# Human Action Recognition using Saliency-based Global and Local Features

**A thesis submitted in partial fulfilment**

**of the requirement for the degree of Doctor of Philosophy**

**Ashwan Anwer Abdulmunem**

**December 2017**

**Cardiff University**
**School of Computer Science & Informatics**

## Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... (candidate)

Date .....

## Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

## Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... (candidate)

Date .....

## Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

**To People you care**
**for their patience and support.**

# Abstract

Recognising human actions from video sequences is one of the most important topics in computer vision and has been extensively researched during the last decades; however, it is still regarded as a challenging task especially in real scenarios due to difficulties mainly resulting from background clutter, partial occlusion, as well as changes in scale, viewpoint, lighting, and appearance. Human action recognition is involved in many applications, including video surveillance systems, human-computer interaction, and robotics for human behaviour characterisation. In this thesis, we aim to introduce new features and methods to enhance and develop human action recognition systems. Specifically, we have introduced three methods for human action recognition. In the first approach, we present a novel framework for human action recognition based on salient object detection and a combination of local and global descriptors. Saliency Guided Feature Extraction (SGFE) is proposed to detect salient objects and extract features on the detected objects. We then propose a simple strategy to identify and process only those video frames that contain salient objects. Processing salient objects instead of all the frames not only makes the algorithm more efficient, but more importantly also suppresses the interference of background pixels. We combine this approach with a new combination of local and global descriptors, namely 3D SIFT and Histograms of Oriented Optical Flow (HOOF). The resulting *S*aliency *G*uided 3D *S*IFT and *H*OOF (SGSH) feature is used along with a multi-class support vector machine (SVM) classifier for human action recognition. The second proposed method is a novel 3D extension of Gradient Location and Orientation Histograms (3D GLOH) which provides

discriminative local features representing both the gradient orientation and their relative locations. We further propose a human action recognition system based on the Bag of Visual Words model, by combining the new 3D GLOH local features with Histograms of Oriented Optical Flow (HOOF) global features. Along with the idea from our first work to extract features only in salient regions, our overall system outperforms existing feature descriptors for human action recognition for challenging video datasets. Finally, we propose to extract minimal representative information, namely deforming skeleton graphs corresponding to foreground shapes, to effectively represent actions and remove the influence of changes of illumination, subject appearance and backgrounds. We propose a novel approach to action recognition based on matching of skeleton graphs, combining static pairwise graph similarity measure using Optimal Subsequence Bijection with Dynamic Time Warping to robustly handle topological and temporal variations. We have evaluated the proposed methods by conducting extensive experiments on widely-used human action datasets including the KTH, the UCF Sports, TV Human Interaction (TVHI), Olympic Sports and UCF11 datasets. Experimental results show the effectiveness of our methods for action recognition.

# Acknowledgements

Walking through the path of knowledge has always been an interesting challenge for me. Despite the ups and downs, the difficulties and disappointments, the happy moments and bad ones, my journey has finally come to its end.

Completing my PhD thesis would have been impossible without the help and support of my supervisor, Dr Yu-Kun. I will always be grateful to him for his guidance and patience in transforming me into a scientist. His excellent professional ethic motivated me to rigorously engage with research in the Computer Vision field. I would also like to thank Dr. Xianfang, for his valuable advise and help during my graduate studies. I was honored to collaborate with him and to learn from his experience. My supervisors have perfectly assisted me to all of my research needs.

Thanks to the Iraqi Ministry of Higher Education and Scientific Research (MHESR) for the financial support of my PhD study in Cardiff.

Special thanks and most important words of love and gratitude are kept for my husband for being so supportive to me and for sharing all of the happiest and stressful days of our lives.

My last thankful words are kept for my parents and my children Naseralhusien, Manar, and Ymama. It is thanks to them I was able to come to the finish line of my PhD. Thank You!

# Contents

# List of Publications

The work introduced in this thesis is based on the following publications.

- Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun, Saliency guided local and global descriptors for effective action recognition, Computational Visual Media, vol. 2, no. 1, pp. 97-106, 2016.
- Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun, 3D GLOH Features for Human Action Recognition, International Conference on Pattern Recognition (ICPR), pp. 805-810, 2016.
- Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun, Action Recognition based on Matching of Deforming Skeleton Graphs, under review.

# List of Figures

# List of Tables

# List of Acronyms

**3D GLOH** 3 Dimensional Gradient Location and Orientation Histogram

**3D SIFT** 3 Dimensional Scale Invariant Feature Transform

**BoVW** Bag of Visual Words

**CNN** Convolutional Neural Network

**CGKs** Context-Dependent Graph Kernels

**CRF** Conditional Random Field

**DAG** Directed Acyclic Graph

**DCT** Discrete Cosine Transform

**DTW** Dynamic Time Warping

**ESURF** Extended Speeded Up Robust Features

**FS-HHMM** Factored-State Hierarchical Hidden Markov Model

**GMKL** Generalized Multiple Kernel Learning

**gSpan** graph-Based Substructure Pattern Mining

**HMM** Hidden Markov Model

**HVT-HMM** hierarchical variable transition HMM

**HOF** Histograms Of Flow

**HOG** Histograms of Oriented Gradients

**HOOF** Histograms of Oriented Optical Flow

**KLT** Kanade-Lucas-Tomasi Feature Tracker

**KTH** Kungliga Tekniska Hogskolan (Swedish, Royal Institute of Technology)

**LBP** Local Binary Patterns

**LoG** Laplacian of Gaussian

**LTP** Local Trinary Patterns

**MBH** Motion Boundary Histograms

**MHI** Motion History Image

**MEI** Motion Energy Image

**NMF** non-Negative Matrix Factorisation

**PCA** Principal Component Analysis

**pLSA** probabilistic Latent Semantics Analysis

**RBF** Radial Basis Function

**ROI** Region of Interest

**RNN** Recurrent Neural Network

**SFGs** String of Feature Graphs

**SGFE** Saliency Guided Feature Extraction

**SGSH** Saliency Guided 3D SIFT- HOOF

**SH** 3D SIFT and HOOF

**SLIC** Simple Linear Iterative Clustering

**SOD** Simplex-based Orientation Decomposition

**STAOG** Spatio-Temporal And-Or Graph

**STIP** Spatio-Temporal Interest Point

**STV** Space-Time Volumes

**SURF** Speeded Up Robust Feature

**SVGS** Spatial Video Graph Set

**SVM** Support Vector Machine

**TVGS** Temporal Video Graph Set

**TVHI** TV Human Interaction

**UCF** University of Central Florida

**VLMM** Variable Length Markov Model

**VCG** Video Cooccurrence Graph

**VSG** Video Successiveness Graph

# *Chapter 1*

# Introduction

## 1.1 Human Activity Recognition

Human activity recognition and analysis, 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 and aims to recognise the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. The challenges of human action recognition come from difficulties such as scaling, occlusion and clutter. Another issue is the large variability in actions. When different subjects are performing the same action, they do not have the same appearance and their movements can be quite different for the same action. Even for a person performing the same action multiple times, each performance can be quite different from the previous one (see Figure 1.1).

There are different types of human activities. Depending on their complexity, they can be reasonably arranged into four distinct levels: gestures, actions, interactions, and group activities [4]. Gestures are basic movements of a person's body part, and are the atomic components depicting the significant movement of a person. 'Opening a hand', and 'raising a leg' are good examples of gestures. Actions are single-person activities that may be composed by multiple gestures organised temporally, such as

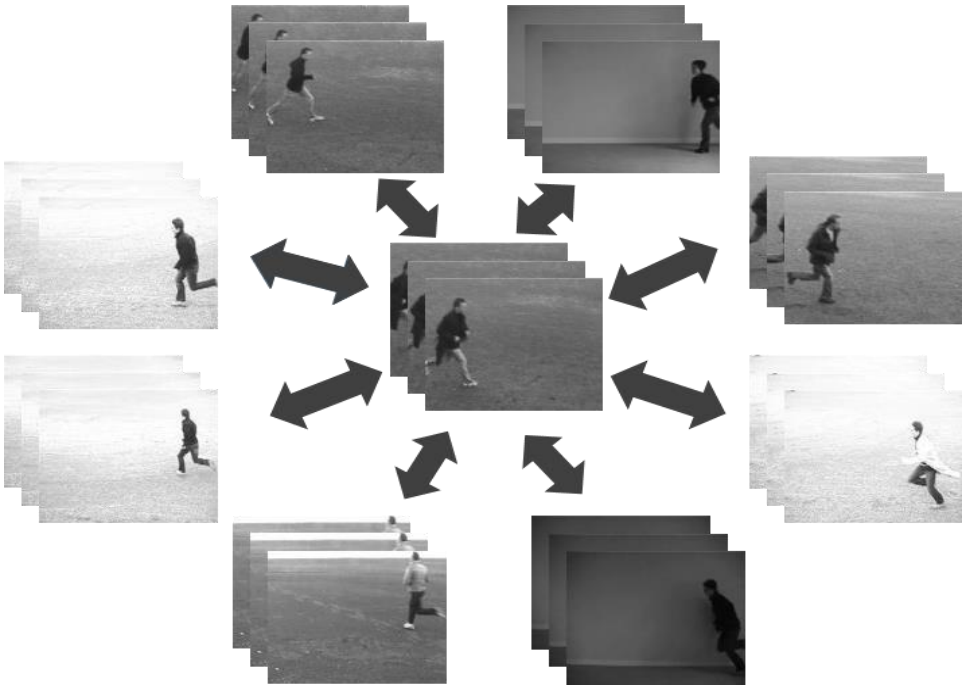


**Figure 1.1: Challenges in human action recognition: different clothes, different illumination, different background and action speed.**

'walking', 'hand-waving', and 'boxing'. Interactions are human activities that involve two or more persons. For example, 'two persons hugging' is an interaction between two people. Finally, group activities are the activities performed by conceptual groups composed of multiple persons: 'a group having a meeting', and 'two groups fighting' are typical examples. In this research, the main focus is to improve the performance and recognition accuracy of single person action and interactions from real-time video sequences.

Action recognition has been extensively researched, although there are still challenges for real-world applications. Earlier work on human action recognition in video [34, 164, 120] employed video datasets with mainly static cameras, simple and homogeneous backgrounds, and humans fully visible such as KTH [142] and Weizmann [179] video datsets. The research focus was to explore classifiers with variations in actors and

actions. In recent years the field of action recognition has in general moved towards less controlled and much more challenging types of data such as sports [137, 146, 119] or movies [86, 126]. For this task, methods that use local and global features such as [99] have shown excellent results. Although many successful methods have been proposed, there are still scopes for improvement, especially for real-world videos which have wide variations in people's posture and clothes, dynamic background and partial occlusions. Therefore, robust classification is still an important issue in the human action recognition problem and it is necessary to develop more robust alternatives. It is possible for humans to identify and distinguish different actions because the brain is capable of both learning new actions and recognising them. However, in computer vision, this same problem has proven to be one of the most difficult and lasting challenges in the field.

## 1.2 Problem statement

This dissertation focuses on the problem of action recognition in real video material, recorded under different environmental conditions varying from a fixed, clean background to complex, cluttered and moving backgrounds. A wide range of human activities have been investigated in this research from single person activities such as walking, running, jogging, etc. to human activities involving complex interaction such as high five, hug, hand shake, etc. We aim to address action recognition issues by introducing new methods for feature extraction, representation and classification to improve the performance and accuracy of human action recognition. We motivate to addressing the action recognition problem from the fact that the foreground region carry more robust information about the action. This helps to suppress the interference of background and thus makes the method more robust to background fluctuation. We will explain the methods in details in the remaining chapters.

## 1.3 Motivation

The development of computer vision has encouraged the occurrence of different novel recognition methods in both images and video sequences. Although it is still challenging to recognise a specific object from a dataset of images due to viewpoint change, illumination, partial occlusions, and intra-class difference and so on, many successful methods have been proposed, including those that are successfully extended from the image domain into video analysis and action recognition. However, current methods still need improvement, especially for real world videos and movies which have wide variations in people's posture and clothes, dynamic background, and partial occlusions. To conquer these deficiencies, a lot of researchers focus on part-based approaches for which only the 'interesting' parts of the video are analysed, rather than the whole video. These 'parts' can be trajectories or flow vectors of corners and spatial temporal interest points. Although part based approaches are promising they still suffer from inaccurate detection and tracking of interesting parts due to background clutter and motion which prevents a clear and informative representation.

The ability to detect, track, recognise and analyse human motion is beneficial for a wide range of high-level applications that rely on representations extracted from visual input. During the past few years, many approaches have been proposed to address these problems [130, 182, 4].

Some examples of applications that could benefit from reliable and efficient human action recognition are:

- Automated surveillance frameworks that monitor the overall public, utilised in places, for example, airport terminals, government buildings and banks. Applications to monitor and recognise suspicious movement without human instructions have yet to be figured out. Having an automated solution limitlessly enhances the recognition of suspicious activities as it decreases the likelihood of human distortion and misconception.

- Safety systems for detecting vulnerable users, most prominently the extremely youthful or the elderly such as systems to monitor users in and around occupied train stations or on cars to caution others of possible danger or for security monitoring.

- Health monitoring and preventative care for patients which have applications to flawlessly detect and track people inside their own environment. For instance, a system to monitor the elderly and alert the neighbourhood healing facility on the off chance that they have a fall or excursion.

This thesis has used several challenging, publicly available datasets designed for human action recognition, which are still very challenging in the field and highlight the ample ongoing room for improvement.

## 1.4 Thesis Contribution

The goal of this dissertation is the recognition of actions in uncontrolled, real video data. The first part of our work is based on saliency to guide local and global features which are employed for action classification. For this, existing approaches to describe local information in videos are investigated and new methods are developed.

The second part of this work introduces a new descriptor for action recognition in videos. We propose a novel effective feature called 3D GLOH (Gradient Location and Orientation Histogram), which describes local spatially varying information for video data. It detects interest points in the video and then describes them in 3D log-polar coordinates.

Thirdly, we propose to extract minimal representative information, namely deforming skeleton graphs corresponding to foreground shapes to effectively represent actions, removing the influence of these typical variations. We propose a novel approach to action recognition based on matching of skeleton graphs.

To summarize, we provide the following main contributions:

- We introduce a novel framework for human action recognition based on saliency guided local and global descriptors, by detecting only keypoints on salient regions and then describing those using 3D SIFT descriptor. This work was published in [1].

- We develop a novel local descriptor for video data based on histograms of gradient location orientations (3D GLOH) [2]. Our approach is based on a log-polar orientations to compute 3D gradients locations histograms for salient keypoints. Descriptor parameters are evaluated in depth and optimized for action recognition using bag-of-features representation.

- We develop a novel combination of local and global descriptors, which outperforms existing descriptors in action recognition with challenging real-world videos.

- We propose to represent actions in video sequences as sequences of deforming skeleton graphs of foreground subjects. The representation has significant advantages of being insensitive to changes of illumination, subject appearance and backgrounds. The proposed method is based on matching of deforming skeleton graphs. Our similarity measure takes into account topological variation, temporal variation and alignment of periodic actions to improve its robustness. Experimental results show that our method purely based on graph matching outperforms state-of-the-art action recognition methods. Moreover, since our method uses compact and highly abstract information, it achieves decent recognition performance with even a single example from each category, which is a very challenging scenario for existing methods. Due to the use of complementary information, we achieve even better recognition performance by fusing our method with an alternative image descriptor based method.

## 1.5 Thesis road map

The remaining chapters of the thesis are organized as follows:

- In **Chapter 2** an overview of the field of articulated human feature representation and recognition is presented. Moreover, evolution of human action recognition in recent years is briefly presented to provide an introduction to different approaches, different features, extraction, representation and classification techniques used by researchers over the last three decades. In addition, a comprehensive review of popular, challenging datasets and their evaluation metrics are also presented.

- **Chapter 3** introduces a novel approach to extracting and representing features for human action recognition. For feature representation and description, a new method has been proposed based on saliency to guide the combined descriptor to describe the video data, where saliency has been extensively researched to represent the importance of image regions. The new descriptor combines two different feature representations, the first one being the 3D SIFT descriptor (a local descriptor) and the other being the HOOF descriptor (a global descriptor), to get benefits from local and global descriptors to build robust and informative descriptor. The pipeline of the proposed method will be illustrated in this chapter.

- **Chapter 4** introduces a new 3D descriptor to better identify spatio-temporal characteristics. A novel 3D extension of Gradient Location and Orientation Histograms will be explained in details in this chapter. 3D GLOH descriptor provides a discriminative local feature representing not only the gradient orientations, but also their relative locations. In addition, a human action recognition system based on the Bag of Visual Words model will be introduced, by combining the new 3D GLOH local features with Histograms of Oriented Optical Flow (HOOF) global features. Along with the idea from Chapter 3 to extract

features only in salient regions, our overall system outperforms existing feature descriptors for human action recognition for challenging real-world video datasets.

- **Chapter 5** presents a method based on a shape-descriptor to extract minimal representative information, namely deforming skeleton graphs corresponding to foreground shapes to effectively represent actions, removing the influence of changes of illumination, subject appearance and backgrounds. In this chapter a framework of a proposed approach to action recognition will be presented. The proposed method based on matching of skeleton graphs combining a static pairwise graph similarity measure using Optimal Subsequence Bijection with Dynamic Time Warping to robustly handle topological and temporal variations. For common periodic actions, we extract a consistent starting frame from each video to temporally align deforming skeleton graphs. Moreover, we further develop a hierarchical matching strategy to significantly improve matching efficiency while keeping recognition accuracy. All these proposed solutions will be shown in this chapter. Comparison with state-of-the-art will be shown where the proposed method outperforms the state-of-the-art methods on standard benchmarks. For effectiveness, the method also has very good generalisability where decent performance can be achieved with only a single example from each action category since our method utilises complementary information to traditional image descriptor based methods (as shown in chapter 3). This chapter further demonstrates that even better performance can be obtained by fusing the output of both methods.

- **Chapter 6** summarises and concludes the thesis, highlights the achievements, discusses the limitations, and points to future research directions.

# *Chapter 2*

# Literature Review in Action Recognition

## 2.1 Introduction

The human action recognition problem has remained a challenging task in computer vision and multimedia content processing for more than two decades. Despite great effort, this task is still challenging as videos are complicated with significant variations even for the same type of action, making robust information extraction difficult. Firstly, the subject under observation can be distinctive in appearance, pose and size. Secondly, moving background, occlusion, non-stationary camera and complex environment can impede the observation. Comprehensive reviews of the literature can be found in many recent research papers [3, 4, 182, 15, 54, 138, 5, 33] addressing different aspects and issues raised in the human action recognition field. Many existing action recognition methods, including both low-level feature extraction and high-level representations, are extended from the text and image domains. Different approaches have been introduced to address the action recognition problem. Successful human action recognition systems have balanced between the recognition accuracy and the efficiency of feature extraction from the computational cost viewpoint. Accordingly, most research tries to find out reliable and robust scheme to extract features and effective classification algorithms to achieve the goal. In this chapter we will review the state-of-the-art methods for action recognition in benchmark video datasets.

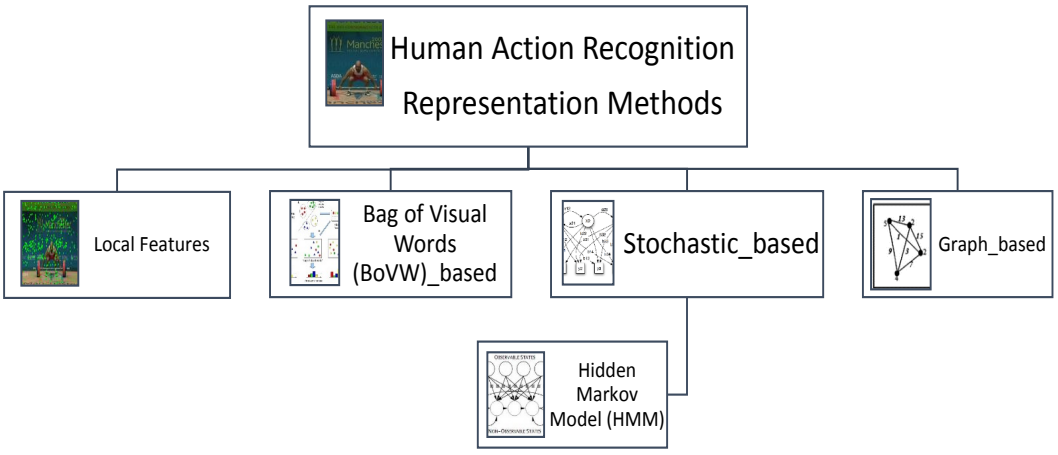



**Figure 2.1: Categorisation of human action recognition representation methods**

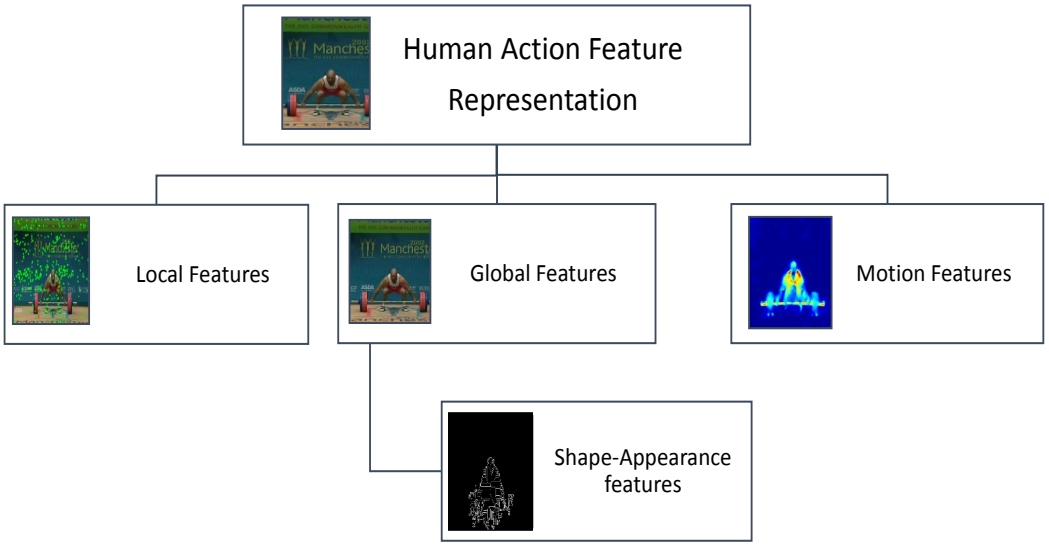



**Figure 2.2: Categorisation of human action feature representations**

## 2.2 Categorisation of Human Action Recognition

Human action recognition systems can be classified based on action representation methods or feature representation methods. In the former categorisation, human action recognition methods can be categorised into four classes: feature representation methods, bag of visual words (BoVW) [166, 180, 104, 131, 121], stochastic methods [116, 136, 38] and graph-based methods [165, 171] (see Figure (2.1)). Regarding feature representations, human action recognition methods can be classified into three classes based on the method of representing or extracting features from video data: local features [77, 100, 32], global features [92, 83], and motion features [16, 37, 129] (see Figure 2.2). Some methods combine different types of features and action representations to improve performance. In addition, recent deep learning based methods exploit large amounts of available training data to achieve human action recognition without hand-crafted features.

Local feature representation-based methods extract local features in the spatial-temporal domain to represent human actions. A set of spatio-temporal features are obtained in a bottom-up structure [77, 143, 180]. In contrast global feature representation methods do not require the localisation of body parts. Rather, global body structure and dynamics are utilised to represent human actions. In general, global approaches employ shape masks or silhouette information, stemming from background subtraction or difference images, to represent actions [121]. As an example, shape-appearance based approaches are based on building models to represent the actions and use these models in recognition. The third category of feature representation is mainly based on optical flow information [32, 26]. Optical-flow approaches depend on calculating the optical flow [19, 60, 102] to encode the energy of the action and represent actions as histograms of optical flow.

In the following sections, we will explain briefly the categorisation of feature representation schemes.

## 2.3 Feature Representation: Local Features

Human action recognition has been extensively researched through methods based on local representations. Methods based on local feature extraction, also known as local methods, encode a video sequence as a collection of local spatio-temporal features (local descriptors). Low-level features play a fundamental role in representations of human actions. In the last decades, many spatio-temporal descriptors have been proposed and shown to be effective for action recognition. These local descriptors are extracted from spatio-temporal interest points (STIPs) which can be sparsely detected from video sequences by detectors [84, 34, 100]. The features extracted from local descriptors are characterised with high dimensionality. As a result, generating codebooks is needed to acquire optimal codebooks with small size. This is usually accomplished by using one of the representation methods as shown in Figure 2.1. More details of these methods will be explained in section 2.7. Local approaches are popular feature extraction methods due to their many advantages:

- Resistance and less sensitivity to the noise in background, partial occlusion, viewpoint, and changes caused by illumination variation.
- Compared to global features, avoidance of some preliminary steps, e.g. background subtraction.
- Flexibility to model the local interactions between multiple features by using a local spatial-temporal feature-based representation.

If the video dataset contains individual actors recorded in clear environment with static camera, global descriptors give acceptable results with low cost but the effectiveness of these descriptors is related to the scenes and the accuracy of localisation or detection of the region of interest (ROI). For example, motion energy image (MEI) features [13] are global features which work by identifying regions with motion as regions of

interest. Conversely, local descriptors can better deal with changes in the environment but usually with higher computational costs.

### 2.3.1 Feature Detectors

In low-level feature representation methods, the key step to extract features from video is to detect interest points considered to be more informative than others, and describe them using some feature descriptors. Many approaches have been proposed to detect interest points. The most popular ones include SIFT detector [100] which works in 2D, space-time interest points detector (STIPs) [84, 85], (which extends the Harries detector [58] to 3D), temporal Gabor filters [17, 34], Hessian detector [112] (based on the determinant of the spatio-temporal Hessian matrix).

Lowe [100] introduced the Scale Invariant Feature Transform (SIFT) detector based on detecting maxima and minima of the difference-of-Gaussian in scale space. For each octave of scale space, the initial image is repeatedly convolved with Gaussians at different scales to produce the set of scale space images. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images. After each octave is produced, the Gaussian image is down-sampled by a factor of 2, and the process is repeated. Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel to its neighbours at the current and adjacent scales. The SIFT detector has the ability to identify a large number of keypoints. These keypoints are robust, informative, and affine and scale invariant.

Laptev [84] extended the Harris [58] and Forstner [45] interest point detectors to 3D detectors. The idea is based on extending the spatial domain of interest points into the spatio-temporal domain by requiring the image values in space-time to have large variations in both the spatial and the temporal dimensions. They proposed that the interest point can be detected at different types of interest point movement. However, the Harris corner detector is sensitive to changes in image scale, as a result it does not

provide a good basis for matching images of different sizes.

The Dollar detector [34] calculated a response function for each keypoint of video sequences. To calculate this response function two distinct linear filters are used. 2D Gaussian kernel filter is the first filter that is applied on the spatial axis and 1D Gabor filter is the second filter that is used for the temporal axis. They applied a spatio-temporal interest point detector to find local regions of interest in space and time (cuboids) which serve as the substrate for action recognition.

### 2.3.2 Feature Descriptors

To capture more information and features from detected keypoints, a description for these key points is needed to represent and encode the video information. Methods have been proposed to describe local interest points, each of which is denoted by $I(x, y, t)$, where $I$ represents input image, $x$ and $y$ indicate the spatial and $t$ indicates time of the point. A local patch is considered around each detected interest point. The detected patches are described to represent the actions. In this section, some descriptors used in human action applications will be introduced.

Many efforts have been made to extract and describe meaningful and robust information. Several feature descriptors have been successfully adapted from the image domain to the video domain to enhance the accuracy of human action recognition. Scovanner *et al.* [143] extended the SIFT descriptor [100] to the spatio-temporal domain.

Willems *et al.* [164] proposed the extended SURF (ESURF) descriptor, which is the generalisation of the SURF descriptor [12] to video by evaluating with changing scales and orientations. Their evaluation however was conducted only on datasets with a single actor and clear recording environments such as KTH.

Klaser *et al.* [77] represented video sequences as a 3D histogram of gradients. They extended the idea of Histogram of Oriented Gradients (HOG) [31] on images to video

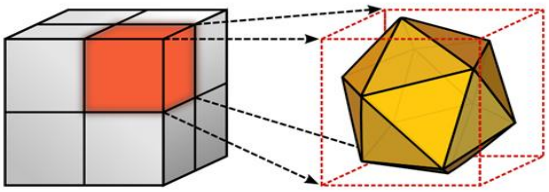


**Figure 2.3: Describing spatio-temporal points (HOG3D descriptor): the support region around a point of interest is divided into a grid of gradient orientation histograms; each histogram is computed over a grid of mean gradients; each gradient orientation is quantised using regular polyhedrons [77].**

to allow dense sampling of the cuboid with different scales and locations in the spatio-temporal representation (Figure 2.3). Laptev *et al.* [86] proposed the combined HOG/HOF descriptor which represents appearance by HOG and local motion by Histogram of Flow (HOF) [32]. A drawback of HOG features is that the local descriptors are extracted at a fixed scale; therefore, the size of the human in the image can have great influence on the performance.

Recently, Zhang *et al.* [180] introduced a 3D feature descriptor called simplex-based orientation decomposition (SOD), and combined it with a BoVW framework to recognise actions. The SOD descriptor is based on decomposing visual cue orientations in the spatio-temporal domain into three angles and transforming the decomposed angles into a simplex space, where the simplex space is a generalisation of the notion of a triangle or tetrahedron to arbitrary dimensions. They used the simplex space in the features representation to construct a compact, representative description of 3D visual features. Then, quadrant decomposition was performed to compute the final feature vector used for classification by combining the decomposed histograms from all quadrants.

Yeffet and Wolf [175] employed a feature descriptor named local trinary patterns (LTP), which was inspired by the local binary patterns (LBP) and successfully used

for action recognition. Every pixel at each frame was encoded as a short string of ternary digits (trits) by comparing this frame to the previous and the next frames. The frame was then divided into (m $\times$ n) regions and the histograms of the trinary strings were computed for each of the m $\times$ n region. These histograms were accumulated every few frames and the vector which contains all concatenated histograms serves as a video descriptor for the video. However, in practice the reliability of the descriptor decreases significantly under large illumination variations [150].

### 2.3.3 Feature Trajectories

Feature trajectories are one of the effective methods for representing video data. Trajectory approaches are recognition approaches that interpret an activity as a set of space-time trajectories [111, 110, 70]. Commonly, trajectories are extracted using Kanade-Lucas-Tomasi (KLT) feature tracker [111, 110]. KLT tracker [154] tracks windows of pixels and identifies windows that contain sufficient texture. Action recognition uses the velocity history of the tracked keypoints or matching SIFT descriptor [147] between two frames.

For more encoded information from video data, researchers proposed to use dense trajectories to describe the features [155, 70, 157, 125]. Wang *et al.* [155] introduced a dense trajectory descriptor represented by tracking interest points (Figure 2.4). Interest points are sampled at spatial-temporal uniform intervals. Tracking is based on displacement information from a dense optical flow field. Based on the work of dense trajectories [155] Jiang *et al.* [70] proposed a method to represent the object relationships by encoding pairwise dense trajectory codewords. Another work improves the trajectories by using dense optical flow to estimate human motion [157].

Because the efficiency of storage and the speed of classification are limited due to the dense samples in the feature space [70], researchers introduced improved dense trajectories by reducing the dimensions and adopting a fast method for classification

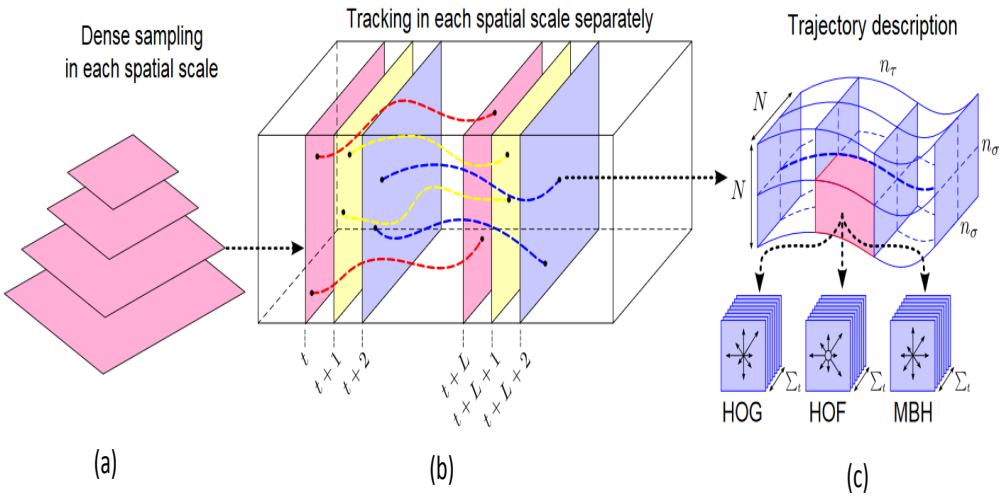



**Figure 2.4: Illustration of dense trajectory description. (a): Feature points are sampled densely for multiple spatial scales. (b): Tracking is performed in the corresponding spatial scale over L frames. (c): Trajectory descriptors are based on its shape represented by relative point coordinates as well as appearance and motion information over local neighbourhood pixels along the trajectory [155].**

such as [170]. In this work, PCA was used to reduce the number of features.

Moreover, Svebor *et al.* [74] treated the human action problem as two steps. The first step is video frame segment extraction and the second step is video frame tracking. The tracking is based on the motion and colour channels. In the second step, every segment is tracked separately both forward and backward in time in the video sequence based on its motion and colour. As a result, the space-time segment is the set of bounding boxes obtained from the tracking process.

Trajectory-based methods have their advantages but also face challenges to cope with self-occlusions, change of appearance, and problems of reinitialisation.

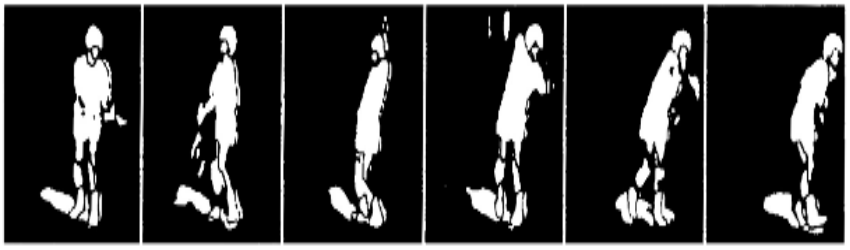


**Figure 2.5: Shape masks for recognising tennis actions [72]**

## 2.4 Feature Representation: Global Features

Methods based on global representations, also called holistic 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 learnt from raw frames in video sequences. Global representations have recently drawn increasing attention [29, 71, 53, 26], because they are able to encode more visual information by preserving spatial and temporal structures of actions occurring in a video sequence. Compared to local methods, global representations encode extracted features as a whole, and are obtained in a top-down manner. Therefore, global descriptors are usually less time consuming to calculate and easier to implement. They give robust results in less challenging scenarios such as those with static background.

### 2.4.1 Shape-Appearance Features

Modelling of human pose and shape has received a great attention from researchers in recent years. Several approaches for action recognition used human shape masks and silhouette information to represent the human body and its dynamics. It is known that action recognition methods based on the human silhouette play an effective role in human action recognition. The shape analysis approaches aim to describe and locate the

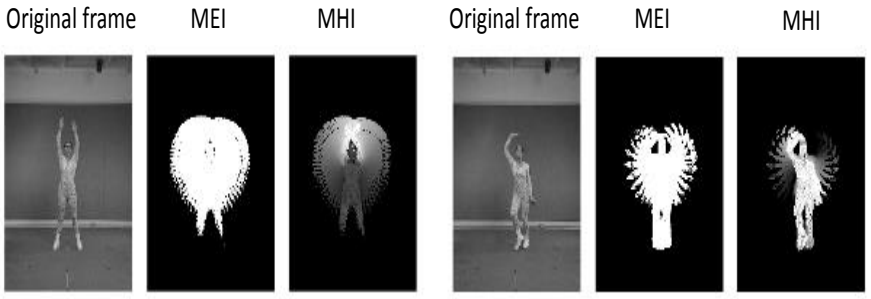



**Figure 2.6: Shape masks from difference images for computing motion history images (MHI) and motion energy images (MEI) [13].**

changes in the human body shape. Shape-based approaches convert video frames into static shape patterns and in the recognition phase compare the patterns with pre-stored ones. In earlier work in this field, Yamato *et al.* [72] are among the first to propose silhouette images (Figure 2.5). They computed a grid representation over the silhouette and computed for each cell the ratio of foreground to background pixels. The grid representations are quantised into a vocabulary, and actions are then learnt as sequences of words using a Hidden Markov Model (HMM). Chomat *et al.* [29] created motion templates and a Bayes classifier was used to perform action recognition. Bobick and Davis [13] used shape masks from difference images to detect human actions. They employed so-called motion energy images (MEI) and motion history images (MHI) as the action representation, as illustrated in Figure 2.6. More precisely, MEIs are binary masks that indicate regions of motion, and MHIs weight these regions according to the point in time when they occur (the higher the weight is the more recent). This approach is the first to introduce the idea of temporal templates for action recognition.

In recent work on shape approaches, Gorelick *et al.* [52] proposed a method to form a 3D spatial-temporal representation by stacking segmented silhouette frame-by-frame. Yang *et al.* [174] treated human pose as latent information and used it to assist the task of action recognition. They represented the action as a model that integrates action recognition and pose estimation. In [71] action interest regions are first localised and

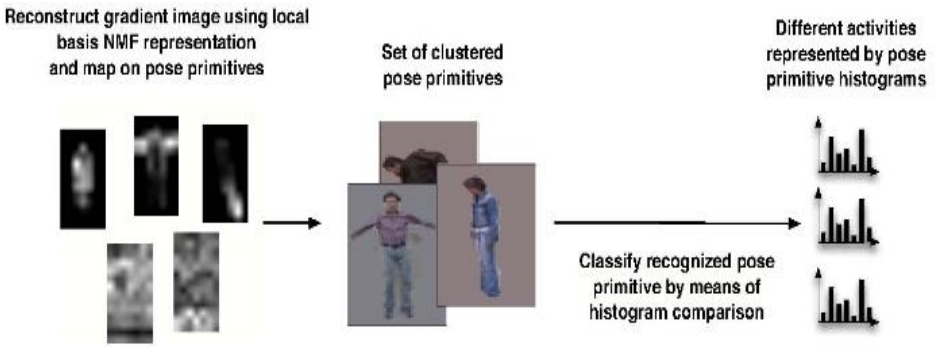



**Figure 2.7: Action representation using histograms of pose primitives [153]**

shape-motion descriptors are computed from them.

For shape descriptors, the histogram of oriented gradient (HOG) [31] was used to encode the shape of each subregion, and then all the histograms were concatenated to form a raw shape feature vector. These features combined with an optical flow descriptor [37] formed the final representation for actions. In other work [153], action classes were represented by histograms of pose primitives using HOG to classify actions. They extended a standard HOG based pose descriptor to better deal with background clutter and articulated poses by exploiting a non-Negative Matrix Factorisation (NMF) basis representation of gradient histograms as shown in Figure 2.7. Ikizler and Duygulu [62] modelled the human body as a sequence of oriented rectangular patches. The authors encoded the video features using BoVW, which they called as bag-of-rectangles. Maji *et al.* [108] introduced a new representation of human pose called "poselet activation vector". The action was represented by estimating the 3D pose of the head and torso, given the bounding box of the person in the image.

## 2.5 Feature Representation: Motion Features

Human-centric approaches based on optical flows and generic shape information form another sub-class of global methods. A number of research works depend on an optical flow technique to transform the motion information from input video to feature vectors. As one of the first works in this direction, Polana and Nelson [129] proposed a human tracking framework along with an action representation using spatio-temporal grids of optical flow magnitudes. The action descriptor is computed for periodic motion patterns. By matching against reference motion templates of known periodic actions (e.g., walking, running, swimming) the final action can be determined.

In another approach purely based on optical flow, Efros *et al.* [37] tracked actions in videos and computed a descriptor on the stabilised tracks using blurred optical flow. Their descriptor separated x and y flows as well as positive and negative components into four different channels, as shown in Figure 2.8. For classification, a test sequence is frame-wisely aligned to a database of stored, annotated actions. The same human-centric representation based on optical flow and human tracks for action recognition was employed by Fathi and Mori [43]. As a classification framework, the authors use a two-layer AdaBoost variant. In the first step, intermediate features are learnt by selecting discriminative pixel flow values in small spatio-temporal blocks. The final classifier is then learnt from all previously aggregated intermediate features.

Rodriguez *et al.* [137] proposed a method to use flow features in a template matching framework. The features are represented by spatio-temporal regularity flow information. The regularity flow shows improvement over optical flow since it globally minimises the overall sum of gradients in the frame sequence. Rodriguez *et al.*'s method [137] learnt cuboid templates by aligning training samples via correlation. For classification, test sequences are correlated with the learnt template using a generalised Fourier transform that allows for vectorial values.

Ali [7] introduced features depending on pure optical flow [59]. These features are

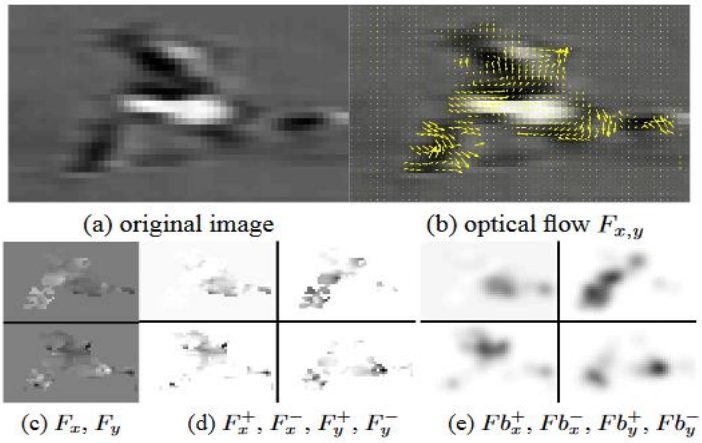



**Figure 2.8: Constructing the motion descriptor based on optical flow [37]. (a) Original video frame, (b) Optical flow $F_{x,y}$, (c) Separating the $x$ and $y$ components of optical flow vectors, (d) Half-wave rectification of each component to produce 4 separate channels, (e) Final blurred motion channels.**

called kinematic features. Each kinematic feature produced three-dimensional information by computing optical flow of a sequence of images to increase spatiotemporal patterns. To reduce the dimensions into a more manageable two-dimensional form, they assumed that the representative dynamics of the optical flow were captured by these spatiotemporal patterns in the form of dominant kinematic trends or kinematic modes. Kinematic features are used to extract different aspects of motion dynamics existing in optical flow which are computed by performing Principal Component Analysis (PCA) on kinematic features to reduce the dimensionality of the features. These features include divergence, vorticity and symmetry. As a result they capture most of the dynamic information and achieved increased recognition. The weakness of kinematic features is that they are view dependent and therefore give different optical flows from different view directions even on the same scene.

The features extracted directly from optical flow are inaccurate [59] because they are affected by noise and illumination environment changes. To improve the features which are selected from optical flow in [132, 141], Ramadass [132] proposed an im-

provement on the optical flow algorithm to increase the features which can be taken from optical flow. The proposed algorithm eliminated irrelevant features by computing Euclidean distance of separation of various features and correspondingly filtering useful low level features for extraction. Despite the fact that good results were achieved by motion descriptors, the methods based on optical flow have limitations due to the difficulty in reliable optical flow estimation, e.g. aperture problems, smooth surfaces, and discontinuities.

## 2.6 Discussions about Feature Representations

As explained in the previous section, different schemes of local and global representations have been proposed to improve the recognition accuracy. However, existing methods still suffer from the limitations for each representation. To overcome these difficulties and benefit from the advantages of different representations, researchers proposed to combine local and global representations to produce a more reliable description for video contents such as [148, 67, 87, 71]. A key advantage of local feature based approaches is their flexibility with respect to the type of video data. Local descriptors represent a video as features extracted from a collection of patches, ideally invariant to environmental clutter, appearance change and occlusion, and possibly to rotation and scale change as well. Global descriptors, on the other hand, treat each video frame as a whole, which is easier to implement and has lower computational costs. Combining features can take the advantages of individual features and provide a trade-off between performance and effectiveness.

## 2.7 Human Action Representation Methods

The common human action representation methods are feature representation, Bag of Visual Words (BoVW), Stochastic-based, Convolution Neural Network, and Graph-

based methods. In brief words, BoVW methods recognise human action by applying a clustering algorithm on feature descriptors to build visual vocabulary [89, 166, 180, 120, 121, 156, 172]. Stochastic-based methods build statistical models to represent human actions (e.g. Hidden Markov Models (HMM)) [136, 124, 83]). Graph-based methods represent an action as a graph [165, 95] to obtain a model used in the classification process. In the following sections we will explain the categorisation of human action systems based on these methods.

### 2.7.1 Bag-of-Visual-Words (BoVW)

A popular representation, based on local features, is the Bag-of-Visual-Words (BoVW) model. It starts from document retrieval applications where orderless strategies are a popular choice for representing textual data. The bag-of-words model was firstly used to represent text documents as recurrence distributions over words and has been applied extensively in this domain [140]. The framework of local spatio-temporal features with Bag of Visual Words (BoVWs) has gained notable achievements and become one of the most popular approaches in the recent work of action recognition [166, 180, 104, 131] and showed a remarkable performance improvement on benchmark datasets.

Generally, a feature descriptor is a vector representation of the features for the local neighbourhood of a given position. To obtain the final representation of an action, the BoVW model is used which is based on mapping local features of each video sequence onto a pre-learnt dictionary. The visual vocabulary (or codebook) is computed by applying a clustering algorithm (e.g. k-means) on feature descriptors obtained from training sequences; each cluster is referred to as a visual word. Descriptors are quantised by assignment to their closest visual words, and video sequences are represented as a histogram of visual word occurrences [166, 180]. The coefficient of each local feature is determined by assigning this feature $x_i$ to its nearest codeword in the codebook vocabulary using a certain distance metric. By using the Euclidean distance, then:

$$u_{i,j} = \begin{cases} 1 & \text{if } j = \arg\min_{j=1,\ldots,M} \|x_i - b_j\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Niebles and Li [121] represented video as spatio-temporal features with bag of visual words. They extracted the interest points and clustered the features, and then modelled actions by using a probabilistic Latent Semantic Analysis (pLSA) to localise and categorise human actions. Laptev and Lindeberg [85] recognised actions based on interest-point features. They first detected interest points using a Hessian detector, and then described the features using scale-invariant spatio-temporal descriptors. Finally, they clustered and recognised the actions based on the similarity of words inside the clusters and the differences among clusters.

Wang *et al.* [156] recognised the action using BoVW framework with an SVM classifier. They represented the video by a combination of several descriptors, which are HOG to describe the appearance, HOF (motion) and trajectories to describe the shape. Moreover, they introduced a descriptor based on motion boundary histograms (MBHs) which relies on differential optical flow. Schuldt *et al.* [142], Dollar *et al.* [34], and Niebles *et al.* [120] proposed using of BoVW in action recognition. For the BoVW representation in videos, feature detectors determine a set of salient positions present in the video sequences.

A non-linear SVM is a popular classifier that is used in different works, e.g. Schuldt *et al.* [142], Dollar *et al.* [34], Laptev *et al.* [86], Willems *et al.* [164], Le *et al.* [90] used non-linear SVMs on a benchmark with different feature descriptors. Such histogram representations have the ability to capture global statistics about the type of descriptors that are present in the video sequence.

### 2.7.2 Stochastic Approaches

There has been a focus on actions in video sequences, where the action can be represented as statistically predictable sequences of states, also called a state model. Stochastic approaches are the methods that represent a human action as a model containing a set of states. The models are statistically trained on feature vectors to generate a general statistical model for action classification. In other words, the statistical model is designed to generate a sequence with certain probability. Existing research has conceived and used many stochastic techniques, such as Hidden Markov Model (HMMs) [116, 124, 136, 83, 44].

In HMM-based methods, a human action is assumed to be in one state at each time frame, and each state generates an observation (i.e., a feature vector). In the next frame, the system transits to another state based on the transition probability between the states. Once transition and observation probabilities are trained for the models, actions are commonly recognised by solving the evaluation problem. The evaluation problem is to calculate the probability of a new input generated by a particular state model. If the calculated probability is high enough, the state model-based approaches are able to decide that the action corresponding to the model occurred in the given input.

Feng and Perona [44] used a static HMM for action recognition where keyposes correspond to states. Lu and Little [101] used a hybrid HMM where one model denotes the closest shape-motion template while the other models position, velocity and scale of the person in the image. Instead of modelling the human body as a single observation, Ikizler and Forsyth [63] introduced 3D trajectories for body parts. They constructed HMMs for the legs and arms individually, where 3D trajectories are the features. For each limb, states of different action models with similar probabilities are linked. This makes training easier, as the combinatorial complexity is reduced to learning dynamical models for each limb individually. However, it leads to the problem of having

to construct a large number of action HMMs, each using a subset of all joints, which results in a large number of weak classifiers.

Several works aimed at improving pose representation by modelling the action and selecting the action class whose corresponding model has the highest probability of generating the observed sequence. Peursum *et al.* [128] used a factored-state hierarchical HMM (FS-HHMM) to jointly model body dynamics per action class. Caillette *et al.* [22] used a variable length Markov model (VLMM) to model observations and 3D poses for each action. Natarajan and Nevatia [117] introduced a hierarchical variable transition HMM (HVT-HMM) which consists of three layers that model composite actions, primitive actions and poses.

A drawback of these models is that they have to make some assumptions in order to be computationally tractable. It can also be hard to learn these models since there are usually many model parameters to be set.

### 2.7.3 Graph-based Approaches

Graph-based approaches have many advantages. They integrate geometrical and topological features of the objects. They are considered as successful shape descriptors for object recognition [139, 163] and matching [173, 35, 20] since graphs are efficient for providing natural description of objects and effective for modeling complex structured data [14, 57]. Based on these advantages, efforts have been undertaken during the last two decades to employ graph characteristics in action recognition. These methods usually differ by the way they construct graphs, features associated with graphs and graph matching methods. Wu *et al.* [165] proposed graph-based action recognition by constructing two graphs to model the action in the video. These graphs are named Video Cooccurrence Graph (VCG) and Video Successiveness Graph (VSG), respectively as shown in Figure 2.9. The vertices in these two graphs correspond to the local features and the edges represent the relationship between the vertices. A family of

Context-Dependent Graph Kernels (CGKs) is used for action recognition.

Wang and Sahbi [161] presented a graph-based action recognition method. They represented an action as a Directed Acyclic Graph (DAG), and used a kernel machine to recognise the action. To construct the graph, dense trajectories are extracted and then clustered using the agglomerative method. The resulted features are mid-level feature components which corresponds to the vertices of the DAG, and the relationship between vertices correspond to the edges of the graph.

Aoun *et al.* [8] introduced an approach for action recognition by constructing a graph based on local features. In their work, the action is modelled by two graph sets: Spatial Video Graph Set (SVGS) and Temporal Video Graph Set (TVGS). The graph-based substructure pattern mining algorithm (gSpan) [171] was then applied to retrieve the spatial and temporal sub-graphs. The histograms of the spatial sub-graphs and temporal sub-graphs are computed. These two types of histograms form the video descriptor, and a bag of sub-graphs method is used to recognise the action in the video sequences.

Recently, Liang *et al.* [95] constructed a model for action recognition in videos with a Spatio-Temporal And-Or Graph (STAOG), which contains four types of nodes: the leaf nodes for representing a batch of local classifiers, the *or*-nodes for specifying an appropriate selection from the leaf-nodes, the *and*-nodes for verifying the holistic appearance of action within the video frame, and the root-nodes for classification and temporal testing. Other recent work [99] represented action as a graph based on Spatio-Temporal Interest Points (STIPs). STIPs are clustered into different labels and each label stands for a kind of movement. Then, all labelled STIPs are defined as nodes of the directed graphs.

Gaur *et al.* [47] modelled the action 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 divided the features into small temporal bins and represented the video as a temporally ordered collection of such feature bins, each bin consisting of a graph-

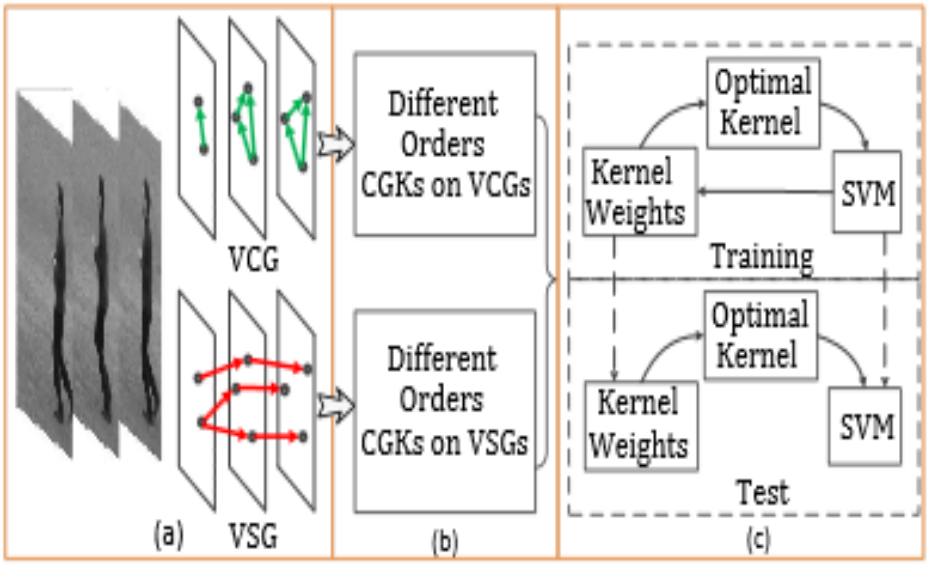



**Figure 2.9: The illustration of the CGKs based human action recognition. (a) A video sequence is represented by VCG and VSG together. (b) Different orders CGKs are computed on both video graphs. (c) Combine the CGKs together using GMKL algorithm and learn action classifiers simultaneously [165].**

ical structure representing the spatial arrangement of low-level features. A video then becomes a string of such graphs and comparing two videos is to match two strings of graphs.

## 2.8 Deep Learning based Approaches

Deep learning has also been used by researchers for detecting and recognising complex events in video sequences. The two main types of techniques in deep learning for action recognition are convolutional neural networks (CNNs) [167, 144, 158, 160, 61, 158] and recurrent neural network (RNN) [168, 118].

Deep learning using convolutional neural networks (CNNs) was introduced by Yann *et al.* [91] in computer vision applications. Convolutional Neural Networks (CNNs) have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval [42, 30, 49]. Motivated by this success of CNNs in image processing applications, researchers are working intensely towards developing CNNs for video processing. The first attempt to use CNNs in human action recognition was introduced by Taylor [152], who introduced a model that learns feature map representations of image sequences from pairs of successive images (Figure 2.10). Baccouche *et al.* [9] proposed to capture the nature of video data based on 3D Convolutional Neural Networks. The network was trained to assign a vector of spatio-temporal features to a small number of consecutive frames (see Figure 2.11). Simonyan [144] introduced an architecture based on spatial and temporal streams which were then combined by fusion. The spatial stream performed action recognition from video frames, whilst the temporal stream was trained to recognise action from motion using dense optical flow (see Figure 2.12).

The second common type of deep learning neural networks is Recurrent Neural Networks (RNNs). RNNs are a class of Neural Networks specialised in sequential processing. While in Feed-Forward Neural Networks the inputs and outputs are fixed in size and independent among samples, RNNs' inputs and outputs can be of arbitrary size and depend on previous observations. One of the main issues that emerges when using RNNs is what is known as the vanishing gradient problem [78]. In Feed-Forward Neural Networks, the gradient is propagated backwards to the input of the model, while in RNNs the gradient is back propagated both within the same neuron to previous time steps, and also to the previous layers. While these deep models are effective and produce promising performance on action recognition, typically the models have millions of parameters and the training of such models requires a large amount of training data. Therefore, such techniques may not perform well if the available training data is limited.

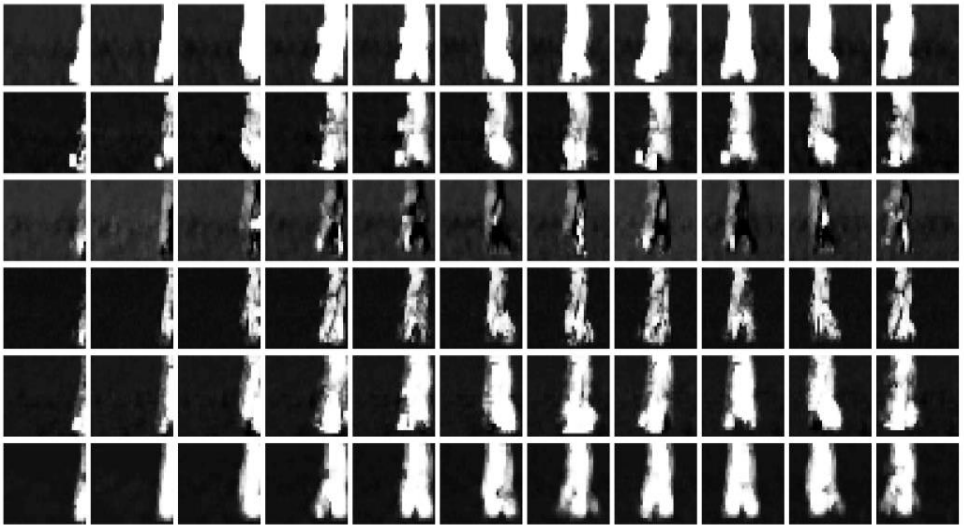


**Figure 2.10: Feature maps inferred from the KTH actions dataset. A subset of 6 (4x4 max-pooled) feature maps 32 in total inferred from sequences of a walking action. Rows correspond to features, and columns correspond to frames [152].**

Unlike images, videos are often much larger in size, which means it is difficult to feed a whole video into deep learning architectures that often have large memory demands. Training a CNN or RNN requires significant computational resources for many iterations. Therefore, researchers try to learn CNNs and RNNs on sampled frames or very short video clips [162]. However, video-level label information can be incomplete or even missing at frame/clip level. This incomplete information leads to the problem of false label assignment.

## 2.9 Video Datasets of Action Recognition

The first step in developing a human action recognition system using machine learning is to acquire an adequate human action database. The dataset should be sufficiently rich in a variety of human actions. Moreover, the creation of such a dataset should

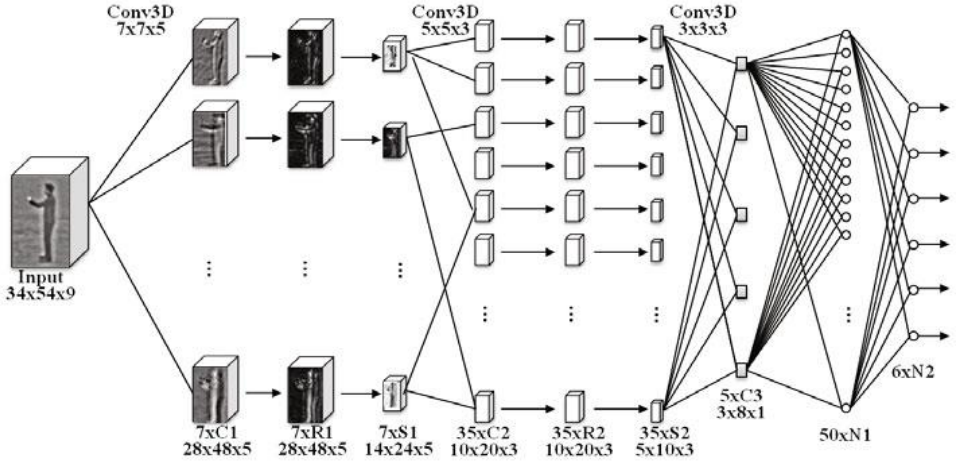



**Figure 2.11: A 3D-Convolutional Network architecture for spatio-temporal feature construction for human action recognition [9].**

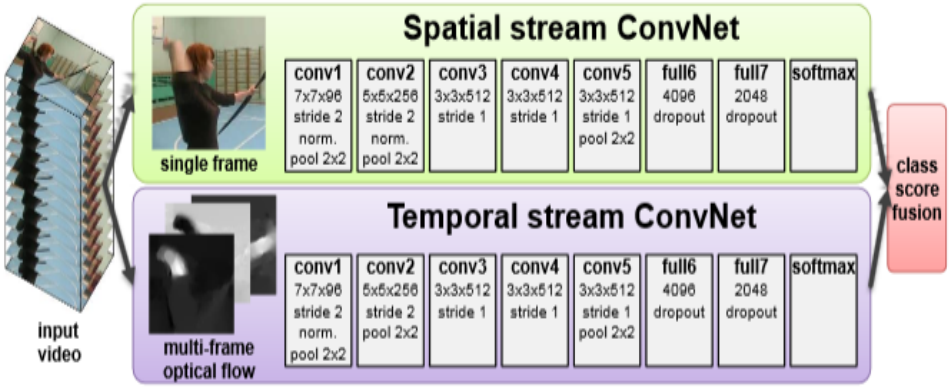



**Figure 2.12: A Two-stream architecture for video action recognition [144].**

correspond to real world scenarios. The quality of the input media that forms the dataset is one of the most important aspects one should take into account. Based on this, researchers introduced different video datasets.

As shown in Figure 2.13, an ideal human action dataset should address the following issues to suit the needs of the target application: (i) the input media should include either still images and/or video sequences, (ii) the amount of data should be sufficient, (iii) input media quality (resolution, grayscale or colour), (iv) number of subjects performing an action, (v) sufficient number of action classes, (vi) sufficient changes in illumination, (vii) large intra-class variations as needed (e.g., variations in subjects' poses), (viii) variations in recording conditions. Based on these, many video datasets have been introduced to address these issues. Generally, video datasets can be classified based on the type of the problem to: controlled action analysis (simple and static background) such as KTH [142] and Weizmann [179] datasets, real-world action analysis (complex and static background) such as UCF-Sports [137], Olympic sports [119], HMDB51 [80] and UCF101 [146], and interaction analysis (real world videos involving interaction) wherevideos were collected from recordings and TV shows such as TVHI [126] and Hollywood [86]. The following survey papers [36, 25] provide extensive discussions about datasets. Based on these assumptions, we evaluate our methods on a range of action recognition benchmark datasets mainly KTH, the UCF-Sports, TV Human Interaction (TVHI), Olympic sports and UCF11 which address these issues. In the following subsections we present the characteristics for each dataset.

### 2.9.1 KTH dataset

The KTH dataset [142] consists of 6 actions *(Boxing, Handclapping, Handwaving, Jogging, Walking and Running*, see Figure 2.14) which were recorded under controlled settings with approximately static motion cameras, clear environment and different scenarios (outdoors, outdoors with different scales, outdoors with different clothes and indoors). Each action was performed by 25 persons and each person was recorded to perform the same action 4 times. The whole video dataset contains 600 video clips with length ranging from 100-700 frames. The standard test setup was provided (training/test separation) to allow fair comparison between different methods. Although this data was recorded under controlled environment, it still remains popular for hu-


**Figure 2.13: An ideal human action recognition dataset**

man action classification, as it provids a good evaluation criterion for many new methods [9, 180, 99].

### 2.9.2 The UCF-Sports dataset

The UCF-Sports dataset [137] contains 10 sport actions (*Diving, Golf swinging, Kicking, Lifting, Horseback riding, Running, Skating, Swinging, Walking*), as shown in Figure 2.15. The video clips in the UCF-Sports action dataset were collected from various broadcast sports channels (e.g. BBC and ESPN), in total composed of 150 videos. The UCF-Sports dataset has large intra-class variation with real world recording environment settings. The standard test setup was provided (leave-one-out testing) to allow fair comparison with different methods.

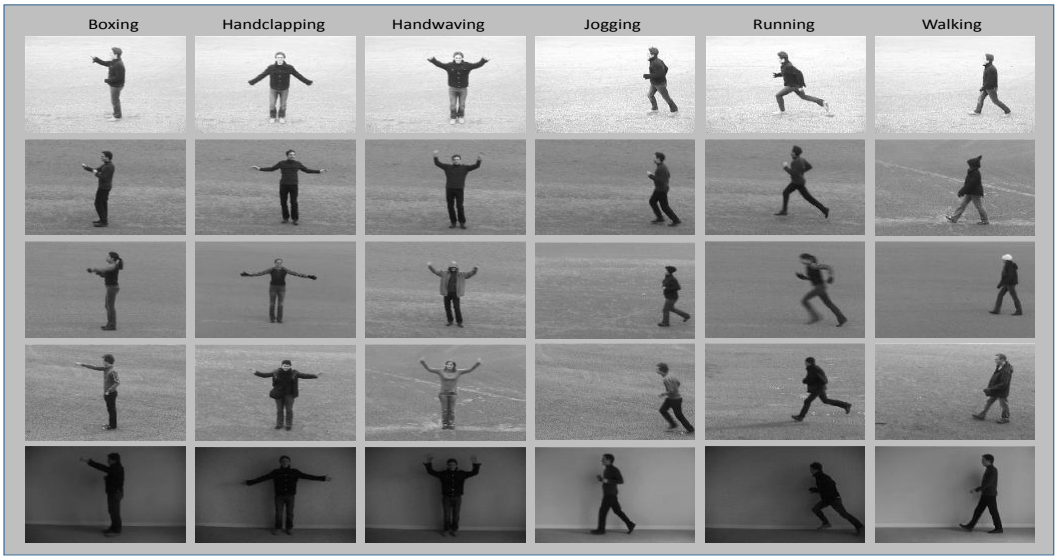



**Figure 2.14: KTH dataset: consists of 6 actions *(Boxing, Handclapping, Handwaving, Jogging, Walking and Running)*.**


**Figure 2.15: UCF dataset: contains 10 sport actions (*Diving, Golf swinging, Kicking, Lifting, Horseback riding, Running, Skating, Swinging, Walking*).**


**Figure 2.16: TV Human interaction dataset: includes 5 action classes (*Handshake, Highfive, Hug, Kiss, and Negative)* where Negative action does not contain any interaction.**

### 2.9.3 TV Human Interaction dataset

The TV-Human Interaction dataset [126] was collected from different TV shows. It includes 300 videos classified into 5 action classes (*Handshake, Highfive, Hug, Kiss, and Negative*, Figure 2.16) where Negative action does not contain any interaction. Two hundred of the clips contain one of the four interaction actions each action appearing in 50 videos. *Negative* examples make up the remaining 100 videos. The length of the video clips ranges from 30 to 600 frames. There is a great degree of variation between different clips and also in several cases within the same clip. The variation consists of the number of actors in each scene, their scales, and the camera angle, including abrupt viewpoint changes at shot boundaries. The dataset is split for training/testing evenly into two groups, each containing videos of mutually exclusive TV shows. Each group contains 25 video clips of each interaction and 50 negative clips.

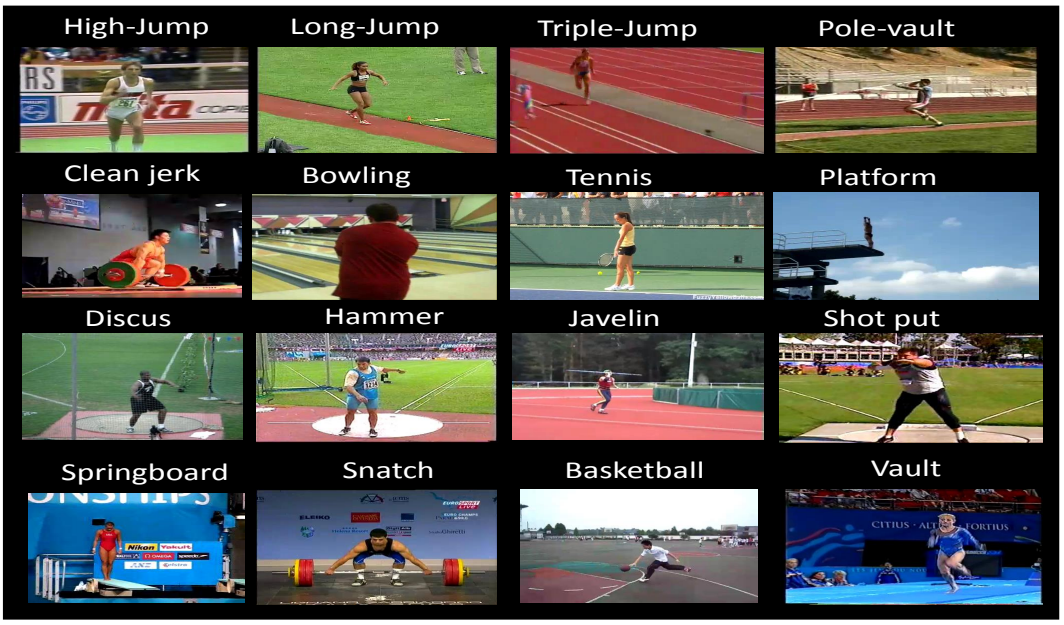



**Figure 2.17: Olymbic sport dataset: consists of 16 actions *such as high-jump, pole-vault, basketball lay-up, discus*.**

### 2.9.4 Olympic Sports dataset

Olympic Sports [119] was collected from sports videos. It contains significant camera motion, which results in a high degree of variation between video sequences. The dataset consists of athletes practising different sports, which were collected from YouTube and annotated using Amazon Mechanical Turk. There are 16 sports actions (such as high-jump, pole-vault, basketball lay-up and discus as shown in Figure 2.17), represented by a total of 783 video sequences.

### 2.9.5 UCF11 dataset

This action dataset [97] contains 11 categories: basketball shooting (b-shooting) ,


**Figure 2.18: UCF11 dataset: consists of 11 actions *such as jumping, diving, horse riding and swinging*.**

volleyball spiking (v-spiking), trampoline jumping (t-jumping), soccer juggling (s-juggling), horse-back riding (h-riding), cycling, diving, swinging, golf swinging (g-swinging), tennis swinging (t-swinging), and walking with 25 subjects. This dataset is characterised with large variations in camera motion, object appearance, object scale, and large intra-class variation in pose, etc. For each category, video clips are put into 25 groups, each with the same subject and similar background, and each group contains more than 4 video clips. These videos were collected from YouTube.

# *Chapter 3*

# Saliency Guided Local and Global Descriptors for Effective Human Action Recognition

## 3.1 Introduction

Local descriptors represent a video as features extracted from a collection of patches, ideally invariant to environment clutter, appearance change and occlusion, and probably to rotation and scale change as well. Global descriptors, on the other hand, treat each video frame as a whole, which are easier to implement and have lower computational costs. Combining features has been shown to be an effective way to improve action recognition performance.

For human action recognition, although the focus is to recognise the action of the subject in video, existing feature descriptor based methods tend to be affected by the background of the video frames. To address this, this chapter presents a novel framework for human action recognition based on saliency object detection and a new combination of local and global descriptors. We propose to first detect salient objects in video frames and only extract features on such objects. We then propose a simple strategy to

identify and process only those video frames that contain salient objects. Processing salient objects instead of all the frames not only makes the algorithm more efficient, but more importantly also suppresses the interference of background pixels. We combine this approach with a new combination of local and global descriptors, namely 3D SIFT and Histogram of Oriented Optical Flow (HOOF). The resulting *S*aliency *G*uided 3D *S*IFT - *H*OOF (SGSH) feature is used along with a multi-class support vector machine (SVM) classifier for human action recognition. Experiments conducted on the standard KTH, the UCF-Sports, TV Human Interaction (TVHI) and Olympic sports action benchmarks show that our new method outperforms the state-of-the-art competing spatio-temporal feature-based human action recognition methods.

To address the action recognition problem, we take a powerful, commonly used Bag of Visual Words (BoVWs) pipeline and focus on the feature extraction step for performance improvement. We propose to extract features on foreground objects identified by *saliency* and use a new combination of *local* and *global* features that provide effective complementary information (see Figure 3.1). Experiments were performed on standard datasets (KTH, the UCF-Sports, TVHI and Olympic sports), which showed that the proposed method outperforms the state-of-the-art features for action recognition. The use of saliency reduces the number of feature descriptors and thus also makes the algorithm faster. More specifically, the major contributions of the proposed method are:

1. Each video frame consists of many interest points, making their descriptions expensive to compute. However, not all the interest points are equally important. We propose to estimate the importance of interest points by salient object detection and only keep those interest points on salient objects for action recognition. This helps to suppress the interference of background and thus makes the method more robust to background fluctuation, while at the same time reduce the running times.

2. We further propose a simple strategy to filter out frames that do not contain

foreground objects, also for the benefit of improved performance and efficiency.

3. We propose a novel combination of local and global descriptors, which has shown good performance in action recognition.

The remaining sections in this chapter are organised as follows. Sections from 3.2 to 3.5 give the details of the proposed approach. The experimental setup and results are discussed in section 3.6. Finally, discussion and conclusions are drawn in section 3.7.

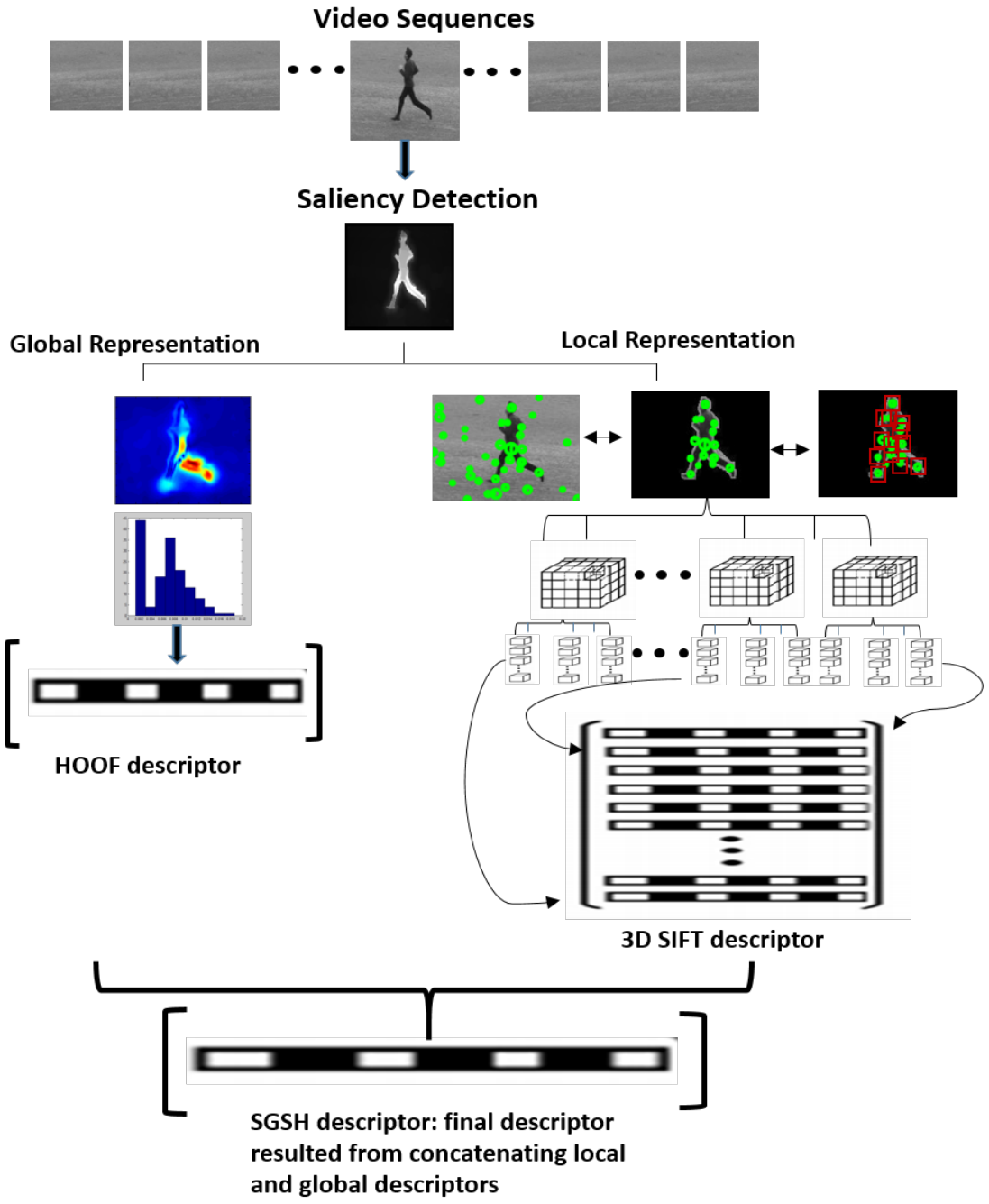



**Figure 3.1: Overview of our novel saliency guided feature extraction pipeline. Given a video sequence, the foreground object pixels are first identified on each frame using a saliency detection method. We then extract a new combination of local and global features guided by saliency, namely 3D SIFT for local features and Histograms of Oriented Optical Flow (HOOF) for global features.**

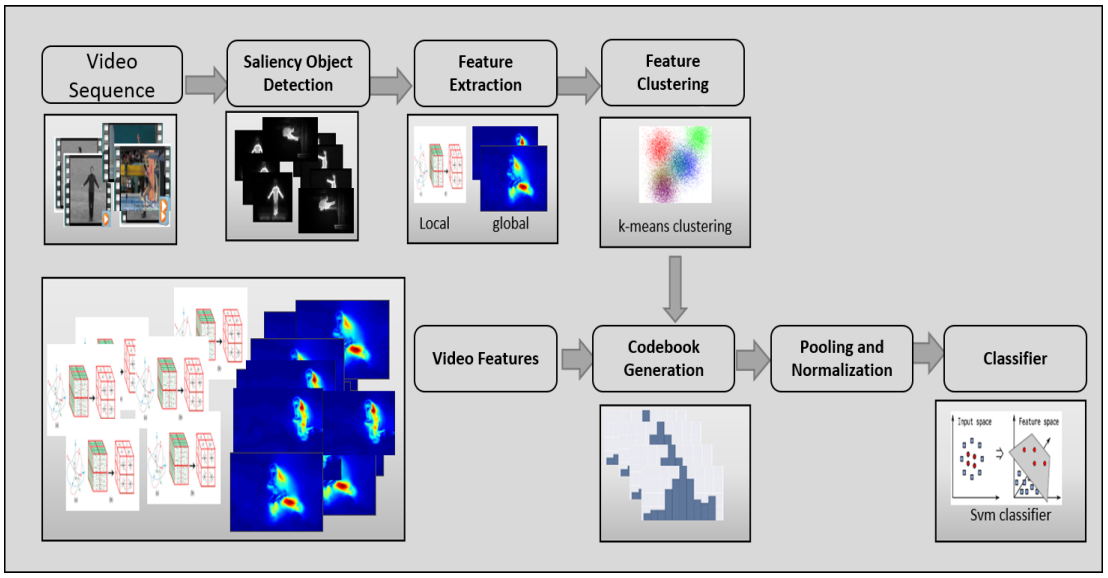



**Figure 3.2: The proposed pipeline of obtaining Bag of Visual Words (BoVWs) representation for action recognition. It mainly contains five steps: (i) saliency guided feature extraction, (ii) feature clustering, (iii) codebook dictionary generation, (iv) pooling and normalisation and (v) classification.**

## 3.2 Proposed Approach

In this section we describe our proposed approach for action recognition. The pipeline of the proposed approach is illustrated in Figure 3.2, which contains the following five main steps. The first step is saliency guided feature extraction, where the salient objects are detected firstly and only interest points on the objects are used. This step involves saliency object detection and feature extraction. With saliency as guidance, local and global features are then extracted to encode video information. In the training step, including feature clustering and codebook generation, features extracted from the training set are clustered to generate visual words. Histograms based on occurrences of visual words in the training set are used as features to train classifiers. Finally, a multi-class SVM classifier is used to achieve action recognition. The following subsections explain each step in details.

## 3.3 Saliency Guided Feature Extraction (SGFE)

### 3.3.1 Detection of Saliency Regions

Analysis and interpretation of image sequences have received a great amount of interest in computer vision for the last few years. Not only because the detection of the most salient region of an image has numerous applications, including object detection [24] and recognition [73], image compression [65], image quality assessment [103], video summarisation [107], and photo collage [50], to name a few, but it can also help to obtain a semantic description of the content of a scene, because we do not need to use all the available information. Therefore, it is not surprising that much work has been done on saliency detection. Different aspects of distinctness have been examined before. Some algorithms look for regions of distinct colour [27, 56]. This is insufficient, as some regions of distinct colour may be non-salient.

Other algorithms [24, 51] detect distinct patterns, such as the boundaries between an object and the background which could lead to missing homogeneous regions of the salient object, whereas [109] combines colour and pattern distinctness to detect salient objects that leads to improve saliency detection results. Figures 3.3, 3.4, 3.5, and 3.6 show the results of applying this algorithm on KTH, the UCF-Sports, TVHI and Olympic sports datasets. From these figures it can be seen that, the algorithm works well with static and dynamic camera motion, where the first set (see Figure 3.3) contains activities with less severe background clutter or motion like boxing, handclapping, and handwaving and second, third and fourth sets (see Figures 3.4 , 3.5, and 3.6) consist of activities with complex background and strong camera motion, clutter, and deformable objects, such as kicking, lifting, skating and horse riding from the UCF-Sport dataset, highfive, hug, and handshake from TVHI dataset, and disc throw and bowling actions from Olympic sports dataset. We also tried alternatives algorithms [94, 66, 169] for saliency detection but they failed to detect the salient regions

precisely, as shown in Figure 3.7.

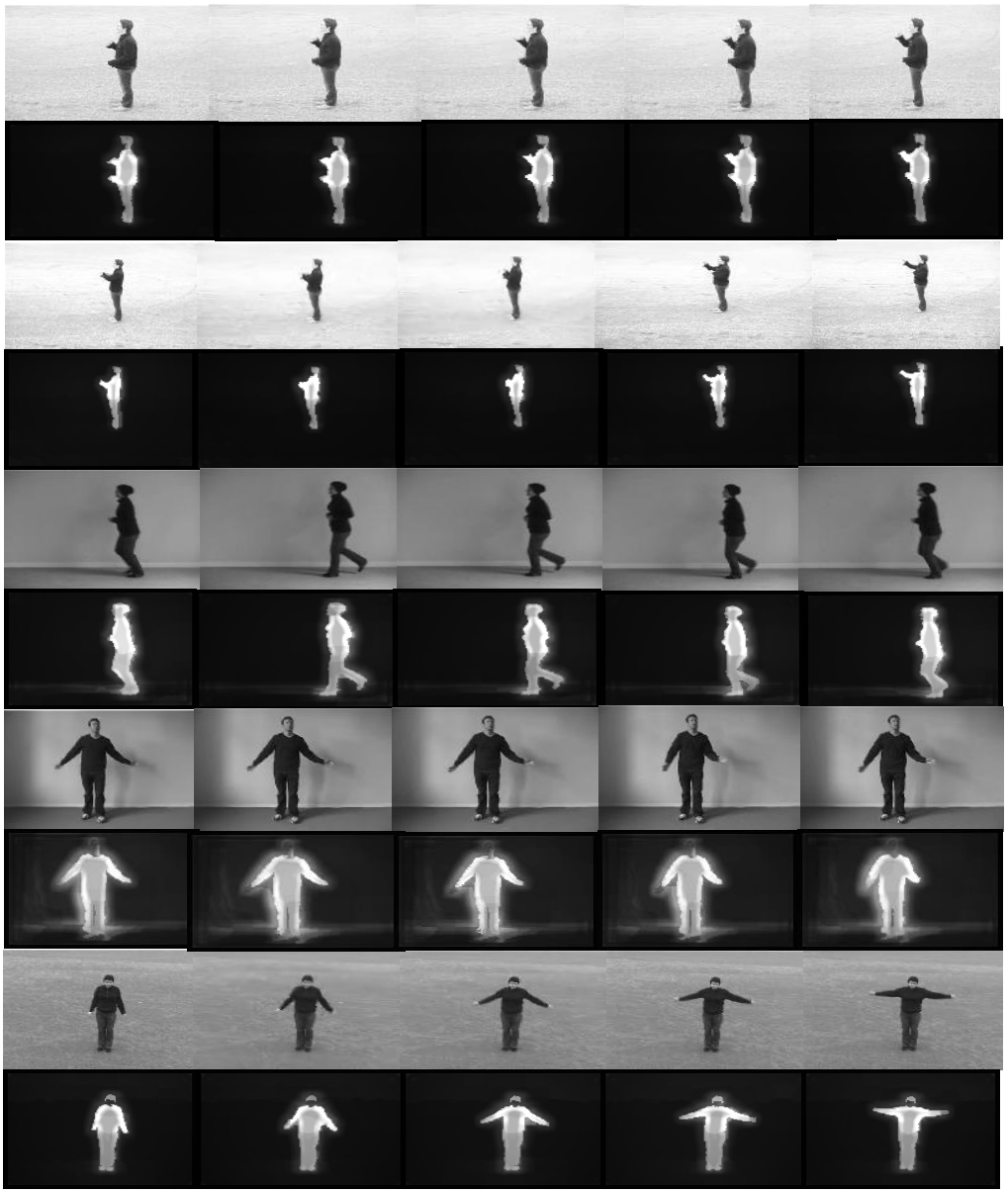


**Figure 3.3: Salient object detection (KTH dataset) using [109]: different actions (e.g. boxing, handclapping, handwaving, and jogging) with different recording environments (indoor, outdoor) and different scales.**

Alternatively, video saliency based methods are based on spatio-temporal mechanism. They detect spatial saliency on single video frames and temporal saliency on inter-

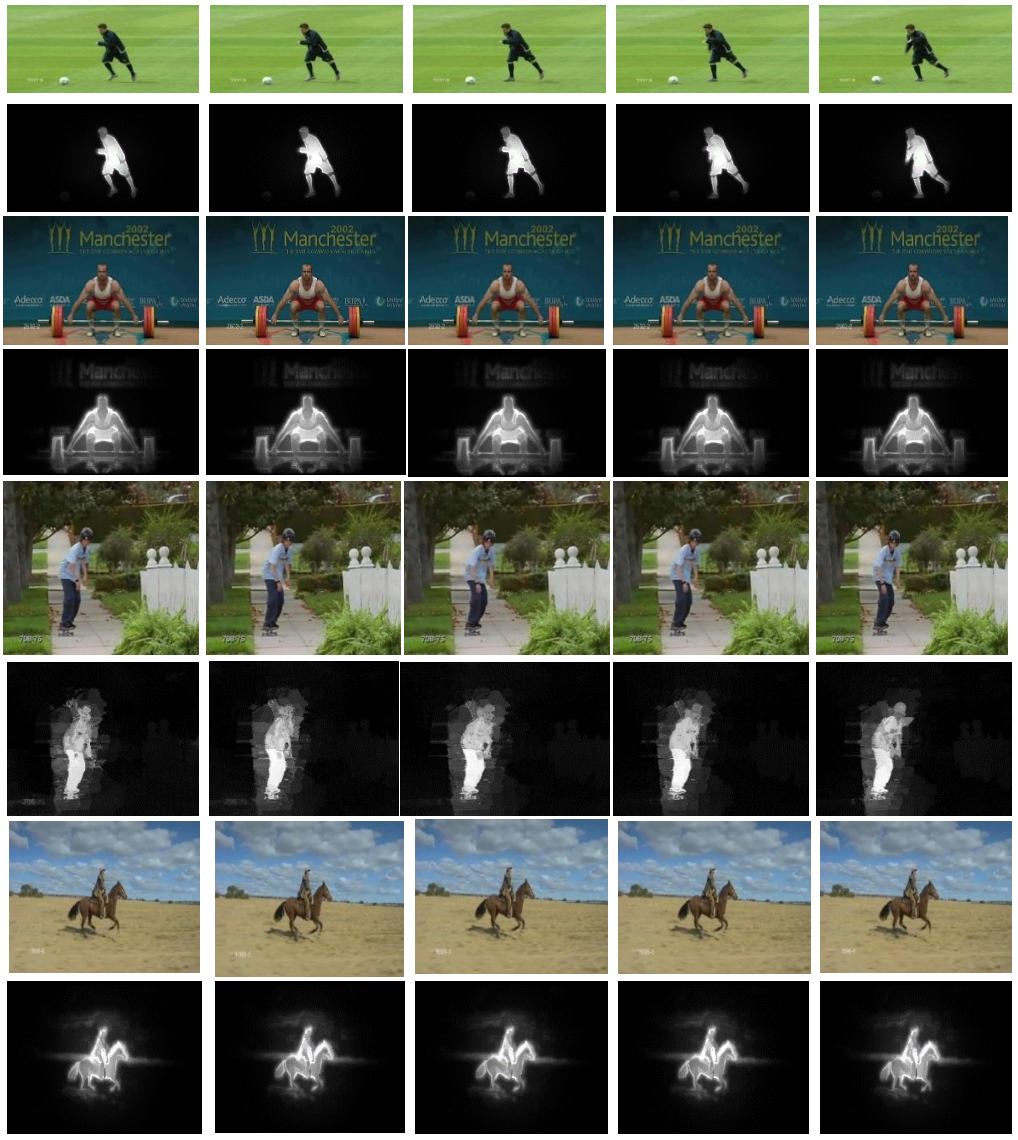



**Figure 3.4: Salient object detection (the UCF-Sport dataset) using [109]: different actions (e.g. kicking, lifting, skating, and horse riding) with real world recording environment.**

frame distinctiveness. The final saliency map is generated by fusing the spatial and temporal saliency maps together and the saliency decision is made by weighting maps. Kim *et al.* [75] introduced a spatio-temporal video saliency detection method. For spatial saliency map detection, edge and colour orientation information is used, while ab-

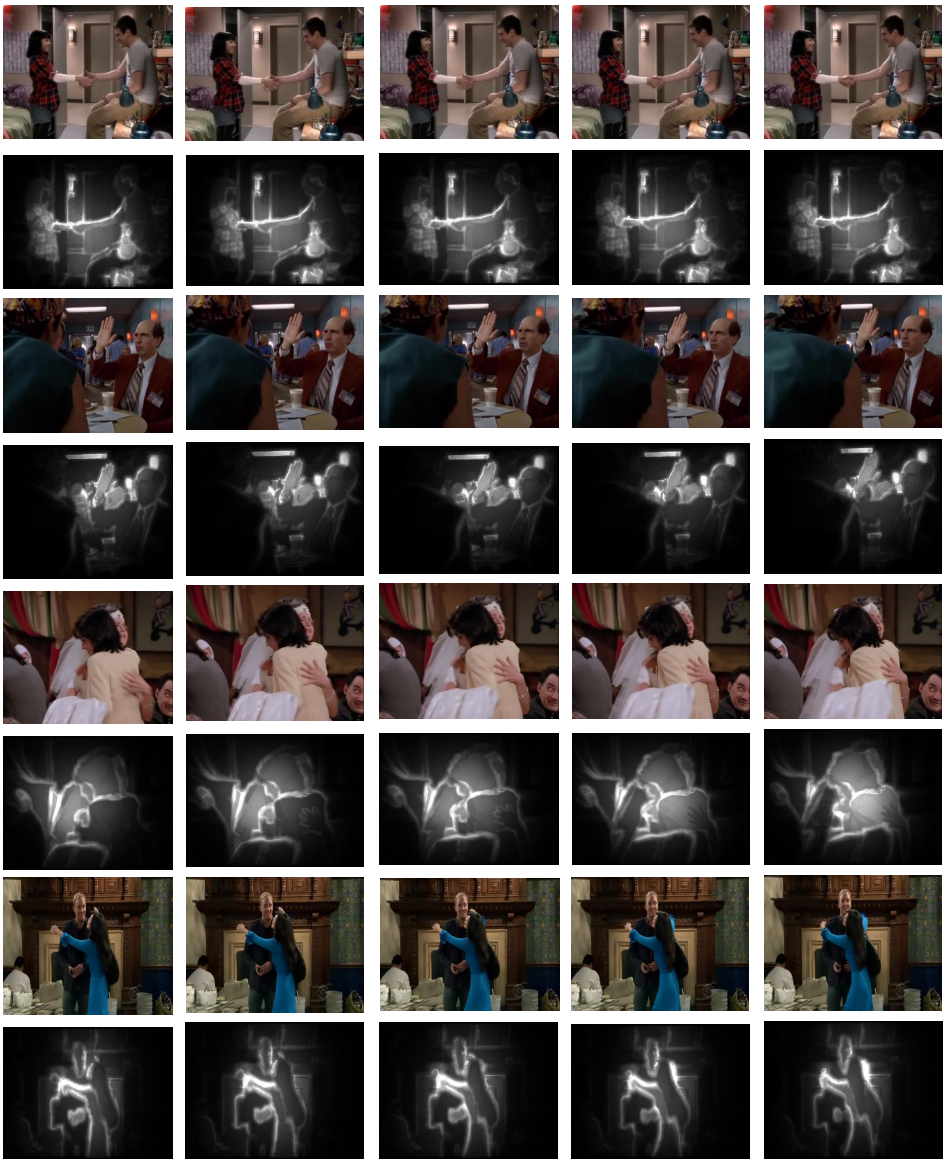


**Figure 3.5: Salient object detection (TV Show Human Interaction TVHI dataset) using [109]: different actions (e.g. handshake, highfive, and hug) with real world recording environment.**

solute inter-frame distinctness information is used for the temporal saliency map. The final saliency map is generated by linearly combining the spatial and temporal saliency maps with fixed weight for each map. In [93] spatial saliency detection was achieved by computing colour information of edge preserving super-pixels, which were extrac-

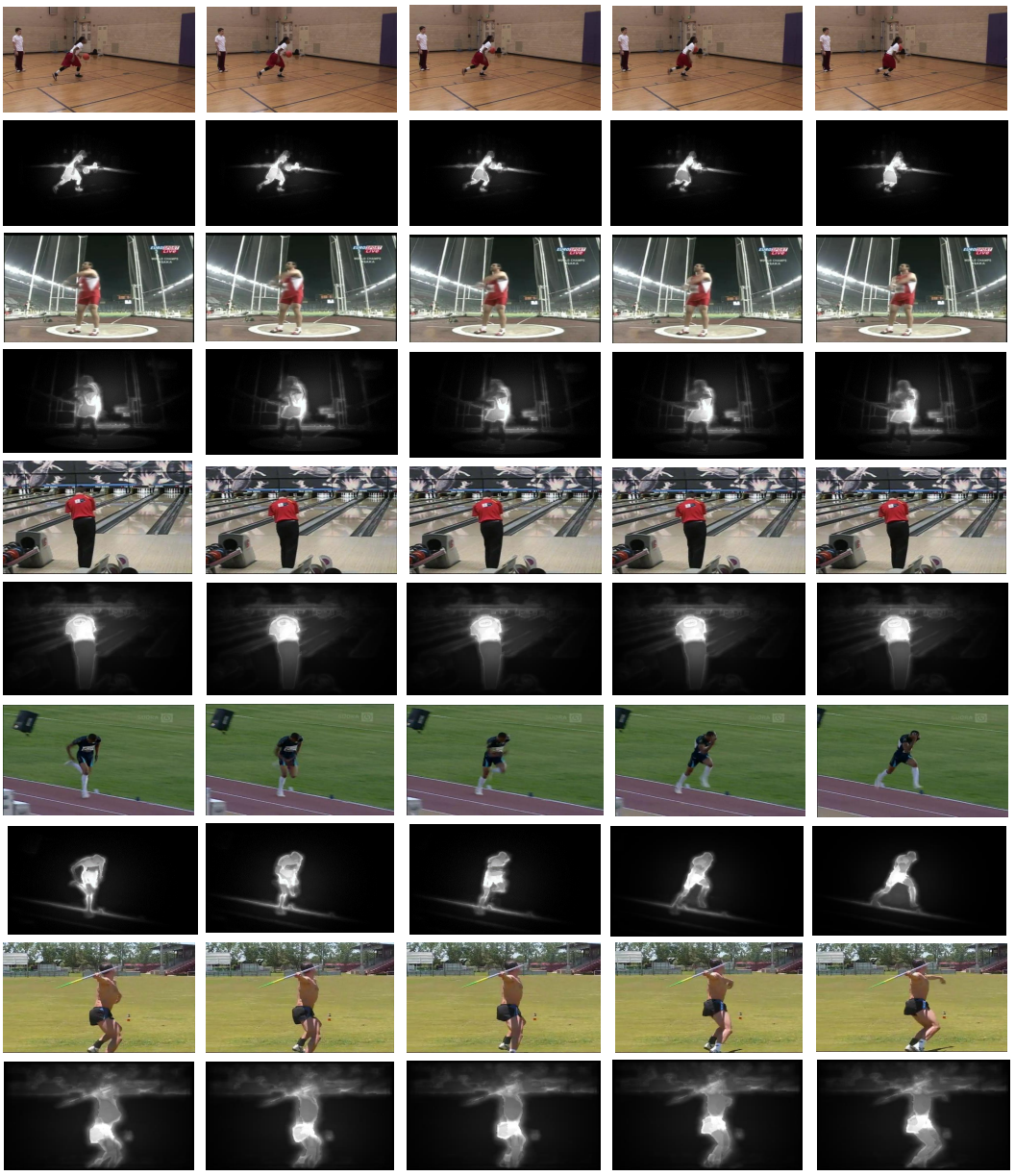


**Figure 3.6: Salient object detection (Olympic sports dataset) using [109]: different actions (e.g. basketball, disc throw, bowling, long jump and javelin throw ) with real world recording environment.**

ted with Turbopixels. For temporal saliency, they used the same mechanism but on dense optical flow information of the video. The spatial and temporal saliency results are then transformed into a conditional random field (CRF) [81] to label each pixel.

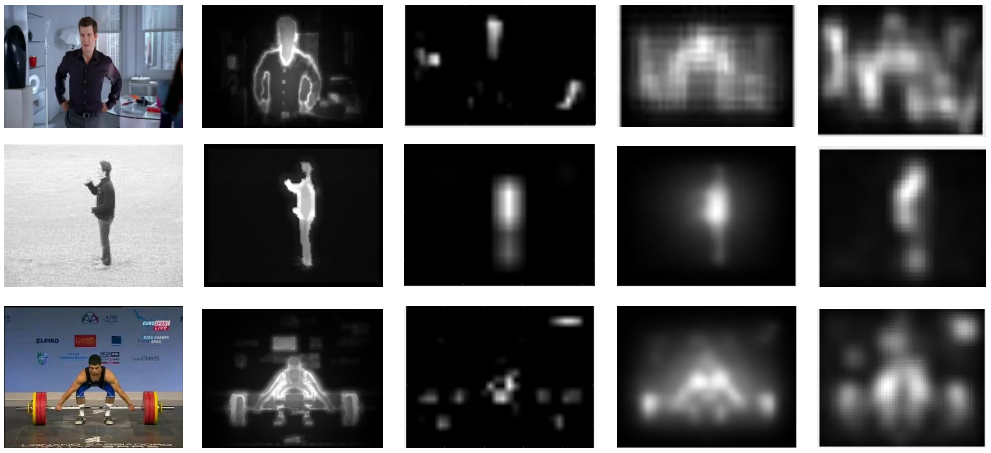



**Figure 3.7: Salient object detection: (a) The original frames from different datasets. (b) Margolin's algorithm [109]. (c) Image signature based on foreground properties [169]. (d) Graph Based Visual Saliency (GBVS) [66]. (e) Hypergraph modelling [94].**

Fang *et al.* [41] measured spatial saliency by extracting intensity, colour, and texture features from Discrete Cosine Transform (DCT) coefficients, then detected temporal saliency using the motion feature in the compressed domain, and designed a new fusion method to obtain the final saliency maps.

Although video saliency methods can also be used, such methods are time-consuming due to the computation redundancy using dense optical flow. Independently computing saliency on every pixel of each frame is redundant, since most neighbouring frames have high similarity. Since the image-based method works sufficiently well, we use it in our methods.

Oikonomopouls *et al.* [123] adapted the idea of saliency region selection in spatial images to the spatio-temporal video space. Salient points are detected by measuring changes in the information content of the set of pixels in cylindrical spatio-temporal

neighbourhoods at different scales. They used a sparse representation of a human action as a set of spatio-temporal salient points that correspond to action-variation peaks to recognise the action. Their method directly uses saliency information as features for action recognition, whereas we use saliency information to guide more general feature descriptors.

### 3.3.2 Description of Saliency Guided Feature Selection (SGFE)

The first step of our pipeline is to detect salient regions in video frames. This provides a fast solution that addresses several key aspects related to action recognition. Firstly, it detects the region of interest (ROI) and attention-grabbing object in a scene. We assume that videos are taken at a reasonable distance and therefore foreground objects do not occupy more than half of the pixels. Secondly, it selects the informative and robust keypoints in the frames. Finally, it reduces the time required to encode the video frame. Saliency detection research has largely been based on images. We use a state-of-the-art image-based algorithm [109] and apply this to each video frame. The main idea of this algorithm is to combine the colour and pattern distinctness. It is inspired by the fact that the neighbouring pixels of each salient object are distinct in both colour and pattern. Colour detection is performed by segmenting a video frame into regions and then determining which region is distinct in colour. The colour distinctness of a region is defined as the sum of $L_2$ distances from all other regions in the colour space.

The pattern distinctness is determined by firstly extracting all $9 \times 9$ patches and computing the average patch. Principal Component Analysis (PCA) is then applied to the collection of patches. After that the pattern distance of a patch is defined as the $L_1$ distance between the patch and the average patch, calculated in the PCA coordinates. Doing so takes not only the difference between a patch and the average patch, but also the distribution of patches into account. Unusual patches based on the distribution will receive a high pattern distinctness. Because objects are more likely to be in the centre of the frame, a Gaussian map surrounding the centre of frame is also generated.

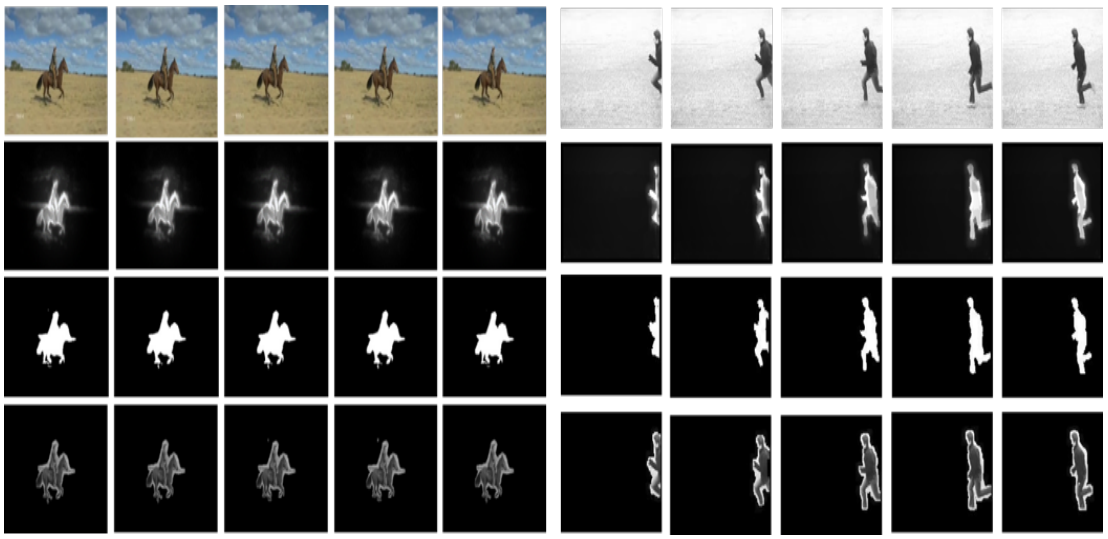


**Figure 3.8: Salient object detection. First row: the original video frames. Second row: the result of saliency detection. Third row: the binary image on the processed frames. Fourth row: foreground objects in video frames. The left five columns contain an example of the UCF-Sports (Horse-riding) and the right five columns contain an example of the KTH dataset (Running).**

The final saliency space map of a single patch $p_x$, $S(p_x)$, is the product of the colour distinctness map, patch distinctness map and the Gaussian map.

After saliency detection, a binary image is generated by thresholding (threshold $0.2$ is used in our experiments) and used as a mask to extract foreground object from the background. Figure 3.8 presents examples on salient object detection for some actions in both KTH and UCF-Sports datasets. Saliency detection works well for both datasets. Note that we have found applying the image based saliency detection technique to individual frames works very well for these datasets, including the UCF-Sports, TVHI and Olympic datasets with complex background. As will be explained later in Section 3.6, histogram-based features are used for classification, which makes the system more robust to inaccuracies of saliency detection in individual frames.

As we will show later, this step improves the performance substantially by selecting only the interest points which are detected on objects and discard others in the

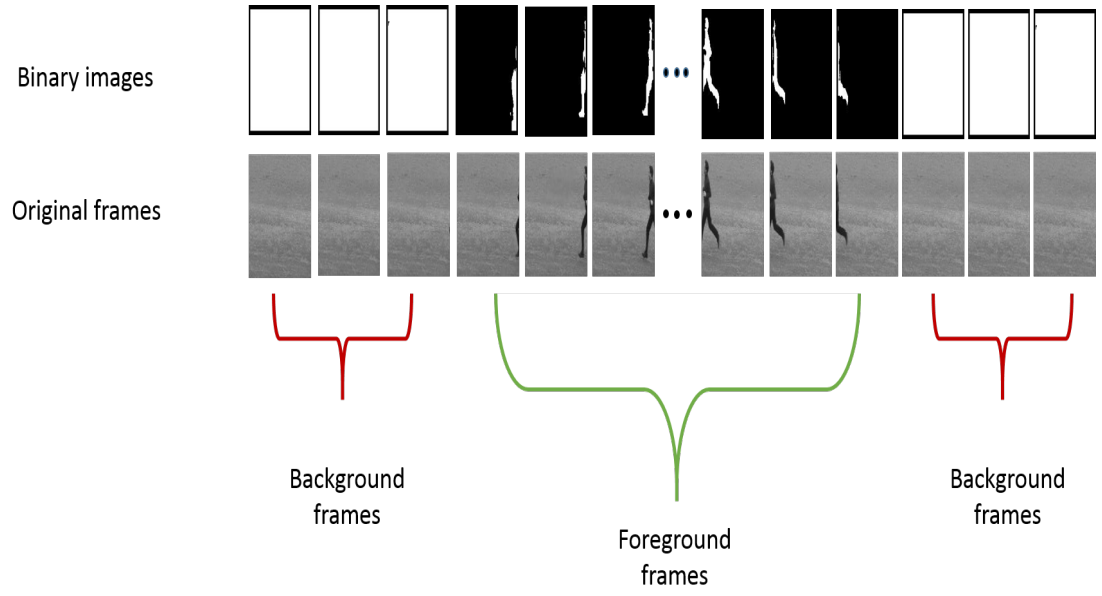



**Figure 3.9: Proposed video frames selection method**

background.

### 3.3.3 Video Frame Selection

We introduce video frame selection method to keep only those video frames containing foreground subjects for further processing. For frames without foreground subjects, the saliency detector tends to classify background areas as salient regions. An example illustrating this is shown in Figure 3.9. Since background usually covers more pixels than the foreground, we select those frames with less than half of the pixels being classified as salient for further processing and discard the remaining frames. This simple heuristic works well for all the datasets tested in this work.

## 3.4 Feature Extraction

We extract two types of descriptors: local and global descriptors represented by 3D-Scale Invariant Feature Transform (3D-SIFT) and Histogram of Oriented Optical Flow

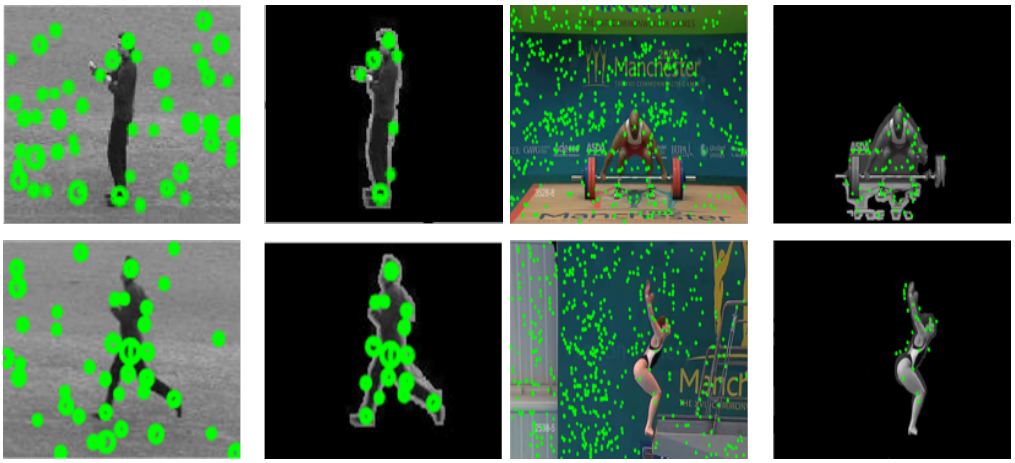


**Figure 3.10: Interest point detection on KTH and UCF-Sports datasets: first and third columns are the original frames, and second and fourth columns are frames with salient object detection.**

(HOOF), respectively as shown in Figure 3.1.

### 3.4.1 Local Features

Local features or interest points provide compact and abstract representations of patterns in a video frame. To encode video data as a local feature, we need to firstly detect interest points in a video frame and describe them effectively to capture video information. Following sections 3.4.2 and 3.4.3 will explain that.

### 3.4.2 Interest Point Detection

In this step, one common approach is to use the Laplacian of Gaussian (LoG) as the response function. We use Lowe's approach to extract interest points [100]. An approximation of the LoG is used based on the difference of the image smoothed at different scales. The scale space of an image is defined as a function $L(x, y, \sigma)$ which is produced from the convolution of a Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.1}$$

We adopt this to detect interest points on video frames. The response function is as follows:

$$D(x, y, \sigma) = (G(x, y, h\sigma) - G(x, y; \sigma)) * I(x, y) = L(x, y, h\sigma) - L(x, y, \sigma) \quad (3.2)$$

where $h = \sqrt{2}$ is a parameter which controls the accuracy of the approximation [100]. We select only the interest points detected on the salient objects. Consequently, we process the most important points in the video frames, which carry robust information of an action. All points detected on the background are discarded. The motivation for this is that the salient interest points are precisely those that maximise the discriminability between actions. Figure 3.10 shows the difference between the interest points detected before and after applying the salient object detection in both datasets *KTH (boxing, running) and UCF-Sports (lifting, diving)*.

### 3.4.3 Local Feature Description

Local representations provide detailed information insensitive to global transformations. Scale Invariant Feature Transform (SIFT) descriptor [100] is one of the most popular local representations due to its invariance to camera movement, robust to noise and scaling. After detecting interest points using image-based SIFT for each video frame, 3D-SIFT descriptor [143, 148] is used to represent local features of interest points, owing to the fact that video frames have a spatio-temporal domain. The 3D SIFT feature is extracted by computing the overall orientation of the neighbourhood centred at an interest point, where the neighbourhood is an $N \times N \times N$ cube. The whole cube will be divided into $M \times M \times M$ multiple sub-cubes. For each sub-cube and each orientation, the orientation histogram with ($b$ bins) is produced. Once this is computed, we can create the sub-histograms which will encode the 3D SIFT descriptor.

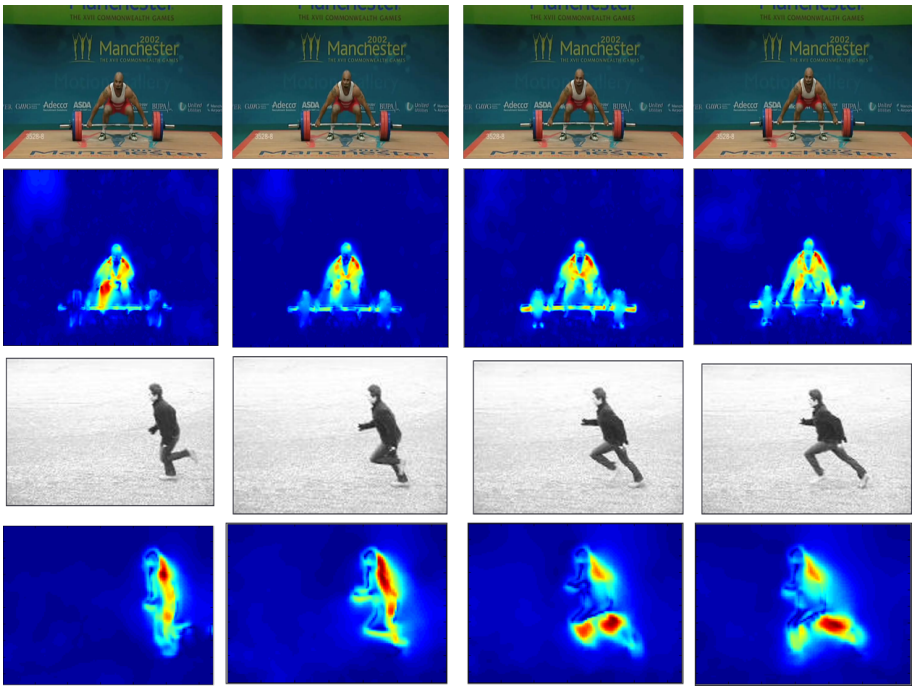


**Figure 3.11: Optical flow calculation by using Brox's method: first two rows for UCF-Sports dataset (Lifting) and last two rows for KTH dataset (Running).**

### 3.4.4 Global Feature

The motion representation as a global descriptor is particularly useful in action representation due to its low computational cost and capability of capturing global motions. In our approach, we describe the motion by the HOOF descriptor [26] for each video frame. For optical flow calculation, we use Brox's method [19] as shown in Figure 3.11. Brox proposed a method to solve problems of discontinuities in the flows field and aperture (the motion direction is ambiguous) [102]. First, optical flow is computed at every frame of the video. Each flow vector is binned according to its primary angle from the horizontal axis and weighted according to its magnitude (see Figure 3.12). Thus, binning according to the primary angle, the smallest signed angle between the horizontal axis and the vector, allows the histogram representation to be independent of the (left or right) direction of motion so the contribution of each optical flow vector to its corresponding bin is proportional to its magnitude. To make

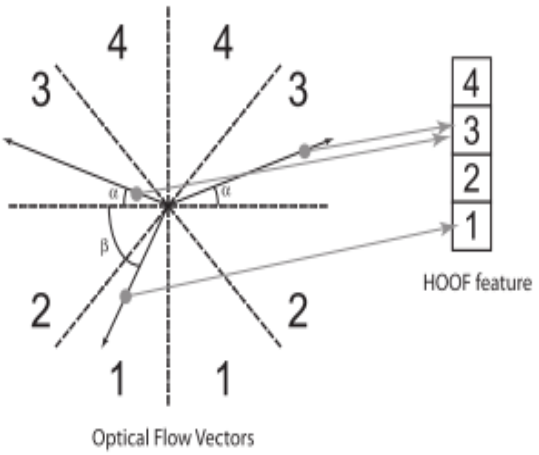



**Figure 3.12: Histgram of Oriented Optical Flow (HOOF) with four bins [26], B = 4.**

the histogram representation scale-invariant, the histogram is normalised to sum up to 1 [26].

## 3.5 Classification

For classification, we use a multi-class support vector machine (SVM) with the Radial Basis Function (RBF) kernel [23]. A bag of visual words approach is used to encode the videos. In clustering we use $k$-means algorithm to generate the vocabulary of visual words. The feature vectors are mapped to closest visual words and a video is then represented as the frequency histogram over the visual words.

## 3.6 Experimental Results

In the following sections, we will show the experimental results and the parameters setup for both local and global descriptors of our approach.

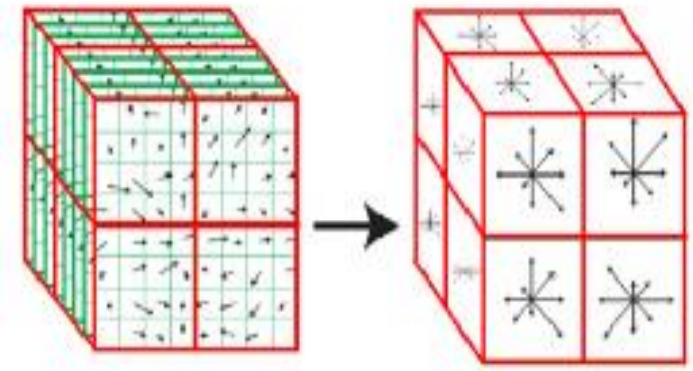


**Figure 3.13: Computation of the 3D SIFT feature descriptor [148]**

### 3.6.1 Parameters Setup for Local and Global Features

The patch size for the SIFT descriptor is a cube of $8 \times 8 \times 8$ and each cube is divided into sub-cubes of size $4 \times 4 \times 4$. Each cube therefore contains 8 sub-cubes. For each sub-cube the orientation histogram with 8 bins is produced. So we have 24 bins for each sub-cube and for the whole cube all these sub-cube histograms are combined to form a 192 (= $24 \times 8$) dimensional feature vector (see Figure 3.13), which is the 3D SIFT feature descriptor. For HOOF descriptor each video frame is represented by a feature vector with 150 bins. In our experiments, vocabularies are constructed with $k$-means clustering with 1000 visual words for 3D-SIFT and 2000 for HOOF. Grid search with $5$-fold cross validation is used to optimise SVM kernel parameters.

### 3.6.2 Experimental Results

Table 3.1 shows our experimental results on the KTH, the UCF-Sports, TVHI and Olympic sports datasets for the cases with and without the saliency guidance (i.e.

**Table 3.1: Action recognition with and without the Saliency Guidance for the combined 3D SIFT and HOOF descriptors (SGSH and SH).**

| Dataset | Descriptor | Accuracy (%) |
|---|---|---|
| KTH | SGSH | **97.2** |
| | SH | 91.2 |
| UCF | SGSH | **90.9** |
| | SH | 85.3 |
| TVHI | SGSH | **70.6** |
| | SH | 65.3 |
| Olympic sports | SGSH | **79.9** |
| | SH | 73.1 |

| | Box | HC | HW | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 0.97 | 0 | 0.03 | 0 | 0 | 0 |
| HC | 0.02 | 0.92 | 0.06 | 0 | 0 | 0 |
| HW | 0 | 0 | 1.00 | 0 | 0 | 0 |
| Jog | 0 | 0 | 0 | 0.94 | 0.03 | 0.03 |
| Run | 0 | 0 | 0 | 0 | 1.00 | 0 |
| Walk | 0 | 0 | 0 | 0 | 0 | 1.00 |

**Figure 3.14: Confusion matrix on the KTH dataset (HC - Handclapping, HW - Handwaving): SGSH .**

*SGSH* and *SH* descriptors) respectively. All the tests were run based on the parameters listed in the above section. From the table, we can see that the SGSH descriptor increases the accuracy by 6% for the KTH dataset, 5.6% for the UCF-Sports dataset, 5.3% for TVHI dataset and 6.8% for Olympic sports dataset. For the KTH dataset, the confusion matrices are shown in Figures 3.14 and 3.15 with and without the saliency guidance. It can be clearly seen that with the SGSH descriptor, the actions *handwaving*, *running* and *walking* in the KTH dataset are recognised fully correctly. Other actions

such as *handclapping* may still be confused, due to the similarity of the movement of the actor between *handclapping* and *handwaving*. With the SH descriptor, confusions exist between these actions including those not confused with SGSH. For example 5% of *walking* actions were recognised as *jogging*. The sensitivity and specificity evaluations for the SGSH descriptor are shown in Table 3.2, where the sensitivity is defined as the proportion of actual positives (relevant actions) correctly identified as such, and specificity means the proportion of actual negatives (irrelevant actions) correctly recognised as irrelevant. From the table we can see that the actions *handwaving*, *running* and *walking* have $100\%$ sensitivity. From the error rate analysis, it can be clearly seen that actions *running* and *walking* have the smallest error rates which means the method is more effective in recognising such actions.

For the UCF sport dataset, the recognition rate for each action is given in the confusion matrices in Figures 3.16 and 3.17), corresponding to SGSH and SH descriptors, respectively. With the saliency guidance the recognition accuracies increase for all the actions in the dataset. The sensitivity and specificity for each action of dataset are shown in Table 3.3. *High bar*, *Horse riding* and *walking* have $100\%$ specificity which means these actions have no false negatives from other actions. For more complex TVHI dataset, SGSH achieves $70.6\%$ accuracy which is significantly better than the state-of-the-art performance of $66.1\%$ reported in the literature [177], while for the Olympic dataset the accuracy increases by $2.6\%$ compared to the best reported performance [18]. Figures 3.18 and 3.19 present the confusion matrices for TVHI and Olympic sports datasets, respectively. The evaluation of sensitivity and specificity is reported in Table 3.4 for TVHI and Table 3.5 for Olympic dataset. For TVHI dataset, all the actions have specificity equal to or greater than $0.8$, while the Olympic dataset, by using SGSH descriptor it can be observed that we have obtained good performance

| | Box | HC | HW | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 0.88 | 0.06 | 0.06 | 0 | 0 | 0 |
| HC | 0.05 | 0.90 | 0.03 | 0 | 0.02 | 0 |
| HW | 0.02 | 0.05 | 0.93 | 0 | 0 | 0 |
| Jog | 0 | 0 | 0 | 0.92 | 0.06 | 0.02 |
| Run | 0 | 0 | 0 | 0 | 0.86 | 0.14 |
| Walk | 0 | 0 | 0 | 0.06 | 0 | 0.94 |

**Figure 3.15: Confusion matrix on the KTH dataset (HC - Handclapping, HW - Handwaving): SH.**

| | Di | Go | HB | Ki | Lf | HR | Ru | Sk | Sw | Wa |
|---|---|---|---|---|---|---|---|---|---|---|
| Diving | 0.93 | 0 | 0 | 0 | 0.07 | 0 | 0 | 0 | 0 | 0 |
| Golf | 0 | 0.88 | 0 | 0.06 | 0 | 0 | 0 | 0.06 | 0 | 0 |
| HB | 0 | 0 | 1.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Kicking | 0 | 0 | 0 | 0.90 | 0 | 0 | 0 | 0.10 | 0 | 0 |
| Lifting | 0 | 0 | 0 | 0 | 0.85 | 0 | 0 | 0 | 0.15 | 0 |
| HR | 0 | 0.17 | 0 | 0 | 0 | 0.83 | 0 | 0 | 0 | 0 |
| Running | 0 | 0 | 0 | 0.08 | 0 | 0 | 0.92 | 0 | 0 | 0 |
| Skating | 0.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0.92 | 0 | 0 |
| Swinging | 0 | 0.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0.95 | 0 |
| Walking | 0 | 0 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0.90 |

**Figure 3.16: Confusion matrix on the UCF-Sports dataset (HB - High bar swinging, HR - Horse Riding): SGSH.**

for most of the actions.

### 3.6.3 Running Time Cost

From a computational cost point of view, SGSH reduces the time required to process the interest points by reducing the number of interest points detected on the video frame and selecting only the informative frames, as shown in Table 3.6 for the boxing action as an example. The number of interest points reduces substantially with saliency guidance. Moreover, with the proposed video frame selection the number of processed frames is also reduced significantly, as shown in Table 3.7 for the running action as an example. In general, on a 2.50 GHz Windows 8 workstation with our current un-optimised implementation the mean CPU-time to process an interest point is 0.090469

| | Di | Go | HB | Ki | Lf | HR | Ru | Sk | Sw | Wa |
|---|---|---|---|---|---|---|---|---|---|---|
| Diving | 0.79 | 0 | 0 | 0 | 0.07 | 0 | 0.07 | 0.07 | 0 | 0 |
| Golf | 0 | 0.87 | 0 | 0 | 0 | 0.07 | 0 | 0 | 0.06 | 0 |
| HB | 0 | 0 | 0.90 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 |
| Kicking | 0 | 0.05 | 0 | 0.90 | 0 | 0 | 0 | 0.05 | 0 | 0 |
| Lifting | 0 | 0 | 0 | 0.16 | 0.68 | 0 | 0 | 0 | 0.16 | 0 |
| HR | 0 | 0.25 | 0 | 0 | 0 | 0.75 | 0 | 0 | 0 | 0 |
| Running | 0 | 0 | 0 | 0 | 0 | 0 | 0.92 | 0 | 0 | 0.08 |
| Skating | 0.08 | 0 | 0.08 | 0 | 0 | 0 | 0 | 0.84 | 0 | 0 |
| Swinging | 0 | 0.11 | 0 | 0 | 0 | 0 | 0 | 0 | 0.89 | 0 |
| Walking | 0.05 | 0 | 0 | 0.05 | 0 | 0 | 0.05 | 0 | 0 | 0.85 |

**Figure 3.17: Confusion matrix on the UCF-Sports dataset (HB - High bar swinging, HR - Horse Riding): SH.**

| | HF | Hug | HS | KS | Neg |
|---|---|---|---|---|---|
| HF | 0.76 | 0.04 | 0.08 | 0.04 | 0.08 |
| Hug | 0.08 | 0.68 | 0.08 | 0.04 | 0.12 |
| HS | 0.16 | 0.04 | 0.72 | 0.04 | 0.04 |
| KS | 0.12 | 0.08 | 0.04 | 0.68 | 0.08 |
| Neg | 0.06 | 0.10 | 0.06 | 0.08 | 0.70 |

**Figure 3.18: Confusion matrix for TV-Human Interaction dataset with our method (with saliency guidance). HF (High Five), HS (Hand Shake), KS (Kiss) and Neg (Negative): SGSH.**

seconds for the 3D-SIFT descriptor and 0.397 seconds for the HOOF descriptor. On average, a 2–5 time speedup is obtained with saliency guidance due to the reduced number of feature points on each frame and reduced number of frames to be detected.

## 3.7 Conclusion

We have proposed a novel video feature extraction method based on saliency detection with a new combination of local and global descriptors. Doing so reduces the time

| | Bask | Bow | ClJe | DcTh | HmTh | HgJu | JavTh | LogJu | DivPl | PoVa | Shou | Sna | SpBo | Tennis | TrJu | Vault |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bask | 0.90 | 0 | 0 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bow | 0 | 0.80 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 |
| ClJe | 0 | 0 | 0.90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0 | 0 |
| DcTh | 0 | 0 | 0 | 0.80 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.10 | 0 |
| HmTh | 0 | 0 | 0 | 0 | 0.87 | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HgJu | 0 | 0 | 0 | 0 | 0 | 0.70 | 0 | 0.10 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0.10 | 0 |
| JavTh | 0 | 0 | 0 | 0 | 0.25 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| LogJu | 0 | 0 | 0 | 0 | 0 | 0.17 | 0 | 0.83 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DivPl | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.80 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0.10 |
| PoVa | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0 | 0 |
| Shou | 0 | 0 | 0 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0.80 | 0 | 0.10 | 0 | 0 | 0 |
| Sna | 0 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.20 | 0 | 0 | 0 | 0.70 | 0 | 0 | 0 | 0 |
| SpBo | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0 | 0 | 0 | 0.87 | 0 | 0 | 0 |
| Tennis | 0.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.85 | 0 | 0 |
| TrJu | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.75 | 0 |
| Vault | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0.75 |

**Figure 3.19: Confusion matrix for Olympic sports dataset with our method (with saliency guidance): SGSH.**

**Table 3.2: Evaluation of the proposed method on the KTH dataset using statistical measures: sensitivity, specificity, and error rate for each action: SGSH.**

| | Sensitivity | Specificity | Error Rate |
|---|---|---|---|
| Box | 0.972 | 0.994 | 0.0093 |
| HC | 0.925 | 1 | 0.0138 |
| HW | 1 | 0.983 | 0.014 |
| Jog | 0.94 | 1 | 0.00925 |
| Run | 1 | 0.994 | 0.005 |
| Walk | 1 | 0.994 | 0.005 |

complexity by processing only the interest points on foreground subjects. We also propose to use video frame selection to discard frames without foreground subjects. As a result, focusing on salience regions provides a powerful mechanism to treat only the attention-grabbing objects in a scene and suppress the influence of the background. Experiments show that the proposed method gives a significant improvement on the action recognition for benchmark datasets (see Table 3.8 for KTH, Table 3.9 for UCF-Sports, Table 3.10 for TVHI and Table 3.11 for Olympic sports) for comparisons with

**Table 3.3: Evaluation of the proposed method on the UCF-Sports dataset using statistical measures: sensitivity, specificity, and error rate for each action: SGSH.**

| | Sensitivity | Specificity | Error Rate |
|---|---|---|---|
| Diving | 0.929 | 0.896 | 0.013 |
| Golf | 0.882 | 0.972 | 0.034 |
| HB | 1 | 1 | 0.001 |
| Kicking | 0.9 | 0.851 | 0.027 |
| Lifting | 0.83 | 0.959 | 0.013 |
| HR | 0.83 | 1 | 0.014 |
| Running | 0.923 | 0.981 | 0.020 |
| Skating | 0.916 | 0.973 | 0.027 |
| Swinging | 0.947 | 0.851 | 0.0139 |
| Walking | 0.9 | 1 | 0.0128 |

**Table 3.4: Evaluation of the proposed method on the TVHI dataset using statistical measures: sensitivity, specificity, and error rate for each action: SGSH.**

| | Sensitivity | Specificity | Error Rate |
|---|---|---|---|
| HF | 0.72 | 0.8697 | 0.126 |
| Hug | 0.518 | 0.8659 | 0.173 |
| HS | 0.7619 | 0.8723 | 0.08 |
| KS | 0.64 | 0.8721 | 0.14 |
| Neg | 0.68 | 0.8 | 0.16 |

best results reported on these datasets. The idea of using saliency guidance to improve action recognition is general and in the future we would like to investigate combining this with alternative features as well as its use in other recognition applications.

**Table 3.5: Evaluation of the proposed method on the Olympic sport dataset using statistical measures: sensitivity, specificity, and error rate for each action: SGSH.**

| | Sensitivity | Specificity | Error Rate |
|---|---|---|---|
| Bask | 0.9 | 0.984 | 0.02 |
| Bow | 0.8 | 0.992 | 0.02 |
| ClJe | 0.9 | 1 | 0.007 |
| DcTh | 0.8 | 0.984 | 0.002 |
| HmTH | 0.75 | 0.984 | 0.029 |
| LonJu | 0.7 | 0.992 | 0.0024 |
| HgJu | 0.75 | 0.977 | 0.0026 |
| JavTh | 0.83 | 0.97 | 0.0029 |
| Divpol | 0.7 | 1 | 0.0022 |
| Pova | 0.875 | 0.992 | 0.0014 |
| Shou | 0.8 | 1 | 0.0014 |
| Sna | 0.7 | 0.984 | 0.0037 |
| SrPO | 0.75 | 0.977 | 0.0037 |
| Ten | 0.877 | 0.992 | 0.0014 |
| TrJu | 0.75 | 0.9849 | 0.002 |
| Vault | 0.875 | 0.992 | 0.0014 |

**Table 3.6: The average numbers of interest points without and with saliency guidance (boxing as an example). The first column is the average number of keypoints detected on the video frames. The second one is the average number of keypoints which are detected only on the object.**

| Interest points/frame | After SGFE |
|---|---|
| 41 | 24 |
| 47 | 23 |
| 39 | 19 |
| 43 | 26 |
| 52 | 31 |

**Table 3.7: Results of the proposed video frame selection approach (running as an example). The first column is the duration of the video, the second column is number of all the frames in the video, and the third column is the number of the frames which contain the foreground object (object on-screen).**

| Duration/Seconds | No of Frames | Obj on-Scr |
|---|---|---|
| 00:00:20 | 500 | 165 |
| 00:00:13 | 345 | 122 |
| 00:00:26 | 666 | 336 |
| 00:00:22 | 570 | 181 |
| 00:00:14 | 248 | 133 |

**Table 3.8: Recognition accuracy comparison on KTH dataset**

| Methods on KTH | Accuracy(%) |
|---|---|
| Ghamdi *et al.* [6] | 90.7 |
| Liu *et al.* [96] | 91.3 |
| Iosifidis *et al.* [64] | 92.1 |
| Baumann *et al.* [11] | 92.1 |
| Kalser [76] | 92.6 |
| Ji *et al.* [69] | 93.1 |
| Wang *et al.* [155] | 94.2 |
| Rapits and Soatto [135] | 94.8 |
| Zhang *et al.* [180] | 94.8 |
| Wang *et al.* [156] | 95.0 |
| Yuan *et al.* [178] | 95.4 |
| **SGSH** | **97.2** |

**Table 3.9: Recognition accuracy comparison on the UCF-Sports dataset**

| Methods on UCF-Sports | Accuracy(%) |
|---|---|
| Raptis *et al.* [134] | 79.4 |
| Ma *et al.* [105] | 81.7 |
| Kalser [76] | 85.0 |
| Everts *et al.* [40] | 85.6 |
| Le *et al.* [90] | 86.5 |
| Yuan *et al.* [178] | 87.3 |
| Zhang *et al.* [180] | 87.5 |
| Wang *et al.* [156] | 88.0 |
| Wang *et al.* [155] | 88.2 |
| Ma *et al.* [104] | 89.4 |
| **SGSH** | **90.9** |

**Table 3.10: Recognition accuracy comparison on the TV-Human Interaction dataset.**

| Methods on (TV-Human Interaction) | Accuracy (%) |
|---|---|
| Patron-Perez [126] | 32.8 |
| Yu [176] | 56.0 |
| Gaidon [46] | 62.4 |
| Yu [177] | 66.1 |
| **SGSH** | **70.6** |

**Table 3.11: Recognition accuracy comparison on the Olympic Sports dataset**

| Methods on Olympic Sports | Accuracy (%) |
|---|---|
| Niebles *et al.* [119] | 62.5 |
| Tang *et al.* [151] | 66.8 |
| Liu *et al.* [96] | 74.3 |
| Wang *et al.* [156] | 77.2 |
| Brendel *et al.* [18] | 77.3 |
| **SGSH** | **79.9** |

# *Chapter 4*

# 3D GLOH Features for Human Action Recognition

## 4.1 Introduction

Many video classification techniques exploit combined spatial and temporal information [34, 77, 143]. In such works, action recognition is based on local or global features extracted from the space-time volume (STV) (see Figure 4.1). The STV is formed by temporally stacking frames over a video sequence as a 3D cuboid of intensity volumes. As an example of a 3D spatio-temporal local feature, Dollar [34] applied a spatio-temporal interest point detector to find local regions of interest in the cuboids of space and time for activity recognition. First, the cuboids of spatio-temporal data surrounding a key point extracted from sample behaviours are clustered to form a dictionary of cuboid prototypes. The histogram of the cuboid types is then used as a feature descriptor for action recognition. For global spatio-temporal feature, Gorelick *et al.* [52] proposed a method to generate 3D spatio-temporal shapes by stacking segmented silhouette frame-by-frame. To improve the recognition rate and capture more detailed information from video contents, inspired by the successful works based on the idea of extending feature descriptors from the image domain to the video do-

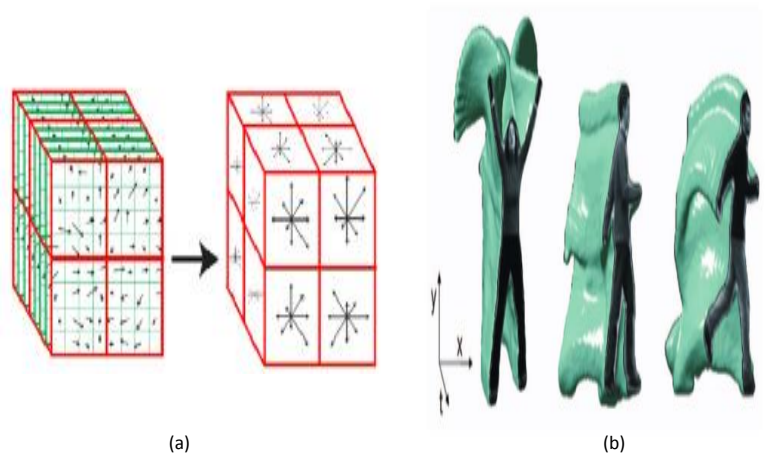



**Figure 4.1: (a) Spatio-temporal local feature descriptor [148]. (b) Spatio-temporal global features (shapes) [52].**

main [143, 77], we propose to extend the Gradient Location and Orientation Histogram (GLOH) descriptor [113] to extract an informative, spatially localised descriptor from video sequences.

In this chapter, we will introduce a novel effective feature called 3D GLOH, which describes local spatially varying information for video data. It detects interest points in the video and then describes them in 3D log-polar coordinates. This descriptor is an extension of the 2D GLOH descriptor [113] and we will demonstrate that it better captures the characteristics of local video information than existing features. Moreover, we propose an action recognition system that uses the 3D GLOH as local features, and the histograms of oriented optical flow (HOOF) [26] as global features. We further employ the idea from our work which introduced in Chapter 3, by extracting features only in salient regions for action recognition. We evaluate the new combined descriptor using a variety of video datasets. The new descriptor outperforms the state-of-the-art descriptors for challenging real-world videos with uncontrolled complicated environment, such as the UCF-Sports , TV-Human Interaction and UCF11 datasets.

The main contributions of this work can be summarised as follows:

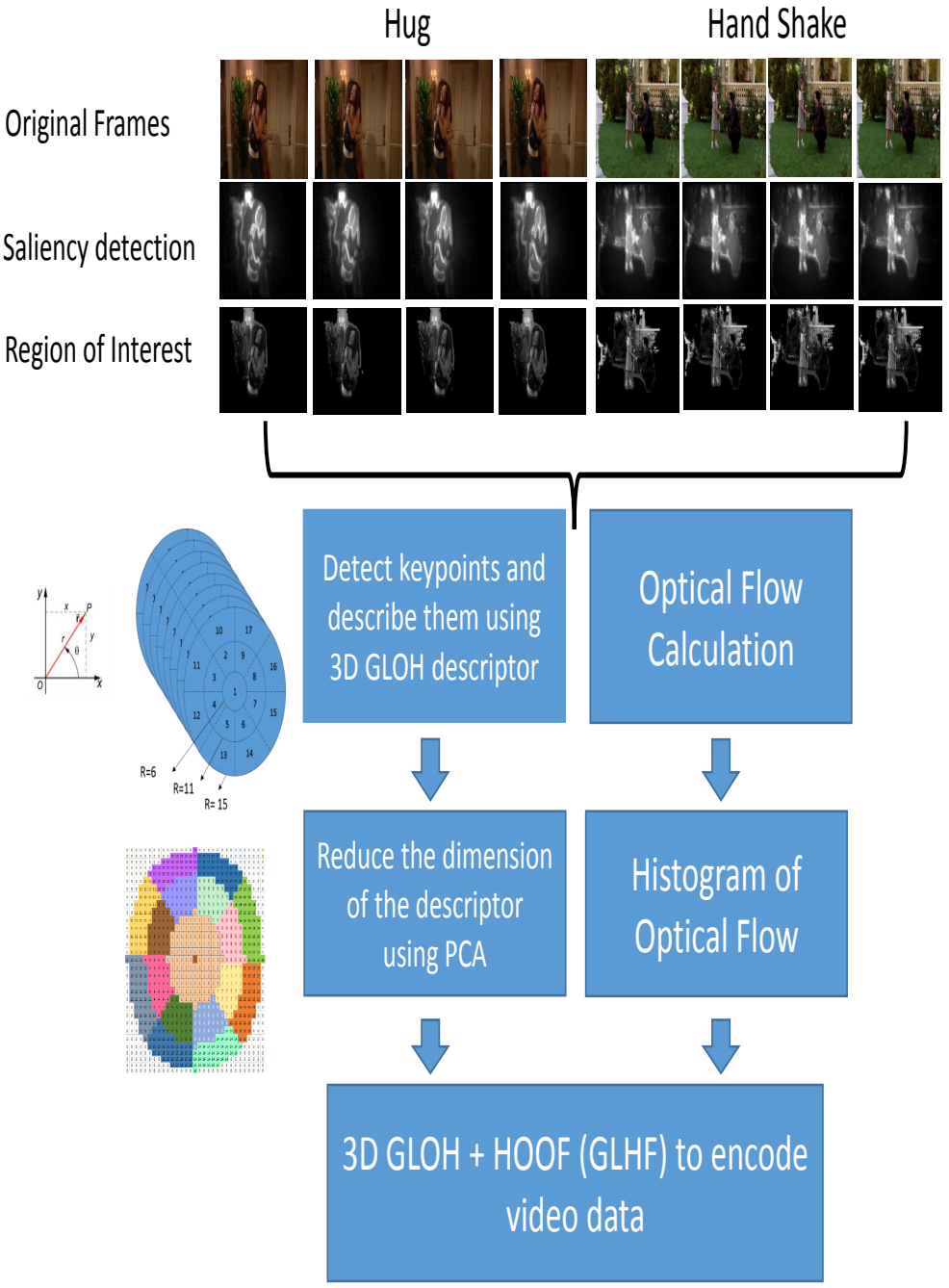



**Figure 4.2: Feature extraction using the S-GLHF (Saliency Guided 3D GLOH and HOOF) descriptor in our action recognition system.**

1. We propose a novel 3D GLOH feature and demonstrate its usefulness for human action recognition.

2. We develop a novel combination of local and global descriptors, which outperforms existing descriptors in action recognition with challenging real-world videos.

The following sections will present the main steps of the proposed descriptor and its implementation and show experimental results on the benchmark datasets.

## 4.2 Proposed Method

The overall framework of our human action recognition system encodes video sequences using a combined local and global representation, along with the Bag of Visual Words (BoVW) framework. The local features are represented by our proposed 3D GLOH from only salient regions in the video frames [1] and the global features are represented using HOOF, we call the combined feature S-GLHF. Figure 4.2 illustrates the main steps of the proposed system for feature extraction. We now describe the system with an emphasis on the novel 3D GLOH descriptor as follows.

### 4.2.1 3D Gradient Location and Orientation Histograms (3D GLOH)

To capture the gradient distribution and localise it in the neighbouring spatio-temporal domain, we extend the GLOH descriptor proposed by Milkolajczky and Schmid [113] to 3D in a log-polar location partitioning. GLOH is designed to increase the robustness and distinctiveness of SIFT [113]. More specifically, we first detect interest points on each frame using the standard 2D SIFT [100]. Common 3D detectors such as Laptev [84] and Dollar [34] typically detect only a sparse set of features since a time-consuming iterative procedure has to be repeated for each feature candidate separately.

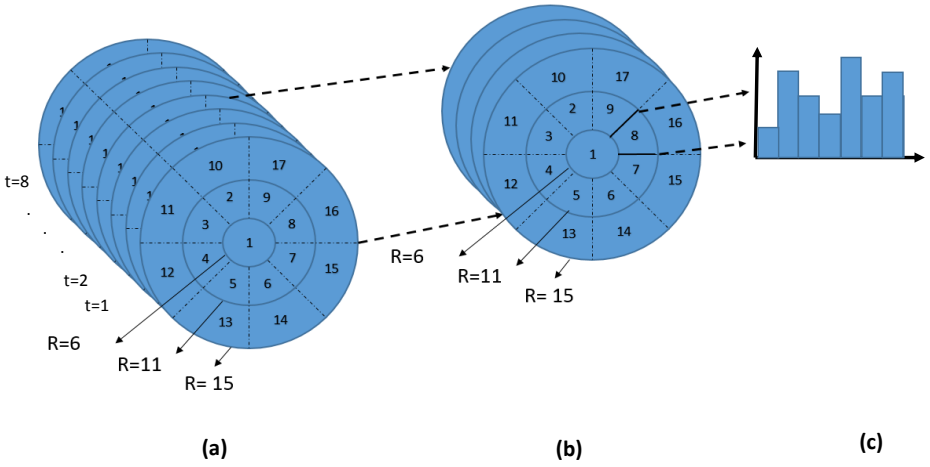



**Figure 4.3: 3D GLOH representation: a) Neighbourhood of the interest point as a cylinder with a diameter of 31 and 8 frames in the spatio-temporal domain. b) Histogram computation over local regions with spatial domain split into 17 log-polar location grid and temporal domain split into two halves. c) Histogram of a local region.**

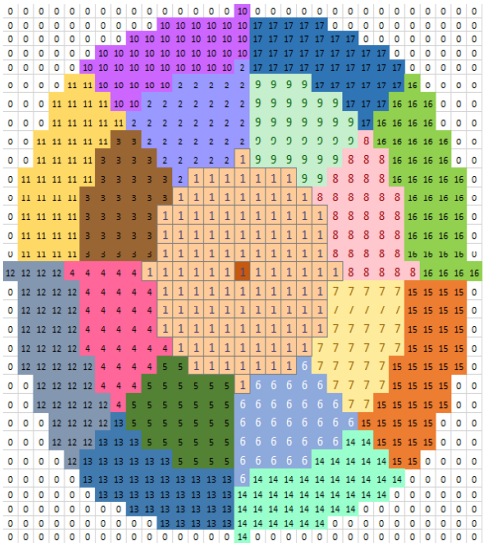


**Figure 4.4: The neighbourhood local region labelling at an interest point used for computing the GLOH descriptor in a log-polar domain.**

Furthermore, the iterative procedure often diverges. As a result, detecting a low number of features is necessary to keep the computation time under control. On the other hand, they claim that direct 3D counterparts to 2D interest point detectors are inadequate for the detection of spatio-temporal feature points, since true spatio-temporal corners are quite rare. They propose to select local maxima over space and time of a response function based on a spatial Gaussian convolved with a quadrature pair of 1D Gabor-filters along the time axis. However, their approaches are not scale-invariant.

For each detected interest point, we consider its neighbourhood as a cylinder in the spatio-temporal volume, with a diameter of 31 pixels in the spatial domain and a height of 8 pixels (frames), 3 frames before the frame of detected interest point and 4 frames after, along the temporal domain. The cylinder is further divided in both the spatial and temporal domains to provide localised distribution. In the temporal domain, the cylinder is split into two halves each with 4 frames (see Figure 4.3). In the spatial domain, following [113] a log-polar location grid is used with three bins in the radial direction (the radii are set to 6, 11, 15) and 8 in the angular direction for each slice, which results in 17 location bins (see Figure 4.4), where the central bin is not divided in angular directions. The Cartesian coordinate system is transformed into the polar coordinate system through the following equations:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, \tag{4.1}$$

$$\theta_i = \tan^{-1}(y_i - y_c)/(x_i - x_c), \tag{4.2}$$

where$(x_i, y_i)$ is the coordinate of pixel in the Cartesian coordinate system, $(r_i, \theta_i)$ is the radius and the angle in the polar coordinate system. $(x_c, y_c)$ is the coordinate of the interest point. This leads to $17 \times 2 = 34$ local regions in the spatio-temporal domain.

For each pixel in a local region, 3D gradients are calculated, similar to 3D SIFT [143, 148]. The 3D gradient orientation for each pixel is described using two angles $\theta$ and $\phi$, which are defined as follows:

$$\theta(x, y, t) = \tan^{-1}(L_y/L_x), \tag{4.3}$$

$$\phi(x, y, t) = \tan^{-1}(L_t / \sqrt{L_x^2 + L_y^2}), \tag{4.4}$$

where $L$ is the intensity of the video frame, $L_x, L_y$, and $L_t$ are partial derivatives, respectively computed using finite difference approximations: $L(x+1, y, t) - L(x-1, y, t)$, $L(x, y+1, t) - L(x, y-1, t)$ and $L(x, y, t+1) - L(x, y, t-1)$. $\theta$ and $\phi$ encode the angles for the 3D gradient direction.

Each gradient orientation angle is quantised into $N$ bins (by default we use $N = 16$). As two angles are used to describe a 3D orientation, the descriptor is a vector of $2N \times 34$ dimension.

The resulting descriptor is high dimensional, which makes computation expensive. For example, when the default $N = 16$ is used, the histogram dimension is $2 \times 16 \times 34 = 1088$. We use Principal Component Analysis (PCA) to reduce the dimensionality. The covariance matrix for PCA is estimated using the training examples in the datasets, and $192$ dominant eigenvectors are used to reduce the dimension to the same level as 3D SIFT features.

### 4.2.2 Human Action Recognition using S-GLHF Descriptor

As we will show later, our proposed 3D GLOH descriptor is particularly effective in describing local spatio-temporal distribution at each interest point. Following our recent work [1], considering only keypoints in the foreground helps suppress the impact of spurious keypoints by incorporating some "semantic" information. The 3D GLOH descriptor is then complemented with a global descriptor namely HOOF [26], which produces a histogram representing the motion in each frame of the video.

To represent the characteristics of a whole video, we employ a Bag-of-Visual-Words framework. We build a vocabulary of visual words for each of the two descriptors (3D GLOH and HOOF) using k-means clustering of features extracted from all the training videos in the dataset. $2000$ visual words are used for each descriptor as it gives a good balance of efficiency and performance. Once this is done, each feature vector is

mapped to the closest visual word in the vocabulary. For each video a feature vector is obtained by concatenating two histograms measuring the distribution of visual words in the video. This combined descriptor (which we call S-GLHF) takes advantages from both local and global representations to describe the information of the video to be more informative and selective. For classification, we use multi-class kernel SVM classifier with Radial Basis Function (RBF) kernels. The SVM kernel parameters are automatically optimised using grid search with 5-fold cross validation.

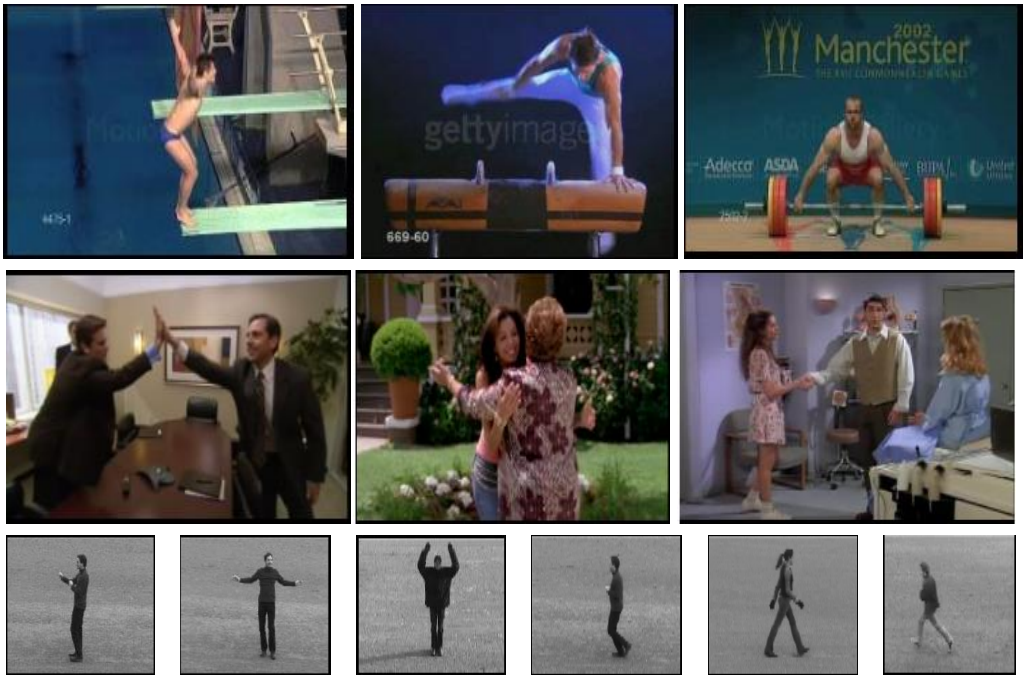



**Figure 4.5: Benchmark datasets used to evaluate our method. Top to bottom: images from videos in UCF-Sports, TV-Human Interaction and the KTH datasets.**

## 4.3 Experimental results

In this section we report results on several benchmark datasets (see Figure 4.5), and discuss how our method behaves with varying key parameters.

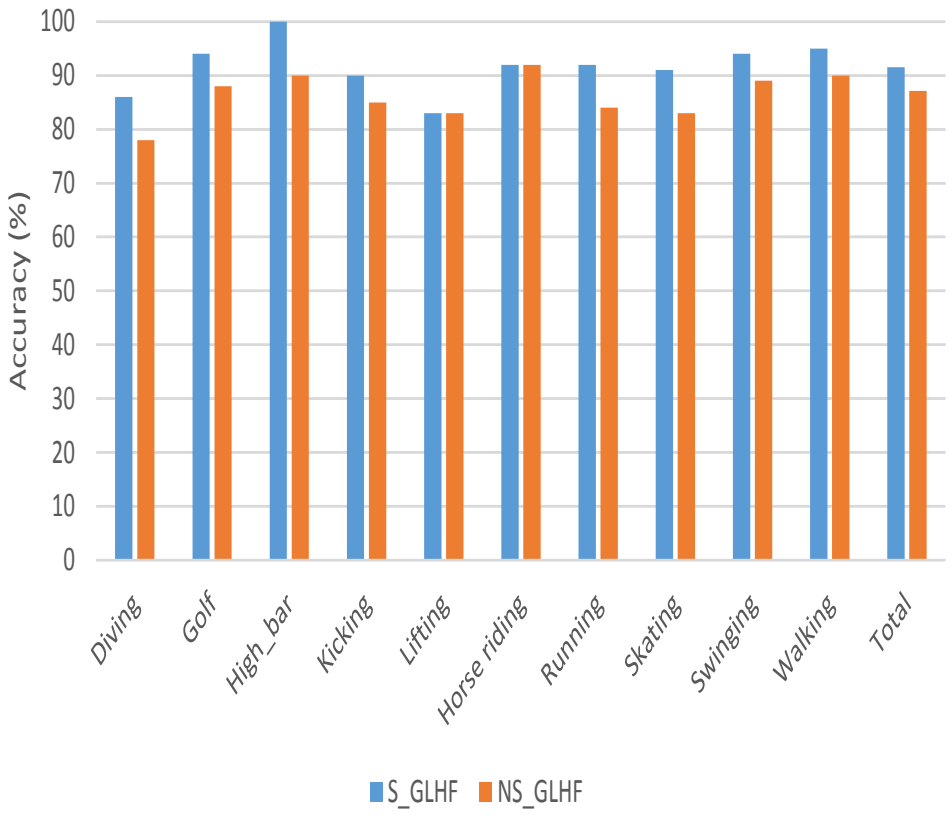



**Figure 4.6: The recognition rates of The UCF-Sports dataset for each individual action and the total accuracy with saliency guidance (S-GLHF) and without saliency guidance (NS-GLHF).**

### 4.3.1 Results and Discussions

We performed extensive experiments using several standard datasets to study the effectiveness of our proposed 3D GLOH descriptor and the human action recognition system.

For the UCF-Sports dataset, Figure 4.6 shows the performance of recognising each class of videos using 3D GLOH and HOOF descriptors. The method works consistently well in all categories, and in particular by using the saliency guidance, the recognition rate increases for every class of videos (blue bars with saliency vs. orange

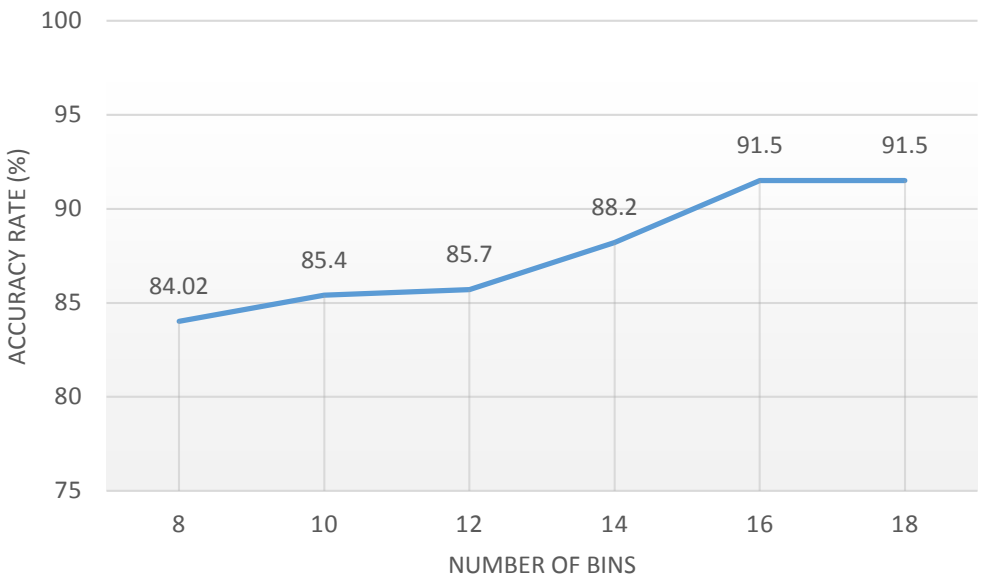



**Figure 4.7: Recognition rate of the UCF-Sports dataset using different numbers of bins for the 3D-GLOH descriptor.**

| | Di | Go | HB | Ki | Lf | HR | Rn | Sk | Sw | Wa |
|---|---|---|---|---|---|---|---|---|---|---|
| Diving | 0.85 | 0 | 0 | 0.15 | 0 | 0 | 0 | 0 | 0 | 0 |
| Golf | 0 | 0.94 | 0 | 0 | 0 | 0 | 0 | 0.06 | 0 | 0 |
| HB | 0 | 0 | 1.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Kicking | 0 | 0 | 0 | 0.90 | 0 | 0 | 0 | 0.05 | 0 | 0.05 |
| Lifting | 0 | 0 | 0 | 0 | 0.83 | 0 | 0 | 0 | 0.17 | 0 |
| HR | 0 | 0.09 | 0 | 0 | 0 | 0.91 | 0 | 0 | 0 | 0 |
| Running | 0 | 0 | 0 | 0 | 0 | 0 | 0.92 | 0 | 0 | 0.08 |
| Skating | 0.09 | 0 | 0 | 0 | 0 | 0 | 0 | 0.91 | 0 | 0 |
| Swinging | 0 | 0 | 0 | 0 | 0.06 | 0 | 0 | 0 | 0.94 | 0 |
| Walking | 0 | 0 | 0 | 0 | 0 | 0 | 0.05 | 0 | 0 | 0.95 |

**Figure 4.8: Confusion matrix for The UCF-Sports dataset with our action recognition system. HB (High bar), HR (Horse Riding).**

bars without saliency). A key parameter in the 3D GLOH descriptor is the number of bins $N$ when histograms are built. To investigate the behaviour of our method with changing $N$, results are reported in Figure 4.7, and it can be seen that $N = 16$ achieves good results and increasing $N$ does not improve the performance further. Thus unless for comparative purpose, we use this setting for all the experiments in the chapter. The confusion matrix of the results obtained using our system is reported in Figure 4.8. We compare our method with the state-of-the-art methods which reported the performance on the UCF-Sports dataset (see Table 4.1). It can be seen that our method (S-GLHF) outperforms the state-of-the-art methods by at least $0.6\%$. It is a significant improvement considering that the current performance has already been over $90\%$.

The TV-Human Interaction dataset is more complicated as it involves interactions between multiple subjects. Figure 4.9 shows the recognition rate of our approach for each action. The highest recognition rate is obtained for the action *High five*, which is $84\%$. Figure 4.10 depicts the confusion matrix of the result obtained using the proposed method. The matrix shows that about $10\%$ of the *High five* action is mistakenly recognised as *Hug* or *Hand shake* actions due to the similarity among these actions. The performance is consistently good, especially with saliency guidance. Compared with existing methods tested on this dataset (see Table 4.2), our method achieves $75.3\%$ accuracy, which improves the accuracy of the proposed method in Chapter 3 $(70.6\%)$ by a significant $4.7\%$, and the improvement is more significant with previously published works reporting their performance on this dataset. The evaluation of the proposed method using statistical measures, namely sensitivity and specificity is shown in Table 4.3. In fact, for each action, our method achieves over $70\%$ accuracy, which is better than the average performance of the method described in Chapter 3. The comparisons of sensitivity (true positive rate) and specificity (true negative rate) are presented in Figure 4.11 and Figure 4.12, respectively, showing that the new descriptor is more effective that SGSH for each action category. The overall error rate is improved from

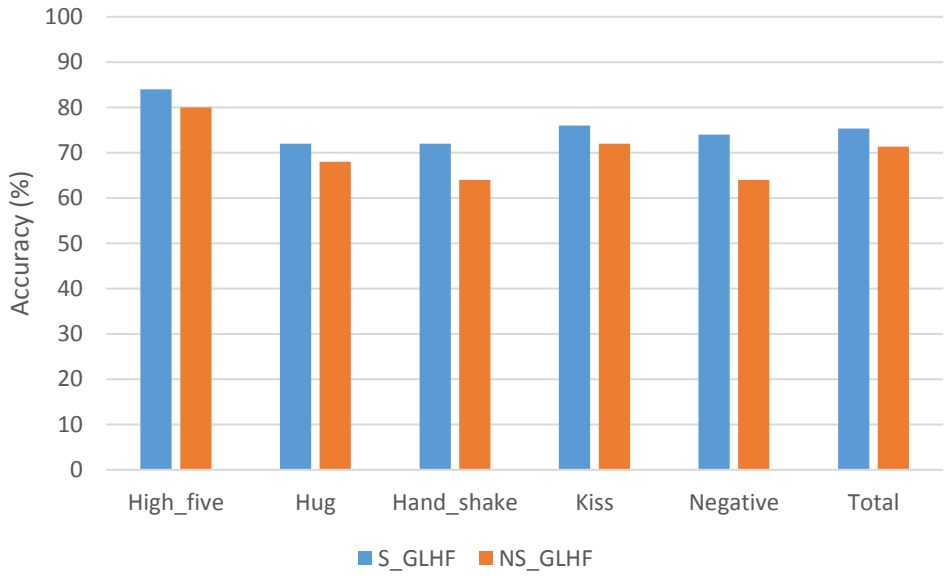



**Figure 4.9: The recognition rate of the TV-Human Interaction dataset for each action using GLHF with and without saliency.**

| | HF | Hug | HS | KS | Neg |
|---|---|---|---|---|---|
| HF | 0.84 | 0.08 | 0.08 | 0 | 0 |
| Hug | 0 | 0.72 | 0.12 | 0.08 | 0.08 |
| HS | 0.08 | 0.12 | 0.72 | 0.08 | 0 |
| KS | 0 | 0.12 | 0.08 | 0.76 | 0.04 |
| Neg | 0.04 | 0.08 | 0.04 | 0.1 | 0.74 |

**Figure 4.10: Confusion matrix for TV-Human Interaction dataset with our method (with saliency guidance). HF (High Five), HS (Hand Shake), KS (Kiss) and Neg (Negative).**

13.58% to 7.68% with S-GLHF descriptor.

For the UCF11 dataset, the dataset is challenging due to large variations in camera

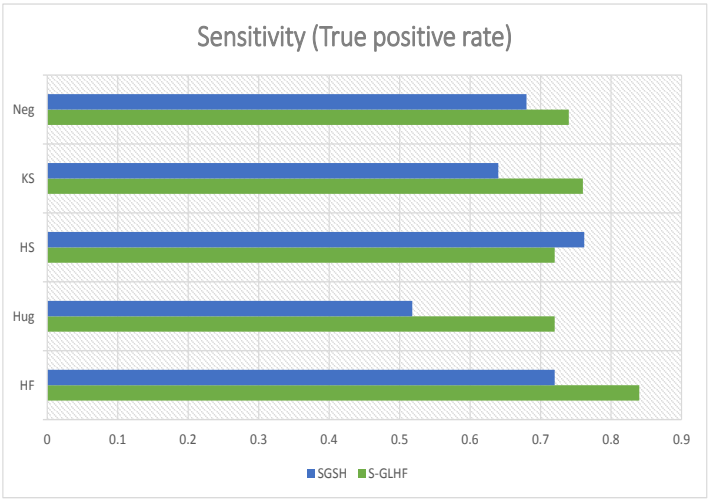



**Figure 4.11: The comparison of sensitivity (true positive rate) for the TVHI dataset using S-GLHF and SGSH features.**

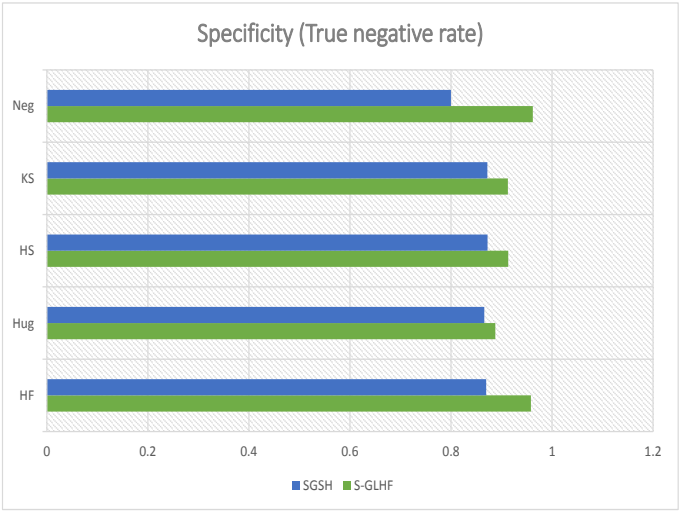



**Figure 4.12: The comparison of specificity (true negative rate) for the TVHI dataset using S-GLHF and SGSH features.**

motion, object appearance and pose, object scale, viewpoint and illumination conditions. From Table 4.4, we can reach the same conclusion that the proposed features (S-GLHF) are effective in representing detailed video contents than SGSH features.

| | b-shoot | spiking | jumping | s-juggling | h-riding | biking | diving | swing | g-swing | t-swing | walking |
|---|---|---|---|---|---|---|---|---|---|---|---|
| b-shoot | 0.85 | 0 | 0 | 0.08 | 0 | 0 | 0 | 0 | 0.07 | 0 | 0 |
| spiking | 0 | 0.90 | 0 | 0 | 0 | 0 | 0.10 | 0 | 0 | 0 | 0 |
| jumping | 0 | 0 | 0.91 | 0 | 0 | 0 | 0 | 0 | 0 | 0.09 | 0 |
| s-juggling | 0 | 0.08 | 0 | 0.84 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0.04 |
| h-riding | 0 | 0 | 0 | 0 | 0.94 | 0 | 0.06 | 0 | 0 | 0 | 0 |
| biking | 0 | 0 | 0 | 0.08 | 0 | 0.92 | 0 | 0 | 0 | 0 | 0 |
| diving | 0.06 | 0 | 0 | 0 | 0.04 | 0 | 0.83 | 0 | 0 | 0.07 | 0 |
| swing | 0 | 0 | 0 | 0 | 0 | 0.13 | 0 | 0.87 | 0 | 0 | 0 |
| g-swing | 0 | 0 | 0.05 | 0 | 0 | 0 | 0 | 0 | 0.85 | 0 | 0 |
| t-swing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.09 | 0 | 0.91 | 0 |
| walking | 0 | 0 | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0.87 |

**Figure 4.13: Confusion matrix for UCF11 dataset with our method (saliency guidance): SGSH.**

| | b-shoot | spiking | jumping | s-juggling | h-riding | biking | diving | swing | g-swing | t-swing | walking |
|---|---|---|---|---|---|---|---|---|---|---|---|
| b-Shoot | 0.89 | 0 | 0 | 0 | 0 | 0 | 0.09 | 0 | 0 | 0 | 0.02 |
| spiking | 0 | 0.91 | 0 | 0 | 0.09 | 0 | 0 | 0 | 0 | 0 | 0 |
| jumping | 0 | 0 | 0.94 | 0 | 0 | 0 | 0 | 0.03 | 0 | 0.03 | 0 |
| s-juggling | 0 | 0.04 | 0 | 0.86 | 0 | 0 | 0 | 0.04 | 0 | 0.06 | 0 |
| h-riding | 0 | 0 | 0 | 0 | 0.95 | 0.05 | 0 | 0 | 0 | 0 | 0 |
| biking | 0 | 0 | 0 | 0.07 | 0 | 0.93 | 0 | 0 | 0 | 0 | 0 |
| diving | 0.06 | 0 | 0.04 | 0 | 0 | 0 | 0.85 | 0 | 0 | 0.05 | 0 |
| swing | 0 | 0 | 0 | 0 | 0 | 0.11 | 0 | 0.89 | 0 | 0 | 0 |
| g-swing | 0 | 0 | 0.05 | 0 | 0.10 | 0 | 0 | 0 | 0.85 | 0 | 0 |
| t-swing | 0 | 0.04 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0.92 | 0 |
| walking | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.12 | 0 | 0.88 |

**Figure 4.14: Confusion matrix for UCF11 dataset with our method (saliency guidance): S-GLHF.**

Figure 4.13 shows the confusion matrix of SGSH descriptor and Figure 4.14 shows confusion matrix of S-GLHF on the UCF11 dataset.

Our 3D GLOH feature exploits the spatio-temporal distribution of gradients to provide a more discriminative descriptor. As a result, our 3D GLOH feature may not be very effective if the video data contains little texture. An example of such kind of data is the KTH dataset (see Figure 4.5). This dataset is relatively easy as it has a clean background and was captured in a controlled environment. However, the images are relatively low-resolution and do not contain much texture. Our method achieves $94.9\%$ accuracy, which is close to some of the recent methods [180] ($94.8\%$) (see Figure 4.15)

| | Box | HC | HW | Jog | Run | Walk |
|---|---|---|---|---|---|---|
| Box | 0.94 | 0 | 0.03 | 0.03 | 0 | 0 |
| HC | 0.08 | 0.89 | 0.03 | 0 | 0 | 0 |
| HW | 0.03 | 0 | 0.97 | 0 | 0 | 0 |
| Jog | 0 | 0 | 0 | 0.92 | 0.03 | 0.05 |
| Run | 0 | 0.03 | 0 | 0 | 0.97 | 0 |
| Walk | 0 | 0 | 0 | 0 | 0 | 1.00 |

**Figure 4.15: Confusion matrix on the KTH dataset (HC - Handclapping, HW - Handwaving): S-GLHF.**

but not as good as SGSH descriptor which achieves $97.2\%$. Figure 4.16 shows the comparison of the recognition rate using SGSH and S-GLHF descriptors. Nevertheless, for more challenging real-world datasets, we have shown that the proposed 3D GLOH descriptor is effective and outperforms existing methods.

## 4.4 Conclusion

In this chapter, we introduce a new local descriptor for video data namely 3D GLOH and propose a human action recognition system using the proposed local descriptor along with a global descriptor. The 2D GLOH descriptor is extended to video frames by partitioning the cylindrical local neighbourhood of an interest point into spatio-temporal bins and calculating 3D histograms of gradients in the local bins. The experimental results show that the proposed the 3D GLOH descriptor is effective in capturing localised spatio-temporal information and the overall system outperforms the state-of-the-art methods in terms of recognition accuracy for challenging real-world datasets,

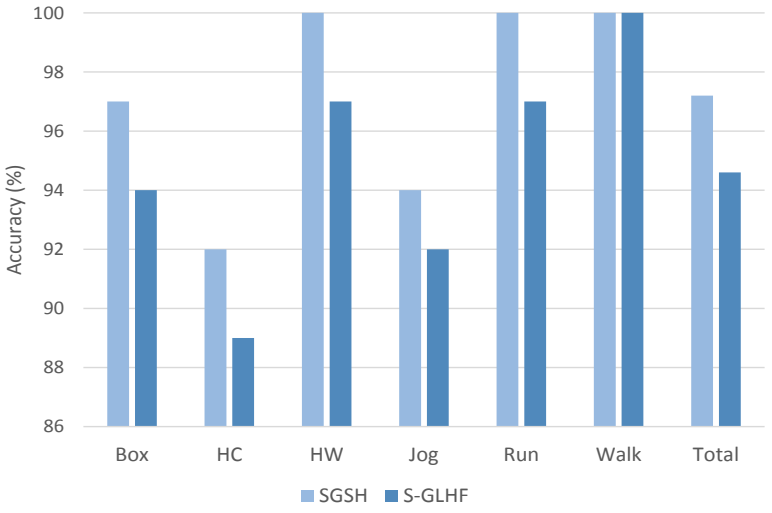



**Figure 4.16: The comparison of recognition rate for KTH dataset using SGSH and S-GLHF features.**

including UCF-Sport, TV-Human Interaction and UCF11 datasets. The proposed 3D GLOH descriptor can be useful for analysing videos, especially for those with rich textures. We would like to further investigate its effectiveness in other video analysis applications.

**Table 4.1: Recognition accuracy comparison on the UCF-Sports dataset**

| Methods on (UCF-Sports) | Accuracy (%) |
|---|---|
| Raptis [134] | 79.4 |
| Ma [105] | 81.7 |
| Kalser [76] | 85.0 |
| Everts [40] | 85.6 |
| Le [90] | 86.5 |
| Zhang [180] | 87.5 |
| Wang [156] | 88.0 |
| Ma [104] | 89.4 |
| SGSH (Chapter 3) | 90.9 |
| **Our Method (S-GLHF)** | **91.5** |

**Table 4.2: Recognition accuracy comparison on the TV-Human Interaction dataset.**

| Methods on (TV-Human Interaction) | Accuracy (%) |
|---|---|
| Patron-Perez [126] | 32.8 |
| Yu [176] | 56.0 |
| Gaidon [46] | 62.4 |
| Yu [177] | 66.1 |
| SGSH (Chapter 3) | 70.6 |
| **Our Method (S-GLHF)** | **75.3** |

**Table 4.3: Evaluation of the proposed method on TVHI dataset using the statistical measures: sensitivity and specificity.**

| | Sensitivity | Specificity |
|---|---|---|
| High five | 0.84 | 0.9583 |
| Hug | 0.72 | 0.8878 |
| Hand shake | 0.72 | 0.9134 |
| Kiss | 0.76 | 0.9126 |
| Negative | 0.74 | 0.9620 |

**Table 4.4: Recognition accuracy comparison on UCF11 dataset**

| Methods on UCF11 | Accuracy(%) |
|---|---|
| Liu *et al.* [98] | 70.4 |
| Liu *et al.* [97] | 71.2 |
| Oikonomopoulos *et al.* [122] | 71.2 |
| Mota *et al.* [115] | 74.5 |
| Everts *et al.* [39] | 78.6 |
| Cho *et al.* [28] | 86.1 |
| **SGSH** | **88.6** |
| **S-GLHF** | **89.9** |

# *Chapter 5*

# Action Recognition based on Matching of Deforming Skeleton Graphs

## 5.1 Introduction

Graph-based methods have achieved great success in image classification [114, 35]. Building on this success, graph-based methods have been generalised from the image to the video domain [99, 161] to represent actions as graphs. Existing methods however still encode local image descriptors at graph nodes and thus suffer from similar limitations of image descriptors. Therefore, we instead propose to extract graphs with *minimal* information, namely deforming skeleton graphs of foreground subjects, to encode actions. Action recognition is then formulated as finding the best matched deforming graphs. By doing so, our method is robust to typical variations such as the appearance of the subject, background, lighting etc. The deforming skeleton graphs can still change topology, due to the pose or imperfect extraction of foreground subjects, and the temporal dynamics (speed of body movement) can vary from person to person. To cope with such *topological* and *temporal* variations robustly, we adopt the Optimal Subsequence Bijection algorithm [88] to measure the similarity between two static graphs, which is combined with Dynamic Time Warping [133] to address tem-

poral change. For periodic actions (e.g. running), video sequences may capture action cycles from different starting points. We further develop a method to automatically identify a representative frame to align such actions. Figure 5.1 illustrates the main steps of the proposed method. Our method can be used to derive a similarity measure between two dynamic skeleton graph sequences. The matching algorithm involves five steps (saliency detection, graph construction, end node matching, graph matching, and action matching). We will explain these steps in details in Section 5.

The main contributions of the work are:

- We propose to represent actions in video sequences as sequences of deforming skeleton graphs of foreground subjects. The representation has significant advantages of being insensitive to changes of illumination, subject appearance and backgrounds.
- We develop a method to recognise human actions based on matching of deforming skeleton graphs. Our similarity measure takes into account topological variation, temporal variation and alignment of periodic actions to improve its robustness. Experimental results show that our method purely based on graph matching outperforms state-of-the-art action recognition methods. Moreover, since our method uses compact and highly abstract information, it achieves decent recognition performance with even a single example from each category, which is a very challenging scenario for existing methods. Due to the use of complementary information, we achieve even better recognition performance by fusing our method with an alternative image descriptor based method.
- To improve efficiency, instead of matching a test video against each training example, we extend our method by clustering training examples and performing matching in a hierarchical manner. This improves the efficiency over 10 times while maintaining the recognition rate.

The remainder of this chapter is organised as follows. Section 5 describes the proposed

method in detail. Experimental results on two benchmark datasets for human action recognition are presented in Section 5.3, and finally the conclusions and future work are discussed in Section 5.4.

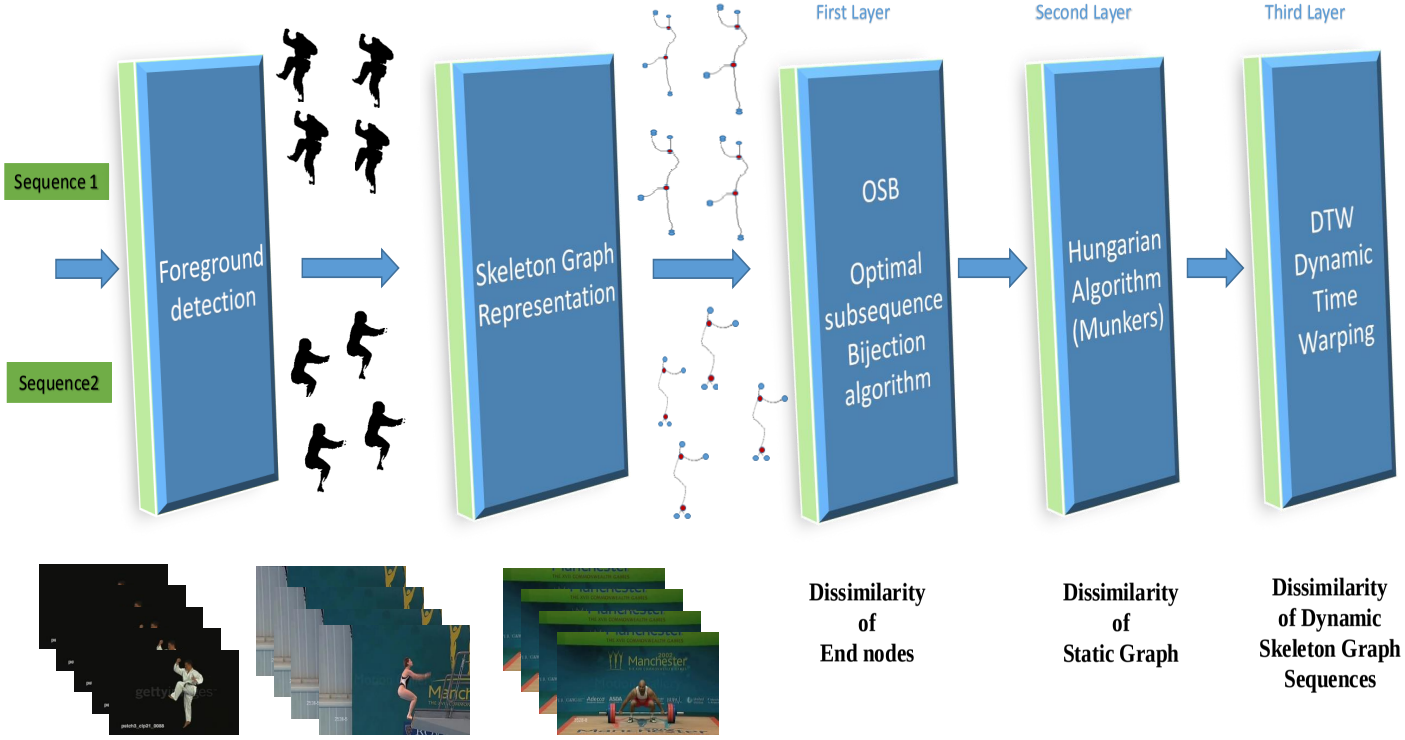



**Figure 5.1: Pipeline of the proposed deforming skeleton graph extraction and matching method, which contains five steps: foreground detection, skeleton graph representation, end nodes dissimilarity computation using OSB, static graph dissimilarity computation using Hungarian algorithm, and graph sequence dissimilarity computation using DTW.**

## 5.2 Proposed Approach

The proposed method for action recognition is based on matching of deforming skeleton graphs. Given two input videos, we work out the dissimilarity measure (distance) between the two deforming skeletons in the following five steps, as shown in Figure 5.1. The first step is to detect the salient regions in the video frames as in Chapter 3. The skeletons (also known as medial axes) are then extracted from the foreground shapes by applying morphological operations (dilation and thinning). Examples of extracted skeleton graphs for some benchmark videos are shown in Figure 5.2. It can be seen that although not perfect, the extracted skeletons well represent the actions regardless of the subject appearance and background. Moreover, skeleton matching has a lower sensitivity to articulation. The extracted skeleton integrates geometrical and topological features of the object, which provides an important shape descriptor for object recognition. The advantage of matching skeleton graphs as opposed to matching skeleton trees such as the Shock Tree [127] is that tree matching techniques require to first convert skeleton graphs to trees. However, this may result in losing important structural information. The third step is to compute for each graph the dissimilarity measure between every pair of the end nodes. This is achieved by using the Optimal Subsequence Bijection (OSB) algorithm [88] as it is known to be robust to topological changes. The dissimilarity measure of two static graphs is then worked out by finding the best matching between nodes that minimises the total costs [10]. This is further generalised to two dynamic skeleton graph sequences, where a Dynamic Time Warping (DTW) [133] based matching algorithm is used to cope with temporal variation. Given an input test video, action recognition is then formulated as finding the video in the training set with the minimal dissimilarity measure. We further consider techniques to choose representative frames and align frames to improve robustness and efficiency. When the training set is large, we also propose a hierarchical matching strategy to

speed up the computation.

### 5.2.1 Graph Representation based on Skeletons of Foreground Regions

This section describes the initial steps of building skeleton graphs to represent actions. To identify foreground regions, similar to Chapter 3, we use a saliency detection method [109] to extract the salient region from each video frame. We then apply morphological thinning and dilation operations [82] to obtain the skeleton for each shape region. The skeleton extraction is applied frame by frame. Although it is possible to exploit coherence between frames, our simple strategy is preferred because some actions (e.g. sports) can be fast-moving, and shapes can change rapidly between adjacent frames. Frame-by-frame extraction can also be beneficial to robustness in that one inaccurate skeleton will not affect adjacent frames. Moreover, matching skeletons using the OSB algorithm allows to deal with holes in the foreground region (i.e. cycles in the skeleton) [10].

For each time step, the skeleton is represented using a graph, similar to [10]. An example is shown in Figure 5.3. The graph nodes consist of *endpoints* where they are connected to only one adjacent skeleton pixel and *junction points* where they have three or more adjacent skeleton pixels. The remaining skeleton points form skeleton branches which are edges in the graph.

So we can define the skeleton graph based on the following definitions.

**Definition 1**: The points in a skeleton graph are classified into three types: 1) endpoint, which is a point having only one adjacent point. It represents an end node in the skeleton graph; 2) junction point, which has three or more adjacent points. It represents a junction node in the skeleton graph; 3) connection point, which is not an endpoint or a junction point. It does not construct a node in the graph.

**Definition 2**: A skeleton branch is a sequence of connection points between two directly connected endpoints and/or junction points.

**Definition 3**: A skeleton path is the shortest path between a pair of end points on a skeleton graph.

**Definition 4**: A path distance is the path dissimilarity between two sequences of end node pairs in two graphs.

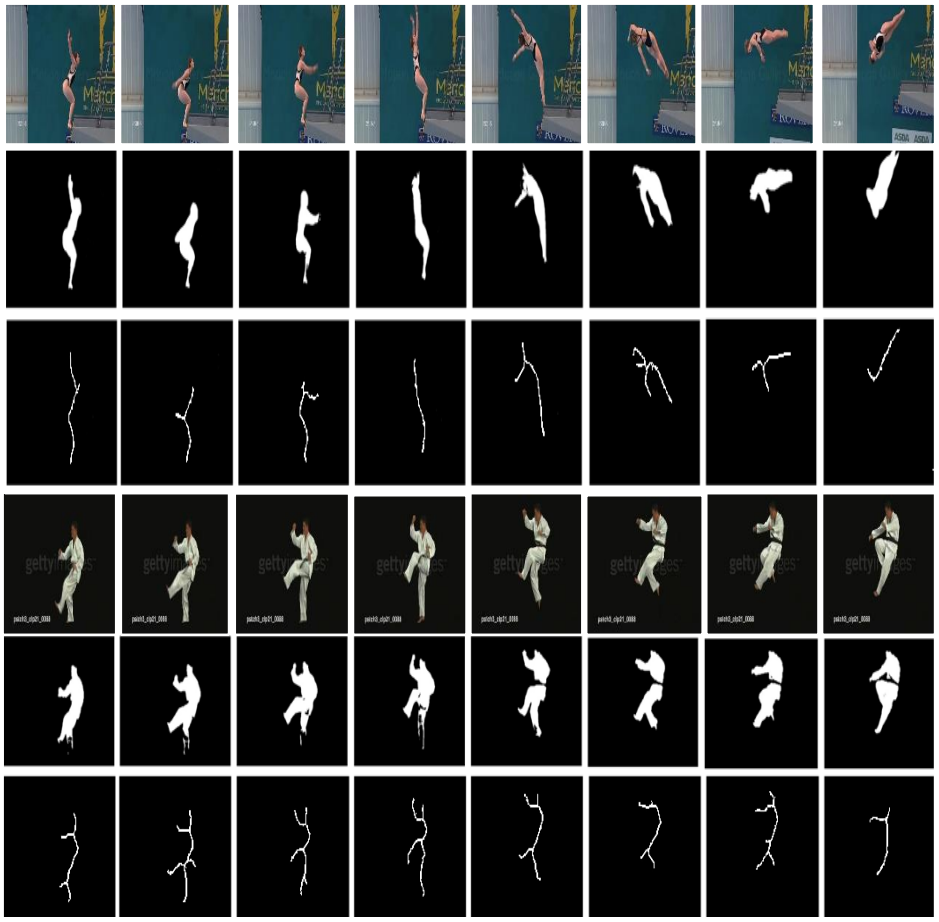


**Figure 5.2: The skeleton graph representations for diving and kicking actions from the UCF-Sports benchmark. Every three rows from top to bottom show the original video frames, the salient regions, and the extracted skeletons from the corresponding salient regions.**

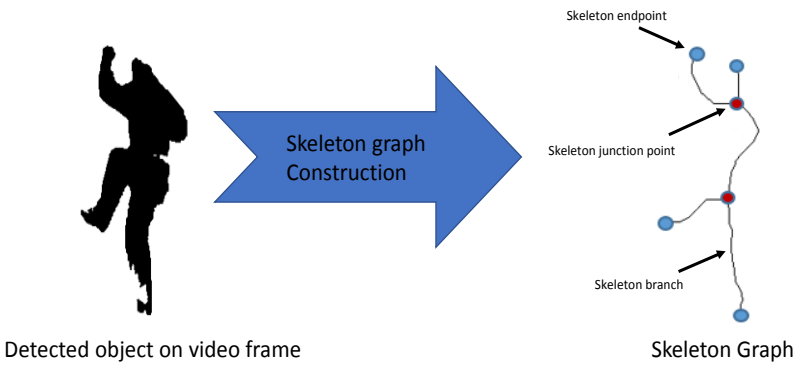



**Figure 5.3: An example of the skeleton graph representation**

A skeleton graph is built in the following way: The end points and junction points are chosen as the nodes of the graph, and the edges of this graph are all the branches between the nodes (see Figure 5.3). We represent the skeleton paths only between end nodes, since these nodes are the salient points on the contour. The proposed graph matching approach is based on the correspondences of the graph end nodes, not considering any junction nodes. The junction points can vary from one graph to another for the same shape, so the graph could be sensitive to distance variation at the junction point, which could result in incorrect graph matching [10].

### 5.2.2 Deforming Skeleton Graph Matching

To measure the dissimilarity of two sequences of deforming skeleton graphs $\mathcal{G}$ and $\mathcal{G}'$, we start with two static skeleton graphs $G$ and $G'$, where $G \in \mathcal{G}$ and $G' \in \mathcal{G}'$. For this purpose, we use the Optimal Subsequence Bijection based method [10], which only matches endpoints, as they correspond more robustly to semantic parts, whereas junction points can be easily affected by e.g. changing of poses [10]. The dissimilarity measure $d(G, G')$ is calculated first for path pairs, then for vertex pairs, and finally

for the graph pairs. Given two endpoint pairs $u, v \in G$ and $u', v' \in G'$, the path dissimilarity $pd(p(u,v), p'(u',v'))$ between path $p$ from $u$ to $v$ on $G$, and path $p'$ from $u'$ to $v'$ on $G'$ is defined as the sum of the differences of skeleton radii at regular sample points along the paths, and the difference of the path lengths, both normalised for scale invariance [10].

$$\left(\{pd_i\}, \{pd'_j\}\right) = \begin{pmatrix} pd(p(v_{i0}, v_{i1}), p(v'_{j0}, v'_{j1})) \ldots\ pd(p(v_{i0}, v_{i1}), p(v'_{j0}, v'_{jN})) \\ pd(p(v_{i0}, v_{i2}), p(v'_{j0}, v'_{j1})) \ldots pd(p(v_{i,0}, v_{i,2}), p(v'_{j,0}, v'_{j,N})) \\ pd(p(v_{i0}, v_{i3}), p(v'_{j0}, v'_{j1})) \ldots\ pd(p(v_{i0}, v_{i3}), p(v'_{j0}, v'_{jN})) \\ \cdot \qquad\qquad \ldots \qquad\qquad \cdot \\ pd(p(v_{i0}, v_{iK}), p(v'_{j0}, v'_{j1})) \ldots pd(p(v_{i0}, v_{iK}), p(v'_{j0}, v'_{jN})) \end{pmatrix}$$

Given a pair of vertices $v_i \in G$ and $v'_j \in G'$, a set of path distances can be worked out for $v_i$ (and $v'_j$) respectively, between $v_i$ and the remaining endpoints in $G$ in the clockwise order starting from $v_i$ (similar applies to $v'_j$) as shown in Figure 5.4, $v_i$ is denoted as $v_{i0}$ and $v'_j$ as $v'_{j0}$ respectively. Denote these two path distance sequences as $\{pd_i\}$ and $\{pd'_j\}$. The distance $d(v_i, v'_j)$ between the vertex pair $v_i$ and $v'_j$ is calculated using Optimal Subsequence Bijection [88] which finds the optimal alignment between the sequences with the option of skipping over elements in the sequences, which is essential for coping with topological changes.

$$d(v_i, v'_j) = OSB\left(\{pd_i\}, \{pd'_j\}\right) \tag{5.1}$$

Taking every vertex pair between $G$ and $G'$, a matrix is obtained. An example is shown in Figure 5.5. To cope with different vertex numbers, virtual vertices are introduced ($N4$ in the example) with the average distance of all pairs assigned. The total distance between two graphs $d(G, G')$ is then defined as the minimum total costs of bipartite matching, which can be efficiently computed using the Hungarian algorithm.

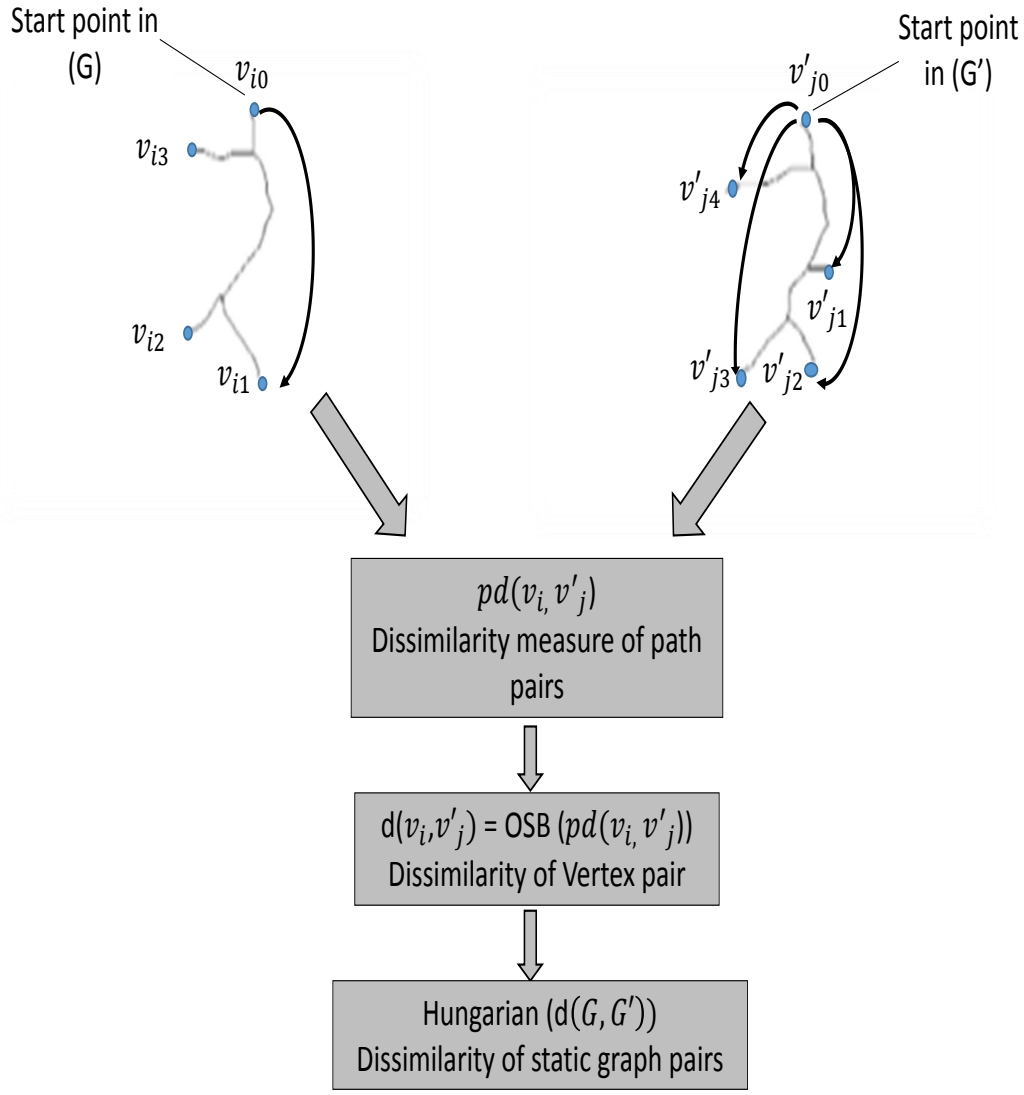
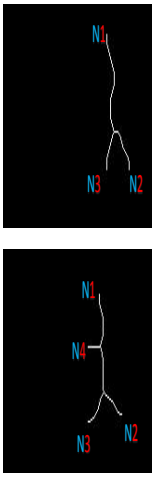



**Figure 5.4: Framework to compute the dissimilarity of end nodes between two graphs using path distance matrix and optimal subsequence bijection algorithm (OSB).**

Given two human action graph sequences $\mathcal{G} : G_1, G_2, \ldots, G_n$ and $\mathcal{G}' : G'_1, G'_2, \ldots, G'_m$, to consider whether two sequences represent the same action, the distance between every pair of graphs $d(G_i, G'_j)$ is first calculated. Dynamic Time Warping (DTW) [133] is applied to find the minimum cost matching $d(\mathcal{G}, \mathcal{G}')$ between two sequences, which helps eliminate the impact of spatio-temporal variations such as walking at different




| | N1 | N2 | N3 | N4 |
|---|---|---|---|---|
| N1 | 0.0322 | 0.4106 | 0.6031 | 0.8272 |
| N2 | 11.0096 | 0.5188 | 5.4467 | 4.2313 |
| N3 | 12.4317 | 6.4898 | 0.3437 | 9.2047 |
| N4 | const | const | const | const |

**Figure 5.5: Distance matrix between vertex pairs from two graphs for a Walking action. The corresponding pairs that contribute to the minimum cost assignment are highlighted.**

speeds.

### 5.2.3 Frame Selection and Action Alignment

In principle, the previously defined distance measure between dynamic skeleton graphs is sufficient to identify closely matched actions. However, a typical video sequence can be quite long (e.g. containing over 100 frames), which makes computation slow. Based on the observations that a relatively short video is usually sufficient to determine the action, and poses usually only change slightly between adjacent frames, we only select one in every $K$ frames from the first $M$ frames. As we will show later in experiments (see section 5.3), the classification performance is stable when $M$ is sufficiently large and $K$ is sufficiently small, which according to our experiments, can be achieved with $M = 50$ (2 seconds for 25 fps videos) and $K = 3$, which means only 17 representative frames are needed. An example is shown in Figure 5.6, where selected frames are highlighted with green borders. This substantially reduces the matching cost while

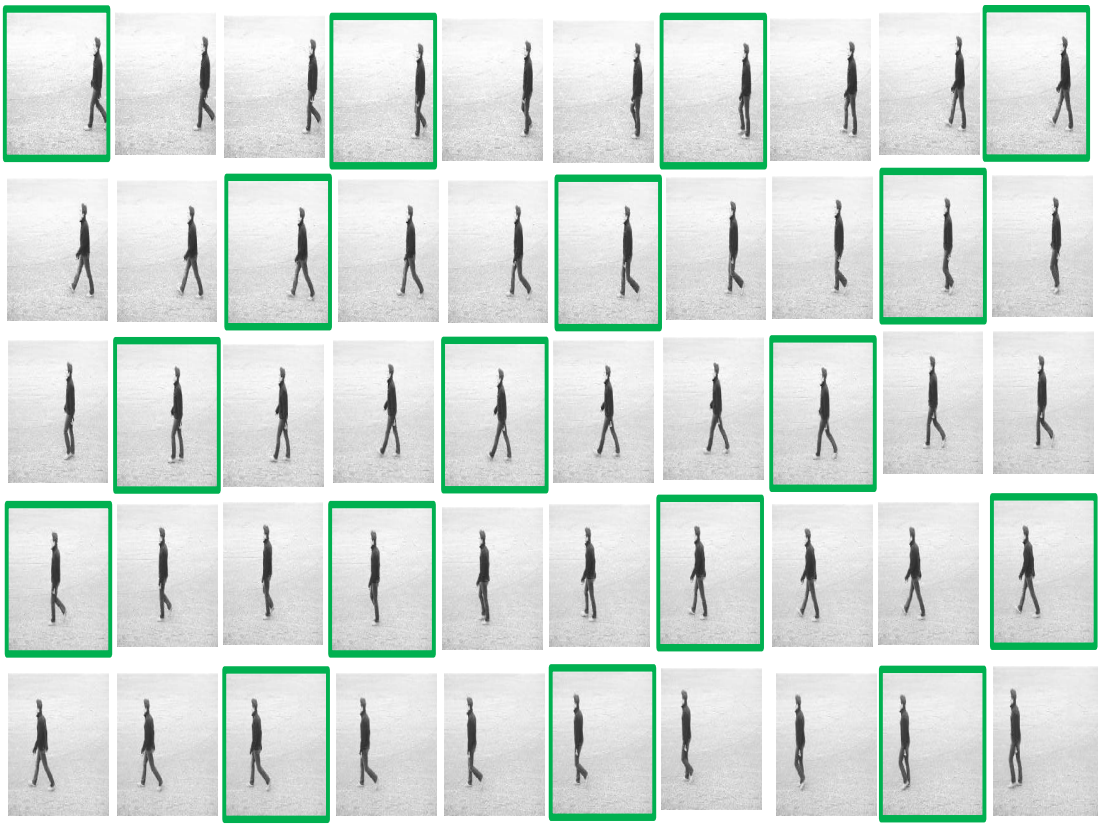


**Figure 5.6: An example of Walking action. Selecting every $K = 3$ frames from the first $M = 50$ frames of the video (giving a total of 17 selected frames shown with green borders) is sufficient to characterise the action for recognition, while substantially reducing the time complexity.**

keeping the recognition rate.

Some common actions are naturally periodic, e.g. walking, running etc. When such actions are captured in videos, however, they may start at any time of the cycle. Such misalignment cannot be effectively coped with using Dynamic Time Warping. A trivial solution would consider all the potential starting frames, and try to find the minimum distance measure. This however is computationally expensive. We instead propose to first detect periodic actions, and automatically choose a *consistent starting* frame to

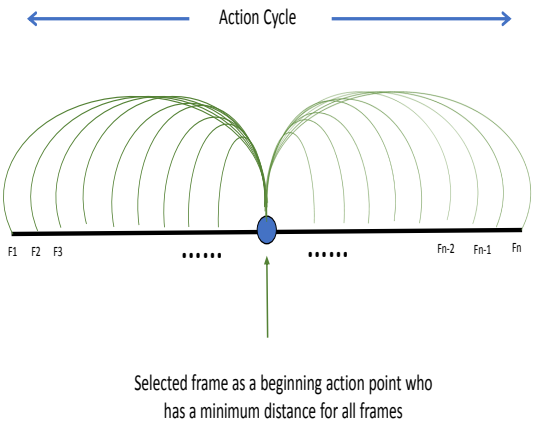



**Figure 5.7: Automatic selection of consistent starting frame for periodic actions based on the total minimum distance. Each video frame is compared with all the other frames in the cycle, and the frame with the minimum total distance is selected as the starting frame.**

make video frames of cyclic actions temporally aligned.

Given an input video, we first need to identify whether this video represents a periodic action, and if so, what is the cycle in terms of the number of frames. To make this possible, we assume that the input video is long enough to have sufficient content of two cycles. To robustly identify cycles, we take a sliding window of $r$ contiguous frames and find the best shift that leads to minimum total distance. The length of the cycle $c^*$ is chosen to be the cycle $c$ that gives the minimum total distance $d^*$:

$$d^* = \min_{c,t} \sum_{i=0}^{r-1} d(G_{t+i}, G_{t+i+c}), \tag{5.2}$$

where $t$ is the first frame of the former window, which is chosen such that frames of both windows for comparison are within the available frames from the video. We further denote $t^*$ as the optimal first frame. We treat a given video as containing a periodic action if $d^* < \delta$. The parameters window size $r$ and $\delta$ are chosen empirically and they are fixed in our experiments. The window size $r = 6$ and $\delta = 0.4$ are used

in our experiments. Table 5.1 shows the distances between two cycles with different window size of $r$ and shift.

Note that this formulation may choose cycles which are multiples of true cycles. This however does not cause a problem as we are only concerned to check whether a given video is periodic, and if so to reliably identify a starting frame.

After finding the cycle of the action, we take one cycle of the action, and identify a consistent starting frame $s^*$, which has a minimum distance to all other frames in the cycle (see Figure 5.7):

$$s^* = \underset{t^* \leq s < t^* + c^*}{\arg\min} \sum_{i=t^*}^{t^*+c^*-1} d(G_s, G_i). \tag{5.3}$$

This is well defined, even if the extracted cycles from different videos are originally misaligned. Figure 5.8 shows two different videos for the walking action with different environment conditions. In the first video the third frame with blue border is considered as the starting frame and in the second video the first frame is treated as the starting frame. The periodic actions are temporally aligned.

### 5.2.4 Hierarchical Matching

When using our dynamic graph matching technique for action recognition, in principle, we only need to compare a test video with each training video to find the one with the minimum distance, the category of which is then assigned to the test video. In practice, however, if there are a large number of training videos, this can be slow. Figure 5.9 shows the comparison between clustered and exhaustive matching in terms of time complexity and number of matching. To speed it up, we propose to classify a given video in a hierarchical manner using $k$-means clustering.

To explain this, in the training stage, we apply $k$-means clustering on the training set to cluster videos of each category into $k$ centres ($k = 5$ is used in our experiments). In the

**Table 5.1: Distance between two cycles with different window size ($r$) and shift (Walking action).**

| Shift | $r$ = 6 | $r$ = 7 | $r$ = 8 | $r$ = 9 |
|---|---|---|---|---|
| 8 | 2.0346 | 2.6213 | 2.6924 | 2.8123 |
| 9 | 4.5679 | 4.6341 | 4.7531 | 4.8179 |
| 10 | 1.7812 | 1.7903 | 2.0381 | 2.2345 |
| 11 | 2.6314 | 2.7631 | 2.9437 | 2.9978 |
| 12 | 1.1691 | 1.3231 | 1.5064 | 1.7103 |
| 13 | 3.4232 | 3.5341 | 3.7021 | 3.9012 |
| 14 | 1.1816 | 1.2949 | 1.6321 | 1.6945 |
| 15 | 0.7318 | 0.8413 | 0.9546 | 0.9934 |
| 16 | 0.3325 | 0.5361 | 0.7653 | 0.8765 |
| 17 | 3.0678 | 3.6539 | 4.5342 | 4.9398 |
| 18 | 4.5612 | 4.9376 | 5.2647 | 6.1523 |

classification stage, for a given video, we first compare its dynamic skeleton graph with all the centres, assuming the minimum distance of all the centres is $\tilde{d}$, then we choose all the videos whose cluster centre has a distance within $\eta\tilde{d}$. $\eta = 1.5$ is used in our experiments as shown in Figure 5.10. We use $\eta\tilde{d}$ as the criterion to consider potential videos which have a distance sufficiently close to the minimum distance. Among all the chosen videos, the one that best matches the test video (i.e. with the minimum graph distance) decides the classification category.

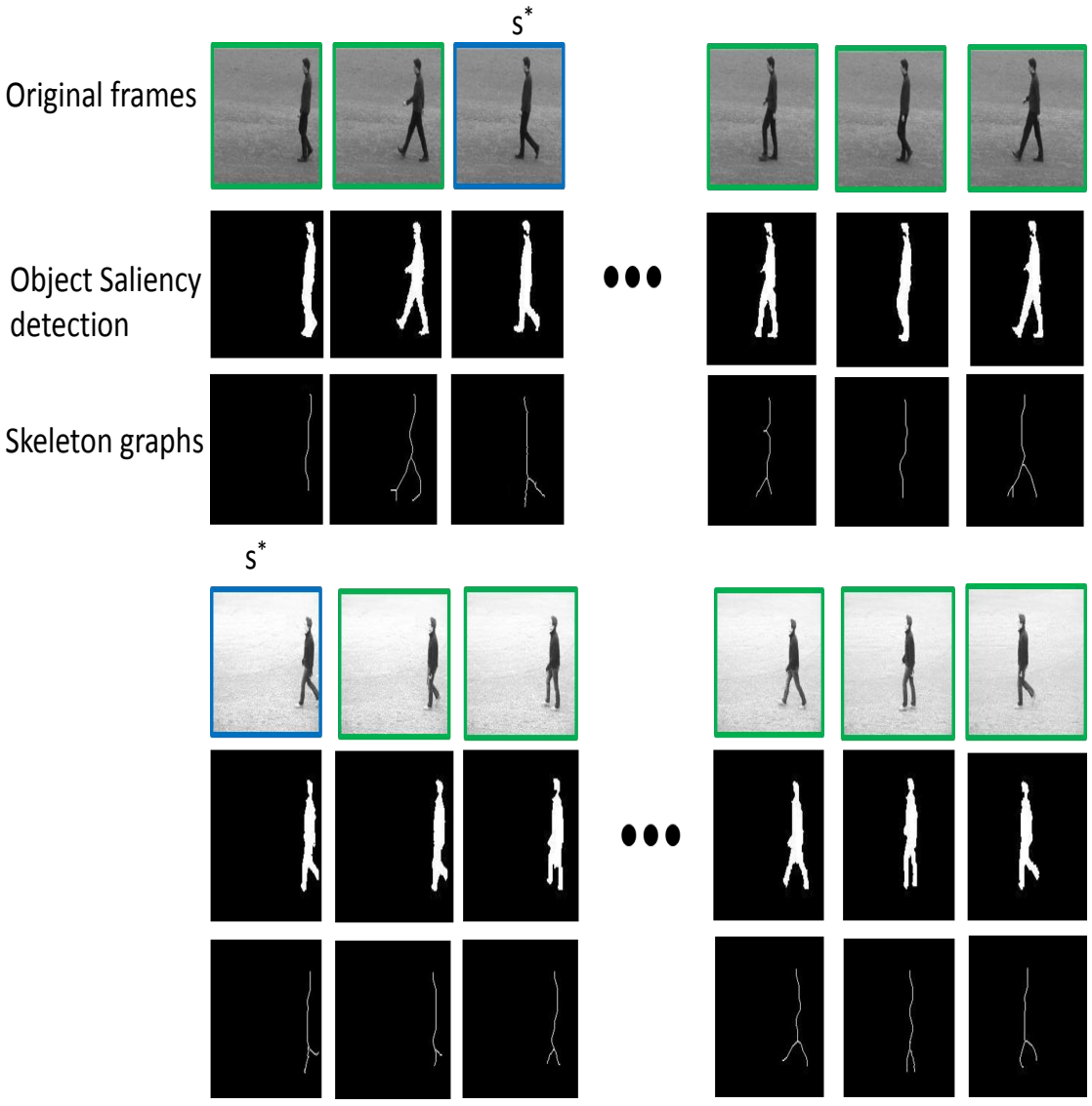



**Figure 5.8: An example of detection of starting frames in the walking action for action alignment. The frame with blue border is the detected starting frame $s^*$ with the minimum total distances to other frames in the same cycle.**

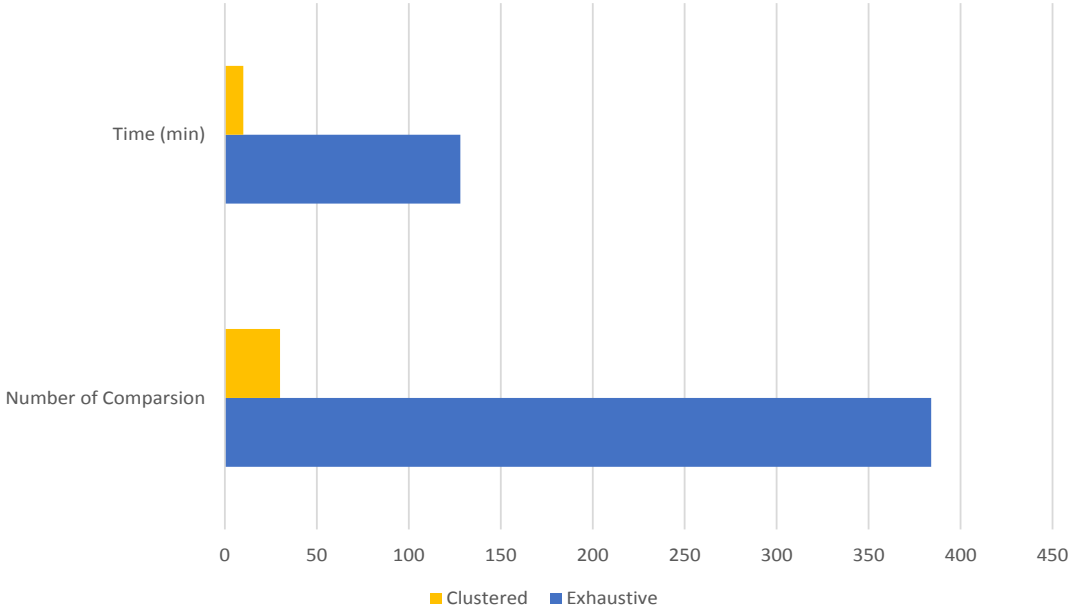



**Figure 5.9: Comparison between clustered and exhaustive matching in terms of time complexity and number of matching.**

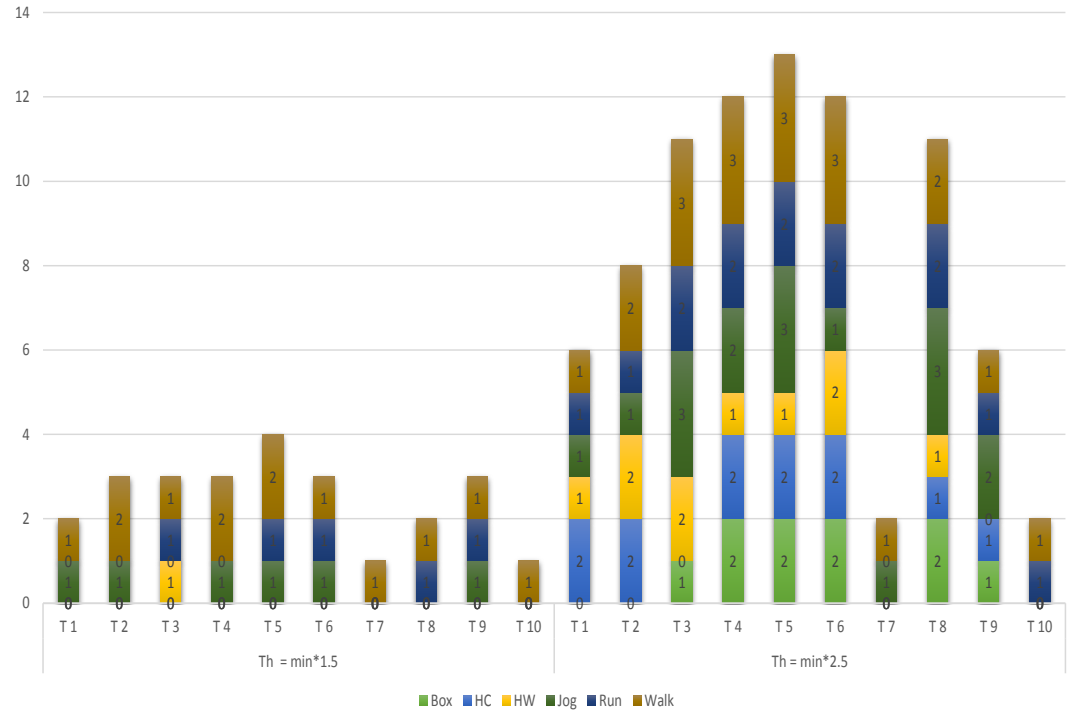



**Figure 5.10: Number of selected members from the cluster with a distance within $\eta\tilde{d}$. $\eta = 1.5, 2.5$.**

**Table 5.2: Experimental results (KTH dataset) by using different number of frames $M$ with $K$ = 3. $T_{pair}$ is the time for calculating dissimilarity between video pair and $T_{test}$ is the time for matching each test video to the training set.**

| $M$, $K$ | **Positive** | **Negative** | **Acc.** | $T_{pair}$ | $T_{test}$ | **#frames** |
|---|---|---|---|---|---|---|
| 100, 3 | 841 | 23 | 0.974 | 3-4 mins | 64 mins | 33 |
| 70, 3 | 841 | 23 | 0.974 | 2-3 mins | 48 mins | 23 |
| **50, 3** | **841** | **23** | **0.974** | **1-2 mins** | **32 mins** | **17** |
| 30, 3 | 808 | 56 | 0.934 | $\leqslant$ 1 min | 16 mins | 10 |
| 20, 3 | 778 | 86 | 0.90 | $\leqslant$ 1 min | 4.8 mins | 7 |

**Table 5.3: Experimental results (KTH dataset) by using different $K$ with $M$ = 50. $T_{pair}$ is the time for calculating dissimilarity between video pair and $T_{test}$ is the time for matching each test video to the training set.**

| $M$, $K$ | **Positive** | **Negative** | **Acc.** | $T_{pair}$ | $T_{test}$ | **#frames** |
|---|---|---|---|---|---|---|
| 50, 2 | 841 | 23 | 0.974 | 2-3 mins | 48 mins | 25 |
| **50, 3** | **841** | **23** | **0.974** | **1-2 mins** | **32 mins** | **17** |
| 50, 4 | 839 | 25 | 0.971 | $\leqslant$ 1 min | 16 mins | 12 |
| 50, 5 | 813 | 51 | 0.941 | $\leqslant$ 1 min | 16 mins | 10 |

### 5.2.5 Feature Fusion Schemes

Fusing multiple features, especially complementary ones, can be an effective method to boost the performance of recognition systems in computer vision [48, 21]. Generally, there are two types of fusion methods [145] namely: fusion at the feature level or early fusion and fusion at the classifier level or late fusion (see Figure 5.11), where figures 5.11 (a) and (b) represent early fusion schemes and figure 5.11 (c) depicts the late fusion scheme. Fusion at the feature level or representation level is performed using simple early fusion technique i.e. concatenation of features one after another. Late fusion is employed in this work to achieve fusion at the classifier level (SVM and graph matching classifiers).

### 5.2.6 Fusion with Image Descriptor based Method

Since our method is a shape-based method and only exploits the dynamics of foreground shapes, it provides complementary information to those widely used image descriptors. Thus it is reasonable to fuse the outputs of both our method and an existing image descriptor based classifier. For this purpose, we use our proposed method [1], which uses multi-class SVM to predict the probability of a test video belonging to each action category. For simplicity, we use $\mathcal{G}$ to represent both the skeleton graph and the video based on the context. Given a test video $\mathcal{G}_t$, let us denote the minimum distance between $\mathcal{G}_t$ and training videos in action category $l$ as $d_l(\mathcal{G}_t)$, the probability of $\mathcal{G}_t$ belonging to category $l$ as $P_l(\mathcal{G}_t)$ [1]. We combine the information in a uniform way to obtain the following preference score for $\mathcal{G}_t$ to belong to category $l$:

$$\tilde{P}_l(\mathcal{G}_t) = \omega \exp\left\{-\frac{d^2(\mathcal{G}_t)}{\sigma^2}\right\} + (1-\omega)P_l(\mathcal{G}_t), \tag{5.4}$$

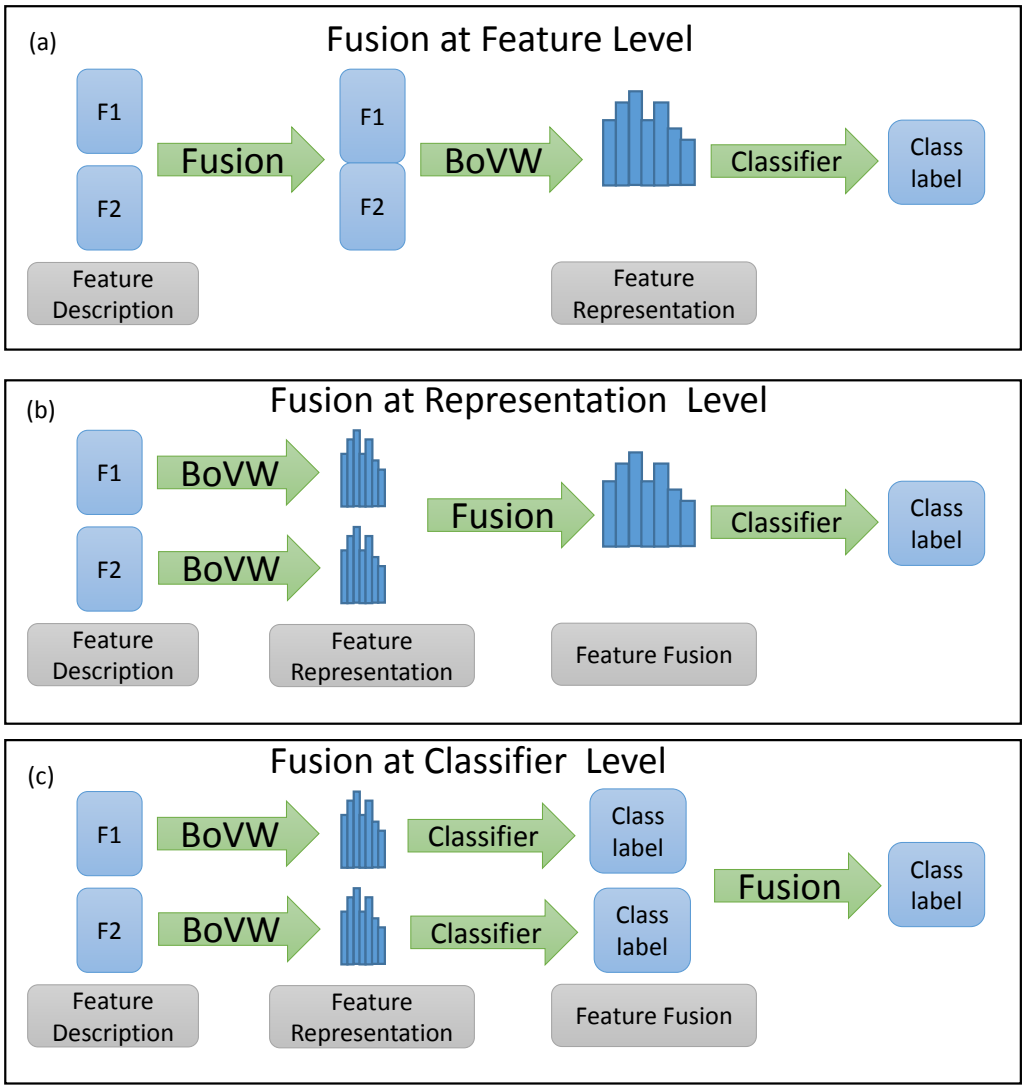



**Figure 5.11: Different Fusion Schemes: Early Fusion (a and b) and Late Fusion (c).**

where $\exp(\cdot)$ makes both terms in the same range of $[0, 1]$, $\sigma$ is a parameter to control the mapping of distance to the probability and $\omega$ is used to balance the importance between shape-based classification and image based classification. In our experiments, the parameters $\sigma$ and $\omega$ are automatically chosen using grid search and cross validation on the training set. As expected, the performance on KTH, the UCF-Sports and

Olympic sports datasets using the combined approach is significantly better, achieving $98.6\%$ accuracy on the KTH dataset, $93.1\%$ on the UCF Sports dataset, and $82.8\%$ on Olympic sports dataset. More discussions will be provided in the next subsection.

## 5.3 Experimental Results

We now demonstrate the performance of the proposed method with comprehensive experiments on the KTH [142], UCF Sports [137] and Olympic sports [119] benchmark datasets.

### 5.3.1 Parameters and Running Times

As explained in Sec. 5.2.3 in the experiments, we take every $K$ frames from the first $M$ frames for matching, to achieve trade-off between the accuracy and the efficiency.

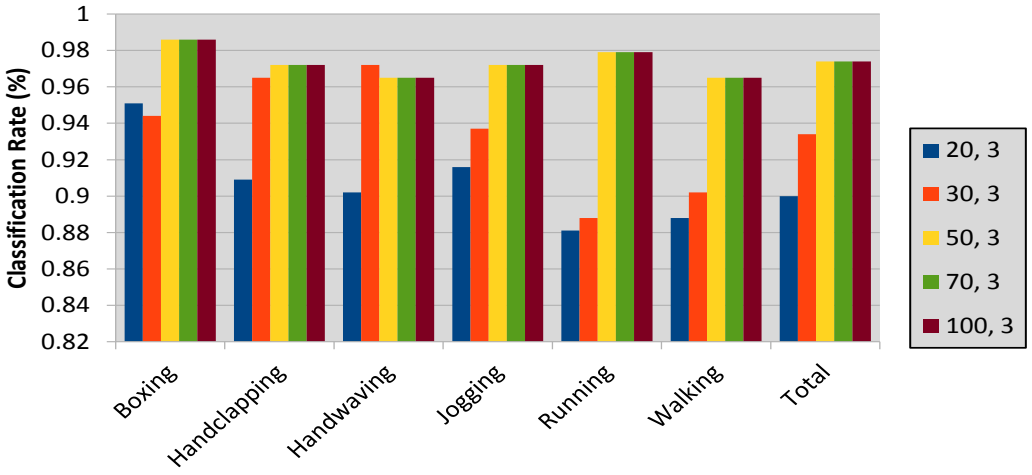



**Figure 5.12: Recognition rate of KTH dataset using different number of $M$ with $K$ = 3.**

To demonstrate the impact of these parameters, we use the KTH dataset and show the classification accuracy with varying $M$ and $K$. As shown in Tables 5.2 and 5.3, the classification rate stays stable when $M \geq 50$ and $K \leq 3$, so we choose $M = 50$, $K = 3$ in the remaining experiments which provides a good balance of accuracy and efficiency. Figures 5.12 and 5.13 depict our experiments to select the key frames. The figures show the recognition accuracy for each action and the overall accuracy of the KTH dataset using different $M$ and $K$. We tested our algorithm using MATLAB on a 2.5GHz Windows PC. With the current unoptimised code, using hierarchical matching reduces the running time for classification of a video for the KTH dataset from about 2 hours to 1-2 minutes, for the UCF Sports dataset from about 4 hours to 3-5 minutes and for Olympic sports dataset from about 8 hours to 6-8 minutes.

### 5.3.2 Performance on Standard Benchmarks

We compare the recognition rates of our method with the state-of-the-art methods on

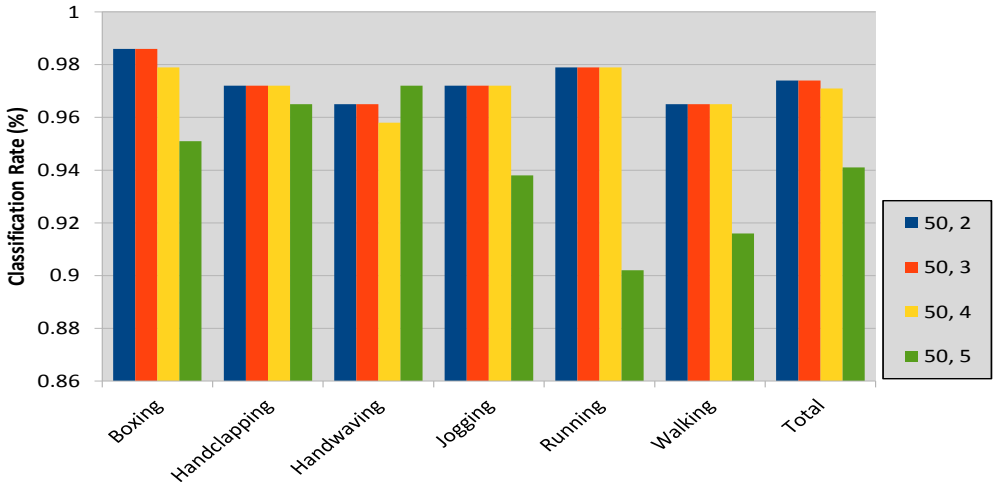



**Figure 5.13: Recognition rate of KTH dataset using different number of $K$ with $M$ = 50.**

**Table 5.4: Recognition accuracy comparison of the proposed graph matching and the fusion with image descriptor based method with the state-of-the-art methods on the KTH dataset.**

| Methods on KTH | Accuracy (%) |
|---|---|
| Somasundaram *et al.* [55] | 90.1 |
| Shuiwang *et al.* [68] | 90.2 |
| Ghamdi *et al.* [6] | 90.7 |
| Liu *et al.* [96] | 91.3 |
| Iosifidis *et al.* [64] | 92.1 |
| Baumann *et al.* [11] | 92.1 |
| Kalser [76] | 92.6 |
| Ji *et al.* [69] | 93.1 |
| Wang *et al.* [155] | 94.2 |
| Baccouche *et al.* [9] | 94.3 |
| Rapits and Soatto [135] | 94.8 |
| Zhang *et al.* [180] | 94.8 |
| Wang *et al.* [156] | 95.0 |
| Wang *et al.* [159] | 95.1 |
| Yuan *et al.* [178] | 95.4 |
| Liu *et al.* [99] | 95.8 |
| SGSH (Chapter 3) | 97.2 |
| **Proposed Graph Matching** | **97.7** |
| **Fusion of Graph Matching with SGSH (Chapter 3)** | **98.6** |

standard benchmarks. Our method using graph matching alone achieves action recognition rates of $97.7\%$ for the KTH dataset, $92.3\%$ for the UCF Sports dataset, and $80.5\%$ for the Olympic sports dataset which outperform state-of-the-art methods (see Tables 5.4, 5.5 and 5.6 for comparison with alternative methods). This shows that our

**Table 5.5: Recognition accuracy comparison of the proposed graph matching and the fusion with image descriptor based method with the state-of-the-art methods on the UCF-Sports dataset.**

| Methods on UCF Sports | Accuracy (%) |
|---|---|
| Raptis *et al.* [134] | 79.4 |
| Ma *et al.* [105] | 81.7 |
| Kalser [76] | 85.0 |
| Everts *et al.* [40] | 85.6 |
| Le *et al.* [90] | 86.5 |
| Somasundaram *et al.* [55] | 87.3 |
| Zhang *et al.* [180] | 87.5 |
| Wang *et al.* [156] | 88.0 |
| Wang *et al.* [159] | 88.6 |
| Ma *et al.* [104] | 89.4 |
| Ma *et al.* [106] | 89.4 |
| SGSH (Chapter 3) | 90.9 |
| **Proposed Graph Matching** | **92.3** |
| **Fusion of Graph Matching with SGSH (Chapter 3)** | **93.1** |

highly abstract method works well for both videos taken in a controlled environment (KTH) and more diverse real-world videos (UCF-Sports and Olympic sports datasets).

### 5.3.3 Performance with Single Training Examples

Our method uses a highly abstract representation, which in principle should be able to characterise an action with very few examples. To verify this, we performed a somewhat extreme test where only one video of each category is used as the training set

**Table 5.6: Recognition accuracy comparison of the proposed graph matching and the fusion with image descriptor based method with the state-of-the-art methods on the Olympic Sports dataset.**

| Methods on Olympic Sports | Accuracy (%) |
|---|---|
| Niebles *et al.* [119] | 62.5 |
| Tang *et al.* [151] | 66.8 |
| Liu *et al.* [96] | 74.3 |
| Wang *et al.* [156] | 77.2 |
| Brendel *et al.* [18] | 77.3 |
| **Proposed Graph Matching** | **80.5** |
| **Fusion of Graph Matching with SGSH (Chapter 3)** | **82.8** |

and all the remaining videos are used for testing. To avoid bias, we always take the first video for training. This test can be useful in real-world scenarios, e.g. to find actions similar to the one example chosen. The user does not need to prepare a comprehensive training set. This test however is challenging, so it is understandable that performance will drop compared with the standard setup. Table 5.7 compares our skeleton graph based method with a state-of-the-art method based on image descriptors. Our method still achieves decent performance: $95.2\%$ accuracy on the KTH dataset, $88.1\%$ accuracy on the UCF Sports dataset and $70.7\%$ accuracy on the Olympic Sports dataset whereas the performance of image descriptor based methods can suffer significantly. Our method achieves accuracies of over $20\%$ (for KTH and UCF Sports) and $10\%$ (for Olympic Sports) better than SGSH (Chapter 3) which achieves state-of-the-art performance in the standard setup. This demonstrates that by using the highly abstracted information, our method has good generalisability. With fewer examples also means our method performs much faster, as for each test video, it only needs to compare with several training videos (one for each category). In this test, our method takes 1-3 seconds for the KTH dataset, 4-6 seconds for the UCF Sports dataset and

**Table 5.7: Recognition accuracy comparison using *single training example* on the KTH, UCF-Sports and Olympic sports datasets.**

| Methods | KTH (%) | UCF Sports (%) | Olympic Sports (%) |
|---|---|---|---|
| **Our Proposed Method** | **95.2** | **88.1** | **70.7** |
| SGSH (Chapter 3) | 72.1 | 65.3 | 58.9 |

10-12 seconds for the Olympic sports dataset to classify an input video.

## 5.4 Discussion and Conclusions

We introduced in this chapter a novel action recognition method based on representing actions as deforming skeleton graphs of foreground subjects and computing similarity measures between them using optimal subsequence bijection based static graph matching and dynamic time warping for temporal alignment. We further developed methods to effectively select representative frames and consistent starting frames for periodic actions. Our method addresses fundamental issues of many existing approaches, such as sensitivity to changes of illumination and clothing. Similar actions are represented by similar graphs regardless of the various conditions in the environment. Our method outperforms state-of-the-art methods on standard benchmarks. It works well with very few training examples, and achieves decent accuracy even when only one training example is provided for each category. We also demonstrated that our shape based method provides complementary information to image descriptor based methods and even better accuracy is achieved when these methods are combined.

# *Chapter 6*

# Conclusions and Perspectives

## 6.1 Key Contributions

This thesis has presented and evaluated several contributions for action recognition in real video data. To conclude our work, we summarise our key contributions and discuss conclusions from our experiments we will then indicate interesting directions for future research in this field.

The thesis aims at developing automatic techniques for recognising actions in uncontrolled, real-world video data. Our first contribution is to use saliency to guide the extraction of local and global features which are then employed for action classification. For this, existing approaches to describe local information in videos are investigated and new methods are developed. In this work, we introduce a novel framework for human action recognition based on saliency guided local and global descriptors, by detecting keypoints only in salient regions and then describing those using 3D SIFT descriptors. We also propose to use video frame selection to discard all the frames without subjects. As a result, this reduces the time complexity since only keypoints on the salient objects need to be processed. Using a combination of local and global descriptors takes advantages of both descriptions to make the final descriptor informative and carry more powerful information about video content. Experiments show that the proposed method gives a significant improvement on the action recognition classification accuracy for benchmark datasets with different characteristics (realistic,

interaction and controlled datasets).

A further contribution is to introduce a new descriptor for action recognition in videos. We propose an effective feature descriptor called 3D GLOH (Gradient Location and Orientation Histogram), which describes local spatially varying information for video data. The 2D GLOH descriptor is extended to video frames by partitioning the cylindrical local neighbourhood of an interest point into spatio-temporal bins and calculating 3D histograms of gradients in the local bins. It detects interest points in the video and then describes them in 3D log-polar coordinates. Our approach is based on a log-polar orientations to compute 3D gradients locations histograms for salient keypoints. Descriptor parameters are evaluated in depth and optimised for action recognition using bag-of-features representation. The experimental results show that the proposed 3D GLOH descriptor is effective in capturing localised spatio-temporal information and the overall system outperforms existing methods in recognition accuracy for challenging real-world datasets, including UCF-Sport, TV-Human Interaction and UCF11 datasets.

Our last key contribution is extracting minimal representative information, namely deforming skeleton graphs corresponding to foreground shapes to effectively represent actions. We propose to represent actions in video sequences as sequences of deforming skeleton graphs of foreground subjects. The representation has significant advantages of being insensitive to changes of illumination, subject appearance and backgrounds to effectively represent actions, removing the influence of these typical variations. The proposed method is based on matching of deforming skeleton graphs. Our similarity measure takes into account topological variation, temporal variation and alignment of periodic actions to improve its robustness. Experimental results show that our method purely based on graph matching outperforms existing action recognition methods. Moreover, since our method uses compact and highly abstracted information, it achieves decent recognition performance with even a single example from each category, which is a very challenging scenario for existing methods. Due to the use of

complementary information, we achieve even better recognition performance by fusing our method with an alternative image descriptor based method.

## 6.2 Future work

The research presented in this thesis have raised more questions that it has answered. There are still many improvements that could be considered from this work which should be pursued.

**Saliency guidance for local and global features**: The idea of using saliency guidance to improve action recognition is general and in the future we would like to investigate combining this with alternative features as well as its use in other recognition applications. An interesting path for future work can be based on salient human tracks for multiple body parts, e.g., for head, upper body, and full body. First, this can help to render the tracking process more robust since additional constraints for relations between the body parts are available. Another direction is to include deep learning architectures in the saliency detection step in the pipeline of the proposed framework. Examples of the deep learning architectures that have been proved successful architectures to detect salient in image content [181], AlexNet [79] and GoogLeNet [149] which can be used in this step. We will investigate which neural network structure is best suited for saliency detection in video sequences and evaluate the performance of the different architectures.

**3D GLOH local features**: The proposed 3D GLOH descriptor can be useful for analysing videos, especially for those with rich textures. We would like to investigate its effectiveness in other video analysis and video quality assessment applications. Moreover, to increase the robustness of features some visual clues can be added such as appearance, motion, structure, and context information.

**Matching of deforming skeleton graphs**: A limitation of the deforming skeleton graph matching method is due to the nature of matching dynamic graph sequences, it

can be time consuming. We address this by developing a hierarchical matching strategy such that the detailed graph matching is only applied to promising candidates. We will exploit this idea further by e.g. identifying candidates using some cascaded filtering strategy. Another limitation is that at the moment we assume the foreground object in each frame can be well represented by a single skeleton. This works sufficiently well for single subjects, but does not work when multiple subjects are involved. In the future we will investigate extending our framework to cope with cases when foreground involves multiple skeleton graphs by exploring the prior knowledge of human skeleton structure.

# Bibliography

[1] A. Abdulmunem, Y.-K. Lai, and X. Sun. Saliency guided local and global descriptors for effective action recognition. *Computational Visual Media*, 2(1):97–106, 2016.

[2] A. Abdulmunem, Y.-K. Lai, and X. Sun. 3D GLOH features for human action recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 805–810, 2017.

[3] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428 – 440, 1999.

[4] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, April 2011.

[5] J.K. Aggarwal and Lu Xia. Human activity recognition from 3D data: A review. *Pattern Recognition Letters*, 48:70 – 80, 2014.

[6] M. Al Ghamdi, L. Zhang, and Y. Gotoh. Spatio-temporal SIFT and its application to human action classification. In *European conference on Computer vision (ECCV)*, pages 301–310, 2012.

[7] 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.

[8] N. Ben Aoun, M. Mejdoub, and C. Ben Amar. Graph-based approach for human action recognition using spatio-temporal features. *Journal of Visual Communication and Image Representation*, 25(2):329 – 338, 2014.

[9] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding (HBU)*, pages 29–39, 2011.

[10] X. Bai and L.J. Latecki. Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 30(7):1282–1292, July 2008.

[11] F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao. Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing*, 173(1):54–63, 2016.

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

[13] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE on Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[14] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. N. Vishwanathan, A. J. Smola, and H. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(1):47–56, January 2005.

[15] K. Bousmalis, M. Mehu, and M. Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203 – 221, 2013.

[16] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.

[17] 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 (CVPR)*, pages 1948–1955, 2009.

[18] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *International Conference on Computer Vision (ICCV)*, pages 778–785, Nov 2011.

[19] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, volume 3024 of *Lecture Notes in Computer Science*, pages 25–36, 2004.

[20] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le, and A. J. Smola. Learning graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1048–1058, June 2009.

[21] Z. Cai, L. Wang, X. Peng, and Y. Qiao. Multi-view super vector for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 596–603, 2014.

[22] F. Caillette, A. Galata, and T. Howard. Real-time 3d human body tracking using learnt models of behaviour. *Computer Vision and Image Understanding*, 109(2):112 – 125, 2008.

[23] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, pages 27:1–27:27, 2011.

[24] K. Chang, T. Liu, H. Chen, and S. Lai. Fusing generic objectness and visual saliency for salient object detection. In *International Conference on Computer Vision (ICCV)*, pages 914–921, 2011.

[25] J. M. Chaquet, E. J. Carmona, and An. FernÃ¡ndez-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633 – 659, 2013.

[26] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1932–1939, 2009.

[27] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu. Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–416, June 2011.

[28] J. Cho, M. Lee, H. Jin Chang, and S. Oh. Robust action recognition using local motion and group sparsity. *Pattern Recognition*, 47(5):1813 – 1825, 2014.

[29] O. Chomat and J. L. Crowley. Recognizing motion using local appearance. In *University of Edinburgh*, pages 271–279, 1998.

[30] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 2843–2851. Curran Associates Inc., 2012.

[31] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893, 2005.

[32] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on Computer vision (ECCV)*, volume 3952, pages 428–441, 2006.

[33] D. Das Dawn and S. H. Shaikh. A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, 32(3):289–306, 2015.

[34] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and perforamnce Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[35] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *International Conference on Computer Vision (ICCV)*, pages 1792–1799, Nov 2011.

[36] M. Edwards, J. Deng, and X. Xie. From pose to activity: Surveying datasets and introducing converse. *Computer Vision and Image Understanding*, 144(Supplement C):73 – 105, 2016. Individual and Group Activities in Video Event Analysis.

[37] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision (ICCV)*, pages 726–733 vol.2, 2003.

[38] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–571–I–578 vol.1, 2003.

[39] I. Everts, J. C. van Gemert, and T. Gevers. Evaluation of color stips for human action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2850–2857, June 2013.

[40] 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 (CVPR)*, pages 2850–2857, 2013.

[41] Y. Fang, W. Lin, Z. Chen, C. M. Tsai, and C. W. Lin. A video saliency detection model in compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(1):27–38, Jan 2014.

[42] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.

[43] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[44] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. In *Proceedings. First International Symposium on 3D Data Processing Visualization and Transmission*, pages 717–721, 2002.

[45] W. Forstner and E. Gulch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop. Interlaken*.

[46] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. *International Journal of Computer Vision*, 107(3):219–238, 2014.

[47] 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 *International Conference on Computer Vision (ICCV)*, pages 2595–2602, Nov 2011.

[48] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *International Conference on Computer Vision (ICCV)*, 2009.

[49] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, June 2014.

[50] S. Goferman, A. Tal, and L. Zelnik-Manor. Puzzle-like collage. *Computer Graphics Forum*, 29(2):459–468, 2010.

[51] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2376–2383, 2010.

[52] 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):2247–2253, Dec 2007.

[53] G. Goudelis, K. Karpouzis, and S. Kollias. Exploring trace transform for robust human action recognition. *Pattern Recognition*, 46(12):3238 – 3248, 2013.

[54] G. Guo and A. Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343 – 3361, 2014.

[55] S. Guruprasad, A. Cherian, V. Morellas, and N. Papanikolopoulos. Action recognition using global spatio-temporal features derived from sparse representations. *Computer Vision and Image Understanding*, 123(Supplement C):1 – 13, 2014.

[56] Z. Yuan T. Liu H. Jiang, J. Wang and N. Zheng. Automatic salient object segmentation based on context and shape prior. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 110.1–110.12, 2011.

[57] Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[58] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[59] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, pages 185–204, 1981.

[60] Berthold P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185 – 203, 1981.

[61] E. Ijjina and K. Chalavadi. Human action recognition using genetic algorithms and convolutional neural networks. *Pattern Recognition*, 59:199–212, 2016.

[62] N. İkizler and P. Duygulu. Human action recognition using distribution of oriented rectangular patches. In *Proceedings of the 2nd Conference on Human Motion: Understanding, Modeling, Capture and Animation*, pages 271–284, Berlin, Heidelberg, 2007. Springer-Verlag.

[63] N. İkizler and David A. Forsyth. Searching for complex human activities with no visual examples. *International Journal of Computer Vision*, 80(3):337–357, 2008.

[64] A. Iosifidis, A. Tefas, and I. Pitas. Discriminant bag of words based representation for human action recognition. *Pattern Recognition Letters*, 49:185 – 192, 2014.

[65] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Transactions on Image Processing*, 13(10):1304–1318, October 2004.

[66] C. Koch J. Harel and P. Perona. Graph-based visual saliency. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2006.

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

[68] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[69] Y. Ji, A. Shimada, H. Nagahara, and R. Taniguchi. A compact descriptor CHOG3D and its application in human action recognition. *IEEE Transactions on Electrical and Electronic Engineering*, 8(1):69–77, 2013.

[70] Y. Jiang, X. Xue, and C. Ngo. Trajectory-based modeling of human actions with motion reference point. In *European conference on Computer vision (ECCV)*, 2012.

[71] Z. Jiang, Z. Lin, and L. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):533–547, 2012.

[72] J.Ohya J.Yamato and Kenichiro. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 379–385, 1992.

[73] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2479, June 2010.

[74] S. Karaman, L. Seidenari, S. Ma, A. Del Bimbo, and S. Sclaroff. Adaptive structured pooling for action recognition. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[75] W. Kim, C. Jung, and C. Kim. Spatiotemporal saliency detection and its applications in static and dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):446–456, April 2011.

[76] A. Kläser. *Learning human actions in video*. PhD thesis, Universitë de Grenoble, Jul 2010.

[77] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference (BMVC)*, pages 995–1004, 2008.

[78] John F. Kolen and Stefan C. Kremer. *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*. 2001.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[80] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre. *HMDB51: A Large Video Database for Human Motion Recognition*, pages 571–582. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[81] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, pages 282–289, San Francisco, CA, USA, 2001.

[82] L. Lam, S. Lee, and C. Y. Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):869–885, September 1992.

[83] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *International Conference on Computer Vision (ICCV)*, pages 2003–2010, 2011.

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

[85] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision (ICCV)*, pages 432–439, 2003.

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

[87] I. Laptev and P. Perez. Retrieving actions in movies. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, Oct 2007.

[88] L. Jan Latecki, Q. Wang, S. Koknar-Tezel, and V. Megalooikonomou. Optimal subsequence bijection. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 565–570, 2007.

[89] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006.

[90] 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 (CVPR)*, CVPR '11, pages 3361–3368, 2011.

[91] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, 2010.

[92] H. Li and M. Greenspan. Multi-scale gesture recognition from time-varying contours. In *10th IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 236–243 Vol. 1, 2005.

[93] W. T. Li, H. S. Chang, K. C. Lien, H. T. Chang, and Y. C. F. Wang. Exploring visual and motion saliency for automatic video object extraction. *IEEE Transactions on Image Processing*, 22(7):2600–2610, July 2013.

[94] X. Li, Y. Li, C. Shen, A. Dick, and A. van den Hengel. Contextual hypergraph modeling for salient object detection. In *IEEE Conference on Computer Vision (ICCV)*, Sydney, Australia, 2013.

[95] X. Liang, L. Lin, and L. Cao. Learning latent spatio-temporal compositional model for human action recognition. *CoRR, 2015*, abs/1502.00258, 2015.

[96] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

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

[98] J. Liu, Y. Yang, I. Saleemi, and M. Shah. Learning semantic features for action recognition via diffusion maps. *Comput. Vis. Image Underst.*, 116(3):361–377, March 2012.

[99] M. Liu, H. Liu, Q. Sun, T. Zhang, and R. Ding. Salient pairwise spatio-temporal interest points for real-time activity recognition. *CAAI Transactions on Intelligence Technology*, 1(1):14 – 29, 2016.

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

[101] W. Lu and James J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. In *European conference on Computer vision (ECCV)*, pages 49–60, 2006.

[102] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. pages 674–679, 1981.

[103] Q. Ma and L. Zhang. Saliency-based image quality assessment criterion. In *Proceedings of the 4th International Conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications - with Aspects of Theoretical and Methodological Issues*, ICIC '08, pages 1124–1133, Berlin, Heidelberg, 2008. Springer-Verlag.

[104] S. Ma, L. Sigal, and S. Sclaroff. Space-time tree ensemble for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[105] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *International Conference on Computer Vision (ICCV)*, pages 2744–2751, 2013.

[106] S. Ma, J. Zhang, S. Sclaroff, N. Ikizler-Cinbis, and L. Sigal. Space-time tree ensemble for action recognition and localization. *International Journal of Computer Vision*, 2017.

[107] Y. Ma, X. Hua, L. Lu, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.

[108] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[109] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1139–1146, 2013.

[110] P. Matikainen, M. Hebert, and R. Sukthankar. Representing pairwise spatial and temporal relations for action recognition.

[111] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 104–111, 2009.

[112] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European conference on Computer vision (ECCV)*, pages 128–142, 2002.

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

[114] A. Morales-Gonzälez, N. Acosta-Mendoza, A. Gago-Alonso, E. B. Garcïa-Reyes, and J. E. Medina-Pagola. A new proposal for graph-based image classification using frequent approximate subgraphs. *Pattern Recognition*, 47(1):169 – 177, 2014.

[115] V. Mota, J. Souza, A. Albuquerque Araújo, and M. Vieira. Combining orientation tensors for human action recognition. In *2013 XXVI Conference on Graphics, Patterns and Images*, pages 328–333, Aug 2013.

[116] P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In *Proceedings of the IEEE Workshop on Motion and Video Computing (WMVC)*, pages 10–, 2007.

[117] P. Natarajan and R. Nevatia. Online, real-time tracking and recognition of human actions. In *IEEE Workshop on Motion and video Computing*, pages 1–8, Jan 2008.

[118] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Computer Vision and Pattern Recognition*, 2015.

[119] J. C. Niebles, C. Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on Computer vision (ECCV)*, pages 392–405, 2010.

[120] J. C. Niebles, H. Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[121] J.C. Niebles and Fei-Fei Li. A hierarchical model of shape and appearance for human action classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[122] A. Oikonomopoulos and M. Pantic. *Human Activity Recognition Using Hierarchically-Mined Feature Constellations*, pages 150–159. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[123] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, And Cybernetics*, 36(3):710–719, 2006.

[124] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:831–843, 1999.

[125] P.-Z. Pan and C.-L. Huang. Human action recognition based on dense trajectories analysis and random forest. *Journal of Electronic Science and Technology*, 14(4):370–376, 2016.

[126] A. Patron, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in TV shows. In *British Machine Vision Conference (BMVC)*, pages 50.1–11, 2010.

[127] M. Pelillo, K. Siddiqi, and S. W. Zucker. Matching hierarchical structures using association graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1105–1120, 1999.

[128] P. Peursum, S. Venkatesh, and G. West. Tracking-as-recognition for articulated full-body human motion analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[129] R. Polana and R. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Nov 1994.

[130] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[131] T. Qu, Y. Liu, J. Li, and M. Wu. Action recognition using multi-layer topographic independent component analysis. *Journal of Information and Computational Science*, 12(9):3537–3546, 2015.

[132] A. Ramadass, M. Suk, and B. Prabhakaran. Feature extraction method for video based human action recognitions: Extended optical flow algorithm. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1106–1109, March 2010.

[133] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood. View-invariant alignment and matching of video sequences. In *IEEE International Conference on Computer Vision (ICCV)*, pages 939–945 vol.2, 2003.

[134] 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 (CVPR)*, pages 1242–1249, 2012.

[135] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *European conference on Computer vision (ECCV)*, volume 6311, pages 577–590. 2010.

[136] N. M. Robertson and I. D. Reid. A general method for human activity recognition in video. *Computer Vision and Image Understanding*, 104:232–248, 2006.

[137] M. D. 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 (CVPR)*, pages 1–8, June 2008.

[138] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores. A survey on ontologies for human behavior recognition. *ACM Comput. Surv.*, 46(4):43:1–43:33, 2014.

[139] P. K. Saha, G. Borgefors, and G. S. di Baja. A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters*, 76:3 – 12, 2016. Special Issue on Skeletonization and its Application.

[140] G. Salton. *Automatic Information Organisation and Retrieval.* McGraw Hill Text, 1968.

[141] H. Scharr. Optimal filters for extended optical flow. In *Complex Motion*, volume 3417 of *Lecture Notes in Computer Science*, pages 14–29. 2007.

[142] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition (ICPR)*, pages 32–36, 2004.

[143] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. In *ACM International Conference on Multimedia*, pages 357–360, 2007.

[144] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014.

[145] C. M. Snoek, M. Worring, and A. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the Annual ACM International Conference on Multimedia (MULTIMEDIA)*, pages 399–402, 2005.

[146] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[147] J. Sun, X. Wu, S. Yan, L. F. Cheong, T. S. Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2004–2011, June 2009.

[148] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshops*, pages 58–65, June 2009.

[149] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[150] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, June 2010.

[151] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1250–1257, June 2012.

[152] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *European conference on Computer vision (ECCV)*, pages 140–153, 2010.

[153] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[154] C. Tomasi and T. Kanade. Detection and tracking of point features. *International Journal of Computer Vision*, 1991.

[155] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176, June 2011.

[156] H. Wang, A. Klaser, C. Schmid, and C. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103:60–79, 2013.

[157] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[158] L. Wang, L. Ge, R. Li, and Y. Fang. Three-stream CNNs for action recognition. *Pattern Recognition Letters*, 92:33 – 40, 2017.

[159] L. Wang, R. Li, and Y. Fang. Power difference template for action recognition. *Machine Vision and Applications*, 2017.

[160] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305–4314, 2015.

[161] L. Wang and H. Sahbi. Directed acyclic graph kernels for action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3168–3175, 2013.

[162] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and . Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36, 2016.

[163] H. Wei, Q. Yu, and C. Yang. Shape-based object recognition via evidence accumulation inference. *Pattern Recognition Letters*, 77:42 – 49, 2016.

[164] G. Willems, T. Tuytelaars, and Luc Van G. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on Computer vision (ECCV)*, volume 5303, pages 650–663. 2008.

[165] B. Wu, C. Yuan, and W. Hu. Human action recognition based on context-dependent graph kernels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2609–2616, 2014.

[166] J. Wu, Y. Zhang, and W. Lin. Towards good practices for action video encoding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2577–2584, 2014.

[167] Z. Wu, Y. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang. Fusing multi-stream deep networks for video classification. *CoRR*, abs/1509.06086, 2015.

[168] Z. Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the ACM International Conference on Multimedia*, MM '15, pages 461–470, 2015.

[169] C. Koch X. Hou, J. Harel. Image signature: Highlighting sparse salient regions. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 34, pages 194–201, 2012.

[170] H. Xu, Q. Tian, Z. Wang, and J. Wu. A joint evaluation of different dimensionality reduction techniques, fusion and learning methods for action recognition. *Neurocomputing*, 214:329 – 339, 2016.

[171] X. Yan and J. Han. gSpan: graph-based substructure pattern mining. In *IEEE International Conference on Data Mining (ICDM)*, pages 721–724, 2002.

[172] X. Yan and Y. Luo. Making full use of spatial-temporal interest points: An adaboost approach for action recognition. In *IEEE International Conference on Image Processing*, pages 4677–4680, 2010.

[173] C. Yang, C. Feinen, O. Tiebe, K. Shirahama, and M. Grzegorzek. Shape-based object matching using interesting points and high-order graphs. *Pattern Recognition Letters*, 83, Part 3:251 – 260, 2016.

[174] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2030–2037, 2010.

[175] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 492–497, Sept 2009.

[176] G. Yu, J. Yuan, and Z. Liu. Propagative Hough voting for human activity recognition. In *European Conference on Computer Vision (ECCV)*, pages 693–706, 2012.

[177] G. Yu, J. Yuan, and Z. Liu. Propagative hough voting for human activity detection and recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(1):87–98, 2015.

[178] 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 (CVPR)*, pages 724–730, 2013.

[179] L. Zelnik-Manor and M. Irani. Event-based video analysis,. Technical report, 2001.

[180] H. Zhang, W. Zhou, C. Reardon, and L.E. Parker. Simplex-based 3d spatio-temporal feature description for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2067–2074, 2014.

[181] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, June 2015.

[182] X. Zhen and L. Shao. Action recognition via spatio-temporal local features: A comprehensive study. *Image and Vision Computing*, 50:1 – 13, 2016.