1	78.5

Table 3.7: Performance of STLPC with different levels of the Laplacian pyramid and different dimensionality reduction techniques on the IXMAS datasets.

of the 3D Gabor filters on the Laplacian pyramid.

3.4.5 Difference of Frames

Difference of Frames (DoF) is an important preprocessing step in the whole framework of our method. To investigate the effect of DoF on the performance of our method, we have performed extensive experiments to evaluate the contribution of DoF. Since for the IXMAS dataset, silhouettes are used so no difference of frames is performed on this dataset. We conducted the experiments on the KTH, UCF and HMDB51 datasets. The results are reported in Table 3.10. Note that to make a fair comparison, we keep all the settings exactly the same for experiments with DoF and without DoF.

As expected, we can see in Table 3.10 that the results with DoF are significantly better than those without DoF on all three datasets. Looking into the results, we

	#level	0	1	2	3	4
S1	DLA	32.8	37.2	36.8	**37.4**	37.0
	PCA	33.0	36.0	37.1	36.5	37.9
S2	DLA	33.0	35.8	37.4	**40.9**	35.3
	PCA	32.1	36.7	36.7	39.1	33.2
S3	DLA	30.2	34.2	**35.4**	34.7	33.7
	PCA	29.1	32.3	34.9	34.2	32.8
Average	DLA	32.0	35.7	36.5	**37.3**	35.3
	PCA	31.4	35.0	36.2	34.6	34.6

Table 3.8: Recognition rates (%) on three training/testing splits (S1, S2 and S3) of a subset (*i.e.* general body movements) of the HMBD51 dataset.

Features	KTH	IXMAS	UCF	HMDB51
3D Laplacian	89.5	79.0	71.7	15.1
3D Gabor	93.3	89.9	92.5	32.0
3D Laplacian + 3D Gabor	**95.0**	**91.8**	**93.4**	**37.3**

Table 3.9: The comparison of the 3D Laplacian, the 3D Gabor filters and the combination of them.

	#level	0	1	2	3	4
KTH	DoF	93.3	94.2	94.3	**95.0**	94.8
	No-DoF	90.9	91.2	91.6	91.5	91.3
UCF	DoF	92.5	93.4	**93.9**	93.4	**93.9**
	No-DoF	77.8	79.3	79.9	79.1	81.2
HMDB51	DoF	32.0	35.7	36.5	**37.3**	35.3
	No-DoF	25.0	27.9	28.9	26.8	27.1

Table 3.10: Performance of STLPC with and without DoF on the KTH, UCF Sports and HMDB51 datasets.

have found that, on the KTH dataset with relatively simple and clear backgrounds, DoF improves the performance, but the improvements are less significant than on the realistic UCF sports and HMDB51 datasets in which the backgrounds are very complicated and cluttered. This is reasonable because, for realistic datasets such as UCF sports and HMDB51, the backgrounds confuse the foreground actions while DoF could effectively suppress the backgrounds.

3.4.6 Dimensionality reduction

Because the final descriptors obtained from the feature extraction steps are of high dimensionality, we perform a dimensionality reduction technique named discriminative locality alignment (DLA) to make it compact and more discriminative. Note that, in the DLA algorithm, principal component analysis (PCA) is first employed for denoising, and we retain 98% of the energy (variance). To evaluate the contribution of DLA to our method, we perform PCA as a baseline for comparison and the same 98% of energy is kept. The results are also shown in Table 3.6, Table 3.7 and Table 3.8. Obviously, DLA outperforms PCA across all the four datasets significantly. Note that, even with PCA, our method can still achieve satisfactory performance and is superior to other methods which again implies that the Laplacian pyramid can encode actions informatively and discriminatively.

3.5 Conclusion

In this chapter, we have introduced the spatio-temporal Gaussian/Laplacian pyramids for multi-resolution video analysis and have proposed a novel global descriptor named spatio-temporal Laplacian pyramid coding (STLPC) for the holistic representation of human actions. In the pyramid model, a sequence with action is decomposed into a series of band-pass filtered components, in which spatio-temporal salient features with various sizes can be well localised and enhanced. Following the Laplacian pyramid, a bank of 3D Gabor filters and max pooling are successively applied to extract discriminative and invariant spatio-temporal features. Because convolving Gabor filtering

and max pooling are all performed over spatial and temporal dimensions, motion and structural information is well preserved in the representation.

In contrast to existing holistic representation methods, most of which depend heavily on accurate and even carefully tuned tracking and localisation algorithms, the proposed method can work well with coarse bounding boxes. The proposed method provides an effective and efficient avenue for holistic human action representation. Evaluations on four increasingly difficult datasets, KTH, IXMAS, UCF sports and HMDB51, suggest that the proposed STLPC is a very promising global descriptor for human action recognition.

Chapter 4

Spatio-temporal Oriented Energies

4.1 Introduction

In Chapter 3, we have proposed a global descriptor, namely the spatio-temporal Laplacian pyramid coding (SPLPC), for holistic representation of actions. STLPC takes the advantages of both multi-resolution and orientational analysis due to the use of Laplacian pyramid and Gabor filters. However, STLPC is relatively computationally expensive because of the 3D convolution in the 3D Gabor filtering. To be more efficient, in this chapter we propose to apply the spatio-temporal steerable filter for the multi-resolution and orientational analysis. Moreover, different from STLPC, spatio-temporal oriented energies are computed for the holistic representation of actions.

4.1.1 Motivations

Low-level features serve as the basis of both mid-level [25, 61, 99, 54] and high-level [95] representations of human actions. Features based on oriented gradients have been widely and successfully extended from the image domain into video analysis and action recognition [61, 99, 54]. Oriented filters play a key role in early vision and in image processing [28, 105, 34]. Freeman and Adelson [28] proposed steerable filters to synthesise filters of arbitrary orientations for linear combinations of basis filters.

Steerable filters can efficiently perform multiple orientation analysis.

Inspired by the success of steerable filters in object classification [79] and video analysis [132], we propose a novel, holistic representation based on the spatio-temporal steerable pyramid (STSP).

A steerable pyramid [34] is non-orthogonal and over-complete, which shows the desirable property of shift and rotation invariance. It is a transform that combines multi-scale decomposition with differential measurements, capturing the oriented structures in spatio-temporal volumes.

We adopt the second-order 3D Gaussian derivative as the steerable basis, which is more efficient than higher orders, *e.g.* the third order, while preserving satisfactory performance. In addition, to obtain robustness to scale variation, we apply the spatio-temporal max pooling operation to the responses of adjacent scales of the steerable filtering, which makes our method distinctive from the previous work [79, 132].

4.1.2 Overview

Given a 3D volume, which in our case can be the intensity volume, optical flow and 3D gradients of a video sequence, a spatio-temporal Laplacian pyramid structure is first constructed. The volume is decomposed into a set of sub-band volumes which can segregate and enhance spatio-temporal features residing in different scales.

To efficiently explore oriented patterns in video sequences, a bank of spatio-temporal steerable filters with different scales is then applied to each level of the obtained Laplacian pyramid. These filters are separable, steerable filters in three dimensions (X-Y-T) and therefore can be computed efficiently (Section 4.3).

Motivated by the previous work [132], we employ a representation based on spatio-temporal local energies which are calculated from the quadrature pairs of responses of the filtering on voxels in each volume (Section 4.4).

Finally, a feature pooling operation, *i.e.*, max pooling, is performed between adjacent scales of the steerable filters and over local spatio-temporal neighbourhoods, which makes the final representation more robust and less sensitive to scaling and shifts. In addition, features become more compact after the max pooling. The

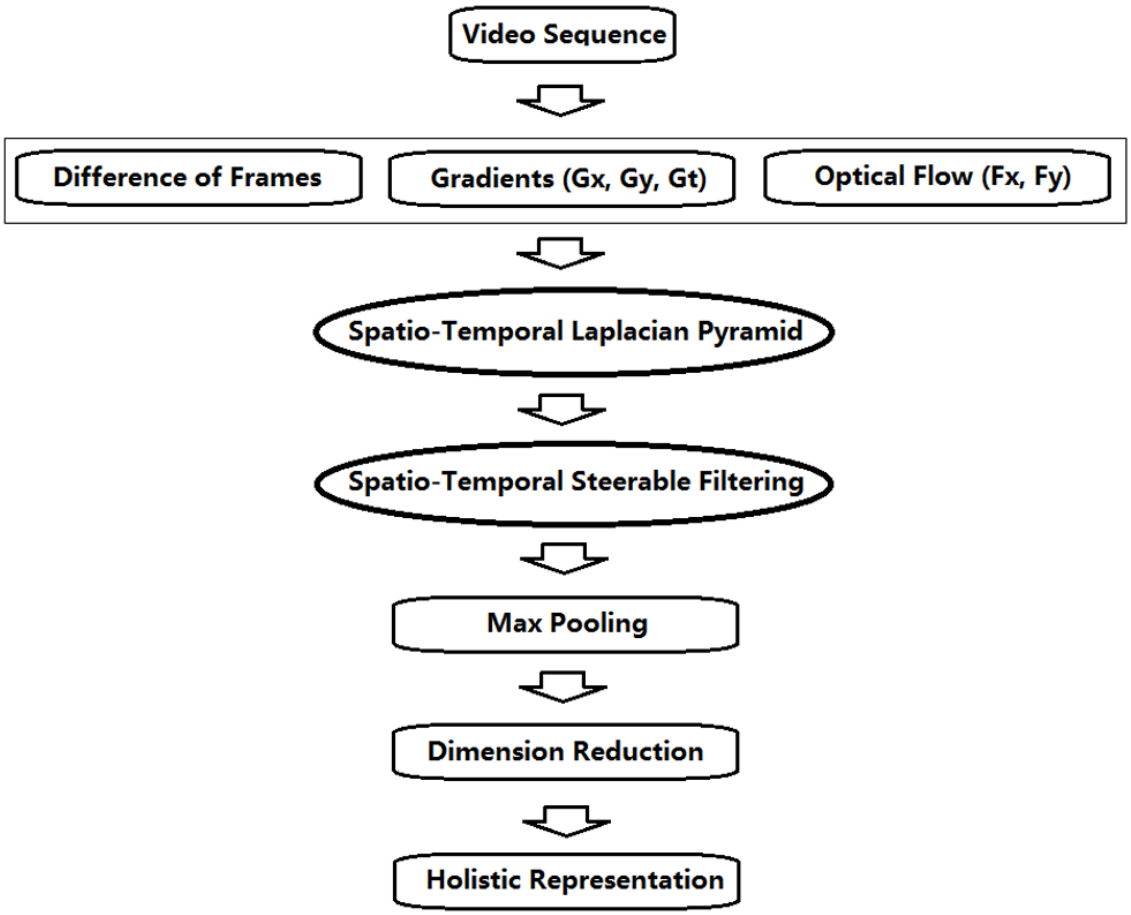

Figure 4-1: The flowchart of feature extraction.

flowchart of feature extraction is illustrated in Fig 4-1 (Section 4.4).

4.1.3 Contributions

The contributions of the proposed method can be summarised as follows:

- A new model based on spatio-temporal steerable pyramid is proposed for action recognition.
- Local oriented energies as spatio-temporal features are first employed for holistic representation of human actions.
- The max pooling operation is adopted into the spatio-temporal steerable pyramid model, which makes features less sensitive to scaling and shifting obtaining a

robust and compact representation.

4.2 Related work

In this section, we review the previous work that is closely related to our method in this chapter, especially focusing on those using steerable filters for video analysis.

By applying the steerable filtered features, Wildes and Bergen [132] presented an approach for qualitative spatio-temporal analysis using an oriented energy representation. This work is deemed as the representational substrate for indexing videos and other spatio-temporal data.

Derpanis and Gryn [22] detailed the construction of three-dimensional separable steerable filters, which extends the construction of two-dimensional separable steerable filters outlined in [28]. The separable and steerable implementations lead to compact and efficient computation.

With the quadrature outputs of the steerable filters, local oriented energy representations have been explored for spatio-temporal grouping [24], efficient action spotting [23] and visual tracking [16].

Derpanis *et al.* [24] adopted an oriented energy representation for grouping raw image data into a set of coherent spatio-temporal regions. This representation describes the presence of particular oriented spatio-temporal structures in a distributed manner to capture multiple oriented structures at a given location. They further designed a descriptor based on the oriented energy measurements for action spotting [23].

Along the same line, Cannons *et al.* [16] proposed a pixel-wise spatio-temporal oriented energy representation for visual tracking. The proposed representation is extremely rich, as it includes appearance and motion information as well as information about how these descriptors are spatially arranged.

Recently, Saraband and Cors [95] presented a high-level representation, *i.e.*, action bank, for human action recognition in which oriented energy features are used to generate action templates for bank detectors and a spatio-temporal orientation

decomposition is realised using broadly tuned 3D Gaussian third derivative filters.

4.3 Spatio-temporal steerable pyramid

Similar to STLPC in Chapter 3, we view a video sequence as a spatio-temporal intensity volume that contains all structural and motion information of the action, including poses of the human figure at any time as well as the dynamic transitions between the poses. Inspired by the success of spatio-temporal Laplacian pyramid (STLP) in Chapter 3, we incorporate the STLP structure as the first step to construct the global descriptor in this chapter. To efficiently explore the orientation features, instead of using Gabor filters, we propose to apply the spatio-temporal steerable filters on each level of the Laplacian pyramid, thus yielding the spatio-temporal steerable pyramid (STSP).

4.3.1 Spatio-temporal steerable filtering

Local oriented structures are important for the representation of spatio-temporal data, especially in motion analysis. From a purely geometric point of view, orientation captures the local first-order correlation structure of a pattern [132]. Motion can be perceived as patterns of appropriate spatio-temporal orientation. Spatio-temporal oriented filters are suitable for analysis of motion because they are able to explore orientation information both spatially and temporally.

A steerable filter [28] is an orientation-selective convolution kernel used for image enhancement and feature extraction that can be expressed via a linear combination of a small set of rotated versions of itself. For any spatio-temporal function $f(x, y, t)$, $f^{\theta}(x, y, t)$ is $f(x, y, t)$ rotated through an angle θ about the origin. This can be formulated as follow:

$$f^{\theta}(x, y, t) = \sum_{j=1}^{M} k_j(\theta) f^{\theta_j}(x, y, t). \tag{4.1}$$

We use the second-order derivative of Gaussian G_2^{θ} with multiple scales as the steer-

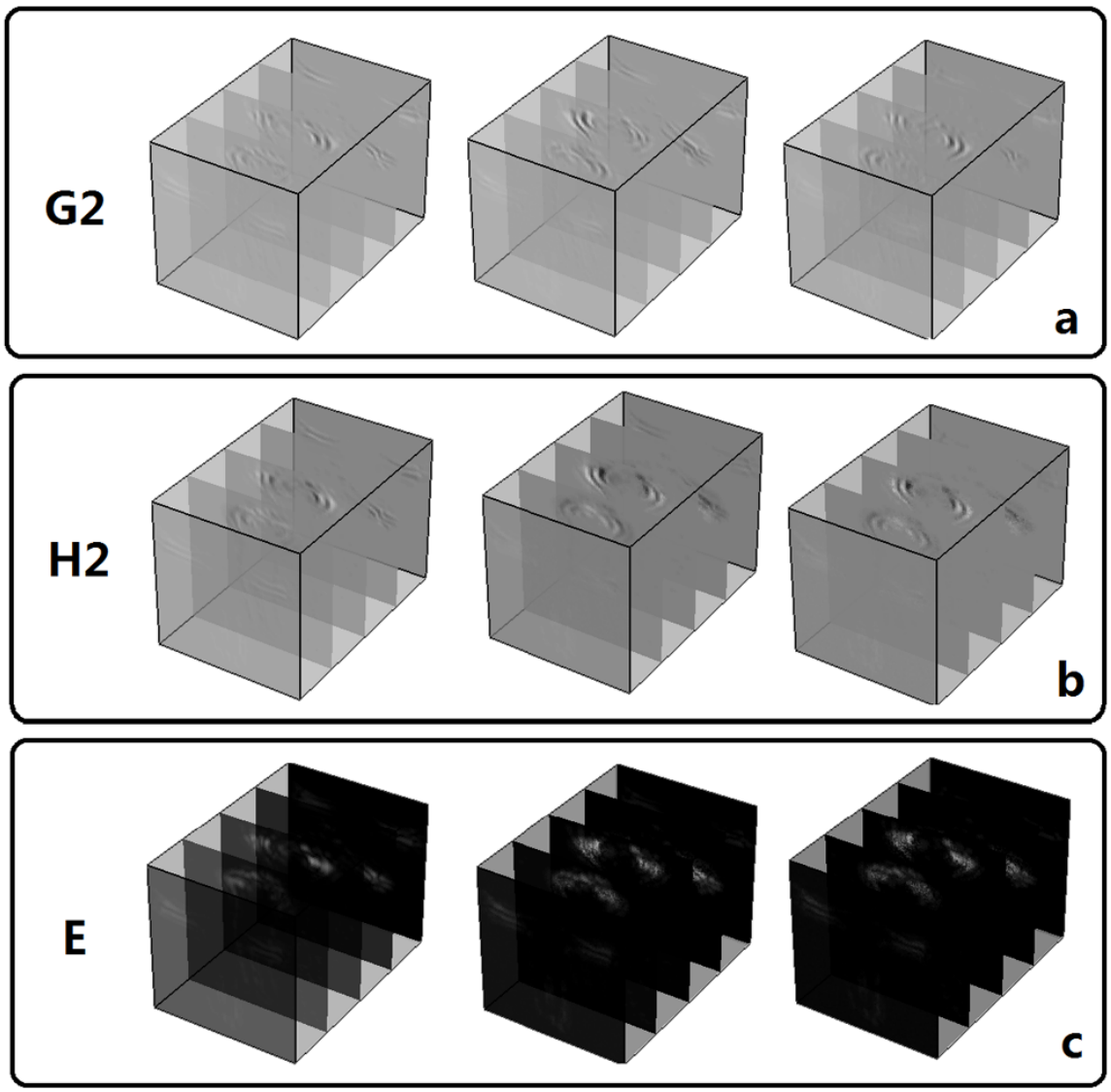

Figure 4-2: (a) and (b): The quadratic pair of the responses from the steerable filters in three orientations; (c): The local energies of the quadratic pairs.

able basis. H_2^θ is its Hilbert transform. The quadrature pair allow for analysing spectral strength independent of phase. They are widely used for motion, texture and orientation analysis [1, 12, 38, 81]. To be computationally efficient, we adopt the three-dimensional separable steerable filter [22]. Fig. 4-2 (a) and (b) illustrates the examples of the quadratic pair of the responses from the steerable filters in three orientations on the first level of the Laplacian pyramid. We can easily see that features in different orientations are enhanced.

4.4 Feature extraction

Actions occurring in a video sequence are mainly composed of appearance and motion. To explicitly exploit and capture them, we build the STSP descriptor on the low-level features including the intensity, gradients and optical flow.

4.4.1 Low-level features

We applied STSP to the spatio-temporal volumes with intensity, gradients and optical flow to extract the features, respectively.

Intensity To capture the appearance of actions, subtraction is performed between adjacent frames in each raw video sequence, obtaining a volume with difference of frames (DoF). The motion-related human body information is enhanced and backgrounds are largely suppressed.

Gradients To extract local intensity changes, we have also applied STSP to the 3D gradients of the volume with DoF. More specifically, for each voxel, we first compute the gradients, G_x, G_y and G_t along X, Y and T, and then perform STSP on two volumes: $G_{xt} = G_t/(|Gx| + 1)$ and $G_{yt} = G_t/(|Gy| + 1)$.

Optical flow With regards to motion, we use the Lucas-Kanade method [73] to estimate the optical flow in horizontal and vertical directions, which is efficient to compute. The volumes with DoF and optical flow will then be fed into the STSP.

4.4.2 Local oriented energies

Actions can be regarded as spatio-temporal patterns with energies in different orientations. In light of the previous work [132], to eliminate the phase variations, we produce a measure of local energy $E(x, y, t)$ within each scale and orientation.

Consider a point (x, y, t) in a video sequence. Its energy for a certain orientation can be obtained by the following formula:

$$E^{\theta}(x, y, t) = [G_2^{\theta} * I(x, y, t)]^2 + [H_2^{\theta} * I(x, y, t)]^2, \tag{4.2}$$

where $I(x, y, t)$ is a spatio-temporal volume.

The local energy is a motion measurement of phase independence. Since local energies are calculated from the quadratic pair of the outputs of the steerable filtering, motion patterns with multiple scales and orientations are efficiently captured. The local oriented energy model provides a robust and efficient representation of actions. The example of the local energies is shown in Fig. 4-2 (c), in which we can see that motion-related features are highlighted and more invariant in the measurement of local energies.

4.4.3 Feature pooling

Inspired by the success of max pooling techniques for spatio-temporal feature pooling in Chapter 3, we also incorporate the spatio-temporal max pooling into STSP to obtain insensitivity to image transforms, more compact representations, and better robustness to clutter [11].

4.4.4 Dimensionality reduction

Based on the experimental results in Chapter 2 and Chapter 3, to obtain a more compact representation, a dimensionality reduction technique named discriminative locality analysis (DLA) [144] has be employed for feature reduction. Principal component analysis (PCA) has also been used for comparison.

#Level	0	1	2	3	4
Intensity	83.5%	86.0%	88.5%	89.3%	87.0%
DoF	87.5%	89.5%	91.3%	92.1%	91.5%
Optical Flow	87.6%	90.1%	91.0%	91.1%	91.1%
Gradients	90.8%	91.5%	92.1%	92.5%	92.3%
DoF + Optical Flow	90.0%	93.5%	93.3%	93.2%	93.5%
DoF + Gradients	90.0%	93.5%	94.1%	**94.5%**	94.3%
DoF + Optical flow + Gradients	91.0%	92.6%	93.8%	94.2%	94.0%

Table 4.1: Performance of STSP with different levels of the Laplacian pyramid on KTH.

4.5 Experiments and results

We evaluate the proposed method, *i.e.*, STSP, on the baseline KTH dataset, the UCF Sports and the newly released HMDB51 dataset. In order to investigate the effect of parameters of the model, we have done comprehensive experiments to evaluate different numbers of levels of the pyramid, and the contribution of discriminative locality alignment (DLA) has also been evaluated by the comparison with principal component analysis (PCA). In addition, as the max pooling is integration component of the model, we have also conducted experiments to explore the performance of max pooling.

4.5.1 Experimental settings

We follow the validation settings in Chapter 3 on all the datasets. We have deliberately used very coarsely extracted bounding boxes or even no bounding boxes for the HMDB51 dataset to demonstrate the effectiveness of our method in realistic scenarios. With advanced person detection and tracking techniques, more accurate bounding boxes are possible and will undoubtedly lead to even better performance of the proposed STSP descriptor. A linear support vector machine (SVM) is used for action classification [18].

4.5.2 Parameter evaluation

The results on the KTH, UCF Sports and HMDB51 datasets are illustrated in Table 4.1, Table 4.2 and Table 4.3. We can see from the tables that the recognition rates increase with the increase of the level number of the pyramid for all the features and their combinations, which shows the effectiveness of the Laplacian pyramid. Note that the best results on KTH and UCF Sports occur with the three-level Laplacian pyramid, while on HMDB51 they happen with the four-level Laplacian pyramid, which implies that more information is needed to encode actions in this dataset because of its complexity and large intra-class variations. Compared with the 3D Gabor filters, 3D steerable filters are computationally more efficient. For a video sequence, the run time of 3D Gabor filters and steerable filters are 34.9s and 335.2s, respectively.

On the KTH dataset in Table 4.1, it can be seen that the use of DoF does improve the performance, which demonstrates that DoF is capable of suppressing backgrounds and validates the use of DoF.

In addition, it is obvious that feature combinations can significantly improve the performance and the best results are achieved by DoF + Gradients. The combination of DoF, gradients and optical flow has achieved better results than each single feature while slightly lower than DoF + Gradients.

The results on UCF Sports shown in Table 4.2 are consistent to those on KTH, namely DoF significantly improves the performance over intensity without DoF. Slightly different from those on KTH, the best results happen with the combination of DoF, optical flow and gradients, which manifests the complementarity of these three features. Also we can see that any combination of two features, *i.e.*, DoF + Optical Flow or DoF + Gradients, outperforms single features.

The results on the HMDB51 dataset are reported in Table 4.3. The HMDB51 dataset is regarded as a very challenging dataset with realistic actions. We use the original video sequences without any bounding boxes to demonstrate the capability of our method on totally unconstrained data. The trends of the performance on this dataset are generally consistent with those on KTH. Multiple levels of the pyramid

#Level	0	1	2	3	4
Intensity	63.6%	65.4%	66.3%	68.4%	67.1%
DoF	64.4%	74.1%	74.8%	73.6%	72.9%
Optical Flow	68.3%	73.8%	73.7%	73.6%	73.6%
Gradients	65.7%	69.0%	69.7%	69.7%	68.3%
DoF + Optical Flow	65.8%	76.0%	77.3%	76.7%	76.7%
DoF + Gradients	65.6%	76.6%	78.0%	78.0%	78.0%
DoF + Optical flow + Gradients	71.0%	79.4%	80.1%	**80.7%**	**80.7%**

Table 4.2: Performance of STSP with different levels of the Laplacian pyramid on UCF Sports.

#Level	0	1	2	3	4
Intensity	18.3%	21.5%	22.5%	21.5%	21.2%
DoF	21.1%	24.1%	24.8%	25.0%	25.6%
Optical Flow	20.5%	23.1%	25.7%	28.1%	28.1%
Gradients	17.9%	19.2%	20.9%	20.5%	21.4%
DoF + Optical Flow	24.6%	27.2%	29.7%	31.0%	**31.7%**
DoF + Gradients	20.4%	24.4%	25.9%	27.0%	27.8%
DoF + Optical flow + Gradients	24.6%	27.1%	29.7%	31.0%	31.6%

Table 4.3: Performance of STSP with different levels of the Laplacian pyramid on HMDB51.

increases the performance compared with a single level. Slightly different to KTH, DoF appears to be more effective on HMDB51, which is reasonable because the background variations and clutter in the realistic dataset are more serious. In addition, the best result -31.7%- is achieved by the combination of DoF and Optical Flow. The result of the combination of DoF, Gradients and Optical Flow is comparable with the best result by DoF + Optical Flow. As we can see from Fig. 4-5, the actions in this dataset are realistic and challenging with complex background variations, while results on the HMDB51 dataset are encouraging in that no bounding boxes and tracking are used in our method.

It would be interesting to look into the results. We plot the confusion matrix of the results for each action category in Fig. 4-3 for the KTH dataset. The results are achieved by the combination of DoF and Optical Flow. We can see from the confusion matrix that STSP can successfully recognise *Boxing*, *HandClappling*, *HandWaving*

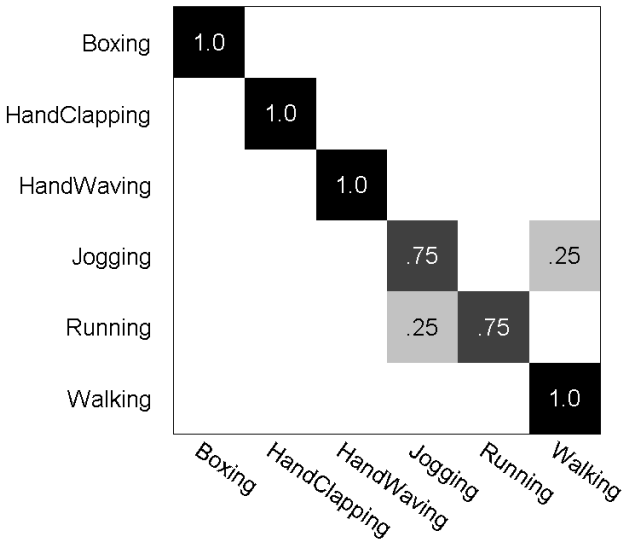

Figure 4-3: The confusion matrix of the results on KTH.

and *Walking*, with recognition rates of 100%. The recognition errors mainly happen on *Jogging* and *Running* which are difficult to distinguish even with the human eye. This is reasonable as these two actions share many similar motion patterns, and are almost the same sequence. Encouragingly, STSP can fully distinguish *Walking* from *Running* and *Jogging*, both of which are quite similar to *Walking* and cause confusion for recognition.

The confusion matrix for the UCF Sports dataset is plotted in Fig. 4-4. The illustrated are the results of the combination of DoF, Optical Flow and Gradients. STSP can successfully recognise the *Kicking* action with 100% recognition rate. The action *SkateBoarding-Front* is severely confused with the action *Golf*. Looking into these two actions, we find that they share similar appearances in many video samples.

For the HMDB51 dataset, the confusion matrix of the average recognition rates over the three splits is plotted in Fig. 4-5. The results are achieved by the combination of DoF and Gradients. In spite of the challenges in this dataset, STSP is still able to recognise a few actions such as *pullup*, *pushup* and *climb* with relatively high accuracies, which demonstrates the potential of STSP for the recognition of unconstrained realistic actions.

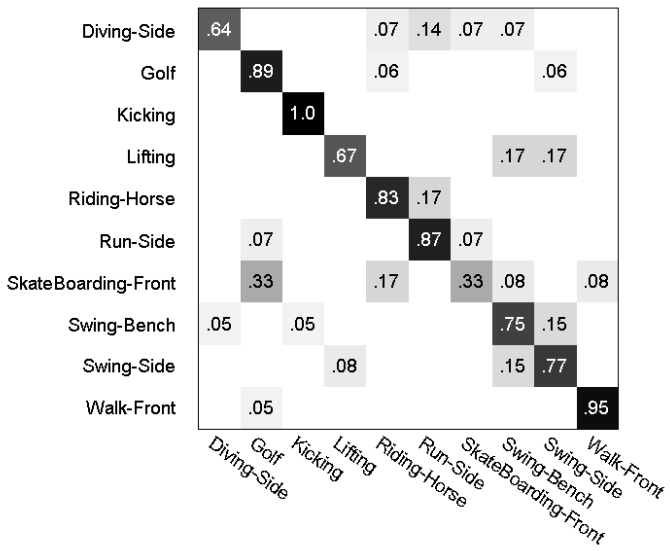

Figure 4-4: The confusion matrix of the results on UCF Sports.

	#level	0	1	2	3	4
KTH	Max Pooling	87.5%	89.5%	91.3%	**92.1%**	91.5%
	Average Pooling	86.3%	88.5%	90.1%	89.6%	89.9%
UCF Sports	Max Pooling	64.4%	74.1%	**74.8%**	73.6%	72.9%
	Average Pooling	66.9%	67.2%	70.1%	67.3%	68.6%
HMDB51	Max Pooling	21.1%	24.1%	24.8%	25.0%	**25.6%**
	Average Pooling	20.9%	21.2%	22.9%	23.5%	23.7%

Table 4.4: Performance of STSP with and without max pooling on the KTH, UCF Sports and HMDB51 datasets. Note that these results are obtained using DoF as the input.

4.5.3 Feature pooling

To investigate the contribution of max pooling to the overall performance of our method, we have conducted experiments to compare the results with and without max pooling. Note that these experiments are carried on the features with DoF. The results are shown in Table 4.4. As expected, the max pooling operation does improve the performance on all the three datasets.

Interesting, the max pooling operation makes a more impact on the realistic datasets, *i.e.*, UCF Sports and HMDB51, than the KTH dataset.

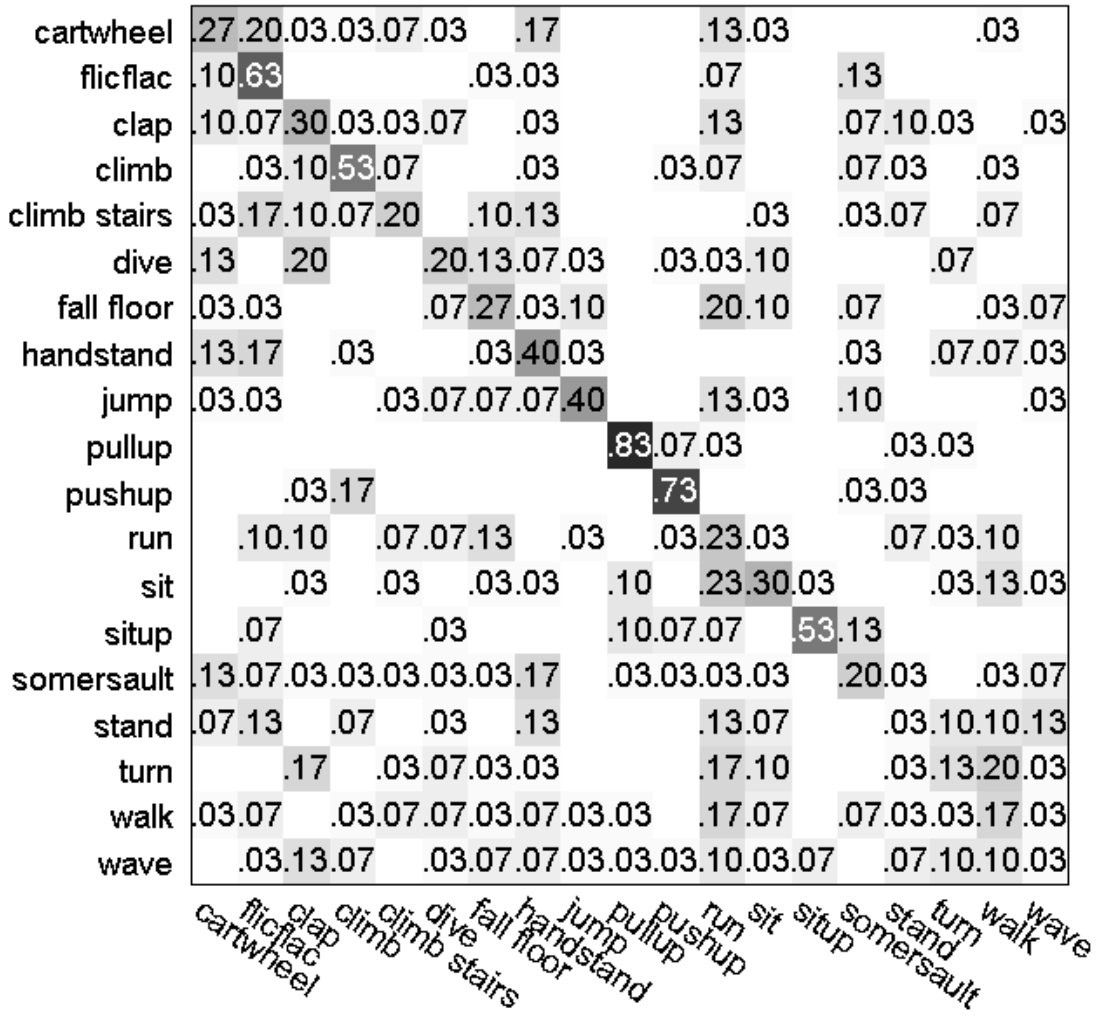

Figure 4-5: The confusion matrix of the results on HMDB51.

#level	0	1	2	3	4
DLA	90.0%	93.5%	94.1%	**94.5%**	94.3%
PCA	88.8%	91.6%	92.3%	91.8%	91.8%

Table 4.5: Performance of STSP with different dimensionality reduction techniques on the KTH dataset. The results are obtained by DoF + Gradients.

#level	0	1	2	3	4
DLA	71.0%	79.4%	80.1%	**80.7%**	**80.7%**
PCA	64.3%	67.5%	70.2%	70.2%	70.2%

Table 4.6: Performance of STSP with different dimensionality reduction techniques on the UCF-Sports dataset. The results are obtained by DoF + Optical flow + Gradients.

4.5.4 Dimensionality reduction

The results of using discriminant locality alignment (DLA) and principal component analysis (PCA) for dimensionality reduction are shown in Table 4.5, Table 4.6 and Table 4.7 for KTH, UCF Sports and HMDB51, respectively. We use the same settings as in Chapter 3 that we keep 98% energy both for PCA and DLA.

On all three datasets, DLA consistently outperforms PCA with all levels of the pyramid, which validates the use of DLA for dimensionality reduction in action recognition. Interestingly, we can find in Table 4.5 and Table 4.7 that the performance does not change significantly with different numbers of levels of the Laplacian pyramid when PCA is used for feature reduction. With DLA, we find that the performance increases with the increase of the level number of the Laplacian pyramid, which indicates that DLA can effectively extract the discriminative information residing in each level of the Laplacian pyramid.

4.5.5 Comparison with state of the art

In Table 4.8, we conduct the comparison of STSP with the state-of-the-art methods on the KTH dataset. To make fair comparisons, we only compare with the published results using holistic representations. STSP outperforms all the holistic methods

	#level	0	1	2	3	4
S1	DLA	26.7%	28.8%	32.5%	33.5%	**33.7%**
	PCA	27.5%	28.1%	27.4%	27.7%	27.4%
S2	DLA	25.8%	28.3%	30.7%	**32.3%**	33.2%
	PCA	23.9%	25.4%	24.4%	24.7%	24.9%
S3	DLA	21.4%	24.4%	25.8%	27.2%	**28.1%**
	PCA	23.9%	24.3%	24.6%	24.9%	24.7%
Average	DLA	24.6%	27.2%	29.7%	31.0%	**31.7%**
	PCA	25.1%	25.9%	25.4%	25.8%	25.7%

Table 4.7: Performance of STSP with different dimensionality reduction techniques on the HMDB51 dataset. S1, S2 and S3 denote the three training/test splits.

Methods	Accuracy
STSP	**94.5**
Jhuang *et al.* [46]	91.7
Schindler *et al.* [97]	90.9
Yeffet *et al.* [139]	90.1
Taylor *et al.* [111]	90.0
Jr *et al.* [47]	90.2

Table 4.8: A longitudinal performance comparison of different methods on the KTH dataset.

listed in the table. Moreover, STSP works more efficiently than other methods such as in [111, 47] in which deep learning, and convolutional neural networks are used, respectively. The comparison with state-of-the-art results on the UCF Sports and HMDB51 datasets are illustrated in Table 4.9 and Table 4.10.

Method	Accuracy
STSP	**80.7%**
Yeffet *et al.* [139]	79.3%
Rodriguez *et al.* [92]	69.2%

Table 4.9: Performance comparison of different methods on the UCF Sports dataset.

Method	Accuracy
STSP	**31.7%**
Kuehne *et al.* [58]	22.8%
Saraband *et al.* [95]	26.9%
Orit *et al.* [55]	29.2%

Table 4.10: Performance comparison of different methods on the HMDB51 dataset.

4.6 Conclusion

In this chapter, we have introduced a compact and efficient holistic representation of human actions. By decomposing a video sequence with a Laplacian pyramid, spatio-temporal salient features with various sizes can be well localised and enhanced. Multi-scale steerable filters can efficiently extract features in multiple scales and orientations. The spatio-temporal max pooling operation makes features more compact but robust.

Extensive experiments have been conducted to investigate the influence of different components and parameters, *i.e.*, the difference of frames (DoF), the max pooling operation and dimensionality reduction techniques. The results validate their effectiveness.

In contrast to existing holistic representation methods, most of which depend heavily on accurate and even carefully-tuned tracking and localisation algorithms, the proposed method can work well with coarse or even no bounding boxes. Furthermore, due to the use of three-dimensional separable steerable filters, the spatio-temporal filtering can be efficiently performed. Evaluations on three increasingly difficult datasets, *i.e.*, KTH, UCF Sports and HMDB51, demonstrate that the proposed STSP is a promising global descriptor for human action recognition.

4.7 Summary of holistic methods

Up until now, we have proposed three global descriptors for holistic representations of human action. We summarise their performance in Table 4.11.

On the KTH dataset, the three descriptors produce comparable results. STLPC

Methods	KTH	UCF Sports	HMDB51
SP+MHI	93.5%	93.9%	-
STLPC	95.0%	93.9%	37.3%
STSP	94.5%	80.7%	31.7%

Table 4.11: The summary of performance of holistic methods.

and STSP slightly outperform SP + MHI. Because both STLPC and STSP take a video as a whole, all the cues of actions including structure and motion features are well preserved and therefore they are more descriptive than SP + MHI.

On the UCF Sports dataset, SP + MHI produces comparable results with STLPC. The reason is that this dataset contains sports actions with backgrounds closely related to the actions, so it could be meaningful by sparsely sampling few frames to represent the video sequence. The appearance of actions in this dataest is more important than motion. This could also explain why SP performs much better than MHI in this dataset. STSP can not yield comparable results with SP + MHI and STLPC which could be due to that the computed optical flow and gradients are not accurate and can not provide complementary information to DoF.

On the HMDB51 dataset (Note that this dataset released after our SP + MHI method.), the performance of STSP is lower than STLPC which is consistent to the results on the KTH dataset.

Chapter 5

A Performance Evaluation on Local Methods

5.1 Introduction

Local methods based on spatio-temporal local features have drawn increasing attention from researchers in visual recognition. In this chapter, we do a comprehensive evaluation of local methods that have demonstrated to be effective and successful in both image/object classification and action recognition.

5.1.1 Motivations

Local features have played an important role in visual recognition. Methods based on local features, *e.g.*, the bag-of-words (BoW) model and sparse coding, have shown their effectiveness for image and object recognition in the past decades. Recently, many new techniques, including the improvements to BoW and sparse coding as well as the non-parametric naive Bayes nearest neighbour (NBNN) classifier, have been proposed and advanced the state-of-the-art in the image domain.

However, in the video domain, the BoW model still dominates the action recognition field. It is unclear how effective the state-of-the-art techniques widely used in the image domain would perform on action recognition. To fill this gap, we aim to imple-

ment and provide a systematic study on these techniques for action recognition, and compare their performance under a unified evaluation framework. Other techniques such as match kernels, which have also demonstrated their potential in handling local features, are also included for a comprehensive evaluation.

5.1.2 Contributions

The contributions of the work in this chapter lie in the following two aspects.

- We transfer some effective techniques including variants of the BoW model and sparse coding (SC), the naive Bayes nearest neighbour (NBNN) classifier and match kernels from the image domain to the video domain.
- We extensively evaluate the basic and widely used methods, *i.e.*, BoW, SC, NBNN and match kernels, for action recognition, which can be taken as a baseline for the feature research.

5.2 Related work

Performance evaluations have gained increasing attention in computer vision with a large number and variety of algorithms being developed. Plenty of evaluation and analysis work has been conducted both in the image domain [79, 142, 17, 117, 20, 19] and on action recognition [125, 101, 21, 110, 27].

A recent work [20] closely related to ours investigated the performance of unsupervised feature learning algorithms with single-layer networks on image classification. Surprisingly, the best performance from their evaluation is obtained by the BoW model with the so-called triangle assignment coding. In addition, Chatifeld *et al.* [19] presented a comprehensive evaluation and deep analysis of the feature encoding methods within the BoW model for image classification.

Two important evaluation works on action recognition were conducted by Wang *et al.* [125] and Shao and Mattivi [101]. They evaluated and compared the performance of different detectors and descriptors as well as their combinations for action recognition. However, both of them used only the standard BoW model for action

representation with a support vector machine (SVM) classifier.

Campos *et al.* [21] have compared the BoW model with spatio-temporal shapes (STS) for action recognition. Two versions of the BoW-based methods, namely spatially-constrained BoW (SBoW) and local-BoW (LBoW), were considered. The 3-dimensional histogram of oriented gradients (HOG3D) [54] was employed as the spatio-temporal descriptor.

Tamrakar *et al.* [110] evaluated low-level features and their combinations for complex event detection. Extensive low-level features, including static visual features and dynamic visual features, are adopted for comparison. Again, the BoW model has been utilised as the final representation in their work.

Recently, Everts *et al.* [27] have done an evaluation on colour STIPs for human action recognition. By incorporating the chromatic representations into the spatio-temporal domain, they reformulated the STIP detectors and descriptors for multi-channel video representation, which are shown to outperform the intensity-based counterparts.

The above evaluations were either in the image domain or centred on the BoW model. In this chapter, we are focused on the evaluation of state-of-the-art techniques on action recognition in the video domain.

5.3 Methods

In this section, we will review the coding methods of local feature to be evaluated in this chapter.

5.3.1 The Bag-of-words (BoW) model

Local features in the training set are first clustered to create a codebook [120]. All the video sequences are represented by coding local features with the visual words in the pre-learned codebook. The coding methods to be used in the BoW model include the hard assignment, the soft assignment [118], the triangle assignment [20] and the localised soft assignment [69].

Before describing the details of all the coding methods, we first define the notations used in both the BoW model and sparse coding (SC). Let $\mathbf{b}_i$ denote a visual word or a basis vector, and $\mathbf{B}_{D\times M}$ denote a codebook or a set of basis vectors, where D is the dimensionality of the local feature vectors and M is the number of codewords or bases. $\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N$ are local features from a video sequence, $\mathbf{u}_i \in R^M$ is the coding coefficient vector of $\mathbf{x}_i$ based on the codebook or basis vectors. u_{ij} is the coefficient associated with the word $\mathbf{b}_j$.

- **Hard assignment coding**

In the hard assignment coding, the coefficient of each local feature is determined by assigning this feature $\mathbf{x}_i$ to its nearest codeword in the codebook using a certain distance metric. If the Euclidean distance is used, then

$$u_{i,j} = \begin{cases} 1 & \text{if } j = \arg\min_{j=1,\cdots,M} ||\mathbf{x}_i - \mathbf{b}_j||_2^2 \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

- **Soft assignment coding**

In the soft assignment coding, The coefficient $u_{i,j}$ is the degree of membership of a local feature $\mathbf{x}_i$ to the jth codeword.

$$u_{ij} = \frac{\exp(-\beta||\mathbf{x}_i - \mathbf{b}_j||_2^2)}{\sum_{k=1}^{M} \exp(-\beta||\mathbf{x}_i - \mathbf{b}_k||_2^2)} \tag{5.2}$$

where β is the smoothing factor controlling the softness of the assignment.

- **Triangle assignment coding**

The triangle assignment coding was proposed in [20]. The coding is defined by the following activation function:

$$u_{ij} = \max\{0, \mu(\mathbf{z}) - z_j\} \tag{5.3}$$

where $z_j = ||\mathbf{x}_i - \mathbf{b}_j||_2$ and $\mu(\mathbf{z})$ is the mean of elements of $\mathbf{z}$. This activation function forces the output to be 0 for any feature $\mathbf{x}_i$ whose distance to the codeword b_j is larger than the average of all distances. As a result, roughly half of the weights will be set to 0.

- **Localised soft assignment coding (LSC)**

By combining the ideas of localisation and the soft assignment coding, Liu *et al.* [69] proposed localised soft-assignment coding (LSC). The activation function takes the form in Eq. (5.2), but with the locality constraint as follows:

$$d(\mathbf{x}_i, \mathbf{b}_j) = \begin{cases} d(\mathbf{x}_i, \mathbf{b}_j), & \text{if } \mathbf{b}_j \in N_k(\mathbf{x}_i) \\ \infty & \text{otherwise.} \end{cases}, \tag{5.4}$$

where $d(\mathbf{x}_i, \mathbf{b}_j) = ||\mathbf{x}_i - \mathbf{b}_j||_2^2$, and N_k denotes the k-nearest neighbours of $\mathbf{x}_i$ defined by the distance $d(\mathbf{x}_i, \mathbf{b}_j)$.

5.3.2 Sparse coding

In sparse coding (SC), a local feature is represented by a linear combination of a sparse set of basis vectors. The coding coefficient is obtained by solving an l_1-norm regularised approximation problem [75]:

$$\mathbf{u}_i = \arg\min_{u \in R^n} ||\mathbf{x}_i - \mathbf{B}\mathbf{u}||_2^2 + \lambda ||\mathbf{u}||_1, \tag{5.5}$$

where λ controls the sparsity of the coefficients.

- **Locality-constrained linear coding (LLC)**

Instead of enforcing sparsity in SC, LLC [127] confines a local feature $\mathbf{x}_i$ to be coded by its local neighbours in the codebook. The locality constraint ensures that similar patches would have similar codes. The coding coefficient is obtained by solving the following optimisation problem:

$$\begin{aligned} \mathbf{u}_i = \arg\min_{u \in R^n} &||\mathbf{x}_i - \mathbf{B}\mathbf{u}||_2^2 + \lambda ||\mathbf{d}_i \odot \mathbf{u}||_2^2, \\ s.t. \quad &\mathbf{1}^T \mathbf{u}_i = 1 \end{aligned} \tag{5.6}$$

where $\odot$ denotes element-wise multiplication, and $\mathbf{d}_i \in R^M$ is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the

input descriptor $\mathbf{x}_i$. Specifically,

$$\mathbf{d}_i = \exp[-\frac{dist(\mathbf{x}_i, \mathbf{B})}{\sigma}] \tag{5.7}$$

where $dist(\mathbf{x}_i, \mathbf{B}) = [dist(\mathbf{x}_i, \mathbf{b}_1), ..., dist(\mathbf{x}_i, \mathbf{b}_M)]^T$, and $dist(\mathbf{x}_i, \mathbf{b}_j)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{b}_j$. σ is used for adjusting the weight decay speed for the locality adaptor. As an approximation of LLC, one can simply use the k nearest neighbours of $\mathbf{x}_i$ as the local bases $\mathbf{B}_i$, and solve a much smaller linear system.

5.3.3 Match kernels

Match kernels between sets of local features have long been exploited [122, 74]. The kernel function is computed to measure the similarity between two images/video sequences represented by sets of local feature vectors.

Given two feature sets, $\mathcal{F}_a = \{\mathbf{x}_1^{(a)}, \ldots, \mathbf{x}_{|\mathcal{F}_a|}^{(a)}\}$ and $\mathcal{F}_b = \{\mathbf{x}_1^{(b)}, \ldots, \mathbf{x}_{|\mathcal{F}_b|}^{(b)}\}$, the summation kernel is defined as:

$$K_S(\mathcal{F}_a, \mathcal{F}_b) = \frac{1}{|\mathcal{F}_a|}\frac{1}{|\mathcal{F}_b|}\sum_{i=1}^{|\mathcal{F}_a|}\sum_{j=1}^{|\mathcal{F}_b|} K_F(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)}), \tag{5.8}$$

where K_F is the kernel of local features.

In [122], a kernel function (the max-sum kernel) for matching local features was proposed:

$$\begin{aligned} K_M(\mathcal{F}_a, \mathcal{F}_b) = &\frac{1}{2}\sum_{i=1}^{|\mathcal{F}_a|}\max_{j=1,\ldots,|\mathcal{F}_b|} K_F(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)}) \\ &+ \frac{1}{2}\sum_{j=1}^{|\mathcal{F}_b|}\max_{i=1,\ldots,|\mathcal{F}_a|} K_F(\mathbf{x}_j^{(b)}, \mathbf{x}_i^{(a)}) \end{aligned} \tag{5.9}$$

This match kernel has been used in object recognition [122] and action classification [60]. Lyu *et al.* [74] has proven it to be a non-Mercer kernel, and proposed a normalised sum-match kernel which satisfies the Mercer condition and is defined as

follows:

$$K_{\mathcal{F}}(\mathcal{F}_a, \mathcal{F}_b) = \frac{1}{|\mathcal{F}_a|}\frac{1}{|\mathcal{F}_b|}\sum_{i=1}^{|\mathcal{F}_a|}\sum_{j=1}^{|\mathcal{F}_b|}[K_F(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)})]^p, \tag{5.10}$$

where p is the model parameter.

5.3.4 Naive Bayes nearest neighbour (NBNN)

Naive Bayes Nearest Neighbour (NBNN) is an approximation of the optimal MAP (maximum a posteriori) Naive-Bayes classifier. Given an image Q represented as a set of local features, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, when the class prior $p(C)$ is uniform, MAP becomes the maximum likelihood (ML) classifier:

$$\hat{C} = arg\max_C p(C|Q) = arg\max_C p(Q|C). \tag{5.11}$$

With the Naive-Bayes assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. given its class C, we have

$$p(Q|C) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n|C) = \prod_{i=i}^{n} p(\mathbf{x}_i|C) \tag{5.12}$$

$p(\mathbf{x}_i|C)$ is further approximated using the Parzen density estimation and when the Parzen kernel keeps only the nearest neighbour and the same kernel bandwidth for all the classes, the resulting classifier takes the following simple form:

$$\bar{c} = arg\min_c \sum_{\mathbf{x}\in\mathbf{X}} ||\mathbf{x} - NN^c(\mathbf{x})||^2, \tag{5.13}$$

where $\sum_{\mathbf{x}\in\mathbf{X}} ||\mathbf{x} - NN^c(\mathbf{x})||^2$ is the image-to-class (I2C) distance from the image $\mathbf{X}$ to the class c, and NN^c is the nearest neighbour of $\mathbf{x}$ in class c.

5.3.5 NBNN kernels

A kernelised version of NBNN has been introduced in [116], which is shown to be complementary to the standard BoW model. The NBNN kernel takes advantage of

the main idea in NBNN, by using the Image-to-Class distance. Instead of directly classifying the image as the class with the minimum I2C distance, they concatenated the I2C distances from all the classes as a vector, which can be regarded as a high-level image representation. A linear support vector machine (SVM) is employed for image classification. The success of the NBNN kernel is largely attributed to the discriminative representation of an image by the I2C distances to its own class but also to classes it does not belong to. This representation gains more discriminative information in contrast to directly using the absolute I2C distance measurement. A similar idea in [143] has been validated that it is the collaborative representation, *i.e.*, using samples from all classes to represent the query sample, that improves face recognition rather than the l_1-norm constraint.

The NBNN kernel is based on the normalised sum match kernel [74], and is formulated as:

$$K(X,Y) = \sum_{c \in C} K^c(X,Y) = \frac{1}{|X||Y|} \sum_{c \in C} \sum_{\mathbf{x} \in X} \sum_{\mathbf{y} \in Y} k^c(\mathbf{x}, \mathbf{y}), \tag{5.14}$$

where $C = \{c\}$ and $k^c(\mathbf{x}, \mathbf{y})$ is the local kernel between local features. In the NBNN kernel, $k^c(\mathbf{x}, \mathbf{y})$ is defined as:

$$k^c(\mathbf{x}, \mathbf{y}) = \phi^c(\mathbf{x})^{\mathrm{T}} \phi^c(\mathbf{y}) = f^c(d^1_{\mathbf{x}}, \ldots, d^{|C|}_{\mathbf{x}})^{\mathrm{T}} f^c(d^1_{\mathbf{y}}, \ldots, d^{|C|}_{\mathbf{y}}) \tag{5.15}$$

Two distance functions have been considered in the original work [116], namely,

$$f^c_1(d^1_{\mathbf{x}}, \ldots, d^{|C|}_{\mathbf{x}}) = d^c_{\mathbf{x}} \tag{5.16}$$

and

$$f^c_2(d^1_{\mathbf{x}}, \ldots, d^{|C|}_{\mathbf{x}}) = d^c_{\mathbf{x}} - d^{\hat{c}}_{\mathbf{x}}, \tag{5.17}$$

where $d^{\hat{c}}_{\mathbf{x}}$ denotes the closest distance to all classes except for c.

5.3.6 Local NBNN

McCann and Lowe [77] developed an improved version of NBNN, named local naive Bayes nearest neighbour (LNBNN), which increases the classification accuracy and scales better with a large number of classes. The motivation of local NBNN is from the observation that only the classes represented in the local neighbourhood of a descriptor contribute significantly and reliably to their posterior probability estimation. Instead of finding the nearest neighbour in each of the classes, local NBNN finds in the local neighbourhood k nearest neighbours which may only come from some of the classes. The "localised" idea is shared with LSC in the BoW model and LLC in SC.

5.4 Experiments and results

We have conducted the experiments on three widely-used datasets including the KTH, UCF YouTube and HMDB51 datasets. We follow the validation settings that are commonly used in most of the previous works [125].

5.4.1 Experimental settings

In this section, we give the implementation details of each method evaluated in our experiments.

- **Spatio-temporal local features**

We employ the periodic detector proposed by Dollár at al. [25] to detect the spatio-temporal interest points from the raw video sequences and follow the parameter settings in the evaluation work of [125]. As in [20], the three-dimensional histogram of oriented gradients (HOG3D) [54] is used to describe each STIP due to its computational efficiency. The chosen detector and descriptor have shown outstanding performance in [125, 101]. For BoW and SC, we randomly select 100000 local features from the training set to learn codebooks and dictionaries.

The spatio-temporal pyramid matching (STPM) [68] can be easily embedded in the methods to encode the structural information and presumably could improve the

performance. As our focus is on the comparison between different methods rather than the overall performance, and we argue that STPM would equally contribute to each method, STPM is not used in our evaluation framework.

- **Feature pooling**

In BoW and SC, a final representation $\mathbf{P} \in R^M$ of an action is obtained by pooling over the coefficients [10]. With average pooling, the jth component of $\mathbf{P}$ is obtained by $p_j = \sum_{i=1}^{N} u_{ij}/N$. With max pooling, p_j is obtained by $p_j = \max_i u_{ij}$, where $i = 1, 2, \cdots, N$.

- **The BoW model**

In the BoW model, the codebooks are created by the k-means clustering algorithm provided in VLFeat toolbox [120]. In LSC, we follow the parameter settings in the original work [69] with β set as 10.

- **Sparse coding**

For sparse coding, we use the open-source optimisation toolbox SPAMS (SPArse Modelling Software) [1]. The dictionary is learned by the algorithm in [75], and the sparse codes are learned using orthogonal matching pursuit (OMP) [75]. The parameter λ in Eq. (5.5) is set 0.15. The number of non-zero coefficients is 10 in the OMP algorithm. For LLC, we use the released code with the same parameter settings.

- **Naive Bayes nearest neighbours (NBNN)**

As NBNN is non-parametric, no parameter is required to be tuned, while for the local NBNN classifier, the single parameter is the number of nearest neighbours k. We have investigated the effect of k in our experiments. With regard to the NBNN kernel, we have experimented with the distance function $f_2^c(d_{\mathbf{x}}^1, \ldots, d_{\mathbf{x}}^{|C|})$ in our implementation.

- **Match kernels**

For the match kernels, we use the linear kernel as the local kernel and the single parameter p in Eq. (5.10) is set as 9 according to the original work [74]. We also use the normalised kernel for building the SVM classifier: $K(x, y) \leftarrow \frac{K(x,y)}{\sqrt{K(x,x)}\sqrt{K(y,y)}}$.

- **Action classification**

[1] http://spams-devel.gforge.inria.fr/

Methods	KTH	YouTube	HMDB
BoW-Hard	87.9%	58.1%	20.0%
BoW-Soft-Average	85.4%	53.5%	19.6%
BoW-Soft-Max	89.2%	61.2%	24.0%
BoW-Triangle-Average	84.1%	52.5%	20.7%
BoW-Triangle-Max	89.8%	61.0%	25.1%
BoW-LSC	92.5%	59.4%	24.6%
SC-Average	91.0%	56.0%	23.3%
SC-Max	91.5%	59.4%	**27.9%**
SC-LLC	91.3%	56.2%	24.1%
NBNN	93.9%	57.8%	19.8%
NBNN Kernel	89.2%	**62.4%**	23.7%
Local NBNN	**94.1%**	60.1%	21.2%
Match Kernel	86.9%	54.5%	13.7%

Table 5.1: The performance of all methods on three datasets, *i.e.*, KTH, UCF-YouTube and HMDB51. Note that the results of the match kernel are obtained by $K_{\mathcal{F}}$.

We use a support vector machine (SVM) [18] classifier for BoW, SC and the match kernels. Note that a linear kernel instead of the χ^2 kernel in [125] is used in BoW and SC to make fair comparisons.

5.4.2 Results

All the final results on the three datasets are shown in Table 5.1. The size of the codebook in BoW and the number of bases in SC are hard to pre-determine while always affect the performance. Therefore, we have investigated the effects and illustrated the results in Fig. 5-1, 5-2, 5-3, 5-4, 5-5 and 5-6.

On the KTH dataset The best result is 94.1% obtained by the local NBNN classifier, which is comparable to state-of-the-art results from more complicated methods. The NBNN classifier achieves the second best result - 93.9%- which is slightly lower than the local NBNN classifier. In addition, the NBNN kernel gives a result of 89.2%, which is still better than the baseline hard assignment coding in BoW.

In the BoW model, LSC achieves an accuracy of 92.5% which is impressive considering its simplicity. The triangle assignment coding with max pooling is better

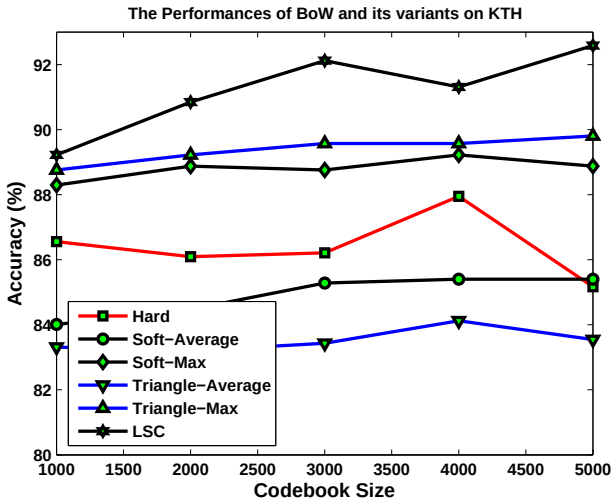

Figure 5-1: The performance of the BoW model and its variants on the KTH dataset.

than both the hard and soft assignment coding techniques, which is consistent with the report in [20]. Note that our implementation of the baseline hard assignment coding is lower than that in [125], which would be due to that a χ^2 kernel is employed in their work. The effects of codebooks' sizes on the BoW model are illustrated in Fig. 5-1. Most of the methods peak around 4000 codewords except for LSC which keeps increasing up to 5000 codewords.

In addition, we find that SC-based methods yield relatively better results compared with the BoW model. The ordinary SC with max pooling achieves even better results than LLC. Both SC and LLC reach the best results with around 3072 bases as shown in Fig. 5-4. Note that LLC with 100 nearest neighbours outperforms those with 5 and 50, and the trend is the same on the UCF-YouTube and HMDB51 datasets.

On the UCF-YouTube dataset The results on the UCF-YouTube dataset are slightly different from those on the KTH dataset. The NBNN kernel produces the best result of 62.4%. The soft assignment coding beats the triangle assignment with max pooling and LSC, obtaining the best result of 61.2% within the BoW family.

In addition, SC with max pooling outperforms LLC obtaining an accuracy of 59.4% which is comparable with the best result. Note that the NBNN kernel classifier outperforms NBNN on this dataset.

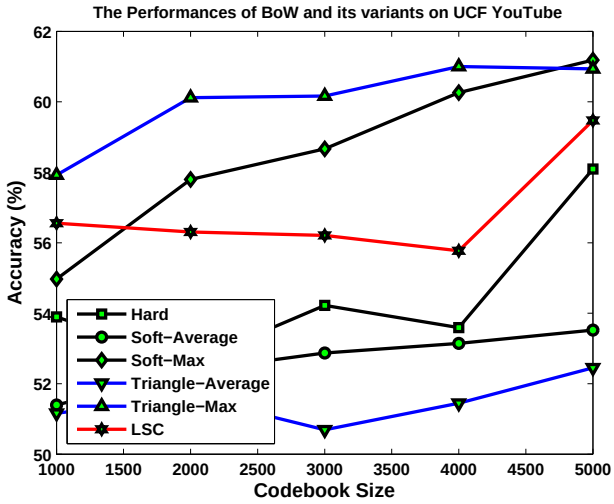

Figure 5-2: The performance of the BoW model and its variants on the UCF-YouTube dataset.

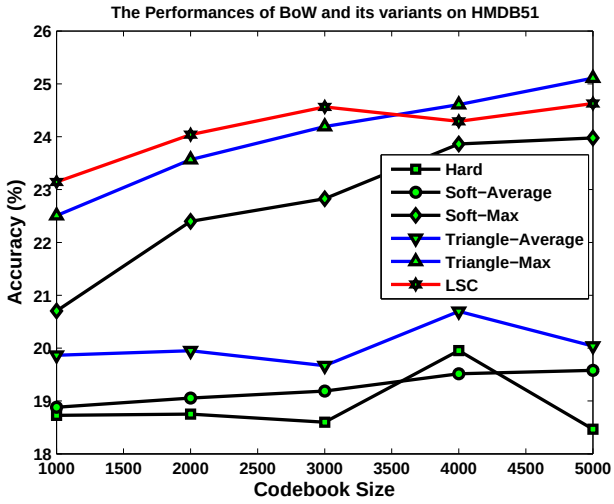

Figure 5-3: The performance of the BoW model and its variants on the HMDB51 dataset.

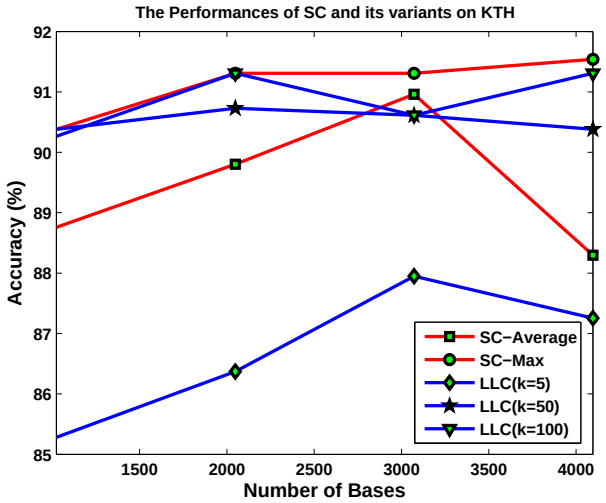

Figure 5-4: The performance of SC and its variants on the KTH dataset.

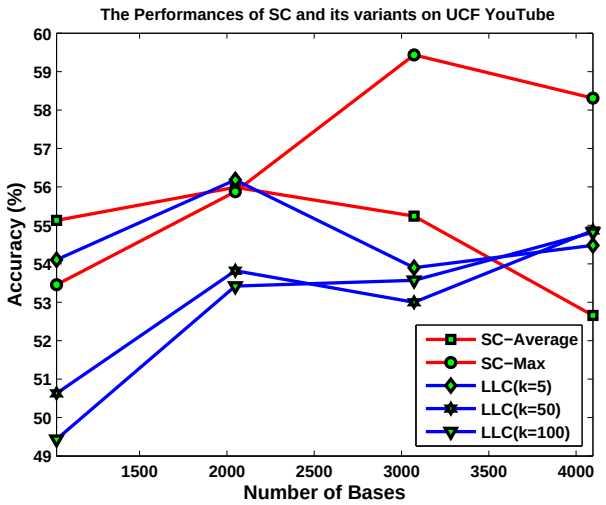

Figure 5-5: The performance of SC and its variants on the UCF-YouTube dataset.

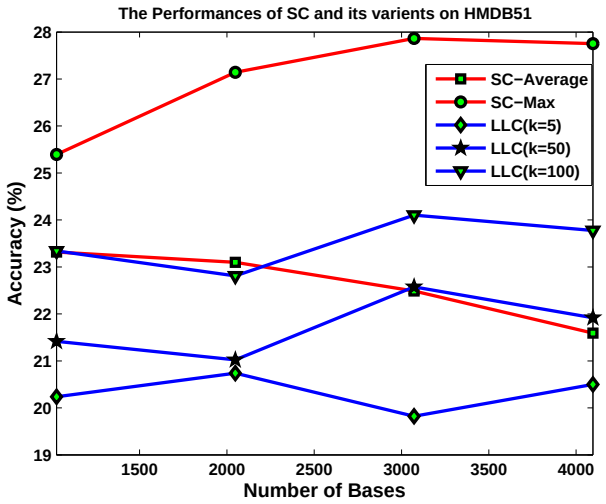

Figure 5-6: The performance of SC and its variants on the HMDB51 dataset.

As shown in Fig. 5-2, the best results happen around 5000 codewords for almost all the methods. As illustrated in Fig. 5-5, most of the best results for SC and LLC occur with 4096 bases.

The performance of the match kernels is inferior in this dataset, producing a low recognition rate of 54.5%.

On the HMDB51 dataset The results on the HMDB51 dataset are similar to those on the UCF-YouTube dataset, however the best result-27.9%-is obtained by SC with max pooling. Again, the triangle assignment coding with max pooling gives the best result within the BoW model. LSC produces a comparable result of 24.6% with the triangle assignment coding. The performance of the NBNN family is similar to that on the UCF-YouTube dataset, where the NBNN kernel is better than either NBNN and local NBNN. The reason would seem that these two datasets contain realistic actions and the NBNN kernel is more robust than NBNN and local NBNN.

In Fig. 5-3, most of the methods under BoW increase with codewords from 1000 to 5000. In Fig. 5-6, both SC and LLC become stable with the number of bases after 2048 with the best results around 3072.

The match kernels fails to provide reasonable results on this dataset.

5.4.3 Summary and discussion

The NBNN family produce impressive results on all the three datasets, with highest recognition rates by the local NBNN classifier on the KTH and UCF-YouTube datasets. This is consistent with the results in image and object recognition [9, 116, 77]. However, we can see from Table 5.1 that the superiority of the NBNN family becomes less significant on more realistic datasets, *i.e.*, HMDB51, with a larger number of action categories. This could be due to the assumption in NBNN that the smoothing parameter, namely the Parzen kernel bandwidth σ, is common to all categories does not fully hold for large category numbers.

We have also evaluated the performance of the local NBNN classifier with different numbers of neighbours k in the local neighbourhood, which, however, only slightly affects the performance with the k ranging from 5 to 30 in our experiments.

Although the bag-of-words (BoW) model has long been criticised for its quantisation errors, the newly proposed techniques such as the triangle assignment coding with max pooling and the localised soft-assignment coding (LSC) significantly improve the baseline hard assignment coding, and achieve the state-of-the-art performance, especially on the KTH dataset. This is mainly because that the information loss during the feature quantisation has been compensated by the sophisticated coding techniques and the powerful classifier, *i.e.*, SVM.

To the best of our knowledge, this is the first time that sparse coding (SC) via spatio-temporal local features is applied to action recognition. With both average and max pooling, SC outperforms most of the BoW based methods, which indicates its potential on action recognition. However, LLC does not outperform SC with max pooling on the three datasets. This is inconsistent with the report on object recognition in [127]. One reason could be that spatio-temporal features in video are much noisier than 2D features, which makes the locality constraint in LLC insignificant.

In addition, we can find in Fig. 5-4, 5-5 and 5-6 that LLC can produce reasonable results with more local neighbours k (over 100) than in the image domain (typically $k = 5$), which could be due to the fact that spatio-temporal local features in the video

domain lie in a higher dimensional space. Therefore, to encode a local feature, more bases would be needed. We have also experimented with k ranging from 5 to 300. The performance remains relatively stable after $k = 100$.

Note that for all the methods using feature pooling, max pooling is significantly better than average pooling both in BoW and SC on the three datasets. This behaviour is consistent with that in image classification [10].

Interestingly, the locality constraint and max pooling have been demonstrated to be more effective in the BoW model, *e.g.*, LSC significantly improves the performance of BoW. Indeed, the local NBNN classifier can also be regarded as imposing the locality constraint on the original NBNN with max pooling if the distance to a neighbour is deemed as the inverse of similarity.

Finally, the recognition rates of the match kernels are relatively low but are comparable to some of the methods in the BoW model such as the hard assignment, the soft and triangle assignments with average coding, especially on the KTH and HMDB51 datasets. With regard to match kernels, we have also experimented the max-sum kernel K_M, however, it performs much worse than the normalised sum kernel $K_{\mathcal{F}}$ and even fails to produce reasonable results on the UCF-YouTube dataset. This could be because it does not meet the Mercer condition and cannot guarantee that the optimisation in SVM training is convex [17].

5.5 Conclusion

In this chapter, we have transferred the state-of-the-art techniques, which have been widely used and shown effectiveness in the image domain, to action recognition. Extensive experiments have been conducted to systematically evaluate and compare these techniques on three benchmark datasets: KTH, UCF-YouTube and HMDB51.

Moreover, we have also provided experimental and theoretical insights into the performance of each method and drawn useful conclusions from findings in the experiments. As many of the techniques are innovated in the image domain and have not yet been applied to action recognition, our work can serve as guidance for future

research in action recognition.

Chapter 6

Discriminant Embedding via Image-To-Class Distances

6.1 Introduction

In Chapter 5, we have done a comprehensive evaluation on local methods for human action recognition. One of most interesting findings is that the image-to-class distance based methods, *i.e.*, naive Bayes nearest neighbour (NBNN), the NBNN kernel and local NBNN, have shown good performance. In this chapter, based on the image-to-class distances, we propose an algorithm for discriminative dimensionality reduction of local feature descriptors.

6.1.1 Motivations

Local features play a key role in visual recognition, *e.g.*, image classification and action recognition. Classification based on local features is still a challenging task due to the large intra-class variances and noisy local features. Widely-used local feature descriptors including SIFT [70], HOG3D [54] and HoG/HoF [61] have shown their effectiveness in image and video domains. However, local features suffer from deficiencies. On the one hand, due to background variation and clutter, local features from backgrounds could be detected as motion-related features leading to less

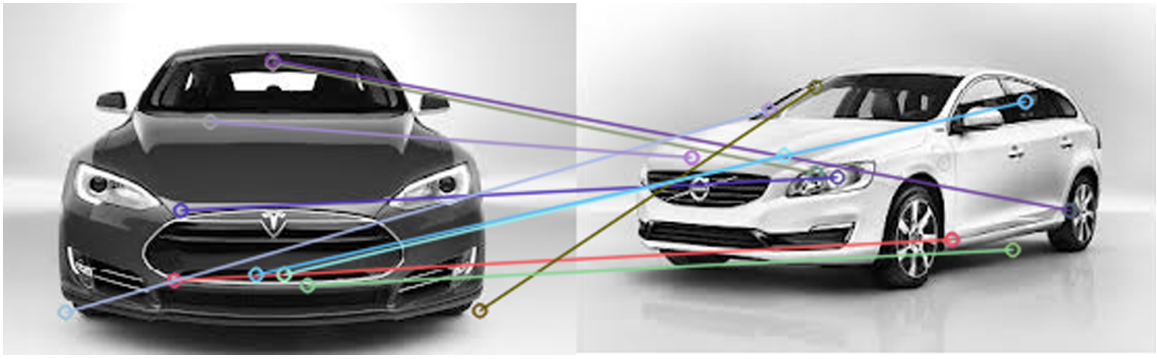

Figure 6-1: Matching by SURF between two images that belong to the same semantic category. The illustrated matched points are those with distances less than a threshold.

discriminative representation of human actions. In addition, similar local features would be shared by different actions which will also make the representation less discriminative. The discriminative ability of local features would greatly influence the performance of later representation and classification. On the other hand, current local feature descriptors, such as HOG3D, Cuboids and HOG/HOF, are always in a space of hundreds even thousands of dimensions, which could be computationally expensive and even intractable when the number of local features is huge.

The bag-of-words (BoW) model and sparse coding have been extensively exploited to encode local features as a global representation. The fact is that even images belonging to the same class would contain quite a large proportion of dissimilar local features which enlarge the intra-class variance, and make directly comparing local features in images not optimal for classification. Fig. 6-1 illustrates that the matched points found by SURF [4] between two images belonging to the same car category are all wrong.

Instead of directly comparing local features from different images, recently, a non-parametric approach named naive Bayes nearest neighbour (NBNN) [9] was proposed for image classification in which the image-to-class (I2C) distance is used. Being conceptually simple, NBNN has achieved state-of-the-art performance comparable with other sophisticated learning algorithms. The success of NBNN is credited to the use of the I2C distance, which has been proven to be the optimal distance to use in

image classification rather than the image-to-image (I2I) distance [9]. It is the I2C distance that effectively deals with the huge intra-class variances of local features.

However, the performance of the I2C-based methods depends highly on the effectiveness of local features because they essentially contribute to the calculation of the I2C distance. The I2C-based methods will be computationally expensive or even intractable with a huge number of local features, especially when the local features are in a high-dimensional space. In addition, the discriminative ability of local features will directly affect the performance of the I2C distance. For instance, local features with noise or from a background would degrade the performance of I2C for classification. Therefore, finding a low-dimensional space to represent the local features becomes very attractive.

Dimensionality reduction techniques such as principal component analysis (PCA) can be used to project the features into a low-dimensional space, which has been exploited in [25, 53] for image classification and action recognition. Unfortunately, PCA is an unsupervised feature reduction method treating each local feature equally without considering the label information of images and therefore suffers from being less discriminative in the low-dimensional space. Unsupervised nonlinear dimensionality reduction (manifold learning) methods such as Locally Linear Embedding (LLE) [93], ISOMAP [112] and Laplacian Eigenmap (LE) [6] suffer from a crucial limitation that the embedding does not generalise well from training to test data due to the out-of-sample problem. Moreover, similar to PCA, as unsupervised learning, their discriminative ability is also limited without using class label information.

In addition, some local features could be visually similar or shared by images in different classes, which is demonstrated in Fig. 6-2. Therefore, the use of conventional discriminant dimensionality reduction techniques, *e.g.*, linear discriminant analysis (LDA), is suboptimal because LDA, when applied to local features, attempts to minimise the within-class variance of different local features and maximise the between-class variance of different local features together.

In this chapter, with the aim of improving the image-to-class distance based methods, we propose a novel dimensionality reduction method by incorporating the I2C

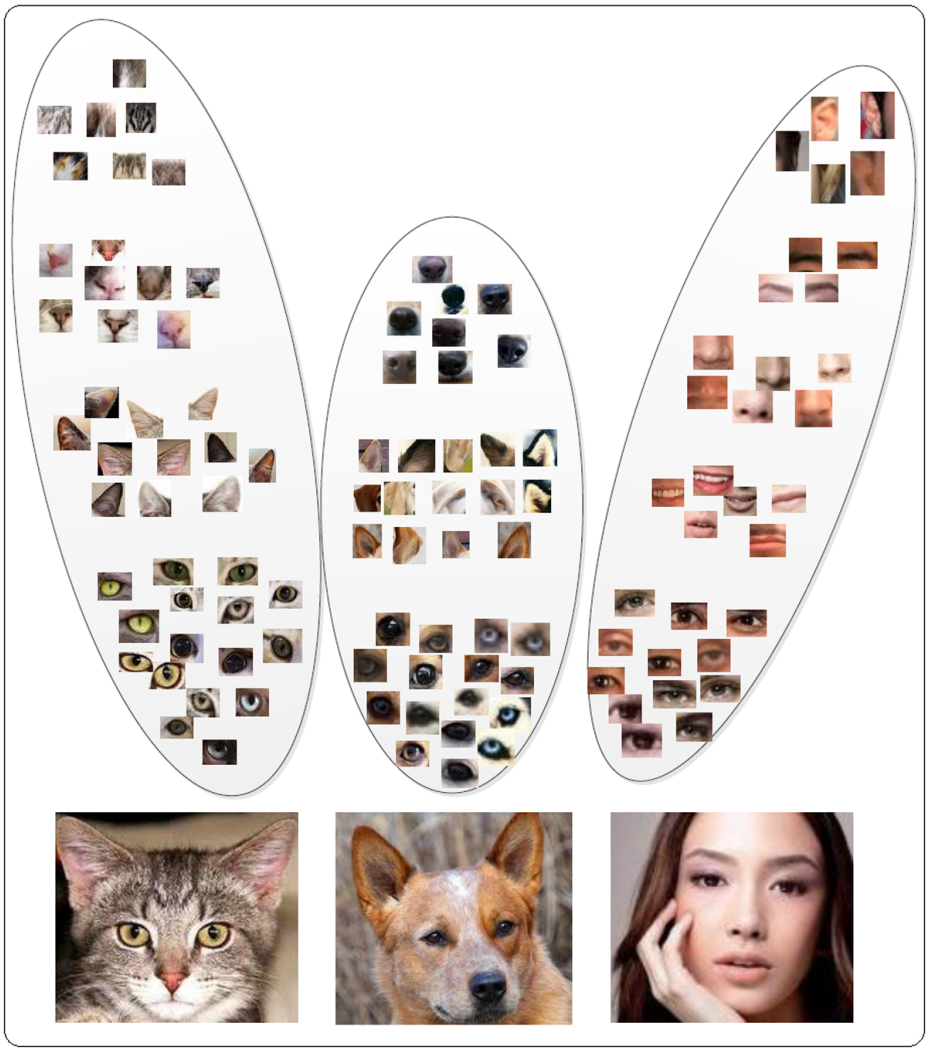

Figure 6-2: Illustration of local patches from different image categories. The local patches 'eyes' from images in different categories can be similar and are close to each other in the feature distribution, while the local patches such as 'eyes', 'noses' and 'ears' are distinctive to each other even though they could be detected from the same image categories.

distance. The use of the I2C distance benefits in two aspects. On the one hand, local features from one image are taken into consideration as a whole and class labels can be directly used for supervised learning. This increases the discriminative capacity of local features. On the other hand, it provides an intuitive and effective avenue to couple the dimensionality reduction of local features with classification, which can improve the performance of classification. In the low-dimensional space, local features from each image are aligned according the I2C distances and the I2C distance to its own class is minimised; the I2C distances to other classes are maximised.

6.1.2 Contributions

Our work contributes in the following aspects:

- A novel discriminative subspace learning algorithm based on the I2C distances is proposed for the dimensionality reduction of local features;
- In the embedded low-dimensional space, I2C-based methods are speeded up, scale well with a large numbers of local features and therefore become tractable in real-world applications;
- We formulate the method as an eigenvector decomposition problem, which can be more efficiently solved with a gradient descent algorithm.

6.2 Related work

In this section, we review the related work including the image-to-class distance based methods and linear dimensionality reduction techniques.

6.2.1 Image-to-class based methods

The image-to-class (I2C) distance was first introduced by Bioman *et al.* [9] in the naive Bayes nearest neighbour (NBNN) classifier. Based on the NBNN, several variants including the NBNN kernel, optimal NBNN and local NBNN have recently been proposed to improve performance. In addition, a metric learning algorithm based

on image-to-class distance has also been explored which is also closely related to our method.

- **Naive Bayes nearest neighbour (NBNN)**

NBNN is a non-parametric algorithm for image classification based on local features. With the naive Bayes assumption, NBNN is simple while in contrast to parametric learning algorithms, NBNN enjoys many attractive advantages. It requires no training stage and can naturally deal with a huge number of classes. Due to the use of the I2C distance calculated on original local features, NBNN can get rid of descriptor quantisation errors. The core of NBNN is the approximation of the log-likelihood of a local feature by the distance to its nearest neighbour, which brings about the image-to-class (I2C) distance. Taking advantage of the I2C distance, several variants of NBNN have been proposed in the past few years to improve the generalisation ability of NBNN.

Under the NBNN framework, to improve the performance of the original NBNN, several variants have recently been proposed including optimal NBNN [5], the NBNN kernel [116], local NBNN [77] and pooled NBNN [89] which have shown their effectiveness for scene/image classification.

- **Image-to-class distance metric learning**

By combining distance metric learning with the I2C distance measurement, Wang *et al.* [129] adopted the idea of a large margin from SVMs and proposed a method named I2C distance metric learning (I2CDML) to learn a distance metric specific to each class. They formulated a convex optimisation problem with the constraint that the I2C distance of each training sample to the class to which it belongs should be less than those to other classes by a large margin. However, as a conventional distance metric learning algorithm, I2CDML suffers from a major drawback that the number of parameters to be learned grows quadratically with the dimensionality of the data, which tends to be intractable with high-dimensional data.

6.2.2 Linear dimensionality reduction

In terms of linear dimensionality reduction, our method is closely related to classical dimensionality reduction techniques, including principal component analysis (PCA), linear discriminant analysis (LDA) and local discriminant embedding (LDE) [40]. We will give brief descriptions of those dimensionality reduction techniques to show the relationship with our method. Given a set of feature descriptors $\{\mathbf{x}_n\}$ with high dimensionality D, where $n = 1, \ldots, N$, dimensionality reduction techniques aim to find a projection to map the feature descriptors into a lower-dimensional space d.

- **Principal component analysis**

Principal component analysis (PCA) is an unsupervised learning algorithm which is widely used for dimensionality reduction. Although there exist various dimensionality reduction algorithms, PCA is still the most popular and very effective linear reduction technique. Without loss of generality, we can consider the projection $\mathbf{w}$ onto a one-dimensional space, namely $d = 1$, where $\mathbf{w}$ is a D-dimensional vector. Indeed, we are only interested in the direction induced by the projection $\mathbf{w}$, so we impose the constraint $\mathbf{w}^T\mathbf{w} = 1$ on the projection vector $\mathbf{w}$. Thereafter, the feature descriptor $\mathbf{x}_n$ can be projected onto a scalar value $\mathbf{w}^T\mathbf{x}_n$.

The variance of the feature descriptors in the projected space (one-dimensional space) can be calculated by

$$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{w}^T\mathbf{x}_n - \mathbf{w}^T\overline{\mathbf{x}}\}^2 = \mathbf{w}^T\mathbf{S}\mathbf{w} \tag{6.1}$$

where $\overline{\mathbf{x}}$ is the mean of the projected descriptors which is given by

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{6.2}$$

and $\mathbf{S}$ is the covariance matrix defined by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T \tag{6.3}$$

To find the projection $\mathbf{w}$, we maximise the variance of feature descriptors in the projected space, namely,

$$\mathbf{w}^{\star} = \max_{\mathbf{w}} \mathbf{w}^T \mathbf{S} \mathbf{w}, \quad s.t. \quad \mathbf{w}^T \mathbf{w} = 1. \tag{6.4}$$

We can solve the maximisation with the Lagrange multiplier method as

$$\mathbf{w}^T \mathbf{S} \mathbf{w} + \lambda (1 - \mathbf{w}^T \mathbf{w}) \tag{6.5}$$

Setting the derivative of Eq. (6.5) with respect to $\mathbf{w}$ equal to 0, we have

$$\mathbf{S} \mathbf{w} = \lambda \mathbf{w}, \tag{6.6}$$

which is a standard eigen-decomposition problem and the variance is just one eigenvalue of $\mathbf{S}$ [7]. So the solution of Eq. (6.5) is the eigenvector of $\mathbf{S}$ corresponding to the largest eigenvalue, which is also known as the first principal component.

If we want the dimensionality of the projected space to be d-dimensional, then the linear projection is composed of d eigenvectors corresponding to the d largest eigenvalues $\lambda_1, \ldots, \lambda_d$.

- **Fisher discriminant analysis**

Fisher discriminant analysis, also known as linear discriminant analysis (LDA), is a well-known classification techniques. LDA is also widely used for supervised linear dimensionality reduction. The primary purpose of LDA is to separate samples of distinct groups which could be associated with their class labels.

With respect to dimensionality reduction, LDA aims to project data points into a lower-dimensional space by taking into account the class labels of the data points. In the projected lower-dimensional space, the between-class separability is maximised while the with-class variability is minimised. The objective of LDA is to find a linear projection that can maximises the ratio of between-class variance to the within-class variance. Similar to the deduction in PCA, we consider a projection vector $\mathbf{w}$ to map the data points $\mathbf{x}_n$ onto scalar values $\mathbf{w}^T \mathbf{x}_n$. The objective function of LDA is to

maximise the ratio, known as the Fisher criterion, and takes the form

$$\mathbf{w}^{\star} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \tag{6.7}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between- and within-class scatter matrices and are defined as

$$\mathbf{S}_b = \sum_{c=1}^{C} N_c (\mathbf{x}_c - \overline{\mathbf{x}})(\mathbf{x}_c - \overline{\mathbf{x}})^T \tag{6.8}$$

and

$$\mathbf{S}_w = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{x}_{c,i} - \overline{\mathbf{x}}_i)(\mathbf{x}_{c,i} - \overline{\mathbf{x}}_i)^T, \tag{6.9}$$

respectively, in which

$$\overline{\mathbf{x}}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{x}_{c,i} \tag{6.10}$$

and

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{c=1}^{C} N_c \overline{\mathbf{x}}_c = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \mathbf{x}_{c,i} \tag{6.11}$$

As is known that the solution of Eq. (6.7) can be found by the following equation

$$\mathbf{S}_b \mathbf{w}^T = \lambda \mathbf{S}_w \mathbf{w}^T \tag{6.12}$$

If $\mathbf{S}_w$ is non-singular matrix and can be inverted, then the Fisher's criterion is maximised when the projection $\mathbf{w}$ is the eigenvector of the $\mathbf{S}_w^{-1}\mathbf{S}_b$ associated with the largest eigenvalue. Note that $\mathbf{S}_w$ is computed by pooling the estimates of the covariance matrix of each class and each covariance matrix is of at most rank $N_c - 1$. Thus the rank of $\mathbf{S}_w$ is at most $N - C$. In addition, $\mathbf{S}_b$ is estimated by C points there will be at most $C - 1$ eigenvectors with non-zero, real eigenvalues, which means the projected space is at most of $C - 1$ dimensions.

- **Linear discriminant embedding**

Dimensionality reduction for local feature descriptors have also been extensively exploited, especially for the task of feature matching. Linear discriminant embedding (LDE) is a non-parametric dimensionality reduction technique for image matching.

LDE is supervised learning algorithm differently from LDA, the labelled training samples S are set of matching/non-matching image patches

$$\mathcal{S} = \{\mathbf{p}_i, \mathbf{p}_j, l_{ij}\}, \tag{6.13}$$

where $\mathbf{p}_i$, $\mathbf{p}_j$ are the input image patches, and l_{ij} is a label equal to 1 if $\mathbf{p}_i$, $\mathbf{p}_j$ constitute a match pair, and 0 otherwise. Similar to PCA and LDA, LDE aims to find the linear projection $\mathbf{w}$. The objective function can be

$$J_1(\mathbf{w}) = \frac{\sum_{l_{ij}=0}(\mathbf{x}_i - \mathbf{x}_j)^2}{\sum_{l_{ij}=1}(\mathbf{x}_i - \mathbf{x}_j)^2} \tag{6.14}$$

which is the ratio of variance between the non-match and match differences along the direction $\mathbf{w}$, where $\mathbf{x}$ is the descriptor associated with an image patch $\mathbf{p}$. $\mathbf{w}$ can be found by solving the maximisation

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} J_1(\mathbf{w}) \tag{6.15}$$

Eq. (6.14) can be rewritten in terms of covariance matrices as

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{A}\mathbf{w}}{\mathbf{w}^T\mathbf{B}\mathbf{w}}, \tag{6.16}$$

where

$$\mathbf{A} = \sum_{l_{ij}=0}(\mathbf{x}_i - \mathbf{x}_j)^2 \tag{6.17}$$

and

$$\mathbf{B} = \sum_{l_{ij}=1}(\mathbf{x}_i - \mathbf{x}_j)^2 \tag{6.18}$$

The solution of Eq. (6.15) is the eigenvector associated with the largest eigenvalue of the generalised eigensystem

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w} \tag{6.19}$$

In the original work, an alternative objective function is also considered

$$J_2(\mathbf{w}) = \frac{\sum_{l_{ij}=1}(\mathbf{w}^T\mathbf{x}_i)^2}{\sum_{l_{ij}=1}(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2} \tag{6.20}$$

In practice, a regularised version of $\mathbf{B}$ is employed in Eq. (6.15).

In [78], Mikolajczyk and Matas independently proposed linear discriminant projections (LDP) for efficient matching of SIFT descriptors. However, LDP was proven to be equivalent to LDE although different methods are used in LDE and LDP [15].

6.3 Embedding based on I2C Distances

We first revisit the image-to-class (I2C) distance based on which we describe our discriminative embedding algorithm. The relationship of our method to other methods [15, 40, 129] is also shown in this section.

6.3.1 Revisit of I2C Distance

The image-to-class (I2C) distance was first defined in the naive Bayes nearest neighbour (NBNN) classifier. NBNN is an approximation of the optimal MAP naive-Bayes classifier under some assumptions.

Given an image Q represented as a set of local features, $\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N$, where $\mathbf{x}_i \in R^D$ and D is the dimensionality of local features. Taking the assumption that the class prior $p(C)$ is uniform, MAP can be simplified as the maximum likelihood (ML) classifier:

$$\hat{C} = arg\max_C p(C|Q) = arg\max_C p(Q|C). \tag{6.21}$$

Under the naive-Bayes assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N$ are i.i.d. given the class C, we have:

$$p(Q|C) = p(\mathbf{x}_1, \ldots, \mathbf{x}_N|C) = \prod_{i=i}^{N} p(\mathbf{x}_i|C), \tag{6.22}$$

where $p(\mathbf{x}_i|C)$ can be approximated using the non-parametric Parzen density estimation.

The Parzen likelihood estimation of the probability of $\mathbf{x}$ from class C is:

$$\hat{p}(\mathbf{x}|C) = \frac{1}{L}\sum_{j=1}^{L} K(\mathbf{x} - \mathbf{x}_j^C), \tag{6.23}$$

where L is the number of local features from class C.

By further assuming that the kernel bandwidths in the Parzen function are the same for all the classes, the likelihood can be simplified using the nearest neighbour. The summation of all the distances from the local features of an image to their corresponding nearest neighbours in each class is defined as the **Image-To-Class (I2C)** distance, which can be calculated by:

$$D_X^c = \sum_{\mathbf{x}\in X} ||\mathbf{x} - NN^c(\mathbf{x})||^2, \tag{6.24}$$

where NN^c is the nearest neighbour of $\mathbf{x}$ in class c. The resulting classifier takes the form:

$$\bar{c} = arg\min_c D_X^c, \tag{6.25}$$

6.3.2 Discriminative Embedding

Our task is to classify a collection of images $\{X_i\}$, each of which is represented by a set of local features: $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}, \ldots, \mathbf{x}_{im_i}\}$, where m_i is the number of local features from image X_i.

Given an image X_i, its I2C distance to class c is computed according to Eq. (6.24) as:

$$D_{X_i}^c = \sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}_{ij}^c)||^2, \tag{6.26}$$

where $\mathbf{x}_{ij}^c$ is the nearest neighbour in class c.

After applying a linear projection $\mathbf{w}$ on the local features, the I2C distance be-

comes:

$$\begin{aligned}\hat{D}^c_{X_i} &= \sum_{j=1}^{m_i} ||\mathbf{w}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^c_{ij})||^2 \\ &= \sum_{j=1}^{m_i} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^c_{ij})^{\mathrm{T}}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_{ij} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^c_{ij}) \\ &= \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}(\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) \end{aligned} \tag{6.27}$$

We introduce ΔX_{ic} as an auxiliary matrix defined as:

$$\Delta X_{ic} = \begin{pmatrix} (\mathbf{x}_{i1} - \mathbf{x}^c_{i1})^{\mathrm{T}} \\ \cdots \\ (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}} \\ \cdots \\ (\mathbf{x}_{im_i} - \mathbf{x}^c_{im_i})^{\mathrm{T}} \end{pmatrix} \tag{6.28}$$

Then $\hat{D}^c_{X_i}$ can be represented as:

$$\hat{D}^c_{X_i} = \mathbf{w}^{\mathrm{T}}\Delta X^{\mathrm{T}}_{ic}\Delta X_{ic}\mathbf{w}, \tag{6.29}$$

Unlike the methods in [40, 15], our aim in the embedded space is to minimise the I2C distances from images to the classes they belong to while simultaneously maximizing the I2C distances to the classes they do not belong to. The objective function we used takes the form as:

$$\begin{aligned}\mathbf{w}^* &= \arg\max_{\mathbf{w}} \frac{\sum_{n=1}^{N_i}\sum_i \mathbf{w}^{\mathrm{T}}\Delta X^{\mathrm{T}}_{in}\Delta X_{in}\mathbf{w}}{\sum_i \mathbf{w}^{\mathrm{T}}\Delta X^{\mathrm{T}}_{iP}\Delta X_{iP}\mathbf{w}} \\ &= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}(\sum_{n=1}^{N_i}\sum_i \Delta X^{\mathrm{T}}_{in}\Delta X_{in})\mathbf{w}}{\mathbf{w}^{\mathrm{T}}(\sum_i \Delta X^{\mathrm{T}}_{iP}\Delta X_{iP})\mathbf{w}}, \end{aligned} \tag{6.30}$$

where ΔX_{iP} is the auxiliary matrix associated with the class that image X_i belongs

to (positive class) and ΔX_{in} is of the negative class that image X_i does not belong to. Note that, given a dataset, the number of negative classes N_i is the same for all images in the dataset.

We can now seek an embedding $\mathbf{w}^*$ to maximise the ratio in Eq. (6.30). The above equation can be rewritten in terms of covariance matrices as:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathrm{T}}\mathbf{C}_N\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{C}_P\mathbf{w}}, \tag{6.31}$$

where

$$\mathbf{C}_N = \sum_{n=1}^{N_i}\sum_{i} \Delta X_{in}^{\mathrm{T}}\Delta X_{in}, \tag{6.32}$$

and

$$\mathbf{C}_P = \sum_{i} \Delta X_{iP}^{\mathrm{T}}\Delta X_{iP}, \tag{6.33}$$

It can be seen that maximizing the objective function in Eq. (6.31) is a well-known eigensystem problem:

$$\mathbf{C}_N\mathbf{w} = \lambda\mathbf{C}_P\mathbf{w} \tag{6.34}$$

The obtained embedding is formed by the k eigenvectors associated with the k largest generalised eigenvalues λ. The whole procedure of the embedding is summarised in Algorithm 1.

Algorithm 1 I2C Distance-based Discriminative Embedding (I2CDDE)

1. Calculate the local features $\{\mathbf{x}_{ij}\}$ for each image X_i in the training set.
2. Find the nearest neighbours of local features: $\{\mathbf{x}_{ij}\}$ in the positive class and negative classes, respectively.
3. For image X_i, compute the auxiliary matrices: ΔX_{in} and ΔX_{iP} using Eq. (6.29).
4. Compute the positive and negative covariance matrices $\mathbf{C}_P$ and $\mathbf{C}_N$.
5. Solve the generalised eigenvector decomposition problem in Eq. (6.34) to find $\mathbf{w}^*$.

6.3.3 Neighbourhood Embedding

Due to the noisy local features, *e.g.*, local features from backgrounds and shared by similar actions, the image-to-class (I2C) distance using the nearest neighbour (NN) would not be reliable. To make the I2C distance more robust and insensitive to noisy features, we further improve the algorithm by incorporating locality (using K nearest neighbours) in the objective function, which could, to some extent, preserve the local structure of features in the reduced space. We will show experimentally that this modification can improve the performance especially on more complex datasets, *e.g.*, HMDB51, in which the backgrounds are quite complicated and local features are extremely noisy. With the neighbourhood embedding, the $D^c_{X_i}$ in Eq. (6.26) is replaced by:

$$D^c_{X_i,K} = \sum_{k=1}^{K} \sum_{j=1}^{m_i} ||\mathbf{x}_{ij} - \mathbf{x}^c_{ij,k}||^2, \tag{6.35}$$

where $\mathbf{x}^c_{ij,k}$ is the k-th nearest neighbour of $\mathbf{x}^c_{ij}$ in the c-th class and K is the number of neighbours. The objective function in Eq. (6.30) needs also to be updated accordingly.

6.3.4 Relation to LDE

Our method is closely related to the linear discriminant projection (LDE) method [40, 15], as both address the dimensionality reduction of local features. In LDE, the objective function is to maximise the ratio of the variance of differently labelled points (unmatched points) to that of identically-labelled points (matched points). The matched and unmatched features vary with different applications. For instance, in image/object classification, matched features could be the points on the objects that are visually similar or that are from the same object category.

The main difference between our I2CDDE and LDE is the obtaining of the covariance matrices. In LDE, the matrices are based on the pairwise descriptor differences while I2CDDE employs the I2C distances. Specifically,

1) LDE deals with the relationship between local features rather than those between images, which does not secure the discriminative ability of local features for

classification. In LDE, ground truth matching/non-matching pairs are needed in the training stage, however these training pairs would be hard to obtain in practice for action recognition. In I2CDDE, the training pairs are not required, which makes it more flexible for classification.

2) I2CDDE treats local features from each image as a whole and copes with the relationship between images and classes. By differentiating the I2C distances to the same class and to different classes, I2CDDE makes the local features as a whole discriminative on an image level and can naturally benefit classification.

6.3.5 Relation to I2CDML

In the Image-To-Class distance metric learning algorithm [129], the squared Euclidean distance in Eq. (6.26) is replaced with the parametric Mahalanobis distance which is to be learned. The I2C distance becomes:

$$D^c_{X_i} = \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}} M_c (\mathbf{x}_{ij} - \mathbf{x}^c_{ij}), \tag{6.36}$$

where M_c is the distance metric learned in [129].

As shown in [44], the Mahalanobis distance metric learning can be considered as learning a linear transformation of the data and measuring the squared Euclidean distance in the transformed space after applying the linear transformation. This can be shown by factorizing the distance matrix M_c in Eq. (6.36) as: $M_c = GG^{\mathrm{T}}$, where G is the linear transformation to be learned. The I2C distance in Eq. (6.36) becomes:

$$D^c_{X_i} = \sum_{j=1}^{m_i} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij})^{\mathrm{T}} GG^{\mathrm{T}} (\mathbf{x}_{ij} - \mathbf{x}^c_{ij}) \tag{6.37}$$

We can see that Eq. (6.37) is equivalent to Eq. (6.27) in terms of linear transformations. The main differences between I2CDDE and I2CDML are summarised as follows:

1) I2CDML adopts the large margin framework from SVMs in the objective function which is solved by gradient descent, while I2CDDE is formulated as an eigenvector

decomposition problem.

2) In I2CDML, multiple distance metrics are learned for all the classes leading to a high computational cost in the high-dimensional space, while I2CDDE learns a unified linear projection, which alleviates the computational burden without compromising the discriminative ability.

6.3.6 Computational complexity

A key deficiency in I2C-based methods is the heavy computational burden resulting from the nearest neighbour search, which is extremely expensive especially when local features are high-dimensional. I2CDDE can greatly reduce the computational cost and at the same time even enhance the discriminative ability of local features.

At the test stage, the computational complexity in the original space is $\mathcal{O}(NMD^2)$, where N is the number of local features from a test sample, M is the total number of local features in the training set and D is the dimensionality of local features in the original space. After the embedding, the computational complexity is reduced to $\mathcal{O}(NMd^2)$, where d ($d \ll D$) is the dimensionality of local features in the embedded space. Take the local descriptor in action recognition for instance. We use the HOG3D descriptor. The dimensionality in the original space is 1000 while in the embedded space it is only tens of dimensions. The computational complexity in the reduced space is $d^2/D^2 = 10^2/1000^2 = 1/10000$ of that in the original space.

6.4 Experiments and results

We comprehensively evaluate I2CDDE for human action recognition on the benchmark KTH dataset, the realistic UCF YouTube and HMDB51 dataset for human action recognition. We compare the performance of I2CDDE with PCA and LDA, and also show the improvement of I2C-based methods including NBNN, local NBNN and the NBNN kernel.

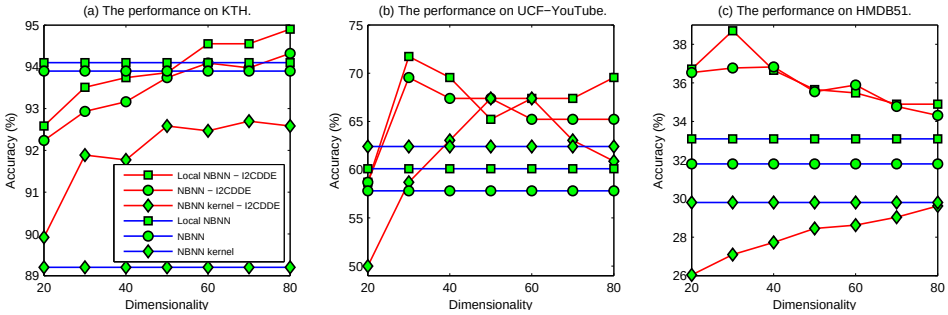

Figure 6-3: The performance of NBNN (**circle**), local NBNN (**square**) and the NBNN kernel (**diamond**) with different dimensions on the three datasets. Blue and red lines denote the performance before and after dimensionality reduction by I2CDDE.

6.4.1 Experimental settings

For action recognition, we utilise Dollár's periodic detector [25] to detect spatio-temporal interest points (STIPs). Three-dimensional histograms of oriented gradients (HOG3D) [54], which are descriptive and relatively compact with 1000 dimensions, are used for the description of STIPs. The code for detection and description of STIPs is available online. The performance of action recognition with different dimensions on the KTH and HMDB51 datasets is plotted in Fig. 6-3 and Fig. 6-4, respectively.

6.4.2 Results

The performance of I2CDDE for action recognition with different dimensions on the KTH, UCF YouTube and HMDB51 datasets are plotted in Fig. 6-3 (a), (b) and (c), respectively. On all the three datasets, we observe that the performance of NBNN, local NBNN and the NBNN kernel has been dramatically improved. On the KTH dataset, the increase on the NBNN kernel is more significant than NBNN and local NBNN, while on the UCF YouTube and HMDB51 datasets, the improvement over NBNN and local NBNN is much more remarkable than that over the NBNN kernel. Note that the superior performance of I2CDDE can be achieved with the local features of less than 60 dimensions, which manifests the effectiveness of I2CDDE for dimensionality reduction of local features.

We have also investigated the effects of different numbers of nearest neighbours on the performance of the neighbourhood embedding. The results on the three datasets

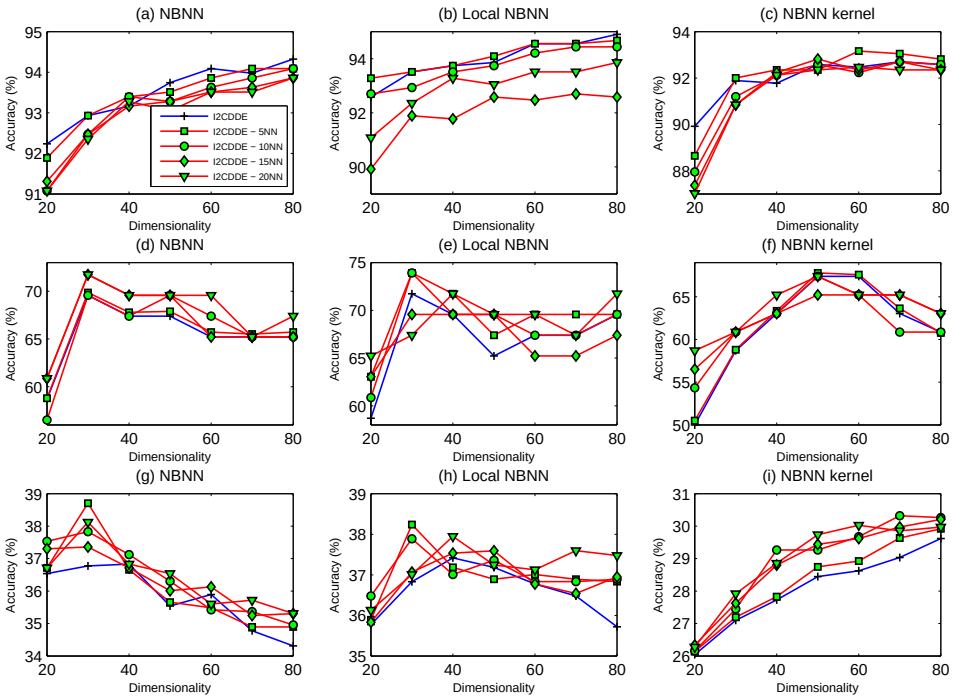

Figure 6-4: The performance of I2CDDE with different numbers of nearest neighbours on the KTH (the top row), UCF YouTube (the middle row) and HMDB51 (the bottom row) datasets. Blue lines denote the performance of I2CDDE with the nearest neighbour (1NN).

Methods	NBNN	Local NBNN	NBNN Kernel
No Reduction	16.4s	8.4s	22685.3s
Reduction	0.9s	0.6s	365.4s

Table 6.1: The run time before and after applying I2CDDE (d=30).

are shown in Fig. 6-4, from which we find that, on the KTH dataset, the performance of the neighbourhood embedding is comparable with the baseline I2CDDE with the nearest neighbour. On the realistic datasets including UCF YouTube and HMDB51, the benefit of incorporating neighbourhood turns to be more significant, especially on HMDB51. This is expected and reasonable because the KTH is relatively easy with simple actions and clear backgrounds, while HMDB51 contains rather complicated actions and clutters in background. Note that NBNN, local NBNN and the NBNN kernel with neighbourhood embedding are all largely improved over the baseline with the nearest neighbour.

6.4.3 Run Time

Since one of the key contributions of I2CDDE is to speed up the I2C-based methods including NBNN, local NBNN and the NBNN kernel, we have compared the run time (in seconds) to classify a test sample before and after using I2CDDE, which is shown in Table 6.1. The I2C-based methods are much faster after dimensionality reduction. The run time after reduction is calculated by setting reduced dimensionality as 30 for each method and experiments are conducted on the KTH dataset.

6.4.4 Comparison with Other Dimension Reduction Techniques

We have also compared I2CDDE with widely used linear dimensionality reduction methods including PCA, LDA, LFDA, LPP and NPE, in Table 6.2. As expected, I2CDDE uniformly outperforms the compared methods. PCA, LPP and NPE are unsupervised without using the label information and therefore tend to be less

		KTH	HMDB51	YouTube
NBNN	I2CDDE	**92.9**	**38.7**	**71.7**
	PCA	91.7	35.6	58.6
	LDA	82.9	31.6	54.3
	LFDA	86.6	29.6	63.1
	LPP	92.8	34.4	56.8
	NPE	91.9	34.8	55.6
	Original	**93.9**	31.8	57.8
LNBNN	I2CDDE	**93.5**	**38.2**	**73.9**
	PCA	91.8	35.7	58.7
	LDA	83.3	31.4	56.5
	LFDA	86.8	28.5	71.7
	LPP	93.3	35.2	60.9
	NPE	92.6	34.9	60.9
	Original	**94.1**	33.1	60.1
NBNN Kernel	I2CDDE	**92.0**	**30.2**	**60.9**
	PCA	89.8	25.8	53.6
	LDA	18.3	13.1	23.9
	LFDA	67.4	10.2	23.9
	LPP	91.0	28.3	58.7
	NPE	91.0	27.9	57.4
	Original	89.2	29.8	**62.4**

Table 6.2: The comparison of I2CDDE with other reduction methods. Note that the results listed in the table are the accuracies (%) achieved by the methods with **30 dimensions** (except for LDA and LFDA).

discriminative for classification. LDA and LFDA discriminatively learn the projections by labelling the local features with the label of the image that it belongs to, which, however, could mislead the classifier as discussed in Section 1. We can see that for the NBNN kernel, they even fail to produce reasonable results for all the three datasets. In I2CDDE, the I2C distance actually creates a bridge between the class labels and local features (by using I2C distance), providing an effective and intuitive way to impose discriminative information on local features, and therefore improve the performance of classification.

6.5 Conclusion

In this chapter,we have proposed a method named image-to-class distance-based embedding (I2CDDE) for dimensionality reduction of local features. The experimental results on the KTH, UCF YouTube and HMDB51 datasets have demonstrated that I2CDDE can significantly improve the performance of previously proposed I2C-based methods including NBNN, local NBNN and the NBNN kernel. More importantly, I2CDDE speeds up these methods, which could boost I2C-based methods for large-scale applications. In addition, I2CDDE uniformly outperforms the classical linear dimensionality reduction techniques such as PCA, LDA, LFDA, LPP and NPE, which further suggests the effectiveness of I2CDDE.

Chapter 7

Locally Gaussian Embedding

7.1 Introduction

In Chapter 6, we have proposed a discriminative dimensionality reduction algorithm, which is based on the image-to-class distances introduced in the naive Bayes nearest neighbour classifier. The image-to-class distance is actually an approximation of the log-likelihood of the local feature descriptor. However, the approximation does not always perform well [5]. In this chapter, we start with investigating and analyzing the theoretical foundation of NBNN. By explicitly modelling the likelihood via local Gaussians, we propose a discriminative dimensionality reduction method for local feature descriptors.

7.1.1 Motivations

NBNN is extremely simple both in theory and in practice. Given an image, one first computes a set of local feature descriptors. Then, one searches for the class that minimises the sum over all feature descriptors of distances to the respective nearest neighbours belonging to that class. In spite of its simplicity and the complete absence of a training phase, NBNN achieves surprisingly good results on standard benchmark datasets such as Caltech101 for image classification [9], being competitive with the state of the art.

As shown by our evaluation work in Chapter 5, the Naive Bayes Nearest Neighbour classifier (NBNN) has also demonstrated impressive performance on human action recognition. The good performance of NBNN on both image/scene classification and action recognition can be largely due to the following two reasons.

- The avoidance of a vector quantisation, which can largely preserve the effectiveness of local feature descriptors.

- The use of the image-to-class distance rather than directly computing the image-to-image distance, which can, to a large extent, tackle the intra-class variations.

The former avoids quantisation errors, which are especially effective for more informative features found in less dense areas of a feature space [116]. The latter enables a good generalisation beyond the provided labelled images. Indeed, when evaluating a test image, NBNN combines a range of information from local feature descriptors.

However, the NBNN framework also suffers from its own limitations.

- Extensive nearest neighbour searches are involved in the NBNN classifier. The computational burden during testing is extremely high, especially when dense sampling of local features is used, which often seems necessary to obtain good results. Moreover, it even could be infeasible when the local features are high-dimensional. The induced long testing time restricts the practical application of NBNN.

- Due to the use of image-to-class distances, similar densities are assumed in the feature space for all classes, such that the same kernel bandwidth can be used for all of them. In practice, however, this assumption is often violated, leading to a strong bias towards one or a few classes.

These two points have been investigated by [5] and [129] respectively, both of which introduce a learning phase on the training samples to tune the parameters. In this chapter, we try to address the problems by a dimensionality-reduction algorithm via a locally Gaussian embedding.

7.1.2 Contributions

In Chapter 6, we have proposed a dimensionality reduction algorithm named I2CDDE based on the image-to-class distances, which is also under the NBNN framework and therefore is still restricted by the assumptions in NBNN. In this chapter, we go beyond the NBNN framework and deal with the above two shortcomings by a discriminative dimensionality reduction algorithm via locally Gaussian embedding (LGE). Although both I2CDDE and LGE are dimensionality-reduction algorithms, they are fundamentally different. In I2CDDE, the image-to-class distances are used for the construction of the objective function while in LGE the objective function is built on the maximum a posteriori (MAP) classifier. Our contributions in this chapter can be summarised in two aspects:

- A discriminative dimensionality reduction algorithm is proposed to project local features into a lower-dimensional space.
- Local Gaussians are incorporated to explicitly model the likelihood of local feature descriptors for dimensionality reduction.

7.2 Related work

In this section, we review two pieces of work that are closely related to the proposed algorithm in terms of improving the original NBNN [5] and dimensionality reduction via local Gaussians [86].

7.2.1 Optimal Naive Bayes Nearest Neighbour

As mentioned above, the assumption underlying the original NBNN is too restrictive and considerably degrades its generalisation ability. It has been observed by Behmo *et al.* [5] that NBNN performs relatively well on certain datasets, but not on others. They have also shown that this performance variability of NBNN could stem from the assumption that the normalisation factor involved in the kernel estimator of the conditional density of features is class-independent. The Parzen likelihood estimation

of the probability of a local descriptor $\mathbf{d}$ from class c was given in Eq. (6.23) as:

$$p(\mathbf{d}|c) = \frac{1}{L}\sum_{j=1}^{L} K(\mathbf{d} - \mathbf{d}_j^c), \tag{7.1}$$

where L is the number of local feature descriptors from class c. If a Gaussian kernel is used, we then have

$$p(\mathbf{d}|c) = \frac{1}{L}\sum_{j=1}^{L} \exp(-\frac{1}{2\sigma^2}||\mathbf{d} - \mathbf{d}_j^c||^2) \tag{7.2}$$

Behmo *et al.* [5] incorporated a learning stage to select parameters in the original NBNN by relaxing the restrictive assumption. In practice, the Parzen estimator does not converge and there is little sense in keeping more than just the first term of the sum. They modelled the likelihood of a feature descriptor $\mathbf{d}$ relative to an image class c as

$$p(\mathbf{d}|c) = \frac{1}{Z^c}\exp(-\frac{\tau^c(\mathbf{d})}{2(\sigma^c)^2}) \tag{7.3}$$

where $\tau^c(\mathbf{d})$ is the Euclidean distance of the descriptor $\mathbf{d}$ to the nearest neighbour in the class c, and Z^c is the normalisation factor and σ^c is the smoothing parameter also called bandwidth, which is associated with class label c.

In the original NBNN, Z^c and σ^c are assumed to be independent of image class while in the optimal NBNN, an optimisation scheme is designed to find the optimal values of the parameters Z^c and σ^c by cross-validation.

7.2.2 Local discriminative Gaussians

Recently, Parrish and Gupta [86] proposed a dimensionality-reduction algorithm based on a local discriminative Gaussian (LDG) criterion, which is a supervised dimensionality reduction technique for classification. The objective function used in LDG is an approximation to the leave-one-out training error of a local quadratic discriminant analysis classifier. LDG acts locally at each training sample with the aim to project the samples into a lower-dimensional space where similar data can be

discriminated from dissimilar data.

Given a set of labelled training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in R^d$ and $y_i \in \{1, 2, \ldots, m\}$ being the i^{th} feature vector and class label, respectively. The goal is to find a matrix $B \in \mathbb{R}^{d \times l}$, $l < d$ such that the reduced-dimensionality feature vectors $\{B^T\mathbf{x}_i\}$ can be separated according to class. The separability is measured by the performance of a generative classifier.

$$\hat{C} = \arg\max_C p(\mathbf{x}_i|C)p(C)$$

The leave-one-out cross-validation error of a maximum a posteriori (MAP) classifier acting on the mapped features measures the separation achieved by B:

$$\sum_{i=1}^{N} I[p(B^T\mathbf{x}_i|y_i)p(y_i) < \max_j p(B^T\mathbf{x}_i|j)p(j)] \tag{7.4}$$

where the indicator function $I(.)$ is one if its argument is true and zero otherwise. $p(\mathbf{x}_i|y_i)$ is the likelihood of $\mathbf{x}_i$ given class y_i, estimated from the other $n-1$ training sample pairs.

The objective function in Eq. (7.1) is difficult to minimise due to the discontinuity of the indicator function. In order to arrive at a smooth, differentiable objective function that approximates Eq. (7.1), a log is substituted for the indicator and a sum for the max.

$$f(B) = \sum_{i=1}^{n} \log\left(\frac{\sum_{j=1}^{m} p(B^T\mathbf{x}_i|j)p(j)}{p(B^T\mathbf{x}_i|y_i)p(y_i)}\right) \tag{7.5}$$

$$= \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} p(B^T\mathbf{x}_i|j)p(j)\right) - \log(p(B^T\mathbf{x}_i|y_i)p(y_i)) \tag{7.6}$$

Bounding Eq. (7.5) from below with Jensen's inequality, replacing the first log term in Eq. (7.6) with

$$\sum_{j=1}^{m} p(j) \log(p(B^T\mathbf{x}_i|j)) \tag{7.7}$$

The final objective function takes the form as:

$$f(B) = \sum_{i=1}^{n}(\sum_{j=1}^{m} p(j)\log(p(B^T\mathbf{x}_i|j)) - \log(p(B^T\mathbf{x}_i|y_i)p(y_i))) \tag{7.8}$$

Imposing the constraint that $B^TB = I$, (7.5) is simplified by making the covariance of the Gaussians in the mapped space independent of B. $p(\mathbf{x}_i|j)$ is assumed as Gaussian, $\mathcal{N}(\mathbf{x}_i; \mu_{i,j}; \Sigma_{i,j})$. In addition, to reduce the model bias of assuming one Gaussian per class, $p(\mathbf{x}_i|j)$ is modelled as a locally Gaussian distribution[29].

The parameters of the Gaussian for point $\mathbf{x}_i$ and class j are estimated by finding the k nearest class j neighbours to training point $\mathbf{x}_i$ by the Euclidean distance and using these points to estimate the Gaussian's maximum likelihood mean and covariance. To reduce estimation variance, we model each covariance matrix as a scaled identity $\sigma_{i,j}^2 I$, where I is the properly-sized identity matrix. Therefore,

$$p(B^T\mathbf{x}_i|j) = \mathcal{N}(B^T\mathbf{x}_i; B^T\mu_{i,j}; B^TB\sigma_{i,j}^2) \tag{7.9}$$

The maximisation of the objective function in Eq. (7.6) is shown to be an eigen-decomposition problem with a closed-form solution.

7.3 Embedding via local Gaussians

We aim to address the limitations of the original NBNN by proposing a discriminative dimensionality reduction method via a locally Gaussian embedding.

- By reducing the dimensions of local feature descriptors, the computational burden of nearest neighbour search is greatly alleviated.
- Through the modelling of likelihood of local feature descriptors as local Gaussians, we naturally avoid the estimation of parameters.

Our method is closely related to LDG [86] in terms of dimensionality reduction and local Gaussian modelling. However, our method is fundamentally distinguished from LDG. Firstly, we address the dimensionality reduction of local descriptors of images/videos while LDG deals with global descriptors. Secondly, the objective function used in our method is essentially different from the one used in LDG. In LDG,

the objective function is based on the minimisation of the leave-one-out training error of a local quadratic discriminant analysis classifier. In our method, the objective function is to maximise the likelihood of an image with respect to the class it belongs to while minimizing the likelihood with respect to the classes it does not belong to.

7.3.1 Problem formulation

We first formulate our problem with the naive Bayes classifier and then provide the solution by modelling the likelihood as a local Gaussian.

Given an image $\mathbf{x}_i$ represented as a set of local features, $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}, \ldots, \mathbf{x}_{im_i}\}$, where $\mathbf{x}_{ij} \in R^D$ and D is the dimensionality of local features. The image $\mathbf{x}_i$ can be classified by the maximum-a-posteriori (MAP) classifier:

$$\hat{C} = \arg\max_C p(C|X_i), \tag{7.10}$$

where the posterior can be calculated by the Bayes' formula:

$$p(C|X_i) = \frac{p(X_i|C)p(C)}{\sum_c p(X_i|c)}. \tag{7.11}$$

Taking the assumption that the class prior $p(c)$ is uniform, MAP can be simplified as the maximum likelihood (ML) classifier:

$$\hat{C} = \arg\max_C p(C|X_i) = \arg\max_C p(X_i|C). \tag{7.12}$$

We now have to deal with the local features from the image X_i. Under the naive-Bayes assumption that $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ij}, \ldots, \mathbf{x}_{im_i}$ are i.i.d. given the class C, we have:

$$p(X_i|C) = p(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i}|C) = \prod_{j=1}^{m_i} p(\mathbf{x}_{ij}|C), \tag{7.13}$$

Taking the log probability of the ML decision rule we arrive at:

$$\hat{C} = \arg\max_C \log p(C|X_i) = \arg\max_C \sum_{j=1}^{m_i} \log p(\mathbf{x}_{ij}|C). \tag{7.14}$$

We aim to find a linear projection $\mathbf{w}$ to map local features into a lower-dimensional space. In the projected space, we expect to maximise the likelihood $p(\hat{X}_i|y_i)$ while at the same time minimizing $p(\hat{X}_i|c), c \neq y_i$, where $\hat{X}_i$ is the counterpart of X_i in the projected space and y_i is the class label associated with the image X_i.

Based on the intuitive idea, we can maximise the following objective function:

$$f(\mathbf{w}) = \frac{\sum_{i=}^{N} p(\hat{X}_i|y_i)}{\sum_{i=1}^{N}\sum_{c=1,c\neq y_i}^{C} p(\hat{X}_i|c)} \tag{7.15}$$

where N is the number of training samples. Taking into account Eq. (7.10) and (7.11), we have

$$f(\mathbf{w}) = \frac{\sum_{i=1}^{N} \frac{1}{m_i}\sum_{j=1}^{m_i} \log p(\mathbf{w}^T\mathbf{x}_{ij}|y_i)}{\sum_{i=1}^{N}\sum_{c=1}^{C} \frac{1}{m_i}\sum_{j=1}^{m_i} \log p(\mathbf{w}^T\mathbf{x}_{ij}|c)} \tag{7.16}$$

To solve this optimisation problem, we now have to estimate the likelihood probability $p(\mathbf{x}|c)$.

7.3.2 Locally Gaussian embedding

Inspired by the work in [86], we model $p(\mathbf{w}^T\mathbf{x}_{ij}|c)$ as a local Gaussian, namely,

$$p(\mathbf{w}^T\mathbf{x}_{ij}|c) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_{ij}; \mathbf{w}^T\mu_{ijc}; \mathbf{w}^T\Sigma_{ijc}\mathbf{w}) \tag{7.17}$$

and the parameters of the Gaussian for $\mathbf{x}_{ij}$ and class c can be approximated by finding the k nearest neighbours from class c.

By substituting Eq. (7.17) into Eq. (7.16), we can find the projection $\mathbf{w}^\star$ by maximizing $f(\mathbf{w})$:

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} \frac{\sum_{i=1}^{N}\sum_{c=1}^{C} \frac{1}{m_i}\sum_{j=1}^{m_i} \frac{1}{2\sigma_{ijc}^2}\Delta_{ijc}^T\mathbf{w}\mathbf{w}^T\Delta_{ijc}}{\sum_{i=1}^{N} \frac{1}{m_i}\sum_{j=1}^{m_i} \frac{1}{2\sigma_{ijy_i}^2}\Delta_{ijy_i}^T\mathbf{w}\mathbf{w}^T\Delta_{ijy_i}} \tag{7.18}$$

s.t. $\mathbf{w}^T\mathbf{w} = 1$, where $\Delta_{ij} = \mu_{ij} - \mathbf{x}_{ij}$.

The objective function can be solved with a single eigen-decomposition. Define

$$A = \sum_{i=1}^{N} \frac{1}{m_i} \sum_{c=1}^{C} \sum_{j=1}^{m_i} \frac{1}{\sigma_{ijc}^2} \Delta_{ij} \Delta_{ijc}^T, \tag{7.19}$$

and

$$B = \sum_{i=1}^{N} \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{1}{\sigma_{ijy_i}^2} \Delta_{ijy_i} \Delta_{ijy_i}^T \tag{7.20}$$

then Eq. (7.9) becomes:

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T A \mathbf{w}}{\mathbf{w}^T B \mathbf{w}} \tag{7.21}$$

s.t. $\mathbf{w}^T \mathbf{w} = 1$.

Maximizing the objective function in Eq. (7.21) is a well-known eigensystem problem:

$$A\mathbf{w} = \lambda B\mathbf{w}. \tag{7.22}$$

The linear projection to be obtained is composed of the eigenvectors of $B^{-1}A$ associated with the first d largest eigenvalues if we want the projected space to be of d dimensions. The whole process of the algorithm is illustrated in Algorithm 2.

Algorithm 2 Locally Gaussian embedding

1. Calculate the local features $\{\mathbf{x}_{ij}\}$ for each video sequence X_i in the training set.
2. Find the k nearest neighbours of local feature descriptors: $\{\mathbf{x}_{ij}\}$ in each class.
3. For video X_i, calculate the parameters μ_{ij} and σ_{ij} of each local feature associated with each class.
4. Compute the auxiliary matrices: A and B using Eq. (7.10) and (7.11).
5. Solve the generalised eigenvector decomposition problem in Eq. (7.13) to find $\mathbf{w}^*$.

7.4 Experiments and results

We evaluate LGE for human action recognition and conduct experiments on the KTH and HMDB51 datasets. We compare the performance of LGE with principal component analysis (PCA) and I2CDDE. As the number of nearest neighbours is the only parameter of LGE, we have also investigated the effects of different values of k

on the performance of LGE.

7.4.1 Experimental settings

Similar to the experimental settings in Chapter 6, we utilise Dollár's periodic detector [25] to detect spatio-temporal interest points (STIPs) and the three-dimensional histogram of oriented gradients (HOG3D) [54], which is descriptive and relatively compact with 1000 dimensions, is used for the description of STIPs. For action recognition, we directly use the maximum a posteriori (MAP) classifier in Eq. (7.14).

7.4.2 Results

Local Gaussians As our algorithm is built on the assumption that local feature descriptors are from multi-modal Gaussian distributions, we would like to look into the distributions of local feature descriptors. We plot the probability density of local feature descriptors from the KTH dataset in Fig. 7-1, which show the descriptors with one dimension (a) and two dimensions (b), respectively. It is clear to see that the local feature descriptors are multi-modal Gaussian distributions.

Results on KTH The comparison results on the KTH dataset are shown in Fig. 7-2. LGE outperforms PCA under all the dimensions and numbers of nearest neighbours with significant margins. After 120 dimensions, both LGE and PCA can produce relatively stable and satisfactory results. Note that both LGE achieves the best results when 20 nearest neighbours are used for the estimation of parameters in local Gaussians.

We have also compared with the algorithm I2CDDE proposed in Chapter 6. The comparative results of LGE and I2CDDE with the NBNN, local NBNN and the NBNN kernel classifiers are shown in Fig. 7-3. With the LGE dimensionality reduction, the performance of NBNN, local NBNN and the NBNN kernel is also improved, especially for the NBNN kernel compared with I2CDDE. With regards to NBNN and local NBNN, LGE and I2CDDE produce comparable results on this dataset.

The comparison of the best results on the KTH dataset produced by the LGE,

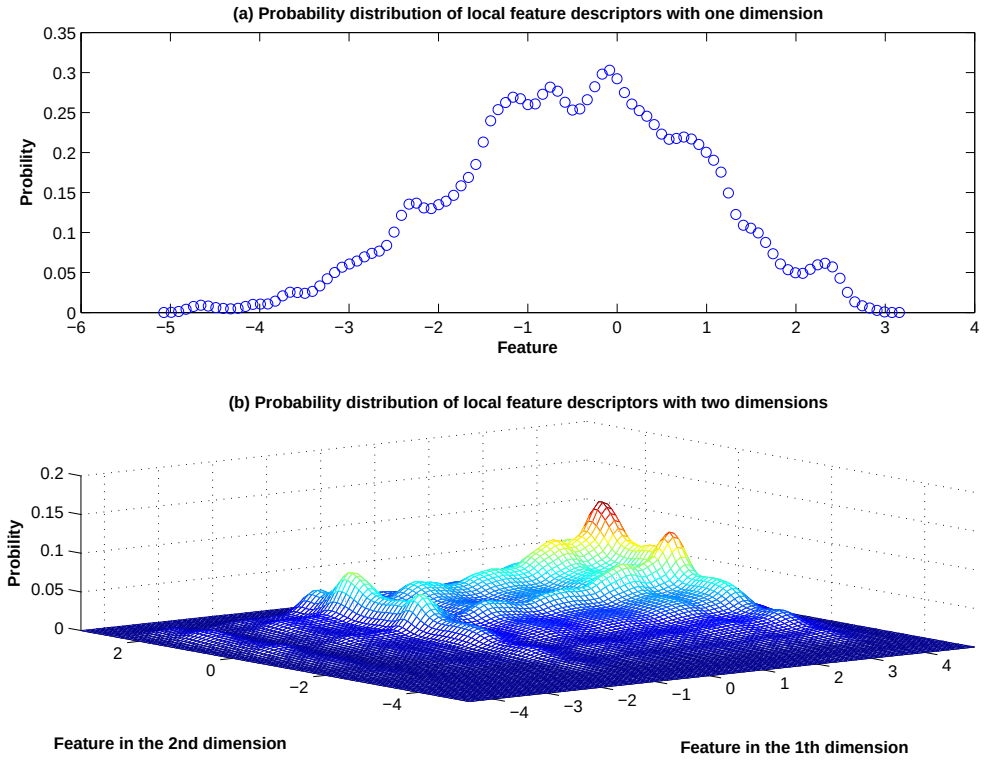

Figure 7-1: The illustration of the probability density of the local feature descriptors with one dimension (a) and two dimensions (b). The local feature descriptors are from the KTH dataset.

Classifiers	LGE	I2CDDE	PCA
MAP	93.5%	92.5%	90.6%
NBNN	94.2%	94.3%	92.9%
Local NBNN	94.4%	94.9%	93.2%
The NBNN kernel	93.5%	93.2%	89.3%

Table 7.1: The comparison of the best results given by LGE, I2CDDE and PCA with different classifiers on the KTH dataset.

I2CDDE and PCA with different classifiers are summarised in Table 7.1. We can see that LGE and I2CDDE achieve comparable results on this dataset and outperform PCA consistently with different classifiers.

Results on HMDB51 The results on the HMDB51 dataset are shown in Fig. 7-3. LGE consistently outperforms PCA with different numbers of nearest neighbours and under different dimensions. Both LGE and PCA achieve relatively stable performance after 50 dimensions. Compared with the results on KTH, the improvement of LGE over PCA is less significant, which would be due to the much noisier local features in the HMDB51 dataset.

Similarly, the comparison of LGE and I2CDDE with the NBNN, local NBNN and NBNN kernel classifiers are illustrated in Fig. 7-5. LGE significantly outperforms I2CDDE with NBNN, local NBNN and the NBNN kernel on the HMDB51 dataset, which manifests that LGE is more robust than I2CDDE especially on realistic datasets.

The comparison of the best results from LGE, I2CDDE and PCA with different classifiers is also shown in Table 7.2. The results of LGE and I2CDDE with NBNN, local NBNN and the NBNN kernel are comparable while LGE with the MAP classifier produces the best results on this dataset. LGE and I2CDDE outperform PCA as well in this dataset.

Note that we have experimented with the original local feature descriptors (without applying dimensionality reduction) with the maximum a posteriori (MAP) classifier. The best results without dimensionality reduction (with different numbers of nearest neighbours) on the KTH and HMDB51 datasets are 51.4% and 20.7%, re-

Classifiers	LGE	I2CDDE	PCA
MAP	40.8%	40.6%	38.7%
NBNN	36.4%	36.8%	33.5%
Local NBNN	37.3%	37.4%	34.4%
The NBNN kernel	31.3%	30.2%	27.9%

Table 7.2: The comparison of the best results given by LGE, I2CDDE and PCA with different classifiers on the HMDB51 dataset.

spectively, which implies the effectiveness of the dimensionality reduction techniques.

7.5 Conclusions

In this chapter, we have proposed a discriminative dimensionality reduction algorithm, named locally Gaussian embedding (LGE), for local features. With the simple maximum a posteriori (MAP) classifier, we have applied LGE for human action recognition on the KTH and HMDB51 datasets. Experimental results show that LGE outperforms PCA significantly, which validates the effectiveness of LGE and also verifies the local Gaussian assumption of local feature descriptors. In addition, the comparison between LGE and I2CDDE shows that LGE is more robust than I2CDDE, which would be due to the explicitly modelling of the likelihood in LGE rather the approximation using the image-to-class distances in I2CDDE.

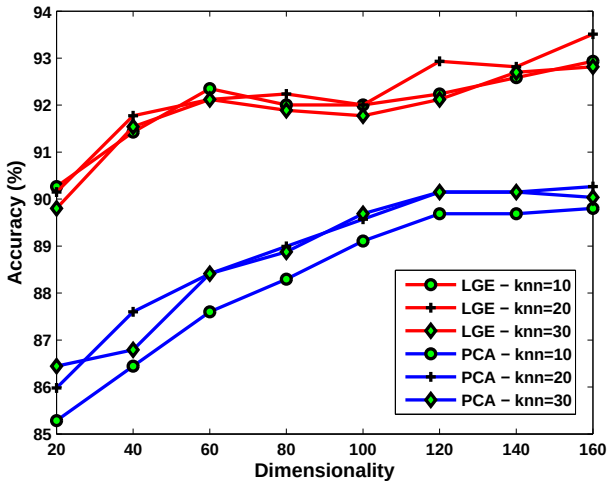

Figure 7-2: The performance comparison between LGE and PCA with different dimensions and numbers of nearest neighbours (knn) on the KTH dataset.

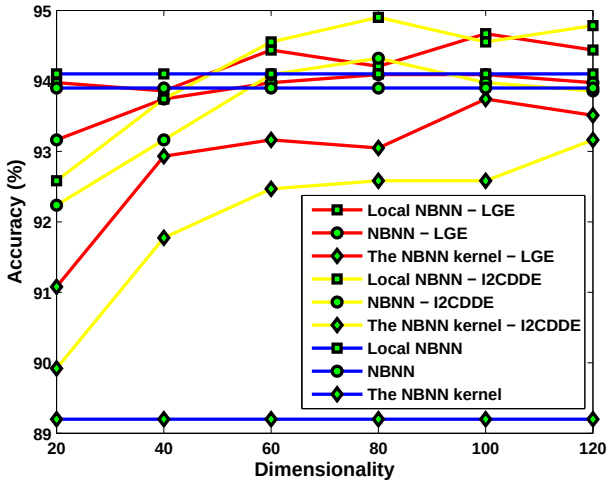

Figure 7-3: The performance comparison of LGE and I2CDDE with NBNN, local NBNN and the NBNN kernel classifiers on the KTH dataset. For LGE, we use knn = 10 in this experiment.

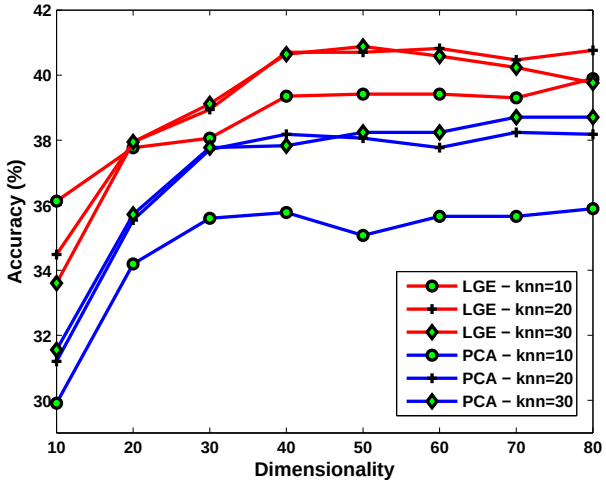

Figure 7-4: The performance comparison between LGE and PCA with different dimensions and numbers of nearest neighbours (knn) on the HMDB51 dataset. The results are the average over three training/test splits.

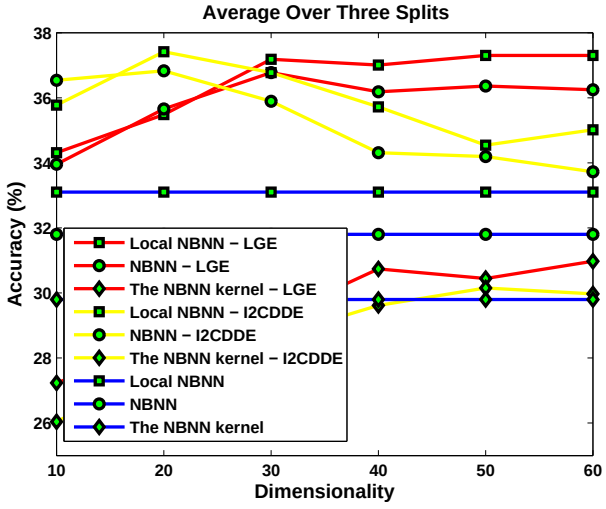

Figure 7-5: The performance comparison of LGE and I2CDDE with NBNN, local NBNN and the NBNN kernel on the HMDB51 dataset. The results are the average over three training/test splits. For LGE, we use $knn = 30$ in this experiment.

Chapter 8

Conclusions and Future Work

In this concluding chapter we summarise the contributions of this thesis, and discuss the important directions of future work.

8.1 Conclusions

The central task of the thesis is human action recognition. We have addressed this problem from the perspective of feature learning for both holistic and local representations. We may carefully draw some conclusions from experimental results of our work on widely used human action datasets.

- Although holistic representations have been regarded to be dependent on preprocessing steps such as background subtraction and tracking, by simply employing a frame differentiae operation, our methods can achieve comparable results with the state of the art. The effectiveness of holistic methods is largely credited to the preservation of structures of actions. With the advance of techniques such as visual tracking, background subtraction and detection with which human figures in the video sequence can be well captured, holistic methods can still find their applications. In addition, feature learning techniques used in holistic representations can be easily tuned for local feature description, which would enhance the importance of holistic methods.
- Local methods have dominated action recognition since the introduction of the

BoW model because of its theoretical simplicity and efficient implementation. Most of the current local methods are based on the BoW model and its variants. The success could be largely due to the use of powerful classifiers such as support vector machines (SUMS). However, the vector quantisation and loss of structural information would be the limitation of local methods based on the BoW model. Moreover, the BoW model would also compromise the effectiveness of local feature descriptors owing to quantisation. It has been shown in our evaluation work and dimensionality reduction methods that local feature descriptors can achieve impressive and comparable performance even with the simple NBNN and maximum a posteriori (MAP) classifiers.

We have also seen that the potential of learning efficient local feature descriptors has long been ignored for action recognition. From the findings of our evaluation work on local methods, the nonparametric naive Bayes nearest neighbour classifier (NBNN) and its variants including the NBNN kernels and the local NBNN, have produced impressive results. Our work on dimensionality reduction of local features have also indicated the importance of effectiveness and efficiency of local feature descriptors.

8.1.1 Holistic representations

In holistic representations, we have proposed three global descriptors for action representations.

- **Motion and structure feature embedding**

Based on the fact that motion and structure features are the main cues of an action, we explicitly extract these features from video sequences by motion history image (MHI) and structure planes, which are actually 2D images called feature maps. Effective biologically-inspired features based on 2D Gabor filters and max pooling are extracted from the feature maps and employed as the final holistic representation of actions.

- **Spatio-temporal Laplacian pyramid coding**

We have extended the idea of Laplacian pyramid from the image domain to the

spatio-temporal video domain and proposed a global descriptor based on spatio-temporal Lapidarian pyramid coding. The spatio-temporal Laplacian pyramid as a multiple resolution technique is firstly adopted into the video domain for human action recognition. Furthermore, the idea of biologically-inspired features has also been transferred for action representation by extending the Gabor filters and max pooling to their 3D versions.

- **Spatio-temporal oriented energies**

Actions in video sequences can be viewed as oriented patterns in spatio-temporal dimensions. To effectively capture the orientation information, we have proposed combining the multiple resolution technique, spatio-temporal Laplacian pyramid and the steerable filters. Another global descriptor based on spatio-temporal oriented energies is obtained for the final representation of actions.

8.1.2 Local representations

In local representations, we have firstly done a comprehensive evaluation of local methods for action recognition. Based on the findings of the evaluation, we have proposed two dimensionality-reduction algorithms via image-to-class (I2C) distances and local Gaussians.

- **Performance evaluation of local methods**

Our evaluation has provided an insight into the performance of local methods based on spatio-temporal interest points. Those methods include feature-coding algorithms both from the image and video domains, some of which were proposed for image/scene classification while have not been applied to human action recognition. We pull all those methods under common experimental settings and compare their performance. Therefore, our evaluation work provides a guideline for the further work on action recognition based on local features.

- **Discriminant embedding via image-to-class distances**

From the experimental results of the evaluation work, we found the methods based image-to-class (I2C) distances, *e.g.*, NBNN, NBNN kernels and local NBNN, show promising results. However, one of the disadvantages of the I2C based methods

is the computational burden due to the nearest-neighbour search of local features. It tends to be impractical if there is a huge number of local features, *e.g.*, dense trajectories, especially when the local features are in a high-dimensional space. We therefore propose a dimensionality-reduction algorithm for local features based on the I2C distances. The criterion used is to minimise the I2C distances of local features to their own classes while maximizing the I2C distances to classes they do not belong to. The I2C distance based methods including NBNN, the NBNN kernel and local NBNN are significantly improved after the embedding.

- **Locally Gaussian Embedding**

Inspired by the success of I2C embedding, we would like to look more deeply into the algorithms based on I2C distances. The I2C distance is actually an approximation of the conditional probabilities of local feature descriptors, which would not be robust due to the existence of noisy local features. We try to avoid the approximation by explicitly model the probability via local Gaussians. A novel discriminant embedding algorithm has been proposed based on local Gaussians for dimensionality reduction of local features. This embedding algorithm has demonstrated to be robust for action recognition, especially on realistic datasets.

8.2 Future work

The directions of future work could be considered in learning of representations both for local and holistic methods.

8.2.1 Holistic methods

The performance of the global descriptors proposed in Chapters 2, 3 and 4 suggests the effectiveness of holistic representations for action recognition. Nevertheless, these global descriptors are all hand-crafted and therefore not flexible. The good performance could be due to the use of discriminative dimensionality-reduction techniques, *e.g.*, discriminant locality alignment (DLA), and powerful classifiers, *e.g.*, support vector machines (SVMs). Although machine learning algorithms including Restricted

Boltzmann Machines (RBMs) [111] and 3D convolutional neural networks (3D CNN) [47] have been exploited for learning spatio-temporal features for action recognition, their performance is still unsatisfactory. A promising direction is to explore more efficient and effective machine learning algorithms to learn spatio-temporal features for holistic representations.

- **Parameter learning**

As shown in our holistic methods, the most important components in extracting a global descriptor are filtering and pooling. However, the filter scales and sizes of pooling regions are experimentally set which could be parameterised and learned by learning algorithms [87, 48, 106, 49].

- **Discriminative deep learning**

The deep learning algorithms such as convolutional RBM and 3D CNN are all unsupervised learning algorithms. However, it has been shown that discriminative learning could improve the performance of deep-learning algorithms in the image domain [31]. Extending discriminative deep-learning algorithms into the video domain for holistic representation of actions would also be an interesting direction.

8.2.2 Local methods

The performance of the dimensionality-reduction algorithms proposed in Chapters 6 and 7 suggests the potential of learning feature descriptors. Deep learning algorithms have also been explored for local spatio-temporal feature learning based on stacked independent analysis (ISA) [62]. The success of deep learning for action recognition also indicates the importance and potential of feature learning. For local methods, two possible directions including dimensionality reduction and local descriptor learning could be considered.

- **Dimensionality reduction of local feature descriptors**

Inspired by the work of I2CDDE and LGE, it would be interesting to consider proposing new algorithms for the dimensionality reduction of local feature descriptors in the following possible directions.

Regularisation The first one is to extend the I2CDDE and LGE algorithms.

As can be seen in our experiments that I2CDDE and LGE perform very well on the KTH dataset which is relatively easy, while being sensitive to noisy local features in realistic datasets such as the UCF YouTube and HMDB51 datasets. The extension could be based on incorporating/integrating regularisation terms into the objective functions of I2CDDE and LGE to make them more robust.

Criteria The second one is to construct novel objective functions for dimensionality-reduction algorithms. The criteria in I2CDDE and LGE are based on the image-to-class distances and likelihood of local feature descriptors, which, however, are restricted in the NBNN framework. Constructing more general objective functions based on other criteria beyond the NBNN framework could be considered.

- **Local feature descriptor learning**

Descriptor learning has recently drawn increasing attention in computer vision. Many machine learning techniques have been applied to learning descriptors in the image domain for feature matching and image retrieval [40, 87, 94]. The following two aspects could be considered for action recognition.

Discriminative descriptor learning From the results in Chapters 6 and 7, we have learned that the reason that our proposed algorithms significantly outperform PCA is the use of label information of samples. With respect to visual recognition, discriminative information such label information can be incorporated into the feature descriptor learning to improve the discriminative ability [108, 114].

Binary descriptor learning With the use of local features, we need local feature descriptors to be not only accurate but also efficient. Binary descriptors [126, 64, 115, 113] are of particular interest as they require far less storage capacity and offer much faster search. With respect to video analysis, currently the widely used descriptors including HOG/HOG and HOG3D are all hand-crafted and of high dimension. It would be promising if descriptors are learned from raw video data. Binary descriptors can be learned based on an intermediate representations such as HOG3D or directly from the 3D cuboids detected from video sequences.

Bibliography

[1] E. H Adelson and J. R Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A*, 2(2):284–299, 1985.

[2] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16, 2011.

[3] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):288–303, 2010.

[4] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. 2006.

[5] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet. Towards optimal naive bayes nearest neighbor. In *European Conference on Computer Vision*, pages 171–184. 2010.

[6] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14:585–591, 2001.

[7] C.M. Bishop and N.M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[8] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2002.

[9] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[10] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2559–2566, 2010.

[11] Y. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, 2010.

[12] Alan C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):55–73, 1990.

[13] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1948–1955, 2009.

[14] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31(4):532–540, 1983.

[15] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):338–352, 2011.

[16] K. Cannons, J. Gryn, and R. Wildes. Visual tracking using a pixelwise spatiotemporal oriented energy representation. *European Conference on Computer Vision*, pages 511–524, 2010.

[17] B. Caputo and L. Jie. A performance evaluation of exact and approximate match kernels for object recognition. *Electronic Letters on Computer Vision and Image Analysis*, 8(3):15–26, 2009.

[18] C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[19] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[20] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[21] T. de Campos, M. Barnard, K. Mikolajczyk, J. Kittler, F. Yan, W. Christmas, and D. Windridge. An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *2011 IEEE Workshop on Applications of Computer Vision*, pages 344–351, 2011.

[22] K.G. Derpanis and J.M. Gryn. Three-dimensional n-th derivative of gaussian separable steerable filters. In *IEEE International Conference on Image Processing*, volume 3, pages III–553, 2005.

[23] K.G. Derpanis, M. Sizintsev, K. Cannons, and R.P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1990–1997, 2010.

[24] K.G. Derpanis and R.P. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–239, 2009.

[25] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[26] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, 2003.

[27] I. Everts, J. C. van Gemert, and T. Gevers. Evaluation of color stips for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[28] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[29] E.K. Garcia, S. Feldman, M.R. Gupta, and S. Srivastava. Completely lazy learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(9):1274–1285, 2010.

[30] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A string of feature graphs model for recognition of complex activities in natural videos. In *IEEE International Conference on Computer Vision*, pages 2595–2602, 2011.

[31] R. Gens and P. Domingos. Discriminative learning of sum-product networks. In *Advances in Neural Information Processing Systems*, pages 3248–3256, 2012.

[32] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):883–897, 2011.

[33] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–53, December 2007.

[34] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and CH Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 222–228, 1994.

[35] T. Guha and R. K Ward. Learning sparse representations for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1576–1588, 2012.

[36] X. He, D. Cai, Y. Shao, H. Bao, and J. Han. Laplacian regularized gaussian mixture model for data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 23(9):1406–1418, 2011.

[37] X. He, M. Ji, C. Zhang, and H. Bao. A variance minimization criterion to feature selection using laplacian regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2013–2025, 2011.

[38] D. J Heeger. Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1(4):279–302, 1988.

[39] G.E. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[40] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[41] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan. View-independent behavior analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(4):1028–1035, 2009.

[42] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[43] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[44] P. Jain, B. Kulis, J.V. Davis, and I.S. Dhillon. Metric and kernel learning using a linear transformation. *Journal of Machine Learning Research*, 13:519–547, 2012.

[45] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *International Conference on Computer Vision*, pages 2146–2153, 2009.

[46] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[47] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In *27th International Conference on Machine Learning*, 2010.

[48] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3370–3377, 2012.

[49] Y. Jia, O. Vinyals, and T. Darrell. On compact codes for spatially pooled features. In *International Conference on Machine Learning*, 2013.

[50] Y. Jiang, Q. Dai, X. Xue, W. Liu, and C. Ngo. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*, pages 425–438. Springer, 2012.

[51] G. Johansson. Visual motion perception. *Scientific American*, 1975.

[52] I. Junejo, E Dexter, I Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *European Conference on Computer Vision*, pages 293–306, 2008.

[53] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–506, 2004.

[54] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Learning Conference*, pages 995–1004, 2008.

[55] O. Kliper-Gross, T. Gurovich, Y.and Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision*, pages 256–269, 2012.

[56] J. Koenderink. The structure of images. *Biological Cybernetics*, 50(5):363–370, 1984.

[57] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2046–2053, 2010.

[58] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, pages 2556–2563, 2011.

[59] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[60] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3):207–229, 2007.

[61] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[62] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[63] H. Lee, R. Grosse, R. Manganate, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, pages 609–616, 2009.

[64] S. Leutenegger, M. Chli, and R.Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision*, pages 2548–2555, 2011.

[65] X. Li, S. Lin, S. Yan, and D. Xu. Discriminant locally linear embedding with high-order tensor data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 38(2):342–352, 2008.

[66] T. Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1-2):225–270, 1994.

[67] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003, 2009.

[68] J. Liu and M. Shah. Learning human actions via information maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[69] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *IEEE International Conference on Computer Vision*, pages 2486–2493, 2011.

[70] D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[71] H. Lu, G. Fang, X. Shao, and X. Li. Segmenting human from photo images based on a coarse-to-fine scheme. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(3):889–899, 2012.

[72] Z. Lu, Y. Peng, and H. Ip. Spectral learning of latent semantics for action recognition. In *IEEE International Conference on Computer Vision*, pages 1503–1510, 2011.

[73] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, 1981.

[74] S. Lyu. Mercer kernels for object recognition with local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 223–229, 2005.

[75] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Internation Conference on Machine Learning*, pages 689–696, 2009.

[76] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition. *European Conference on Computer Vision*, pages 508–521, 2010.

[77] S. McCann and D. G. Lowe. Local naive bayes nearest neighbor for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3650–3656, 2012.

[78] K. Mikolajczyk and J. Matas. Improving descriptors for fast tree matching by optimal linear projection. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[79] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[80] K. Mikolajczyk and H. Uemura. Action recognition with appearance–motion features and fast search trees. *Computer Vision and Image Understanding*, 115(3):426–438, 2011.

[81] M Concetta Morrone and RA Owens. Feature detection from local energy. *Pattern Recognition Letters*, 6(5):303–313, 1987.

[82] J. Mutch and D.G. Lowe. Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, 80(1):45–57, 2008.

[83] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 79(3):299–318, 2008.

[84] H. Ning, Y. Hu, T. Huang, and N. Avenue. Searching human behaviors using spatial-temporal words. In *IEEE International Conference on Image Processing*, 2007.

[85] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(3):710–719, 2005.

[86] Nathan P. and Maya R.G. Dimensionality reduction by local discriminative gaussians. In *ACM Internatinal Conference on Machine Learning*, 2012.

[87] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Descriptor learning for efficient retrieval. In *European Conference on Computer Vision*, pages 677–691. 2010.

[88] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1242–1249, 2012.

[89] K. Rematas, M. Fritz, and T. Tuytelaars. The pooled nbnn kernel: beyond image-to-class and image-to-image. In *Asian Conference on Computer Vision*, pages 176–189, 2012.

[90] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[91] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature neuroscience*, 3 Suppl:1199–204, November 2000.

[92] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[93] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[94] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *IEEE International Conference on Computer Vision*, pages 2564–2571, 2011.

[95] S. Sadanand and J.J. Corso. Action bank: A high-level representation of activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241, 2012.

[96] S. Savarese, A. DelPozo, J.C. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and video Computing*, pages 1–8, 2008.

[97] K. Schilder and L. Van Goo. Action snippets: How many frames does human action recognition require? In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[98] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[99] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *15th International Conference on Multimedia*, pages 357–360, 2007.

[100] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 994–1000, 2005.

[101] L. Shao and R. Mattivi. Feature detector and descriptor evaluation in human action recognition. In *ACM International Conference on Image and Video Retrieval*, pages 477–484, 2010.

[102] L. Shao, X. Zhen, Y. Liu, and L. Ji. Human action representation using pyramid correlogram of oriented gradients on motion history images. *International Journal of Computer Mathematics*, 88(18):3882–3895, 2011.

[103] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: an application to weakly supervised action classification. In *European Conference on Computer Vision*, pages 55–68. Springer, 2012.

[104] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):300–312, 2007.

[105] E.P. Simoncelli and W.T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE Conference on Image Processing*, volume 3, pages 444–447, 1995.

[106] K. Simonyan, A. Vedaldi, and A. Zisserman. Descriptor learning using convex optimisation. In *European Conference on Computer Vision*, pages 243–256. 2012.

[107] D. Song and D. Tao. Biologically inspired feature manifold for scene classification. *IEEE Transactions on Image Processing*, 19(1):174–184, 2010.

[108] C. Strecha, A. M. Bronstein, M. M. Bronstein, and P. Fua. Ldahash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):66–78, 2012.

[109] J Sun, X. Wu, S. Yan, L.-F. Cheong, T-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2004–2011, 2009.

[110] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3681–3688, 2012.

[111] G. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *European Conference on Computer Vision*, pages 140–153. 2010.

[112] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[113] T. Trzcinski, C. M. Christoudias, P. Fua, and V. Lepetit. Boosting binary keypoint descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[114] T. Trzcinski, M. Christoudias, V. Lepetit, and P. Fua. Learning image descriptors with the boosting-trick. In *Advances in Neural Information Processing Systems*, pages 278–286, 2012.

[115] T. Trzcinski and V. Lepetit. Efficient discriminative projections for compact binary descriptors. In *European Conference on Computer Vision*, pages 228–242, 2012.

[116] T. Tuytelaars, M. Fritz, K. Saenko, and T. Darrell. The nbnn kernel. In *IEEE International Conference on Computer Vision*, pages 1824–1831, 2011.

[117] K. E. A. Van De Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[118] J. C. van Gemert, C.J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligance*, 32(7):1271–1283, 2010.

[119] M. Varma and R. Babu. More generality in efficient multiple kernel learning. In *International Conference on Machine Learning*, pages 1065–1072, 2009.

[120] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *International Conference on Multimedia*, pages 1469–1472, 2010.

[121] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European Conference on Computer Vision*, pages 84–97, 2012.

[122] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *IEEE International Conference on Computer Vision*, pages 257–264, 2003.

[123] H. Wang, A. Kläser, C. Schmid, and C-L. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

[124] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Learning Conference*, 2009.

[125] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

[126] J. Wang, S. Kumar, and S. Chang. Sequential projection learning for hashing with compact codes. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1127–1134, 2010.

[127] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3360–3367, 2010.

[128] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann. Action recognition by exploring data distribution and feature correlation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1370–1377, 2012.

[129] Zh. Wang, Y. Hu, and L-T. Chia. Image-to-class distance metric learning for image classification. In *European Conference on Computer Vision*, pages 706–719, 2010.

[130] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *IEEE International Conference on Computer Vision*, pages 1–7, 2007.

[131] D. Weinland, M. Özuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*, pages 635–648, 2010.

[132] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. *European Conference on Computer Vision*, pages 768–784, 2000.

[133] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*, pages 650–663. Springer, 2008.

[134] H. Wilson and J. Bergen. A four mechanisim model for threshold special vision. *Vision Research*, pages 19–31, 1979.

[135] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[136] P. Yan, S. Khan, and M. Shah. Learning 4D action feature models for arbitrary view action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[137] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, 2009.

[138] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2061–2068, 2010.

[139] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *IEEE International Conference on Computer Vision*, pages 492–497, 2009.

[140] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 984–989, 2005.

[141] C. Yuan, X. Li, W. Hu, H. Ling, and S. Maybank. 3d r transform on spatio-temporal interest points for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[142] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[143] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *IEEE International Conference on Computer Vision*, pages 471–478, 2011.

[144] T. Zhang, D. Tao, X. Li, and J. Yang. Patch alignment for dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1299–1313, 2009.

[145] Y. Zhang, X. Liu, M-Ch. Chang, W Ge, and T. Chen. Spatio-temporal phrases for activity recognition. In *European Conference on Computer Vision*, pages 707–721, 2012.

[146] Z. Zhang, Y. Hu, S. Chan, and L.T. Chia. Motion context: A new representation for human action recognition. In *European Conference on Computer Vision*, pages 817–829, 2008.

[147] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *IEEE Transactions on attern Analysis and Machine Intelligence*, 34(3):436–450, 2012.

[148] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition)*, 8(4):406–424, 2004.

[149] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[150] X. Zhen and L. Shao. Spatio-temporal steerable pyramid for human action recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2013.

[151] X. Zhen, L. Shao, D. Tao, and X. Li. Embedding motion and structure features for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7):1182–1190, 2013.