Representation and recognition of human actions in video

Bregonzio, Matteo

Representation and Recognition of Human Actions in Video

Matteo Bregonzio

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2011

Abstract

Automated human action recognition plays a critical role in the development of human-machine communication, by aiming for a more natural interaction between artificial intelligence and the human society. Recent developments in technology have permitted a shift from a traditional human action recognition performed in a well-constrained laboratory environment to realistic unconstrained scenarios. This advancement has given rise to new problems and challenges still not addressed by the available methods. Thus, the aim of this thesis is to study innovative approaches that address the challenging problems of human action recognition from video captured in unconstrained scenarios. To this end, novel action representations, feature selection methods, fusion strategies and classification approaches are formulated.

More specifically, a novel interest points based action representation is firstly introduced, this representation seeks to describe actions as clouds of interest points accumulated at different temporal scales. The idea behind this method consists of extracting holistic features from the point clouds and explicitly and globally describing the spatial and temporal action dynamic. Since the proposed clouds of points representation exploits alternative and complementary information compared to the conventional interest points-based methods, a more solid representation is then obtained by fusing the two representations, adopting a Multiple Kernel Learning strategy. The validity of the proposed approach in recognising action from a well-known benchmark dataset is demonstrated as well as the superior performance achieved by fusing representations.

Since the proposed method appears limited by the presence of a dynamic background and fast camera movements, a novel trajectory-based representation is formulated. Different from interest points, trajectories can simultaneously retain motion and appearance information even in noisy and crowded scenarios. Additionally, they can handle drastic camera movements and a robust region of interest estimation. An equally important contribution is the proposed collaborative feature selection performed to remove redundant and noisy components. In particular, a novel feature selection method based on Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA) is introduced. Crucial, to enrich the final action representation, the trajectory representation is

adaptively fused with a conventional interest point representation. The proposed approach is extensively validated on different datasets, and the reported performances are comparable with the best state-of-the-art. The obtained results also confirm the fundamental contribution of both collaborative feature selection and adaptive fusion.

Finally, the problem of realistic human action classification in very ambiguous scenarios is taken into account. In these circumstances, standard feature selection methods and multi-class classifiers appear inadequate due to: sparse training set, high intra-class variation and inter-class similarity. Thus, both the feature selection and classification problems need to be redesigned. The proposed idea is to iteratively decompose the classification task in subtasks and select the optimal feature set and classifier in accordance with the subtask context. To this end, a cascaded feature selection and action classification approach is introduced. The proposed cascade aims to classify actions by exploiting as much information as possible, and at the same time trying to simplify the multi-class classification in a cascade of binary separations. Specifically, instead of separating multiple action classes simultaneously, the overall task is automatically divided into easier binary sub-tasks. Experiments have been carried out using challenging public datasets; the obtained results demonstrate that with identical action representation, the cascaded classifier significantly outperforms standard multi-class classifiers.

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Some parts of the work have previously been published as:

1. M. Bregonzio, S. Gong and T. Xiang, Recognising Action as Clouds of Space-Time Interest Points, *IEEE Conference on Computer Vision and Pattern Recognition* 2009, Miami.

2. M. Bregonzio, S. Gong and T. Xiang, Action Recognition with Cascaded Feature Selection and Classification, *International Conference on Imaging for Crime Detection and Prevention* 2009, London.

3. M. Bregonzio, Jian Li, S. Gong and T. Xiang, Discriminative Topics Modelling for Action Feature Selection and Recognition, *British Machine Vision Conference* 2010, Aberystwyth.

4. M. Bregonzio, T. Xiang and S. Gong, Fusing Appearance and Distribution Information of Interest Points for Action Recognition, Accepted, subject to revision, *Journal of Pattern Recognition* 2010

Matteo Bregonzio.

London, 9 July 2011

Acknowledgements

First and foremost I would like to thank both my supervisors Professor Shaogang Gong and Doctor Tao Xiang for their stimulating input, rewarding discussion, encouragement in the completion of this research and thesis. Thanks also to Professor Andrea Cavallaro for helping and following my research steps.

I would like to thank and acknowledge members of the academic staff and PhD students past and present for their enlightening discussions and general help in all manner of ways. In no particular order: Samuel Pachoud, Jian Li, Chen Change Loy, Bryan Prosser, Khalid Bashir, Parthipan Siva, Dave Russell, Emanuel Zelniker, João Fayad, Ravi Garg, Marco Paladini, Javier Orozco, Milan Verma, Stuart Battersby, Colombine Gardair, Jonathan Heusser, Prathap Nair, Fabrizio Smeraldi, Michael Terrell, Murtaza Taj and William Marsh.

Many thanks also to members of the Systems Support department: Lukasz Zalewski, David Hawes, Matt Bernstein, Tim Kay, Tom King and Keith Clarke, for quickly solving any type of problem.

I would like to express my gratitude to members of the departmental administrative staff, without whose help I would frequently have been lost: Julie Macdonald, Colin Powell, Sue White, Rupal Vaja and Sharon Cording.

A special acknowledge also goes to all the friends I meet in London: Moritz , Heidar , John and Eva, Gosia, Nila, Andrea, Perry, Carlo and Mariateresa.

I am very grateful to the Engineering and Physical Sciences Research Council (EPSRC) for sponsorship of the work described herein.

Last, but certainly not least, my warmest thanks go to my girlfriend, Eva, and my family in Italy. They always believed in me and in the success of my work. Thank you for encouraging and helping me in every situation. This work is dedicated to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Scope

Video-based human action recognition aims to automatically classify human actions by observing frames from a video sequence. It is important to clarify at the beginning the meaning of *action*, since it may have different levels of abstraction. Moeslund et al. (2006) suggest that general human movements can be divided into three levels: primitives, actions and activities. Primitives are defined as a basic movement that can be described at the limb level. For instance, left leg forward, right arm upward or head twist. *Actions* consist of a sequence of primitives and involve part or the whole body movement such as "*walking*", "*running*", "*clapping*"or "*jumping*". Fig 1.1 shows some examples. Finally, activities are defined as a routine of subsequent actions. An example of activity is high hurdles, which contains starting, jumping and running. A more complex example is cooking which involves selecting and cutting the ingredients, boiling them, waiting and serving the food.

While primitives movements are limited and specific in describing a single human movement, actions provide a compact and detailed representation of human intentions or dynamics. For this reason, actions may be interpreted as atomic samples of human life (Yu-Ming et al. 2007). Differing from primitives and activities, actions have the proper spatial and temporal resolution to capture and discriminate human behaviours, which is crucial for understanding human intentions and interactions.

Automatic human action recognition attempts to group actions into different classes accord-

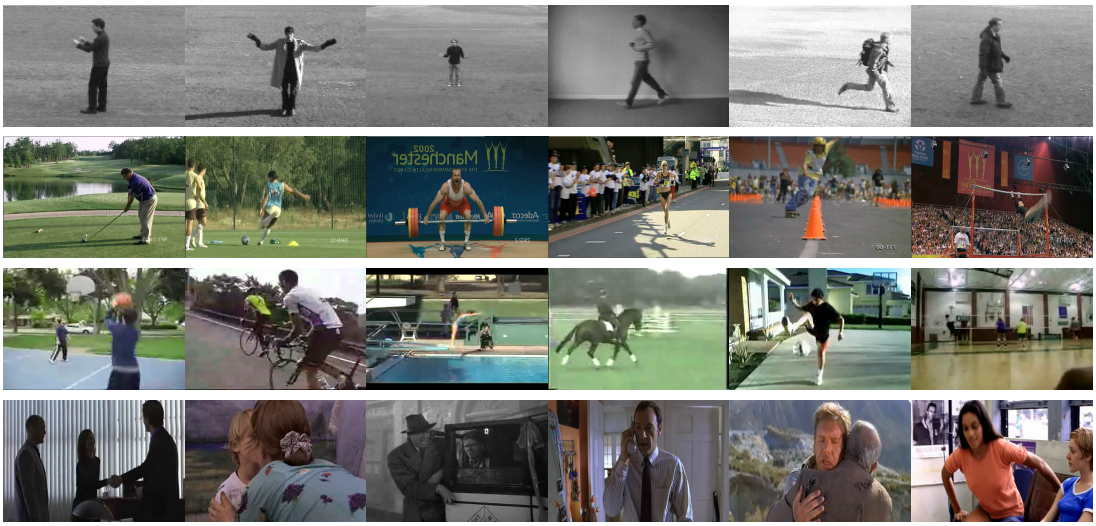


Figure 1.1: Examples from common datasets of human actions, from top to bottom: KTH dataset (Schüldt et al. 2004), UCF Sport Actions (Rodriguez et al. 2008), YouTube datasets (Liu et al. 2009a) and Hollywood dataset (Laptev et al. 2008).

ing to their visual similarity. This process can be seen as a perceptual grouping problem (Boyer and Sarkar 1999). In its most general formulation, action recognition is composed of two major phases: representation and classification. The former starts with the extraction of low-level features from videos, like colour, texture and optical flow, for example. Next, these features are usually mapped into a multidimensional space achieving a more compact and descriptive representation. With this information, it is then possible to design a model capable of separating actions into different classes. The second phase, classification, takes place when unlabelled videos are analysed and matched against the previously built action model, so that those videos can be associated with one of the known action classes.

The overall scope of this work is to research and define innovative methods to perform human action recognition in a robust manner, exploring both the representation and the classification phases.

1.2 Why automated human action recognition

Understanding the meaning of an action is an essential aspect of human social communication (Decety et al. 1997). Furthermore, it can be said that a large part of our daily life is spent watching and interpreting the actions of others (Barresi and Moore 1996). Many studies indicate that recognition of actions is a highly developed ability in humans and non-human primates (Premack and Woodruff 1978). This highly developed ability permits us to recognise actions

even when only a limited number of cues are available (Johansson 1973). In the context of an autonomous machine designed to interact with human beings, recognising human actions plays a critical role. Specifically, by employing an action recognition framework, a machine will be able to automatically learn and respond to outside inputs, as well as to monitor and detect human behaviours. Ultimately, intelligent machines could perhaps achieve the ability to interact and automatically move within the society, performing helpful and complex tasks without human supervision.

The development of intelligent machines has interested humans since ancient times (Koch and Mathur 1996), and today with the advances in computers and sensors, such developments are attracting an increasing interest from both research and commercial companies. Without a doubt, the study of automated vision is receiving considerable attention. Vision is a key-ability for intelligent machine to interpret information and deal with the surroundings effectively. By means of vision, it becomes possible to learn, understand and recognize scenarios as well as to communicate and interact.

Within this domain, recognising human actions is probably one of the most important factors, because it lets machines directly and naturally interact and understand humans without the need for any specific hardware interface. Practically, connecting a video camera to an intelligent machine and developing algorithms that replicate the human understanding, it becomes possible to create an intelligent machine that utilises visual information similarly to humans.

Automated action recognition has many potential applications, including but not limited to medical surgery, security, education, media, and the military sector, thus improving and simplifying human jobs. Although methods for recognising human actions are still to be fully developed, there are already applications that exploit this technique to address practical problems. A good example is smart video surveillance systems, which aim to detect suspicious behaviours automatically (Haritaoglu et al. 2000). Moreover, within the same context human action recognition can allow one to search for specific events in recorded surveillance videos. The analysis of sports videos is another important application (Efros et al. 2003). It may involve the classification of video segments between play and break intervals to summarize a video. Also soccer games can be analysed (Xie et al. 2002). Player's activities are recognized and used to help coaches in tactical analysis or TV commentary. Human-computer interaction systems can also benefit from the ability of recognizing actions (Bobick et al. 1999). For instance, an intelligent system can

(a) Intra-class variation

(b) Inter-class similarity

Figure 1.2: From top to bottom: action *walking* performed by and old man and a boy. Action *riding a bike* recorded frontally and sideways. Action *hugging* a person and *kissing* look similar if recorded from a particular angle. Action *jogging* and *running* look very similar

interact with children by interpreting and reacting to specific actions or needs. In the educational environment (smart classroom), the actions performed by a teacher are recognized to allow automatic camera motion and virtual mouse movement (Ren and Xu 2002). In robotics, interpretation of actions can be used either for reaction to the recognized action or for learning and imitation (Kruger et al. 2007). Finally, in medical treatment human motion analysis and recognition can aid diagnosis of motor problems by comparing patient motion to normality patterns as well identify progress over time (Branzan et al. 2007). Another possible medical application is to provide remote assistance to elderly people (Kosta and Benoit 2008; McKenna and Nait-Charif 2004).

1.3 Challenges and Motivations

Recognising human action is a challenging problem because it requires representing and clustering complex motions of an articulated human body. Additionally, the problem is further complicated because some actions may be visually very similar and in an unconstrained environment the presence of noise substantially increases the ambiguity.

Early studies focused on action recognition in heavily constrained motion capture environments, avoiding realistic challenges such as significant intra-class variations, inter-class similar-

ity, occlusion, and dynamic background (Lopes et al. 2010). Moreover, in order to obtain reliable features, most of the early works made a number of strong assumptions about the videos, such as the availability of reliable human body tracking, slight or no camera motion, and a limited number of viewpoints.

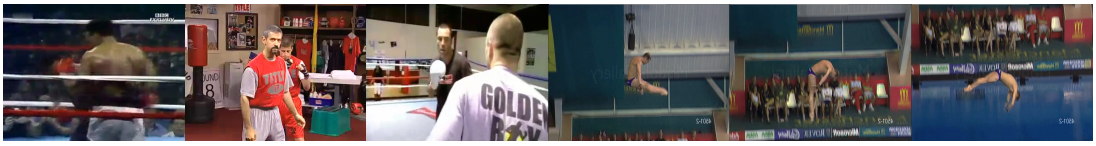
This thesis instead, studies the action recognition problem in a more realistic and unconstrained environment. By relaxing all the above-mentioned assumptions and constraints, the action recognition process becomes increasingly complex, and is required to face new challenges. These challenges are originated from a number of different factors which can be broadly categorized as following:

Intra-Class Variations - Actions are often performed by subjects of different age, size and appearance. Action speed, duration and spatio-temporal dynamics can differ as well. For instance, Figure 1.2 (a) shows how an action "*walking*" performed by a boy or an old man can appear different. Similarly, the action "*riding a bike*" recorded frontally or sideways has a complete different spatio-temporal characteristic highlighting a significant intra-class variation. To handle intra-class variations, an action representation able to generalize over these variations is required, capturing invariant features.

Inter-Class Similarity - The complementary problem to intra-class variations is known as inter-class similarity; it occurs when different actions look similar. Some examples are reported in Figure 1.2 (b). By observing the actions "*kissing*" and "*hugging*" or "*jogging*" and "*running*" from a specific angle, it can be noticed that they share numerous similarities in both spatial and temporal aspects. Consequently, action representation needs to take into account both intra-class variation and inter-class similarity to minimise misclassification. Action representation needs to be as invariant as possible and at the same time be able to capture discriminative and reliable features. This problem is more evident for an increasing numbers of classes, where different actions can be easily confused because they share similar primitive components.

Occlusion - In realistic scenarios a person may be partially occluded by other objects or partially self-occluded (for instance one leg is occluding the other). In the presence of occlusion the extracted features may be incomplete and misleading, causing classification errors. For instance, the action "*running*" generates key features associated with the fast legs movements. Having an object occluding the legs, prevents extraction of important information fundamental to classify the action. Similar problems appear for the action "*boxing*" when the arms are occluded or self-

(a) View Point changes



(b) Occlusion (c) Camera motion

Figure 1.3: From top to bottom: action *swinging* observed from different angles and zooms. Action *boxing* characterized by self and partial occlusion. Strong camera motion observed in a *diving* sequence, the camera follows the athlete.

occluded as presented in 1.3(b). Additionally, in this last example it may happen that the target is temporarily totally occluded causing problems with the target identification. Thereby, occlusion also affects the target identification and localization.

View Point - Observing the same action from different viewpoints may lead to extracting completely different action representations. For instance, as presented in Figure 1.2 (b), the action "*riding a bike*" appears very different if recorded frontally or sideways. Similar representation distortions can be observed when the action is recorded at different camera distances. A representative example is presented in Figure 1.3(a), where the action "*swinging*" is recorded with different viewpoint orientation and zoom.

Camera Motion - A fundamental action component consists of its motion, which is strongly modified and distorted in the presence of camera motion. Moreover, in this circumstance new background components irrelevant to the action are added. As such, the same action observed in still or moving camera scenarios generates significantly different representations. This difference may be partially reduced by separating motion components associated with the action from the background one. To this end, a pre-processing step can be employed aiming to compensate the camera movements. Unfortunately, this is not always doable especially for video containing drastic and variable camera movements. A representative example of this challenge is reported in Figure 1.3(c), where the camera follows a "*diving*" athlete.

Dynamic Background - Realistic scenarios may contain multiple persons or objects moving at the same time. As consequence, the background scene constantly changes. These changes influence the action recognition process in two ways: the target identification becomes more com-

plex and partial or total occlusion may occur. Secondly, the background components separation from the foreground's becomes complicated, thus the action representation may be influenced by strong background noise.

Other Environmental Conditions - Recording setting and scenario variations also play a role in action perception. For instance, shadows, lighting changes and crowdedness further increase action ambiguity and complicate the recognition process.

Due to the above-mentioned challenges, it is fundamental to develop a solid action representation invariant to changes in environment conditions and recording set-up. And at the same time make sure the representation is discriminative enough to clearly separate different actions. In light of this, it is important to capture robust action features aiming to minimizing the misclassification error. Eventually, multiple sources of information, such as optical flow, shape, action dynamics or context, can be simultaneously exploited and fused together. Then, to reduce both intra-class variations and inter-class similarity, feature selection should be applied to filter out redundant and misleading components. Feature selection also improves the classification phase, reducing the feature space ambiguity. Finally, in order to take into account challenges such as occlusion, dynamic background and camera motion, a specifically designed classification method is motivated.

1.4 Approach

This thesis focuses on the problem of automated action recognition from videos and covers four main areas: robust action representation, feature selection, feature fusion and iterative feature selection, and action classification. Action recognition is performed off-line in both training and testing scenarios.

1.4.1 Robust Action Representation

As an intuitive starting point, it can be said that human actions involve movements in space and time and these movements are characterised by a well defined spatio-temporal dynamic. Consequently, these spatio-temporal dynamics can be used to represent and discriminate between different actions. Additionally, it can be observed that each action is further characterised by dominant primitive movements, which differ from the rest. Typically these dominant primitives are short-term and fast-motion components (such as legs movement for running or arms move-

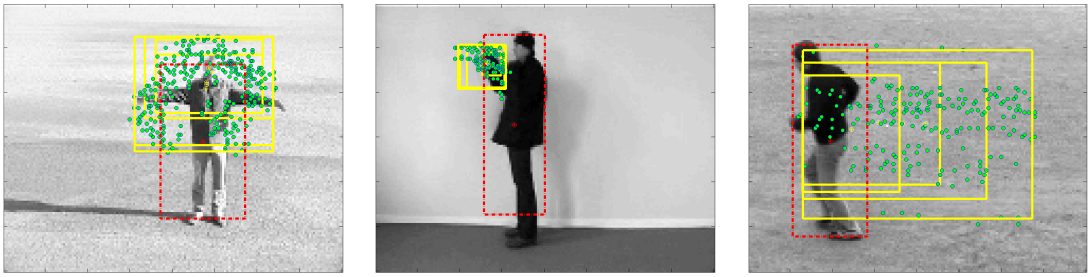


Figure 1.4: Examples of the proposed Clouds of Points method showing the point distribution generated by different actions, from left to right: *hand waving*, *boxing*, and *running*

ment for boxing). These components are very useful in overcoming problems of ambiguity.

It needs also to be mentioned, as discussed in Section 1.3, that action recognition performed in realistic scenarios is notably influenced by noisy components and distortions. Consequently, the action representation has to be designed taking into account these issues.

In the light of these observations, this thesis aims to represent actions as spatio-temporal dynamics with specific attention to modelling unique primitive components in a robust way. To this end, a space-time interest points based representation, capable of globally and explicitly describing the action dynamic in both space and temporal domains, has been chosen. Compared with alternative methods, such as shape analysis, tracking or optical flow, this representation is more robust to noise, small camera movements, and low-resolution inputs. Moreover, it does not require extraction of highly detailed silhouettes or target tracking. Conventional interest points based methods rely primarily on the discriminative power of individual local space-time descriptors, thus information about the spatial and temporal points distribution is lost. In contrast, this thesis initially presents a novel approach, named Clouds of Points, which exploits only the global spatio-temporal information about the point distribution, without the need to represent the detected interest points using local descriptors and visual vocabulary. In addition, this model is novel in capturing information about global spatio-temporal distribution of interest points explicitly and at different scales. The Clouds of Points approach does not require any assumptions on object shape, position or motion behaviour, thus avoiding tracking and segmentation problems. As a result, the formulated Clouds of Points approach appears to be more discriminative compared with existing interest points based method, and more invariant to local distortions, outliers and environments changes. The main idea is to collect reliable information by observing the clouds of points generated by the action over different temporal scales, then a robust multidimensional descriptors is computed to store the action information. Some example frames of the

Clouds of Points method are shown in Figure 1.4.

Furthermore, an innovative interest points detector designed to capture samples specifically located where the dominant primitive movements occur is also introduced.

Despite comparable performance with the best state-of-the-art methods, the proposed representation appears inadequate to deal with action containing drastic camera movements and dynamic background. In these conditions, the majority of the detected interest points are associated with the background leading to a mistaken action representation.

1.4.2 Feature Fusion and Selection

Feature fusion has been largely used to enrich action representations, usually by merging meaningful sources of information. The basic idea is to capture different action aspects or dynamics, then merging them to achieve less sensitivity to distortions and noise. Since standard Bag of Words interest points based methods and the proposed Clouds of Points exploit complementary and alternative action features, it is convenient to fuse these representations in order to build a solid action description. To this end, a Multiple Kernel Learning strategy is employed (Sonnenburg et al. 2006). Despite the obtained representation appearing more robust, it is still inadequate to handle realistic challenges such as camera movements or crowded background.

To tackle this problem, a novel action representation based on key-point trajectories is formulated. This representation is principally motivated by the fact that interest point based methods fail in the presence of shaky and constant camera movement. In contrast, trajectories based representations are less sensitive to camera movements, since the trajectories associated with the background can be properly filtered out leading to a coherent region of interest definition. Subsequently, within this region of interest key-points trajectories descriptors are exploited to represent the action. Unlike the trajectories method, interest points based methods are limited in temporal scalability. They only capture short movements within a short temporal window and, therefore, are inadequate for describing longer-term and more complex movements. Alternatively, long-term motion characteristics can be extracted from trajectories through tracking key points. To deal with large ranges of variation, a set of novel descriptors are also introduced. Critically, these descriptors are invariant to changes in scale, action direction, and frame resolution providing more discriminative description compared to existing methods. Some example frames presenting the proposed idea are shown in Figure 1.5.

In order to enrich the action representation and raise the recognition robustness, an adaptive

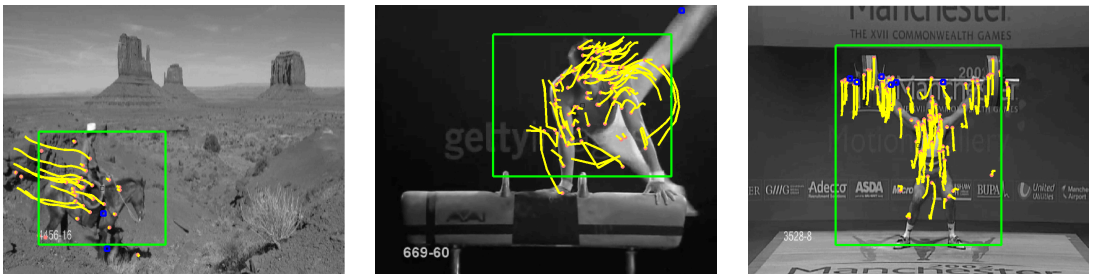


Figure 1.5: Examples of the proposed key-point trajectory method showing the components generated by different actions, from left to right: *horse riding*, *bench swing* , and *weight lifting*

feature fusion method combines the proposed trajectory based representation with an interest points based representation. This adaptive approach, in accordance with the camera movements detected, selects the optimal fusion strategy in order to cope with drastic changes in motion.

The general problem of unconstrained environments is the strong presence of noisy components, which affect the recognition performance. Moreover, the recorded actions may contain viewpoints variation, occlusion, and multiple subjects thus necessitating a feature selection approach to reduce the intra-class variation and inter-class similarity. To this end, a novel multi-class Delta Latent Dirichlet Allocation model based on (Blei et al. 2003) for feature selection is proposed. Specifically, the most informative features are selected collaboratively, rather than independently.

The proposed action recognition framework has been tested on different datasets and results comparable with the state-of-the-art have been observed.

1.4.3 Cascade Feature Selection and Action Classification

As stated above, action recognition in unconstrained environments is a very challenging problem due to the strong presence of noisy components. In addition, it can be further complicated if the observed video sequences are highly ambiguous and the available training set is noisy and sparse. (In this context sparse refers to a training set having few samples per class). These extreme conditions place new challenges on performing action recognition, principally in both the feature selection and action classification phases. More specifically, in the presence of sparse training data, high intra class variation and high inter class similarity, standard feature selection methods become extremely inefficient. Similarly, a noisy and sparse training set is inappropriate to train standard multi-class classifiers. In this context, it is difficult to simultaneously estimate the optimal decision boundaries that separate multiple action classes. Furthermore, the action

recognition task is particularly complicated when different action classes are visually similar due to the shared primitive action components. For instance, as presented in Figure 1.2, "*running*" and "*jogging*" would involve mostly the same body parts moving in a very similar ways. "*Hugging*" and "*kissing*" may look identical at the beginning of the action sequences, so it is therefore critical to perform feature selection in order to identify the most discriminative features per inter-class before classification. However, different feature sets are useful for separating different groups of actions, and there will rarely be features that are universally informative for separating all classes simultaneously. Therefore, the proposed solution to this problem is to select different sets of features for classifying different subsets. This unconventional, but necessary feature selection requirement is not well met by deploying standard multi-class classifiers.

The proposed framework aims to recursively decompose the classification task in subtasks, and each subtask is then addressed optimizing simultaneously the selected feature set and the learned classifier. The basic idea is to iteratively redesign the classification task in accordance with a specific context analysed. To this end a cascade of feature selections and binary classifiers is formulated, which seeks to optimise the feature selection and classification in each classification subtask.

1.4.4 Contributions

In order to formulate a solid action recognition approach, this thesis studies the problems of robust action representations, feature selection, feature fusion and classification. In the presence of realistic scenarios, characterized by noise and data ambiguity, the action recognition problem is further complicated. Thus, innovative approaches have been formulated. The main contributions are:

- A new **space-time interest points detection** method is developed to extract denser and more informative interest points compared to the existing methods (Dollar et al. 2005; Schüldt et al. 2004). In particular, primitive dominant components are highlighted while spurious detections in the background area and highly textured foreground areas, irrelevant to the action, are avoided (Bregonzio et al. 2009a). The extracted interest points are then used in the proposed **Clouds of Points action representation**. The idea of this representation is to describe actions as clouds of interest points accumulated at different temporal scales. Holistic features are then computed from these point clouds capturing explicitly

and globally the spatial and temporal dynamics of the action. In contrast with existing methods (Dollar et al. 2005; Schüldt et al. 2004; Liu and Shah 2008), which rely on single point descriptors, the proposed representation exploits the discriminative power of the point distribution (Bregonzio et al. 2009a).

- To address the problem of action recognition in unconstrained environments, a novel **trajectory based representation** is formulated (Bregonzio et al. 2010). An advantage of the proposed representation is to be able to simultaneously retain motion and appearance information even in noisy and crowded scenarios. Compared with interest points based methods, this approach can handle drastic camera movements allowing a robust region of interest estimation. Moreover, it describes actions with a large range of temporal and spatial scales, impracticable for points based representations. The used descriptors are invariant to changes in scale, action direction, and frame resolution providing a more robust description compared to existing methods (Sun et al. 2009a).

- To select more informative and discriminative features from a large feature set, a novel **feature selection method based on Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA)** is developed (Bregonzio et al. 2010). The idea behind this model is to collaboratively select features observing the shared feature patterns within different action classes. Compared to mutual information based methods, MC-ΔLDA returns better results.

- Aiming to enrich the action representation, an **adaptive feature fusion strategy** is formulated to merge trajectory based and interest points based representations (Bregonzio et al. 2010). These representations observe different action aspects and exploit complementary information, thus optimal to be fused. The experimental results confirm the benefit of the presented fusion strategy.

- To deal with highly ambiguous sequences and noisy and sparse training set, an **iterative feature selection and action classification approach** is introduced (Bregonzio et al. 2009b). The proposed approach employs a cascade structure to iteratively simplify the classification problem and simultaneously redesign both feature selection and classification task according with the actual context. Specifically, instead of separating multiple action classes simultaneously, the overall task is automatically decomposed into easier binary subtasks and the optimal feature set and classifier are employed.

1.5 Thesis Structure

The arguments of the thesis are presented in the following chapters, the breakdown of which is as follows:

- **Chapter 2** reviews literature relevant to the proposed lines of research, and provides insight into the reasons underlying their choice.
- **Chapter 3** initially introduces an alternative interest point detector, then the interest points based representation named Clouds of Points is formulated and validated over different benchmarks. The proposed representation exploits the information associated with the global points distribution captured at different temporal scales. In the experiment section, an extensive comparison with state-of-the-art approaches is presented.
- **Chapter 4** formulates a fusion strategy to combine the proposed Clouds of Points representation with a conventional Bag of Word representation. As a result, a more robust representation is achieved. Then, aiming to handle action recognition in unconstrained environments (especially in the presence of drastic camera movements and dynamic backgrounds), a trajectory based representation is formulated. In order to improve the recognition, the formulated trajectory representation is adaptively fused with an interest points representation. Finally, to reduce the effects of intra-class variation and inter-class similarity, a collaborative feature selection approach is derived. The experiment section validates the method over very challenging datasets and highlights the advantages obtained by using the proposed feature fusion and feature selection approaches. The action recognition results reported are comparable with the state-of-the-art.
- **Chapter 5** formulates and describes the cascaded feature selection and classification approach designed to handle action recognition in extreme conditions. Extreme conditions refer to a feature space characterized by a noisy and sparse training set, high intra-class variation and inter-class similarity. The proposed cascade approach simplifies the multi-class decision process in binary subtasks where only the best performing features are used for classification. Extensive experiments show that the proposed method has superior performance when compared to standard feature selection methods and multi-class classifiers.
- **Chapter 6** concludes the thesis, summarizing results and ground covered in the research,

and suggests various promising directions for future studies based on the results achieved so far.

Chapter 2

Literature Review

Action recognition in video involves a large number of steps and issues, from the initial low-level feature extraction to the final action labelling. This chapter reviews the principal contributions relevant to action representation, feature selection, feature fusion and classification. A recent survey of the most commonly used techniques can be found in (Poppe 2010).

2.1 Action Representation

Given a set of low-level observations directly extracted from a video sequence, action representation aims to map these observations in a multidimensional feature space (descriptor). This multidimensional space mapping is usually automatically performed and attempts to reshape the low-level observations in a richer and more convenient space where machine learning techniques (e.g. classifiers) can optimally work. Typically, action representation has two essential requirements. Firstly it needs to be invariant and generalized over small variations, for instance the same action may look slightly different due to: execution speed (action performed by an old man or young boy), view angle (action observed frontally or at 30 degrees left), light changes (afternoon or evening time) and clothes (wearing a jacket or a t-shirt). Secondly, the representation should be discriminative and non-ambiguous to allow a robust classification. This implies that representations belonging to different classes should be clearly separable even if the associated actions may look similar (*jogging* and *running*). The practical interpretation of these constraints yields to an action representation that is ideally invariant to: person appearance changes, dynamic background, viewpoint, and action execution variations. To satisfy these requirements it is crucial

to perform an appropriate low-level observation extraction (pixel-level measures) and mapping, which ensure a reliable action representation.

To this end alternative methods can be found within the existing literature, and they can be broadly divided into four categories: 1) spatio-temporal shape template based, 2) optical flow based, 3) interest points based, and 4) trajectories based. The former two methods use a global representation that encodes the visual observation as a whole. Global representation is obtained in a top-down fashion: a person is localized first in the image using background subtraction or tracking. Then, the region of interest is encoded as a whole and mapped in a descriptor. The representations are powerful since they encode much of the available information. However, they rely on accurate localization, background subtraction or tracking. They are also particularly sensitive to viewpoint, background noise, occlusions and inconstant frame rate. Differently, interest points and trajectories based methods use local representations derived by sample observations. The calculation of local representations proceeds in a bottom-up fashion: spatio-temporal interest points or key-point trajectories are detected first. Next, descriptors are calculated at either sample or distribution level. Sample level observes a single measurement at a time, while distribution level observes clouds of measurements. Compared to global representation, local representations are less sensitive to noise, partial occlusion and do not strictly require background subtraction or tracking. However, they do depend on the extraction of a sufficient number of relevant samples, and pre-processing is sometimes required to compensate for camera movements.

2.1.1 Spatio-Temporal Shape Template based Representation

Spatio-temporal shape template based approaches have been one of the early attempts to address action representation. Essentially, they treat action recognition as an object recognition problem by representing the action classes as a collection of spatio-temporal templates. The recognition is performed by matching known spatio-temporal templates with the testing query. The technique requires highly detailed silhouettes that can be computed using background subtraction. In the presence of camera movement, shadow, occlusion and multiple targets, clear silhouettes extraction is impracticable. Consequently, these approaches are problematic if applied to real-world videos.

One of the earliest examples of this line of research is presented by Bobick and Davis (2001), where shape templates are generated using silhouettes only information. Specifically, they extract silhouettes from a single view and aggregate differences between subsequent frames of an action

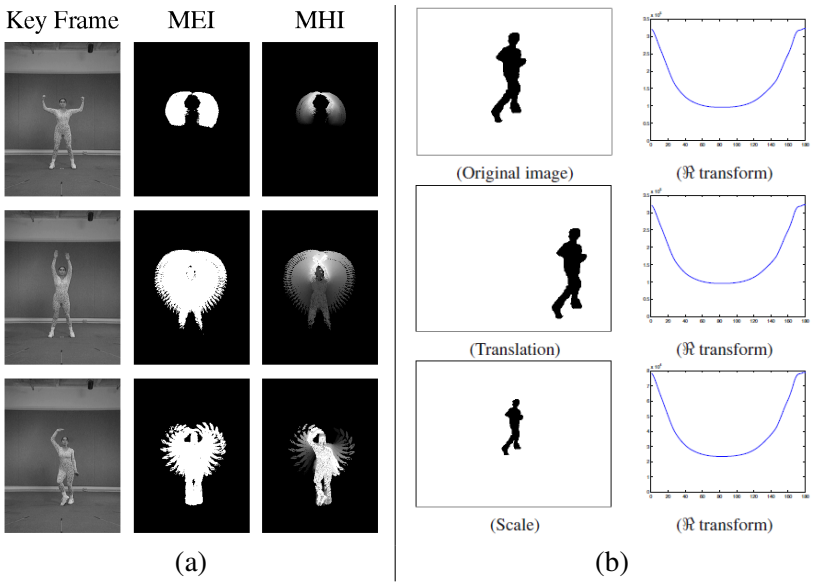


Figure 2.1: (a) Comparison of MEI and MHI representation presented by (Bobick and Davis 2001). (b) ℜ transforms of the same human silhouette which has been translated and scaled (Wang et al. 2007b).

sequence. This results in a binary motion energy image (MEI), which indicates where motion occurs. Also, a motion history image (MHI) is constructed where pixel intensities are functions of the silhouette sequence. Some representative frames are shown in Fig. 2.1(a). The template comparison is done using Hu moments, which permits a reasonable shape discrimination in a translation and scale invariant manner.

To further increase the robustness Wang et al. (2007b) apply a ℜ transform to the extracted silhouettes. This results in a translation and scale invariant representation that can be generalized over local appearance variations. Examples of silhouette variations and the correspondent ℜ transforms are reported in Fig. 2.1(b). To enrich the action representation Wang and Suter (2006) used both silhouette and contour descriptors to capture the global and local body parts motion properties. Given a sequence of frames, an average silhouette is formed by calculating the mean intensity over all centred frames. Similarly, the mean shape is formed from the centred contours of all frames.

Human body localisation, silhouette extraction and low template generalization still remain the major drawbacks of the above methods. Despite the rich information associated with the body shape, no explicit motion components are explored. Moreover, video sequences containing partial occlusions, crowd background and viewpoint changes are not handled by these representations.

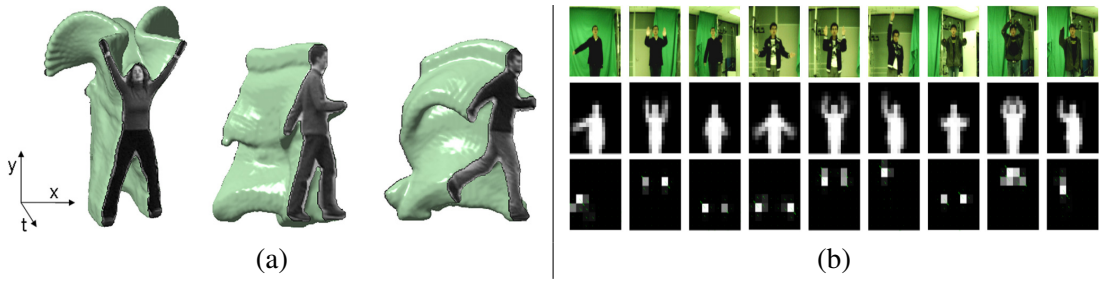


Figure 2.2: (a) Space-time shapes of *jumping-jack*, *walking* and *running* actions (Blank et al. 2005). (b) Examples of frame-to-prototype matching (Lin et al. 2009), from top line to bottom: original frame, shape components, motion components.

To address some of the mentioned limitations, spatio-temporal shape templates have been extended attempting to explicitly include also the motion information. (Blank et al. 2005) is an earlier example of work on this research line. They proposed a representation based on 3D spatio-temporal volumes formed by stacking silhouettes over a given sequence, examples of this representation are reported in Fig. 2.2(a). Then, the solution of the Poisson equation is used to derive local space-time saliency and orientation features. Global features for a given temporal range are obtained by calculating weighted moments over these local features. The obtained templates are able to incorporate the pose of the human body as well as dynamic information such as global body motion and local limb motion. Although the representation appears robust to partial occlusions, non-rigid deformations, changes in scale and viewpoint, the method does require accurate localization, alignment and solid background subtraction. A similar line of research, which attempts to simultaneously represent both shape and motion, is introduced by Yilmaz and Shah (2005). They initially generate a spatio-temporal volume by matching the subject contours over consecutive frames. Then, information such as speed, direction and shape are extracted by analysing the differential geometric properties of the spatio-temporal volume. An interesting property of this representation is that it is invariant to viewpoint changes. This is achieved by relying on the maxima/minima contour extremes which have been demonstrated to be view invariant. Basically, the idea is that by observing a 2D subject contour from different views, the curvature maxima and minima are invariant across different the views, thus they can be used to re-project the action. Finally, action recognition is achieved as 3D spatio-temporal volumes object matching.

Aiming to reduce the constraints about where the action takes place, Lin et al. (2009) propose a hybrid system able to deal with appearance variations, camera motion, dynamic background

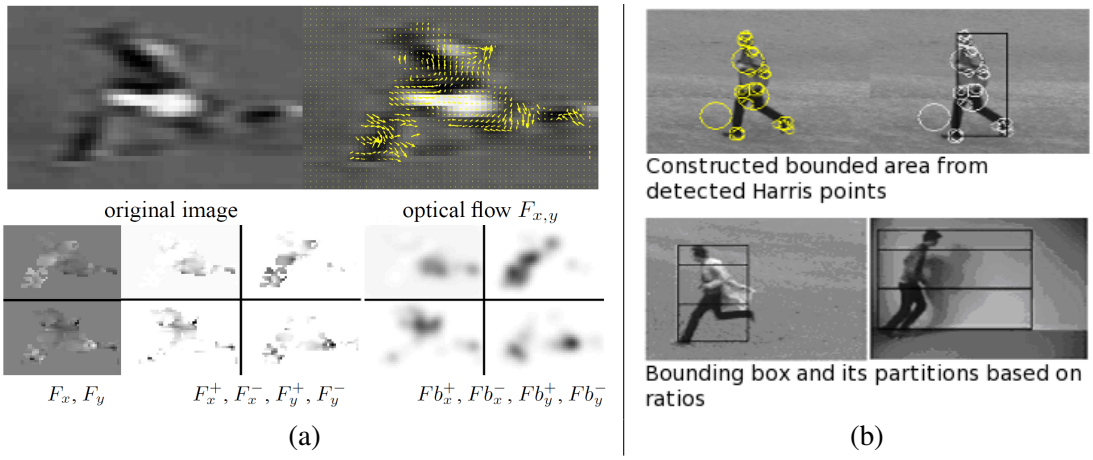


Figure 2.3: (a) Constructing the motion descriptor. The global optical flow vector is computed and then the x and y components are separated; finally four motion channels are defined and subsequently smoothed (Efros et al. 2003). (b) Harris corners are detected and used to construct the bounded area around the subject. Then the bounding area is partitioned into head, torso and legs (Danafar and Gheissari 2007).

and partial occlusion. This approach involves a pre-processing step that automatically compensates the camera movements and at the same time localizes and tracks the human body. Next, both shape and motion cues are explored by creating action prototypes designed to capture the correlations between action shape and motion. The shape information is computed using a grey-scale mask obtained by background subtraction, while the motion components are obtained using optical flow mapped into a grey-scale mask. Fig. 2.2(b) shows some examples of the used shape and motion templates. Finally, in order to rapidly match prototypes a decision tree structure is used. This research emphasized the fact that the shape component plays an important role in the recognition process, especially in the presence of dynamic background.

Clear silhouette extraction and subject localization still remain the principal limitations of these methods. Hence, none of these have been tested on realistic datasets (e.g. YouTube, Hollywood or UCF Sport actions). The problem stems from the inability to handle constant camera movements, low image resolution and cluttered background.

2.1.2 Optical Flow based Representation

Action representation based on optical flow tries to describe the human body as a whole and recognize action based on its dominant motion. This approach relaxes some constraints imposed by the shape based representation (such as does not require a detailed silhouette extraction) resulting in a more robust representation that is invariant to environmental variations such as light

changes, partial occlusion, and viewpoint changes.

Efros et al. (2003) pioneered work in this area. Their aim was to recognise action from a medium distance where both human bodies and the camera may be moving fast. The videos contain low-resolution frames with people whose images are only 30 pixels tall. First, by tracking the person and stabilizing the image in the middle of a tracking window, a sequence of spatial-temporal volumes are generated. Secondly, a descriptor based on blurred optic flow is extracted from each volume. Specifically, the motion channels are computed and then decomposed in four dominant components along the principal directions (left, right, up, down), next smoothed and normalized. A schematic description of this method is reported in Fig. 2.3(a). Recognition is performed in a nearest neighbour framework by observing the spatio-temporal correlation between descriptors. Although silhouette and contours are not required this method still depends on: robust subject segmentation, image stabilization and tracking; moreover the representation is strongly influenced by target occlusions and changes in size which notably modify the optical flow model.

Fathi and Mori (2008) improved this idea extending the optical flow descriptor to a two-fold framework. The low-level motion descriptors proposed by Efros et al. (2003) are initially extracted and used in the second step to derive the mid-level motion features. In practice, the low level features represent the weak classifier in the AdaBoost training algorithm to extract informative mid-level motion descriptors. The method produces interesting results on well known benchmarks. In addition, it appears robust to clutter and tolerant to both scale and viewpoint changes.

An alternative idea aiming to represent optical flow not as a single global component but as a set of local contributions is presented by Danafar and Gheissari (2007). Here the optical flow is evaluated using a grid-based representation. Specifically, a region of interest is found around the subject and divided in three horizontal sections (head, torso and legs) as shown in Fig. 2.3(b). Then, for each section two histograms of horizontal and vertical optic flow components are computed and used as descriptors. The histogram representation turns out to be robust to environmental noise, changes in illumination and viewpoint. Furthermore, the authors report an encouraging recognition rate on a popular benchmark . On the other hand, high performances are achieved only if region of interest and sections division are precisely estimated.

A similar approach of representing optical flow information via histograms is adopted in the

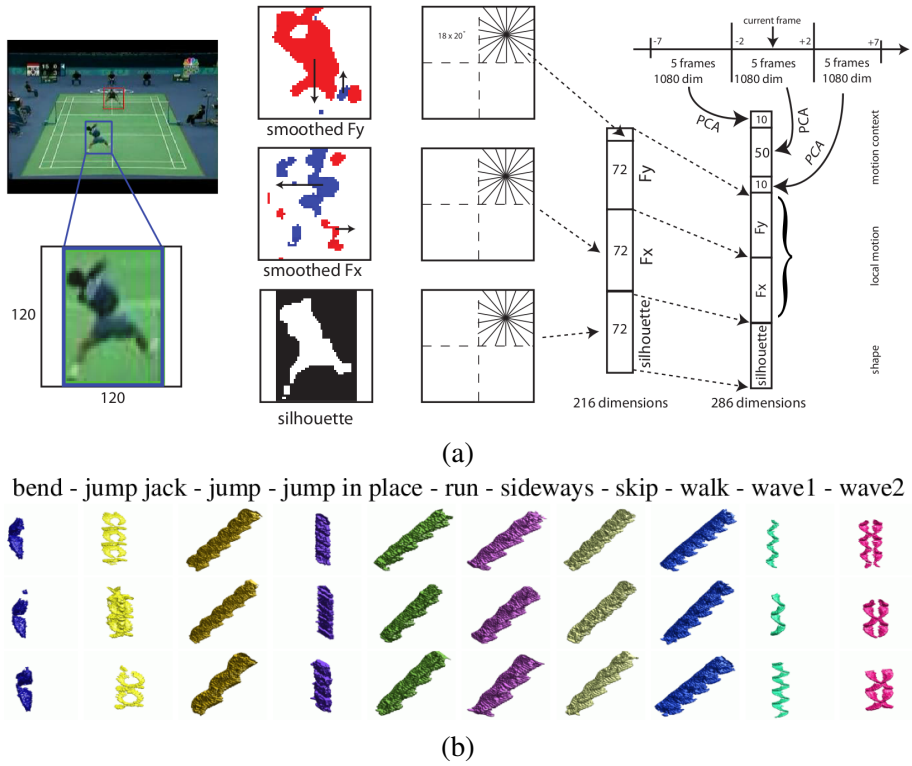


Figure 2.4: (a) Metric learning structure, the three information channels (vertical flow, horizontal flow and silhouette) are mapped in a 268-dimension histogram (Tran and Sorokin 2008). (b) Illustration of the 3D sable optical flow fields from different action sequences. Note how each action group has a unique flow volume surface (Riemenschneider et al. 2009)

Tran and Sorokin (2008) work. In contrast to the early work, this method appears robust to low resolution images, camera movements and imprecise subject localization. The authors also report high performance results on realistic sequences such as badminton games. The basic idea behind the method involves representing action using histograms of silhouettes and optical flow, which are computed inside a normalized bounding box located around the target. The bounding box is then divided into 2×2 sub-windows and in each sub-window the histograms are computed. The final descriptor is derived by merging the information collected over a window of 15 frames. Fig. 2.4(a) summarizes the principal steps of this representation.

More recently, Riemenschneider et al. (2009) proposed a different approach to explore optical flow information. They observed that different actions generate a unique spatio-temporal optical flow volume that can be stabilized and sampled as shown in Fig. 2.4(b). Thus, instead to directly analyse the raw optical flow, they first derive a more stable and continuous optical flow volume. Later, 3D interest points are extracted on the volume surface and the bag of words framework is applied for recognition. Although the paper reports promising results on a clean

dataset, the authors do not address more complex and realistic scenarios. It should be emphasized that the method strongly relies on still camera and clear background. In the presence of camera movements or crowd background the optical flow volume may assume a different shape, which cannot be handled by the proposed representation.

2.1.3 Interest Points based Representation

Interest point based representations involve two consecutive steps. Firstly, a set of spatio-temporal samples is captured: a process referred to as interest point sampling. Secondly, these samples are used to create the action descriptor: a process referred to as point representation. Among the recent techniques proposed in human action recognition literature, the interest points approach is the most popular representation. In contrast with previous model-based representation, an important advantage of interest points representation is to be model-free. This means that it does not require the presence of moving subjects under specific conditions or the observations matching with a predefined model.

Interest Points Sampling - Interest points are spatio-temporal locations where salient movements associated with the action occur. The goal of these points is to sample movements that are crucial to characterize human actions such as motion discontinuities. These interest points need to be stable with respect to perspective transformation and temporal periodicity.

Different ideas have been proposed to detect interest points. One of the simplest methods detects corners on the image plane employing the standard 2D Harris corner detector (Harris and Stephens 1988). However appearance information is captured, 2D interest points totally ignore temporal variations. Since the actions are defined in a spatio-temporal domain, the temporal cue plays a fundamental role in the action description. Hence, 3D spatio-temporal interest points appear to be the best alternative because able to simultaneously capture visual appearance and short temporal changes or dynamics. Along this line, Laptev and Lindeberg (2003) extended the 2D Harris corner detector to a 3D spatio-temporal domain. They define space-time interest points as those where the local neighbourhood has a significant variation in both spatial and temporal domain. The space-time scale of the neighbourhood is automatically selected. Drawbacks of this method include sparse samples and poor region of focus. Specifically, in the presence of smooth movements or low texture foreground a relatively small number of stable interest points may be detected and frequently the majority of the points are associated with the background, which creates misleading observations.

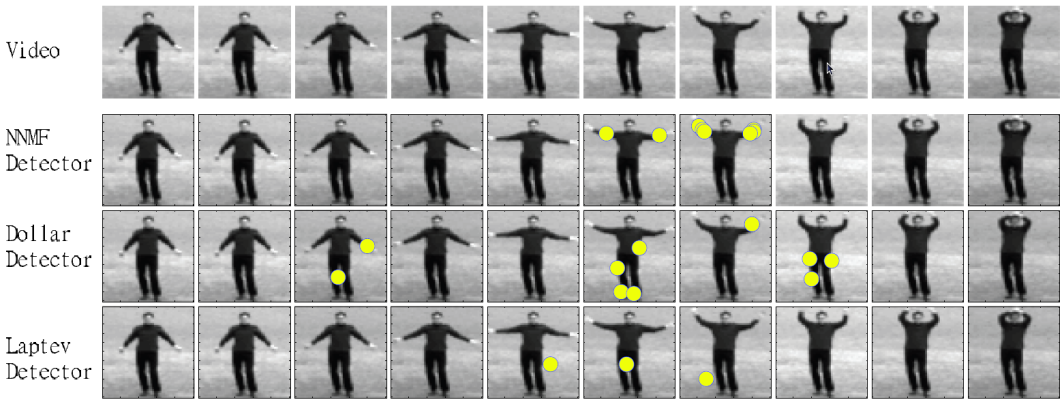
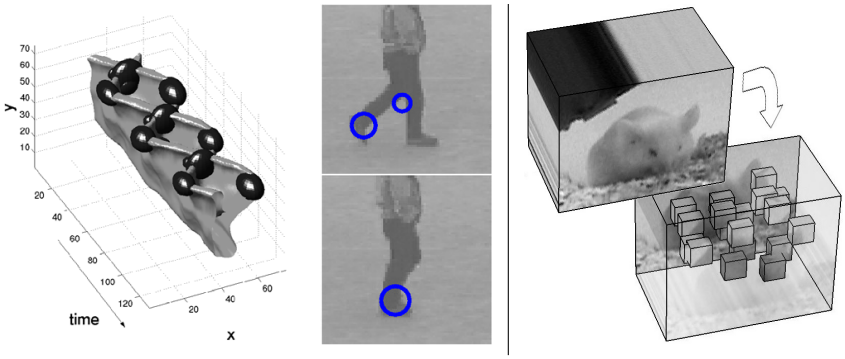


Figure 2.5: Interest point detection on *hand-waving* action. The first row shows the input image sequence, the second, third and fourth rows show the interest points detected by (Wong and Cipolla 2007), (Dollar et al. 2005) and (Laptev and Lindeberg 2003) methods. This comparison image has been extracted from (Wong and Cipolla 2007)

The sparse samples issue is well addressed by the Dollar et al. (2005) detector. They separately extract spatial and temporal information using a 2D Gaussian kernel and a 1D Gabor filter respectively. Then, the two responses are combined and interest points located in correspondence at the local maxima. The number of interest points is manually adjusted by changing the scale of the two filters. Despite its popularity, the Dollar et al. (2005) detector has a number of shortcomings. As mentioned by the as a moving, smoothed edge will cause only a gradual change in intensity at a given spatial location. Areas without spatially distinguishing features cannot induce high response. The detector is also prone to generate spurious detections in highly textured background areas. Additionally, the points are extracted at a single spatio-temporal scale.

Oikonomopoulos et al. (2006) propose a more sophisticated detector which addresses some of the above mentioned limitations. Saliency points are localized using the entropy information computed on the optical flow field. (Optical flow is compensated to reduce camera movements noise). Moreover, for each point an optimal spatio-temporal scale is automatically associated. Finally, each detected point is represented using both optical-flow and spatial-gradient descriptors. To be more robust against noise, they also introduced a clustering algorithm which removes points with low saliency value and creates clusters that are well localized in space, time, scale and sufficiently distant from each other. The method has been only tested on still camera scenarios and it appears unstable in the presence of dynamic background. This is because no region of focus is deployed to drive the detector.

This issue has been taken into account by Wong and Cipolla (2007) who propose a detector



(a) (b)

Figure 2.6: (a) Local space-time features detected for a *walking* pattern: 3-D spatio-temporal plot of leg motion (upside down) and corresponding features (Schüldt et al. 2004). (b) Visualization of spatio-temporal cuboids of mouse footage (Dollar et al. 2005).

driven by a region of focus named non-negative matrix factorisation (NNMF). Practically, the original 3D interest point detection problem is broken down into two filtering steps. Initially, 2D interest regions are detected observing motion components (regions of focus). Next, 1D points are detected as local maxima of the previous step's response. The authors compare the method with the Dollar and Laptev detectors; an example is reported in Fig. 2.5. In terms of performance, NNMF appears superior on three benchmark datasets. Despite the proposed detector partially addresses the problem of false detection reduction. Points associated with the background in the presence of a moving or zooming camera the model fails due to the incorrect motion components estimation.

To further improve the interest point detection, Chapter 3 presents an innovative method capable of extracting denser and more informative points compared to the above-mentioned detectors. In particular, our model avoids spurious detection in both background areas and highly textured static foreground areas. More specifically, our interest point detection method consists of two steps: 1) frame differencing to select the region of interest and 2) 2D Gabor filtering to select motion components. In both steps, saliency detection in temporal and spatial domains is exploited for the interest point detection.

Interest Points Descriptor - The interest points representation aims to summarize the 3D event observed by the interest point using a compact descriptor. Ideally it is invariant to background clutter, appearance, occlusions, and possibly to rotation and scale. In practice, a 3D cuboid of neighbouring pixels is extracted around each interest point and mapped in a feature vector (descriptor). Examples of detected interest points are presented in Fig. 2.6. Interest point

representation can be seen as a direct scheme for event detection and interpretation that does not require feature tracking, segmentation or computation of optic flow; only point detection and point description are involved.

Schüldt et al. (2004) did early studies in this area. In the first step of their approach interest points are found by using the Laptev and Lindeberg (2003) detector; an example of interest point extraction from a *walking* sequence is reported in Fig 2.6(a). Next, by observing the 3D cuboids both motion and appearance are captured with a normalized derivative descriptor. In the training phase all the point descriptors are collected and a K-means clustering is used to originate a visual-word vocabulary. Finally, each video sequence is represented as a histogram of visual-word co-occurrence. The recognition phase follows the standard Bag of Words framework and SVM is used as final classifier. Although the representation is designed to be invariant with respect to relative camera motions, the principal limitation of this method is the low discriminative power between similar actions. For instance, actions such as *jogging* - *running* or hand clapping - hand waving are easily misclassified. Such errors occur because of the low level of detail provided by the used descriptor.

Improvements in performance and representation are presented by Dollar et al. (2005). This work introduces novelties in both interest point detector and point representation. Specifically, they employ the (Dollar et al. 2005) detector and test three different descriptors named: normalized pixel based descriptor, brightness gradient based descriptor and optical flow based descriptor. The authors report that the best performance is achieved with the brightness gradient based descriptor. In addition they use PCA to reduce the descriptor dimensionality. For the recognition process, the method follows the standard Bag of Words framework with K-means clustering and 1-Nearest Neighbour as the final classifier. The recognition rate is reported on a standard benchmark and it is comparable with the state-of-art. Fig. 2.6(b) shows spatio-temporal cuboids extracted from a video sequence.

Similarly, Niebles et al. (2008) employ the Dollar et al. (2005) detector with gradient descriptors but applies smoothing before reducing the descriptor dimensionality using PCA. They achieve slightly better performance employing an unsupervised learning strategy.

Encoding Spatio-Temporal Distribution Information - These methods tend to rely on the discriminative power of a single point, ignoring the information associated with the global spatio-temporal distribution. Additionally, representing action based on single points, they are unable

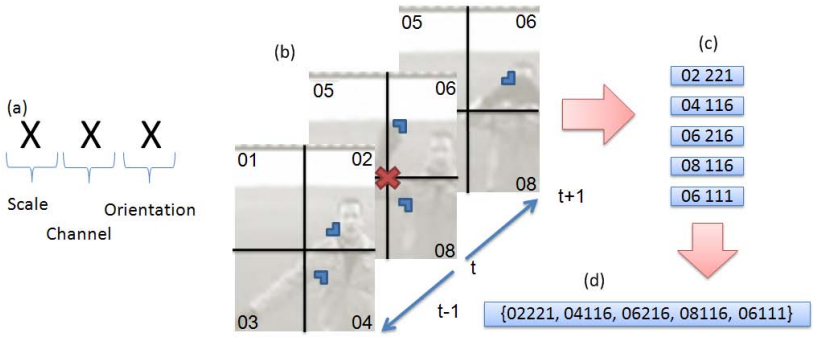


Figure 2.7: (a) The three parts that make up the (Gilbert et al. 2008) local feature descriptor. (b) A close-up example of a 2x2x2 neighbourhood of an interest point, with five local features shown as corners. (c) The spatial and temporal encoding applied to each local feature. (d) Concatenating the local features into a coded vector for the observed interest point.

to capture spatial relationship and global action dynamics. Aiming to address these limitations, a number of recent methods attempt to describe actions as a spatio-temporal distribution, instead of relying on single point descriptors. Simultaneously modelling the point itself and its neighbours generates a global representation of the action.

Along this vein, Liu and Shah (2008) exploit the spatial distribution of interest points using a modified correlogram. The idea behind the correlogram is to provide the local correlation in terms of the spatial location between interest points. The performance enhancements reported by the authors underline the importance of spatial information.

Gilbert et al. (2008) encode spatial and temporal position of the neighbouring interest points through a grid. Each single interest point descriptor is extended with additional information from the nearby interest point. A schematic overview of this method is proposed in Fig. 2.7. In a similar fashion, Zhang et al. (2008) introduce the concept of motion context to capture both spatial and temporal distribution of neighbouring interest points. Oikonomopoulos et al. (2009) propose an alternative approach where the spatial and temporal locations are encoded taking into account the co-occurrences of visual words pairs in relation to the object center. The object center is manually estimated in the training phase and automatically computed in the testing, allowing action localization.

All these models, however, still suffer from some of the flaws of the original Bag of Words method, in that ad hoc and arbitrary processes are needed for selecting a data dependent space-time descriptor, clustering algorithm for constructing a codebook, and codebook size. In addition, spatial and temporal information about the distribution of interest points is only exploited implic-

Figure 2.8: Action localization and classification (Liu et al. 2009a). ”M”, ”S” and ”H” in the images means the following judgments are made on the ”motion”, ”static” and ”hybrid” features, respectively.

itly, locally, and at a fixed temporal scale. In contrast, Chapter 3 proposes a model which exploits spatio-temporal information explicitly and at multiple temporal scales therefore capturing both local and global temporal information about interest point distributions. Moreover, it avoids data specific parameter tuning.

More recently action recognition has been applied to more realistic video sequences captured under uncontrolled conditions, such as: videos recoded by an amateur using a hand-held camera, YouTube, broadcast TV and personal video collections. This type of video generally contains significant camera motion, background clutter, and changes in object appearance, scale, illumination conditions, and viewpoint. These video sequences arise new challenges for the interest points based representations, mainly because in the presence of constant camera movements, multiple targets or background clutter, the detected points are misleading and unrepresentative. Hence, an additional pre-processing is required to filter out the noisy points and remove the components associated with the background. To this end, the action representation embeds both regain of interest estimation (target localization) and feature selection to tackle the problem. Region of interest is used to select the main subject in the video and filter out the points generated by others, whereas feature selection analyses the action descriptor and removes noisy and ambiguous features.

Promising results in this unconstrained context have been reported by Liu et al. (2009a). The goal is to initially detect a very dense set of interest points at multiple scales by using different detectors. Then, a region of interest is estimated and points not relevant to the action are filtered out. The representation is composed by a static contribution, extracted using SIFT (Lowe 2004), and a motion contribution, computed using gradient descriptors on 3D cuboids. Both contributions are then merged in a single hybrid feature representation. To further reduce the noise component, they implement a pruning framework based on Page Ranking (Kim et al. 2008): the goal is to select features according to their consistency. Only features that appear in the

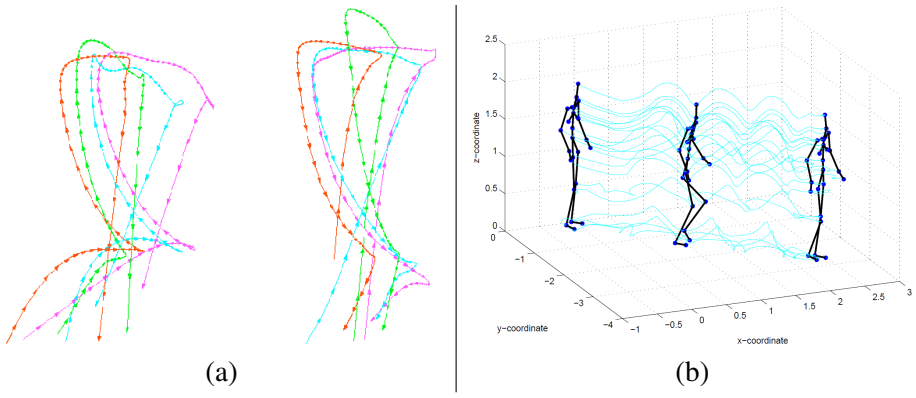


Figure 2.9: (a) Several trajectories used by (Rao and Shah 2001) to describe the actions opening overhead cabinet and closing overhead cabinet. (b) Representation of an action in 4-space employing the (Sheikh et al. 2005) representation.

video consistently and have similar behaviour are kept. At this stage, a very solid set of features is used in a Bag of Worlds recognition schema. Fig. 2.8 shows some example frames where the estimated region of interest and the action classification performed using different feature contributions are highlighted. The authors validate the method on standard datasets reporting results comparable with the state of the art. Despite the ability to reliably select features in static background, in the presence of constant motion or zooming camera, feature pruning loses its selectivity power. Due to camera movements, interest points are equally spread all over the image making region of interest estimation difficult and page ranking pruning much less efficient.

This section shows how action representation based on interest points has been successfully employed in different works. The achieved performance is comparable with the state-of-the-art. Although interest points are reliable in constrained scenarios, a number of limitations have been highlighted in more challenging circumstances. For instance, interest points describe the scene only locally ignoring middle and long term information. Moreover, they usually sample the sequence with a fixed spatial-temporal scale, which may be not optimal in describing different human dynamics. Finally, as mentioned above, it is difficult to isolate interest points associated with the background, mainly in the presence of dynamic background. To address some of these limitations it possible to track the interest points over different frames and use these point-trajectories to build a more robust action representation. This alternative action representation is presented in the following section.

2.1.4 Trajectories based Representation

Trajectories based representations have been extensively studied in the past few years. The principal concept in trajectories based representation consists of encoding the dynamics and behaviours of trajectories in a robust space-time descriptor. Compared to interest points based representation, the significant advantages of this representation lie in its ability to simultaneously capture the spatial and temporal dynamics involved in an action, to be more robust to dynamic background, and finally to offer a variable spatial-temporal scale. On the other hand, a general problem with trajectories derives from the fact that they are projected from an original 3D space on a 2D image plane. Consequently, different points of view generate different trajectories resulting in a viewpoint dependent representation. Therefore, the representation tends to capture view-dependent characteristics, which increase the action ambiguity. This shortcoming is more noticeable in older research that used long trajectories. Instead more recent works employ trajectory segments together with other cues such as local appearance, shape and distribution. Since trajectory segments are relatively invariant to viewpoint, this results in a robust representation capable of handling distortions in viewpoint changes, dynamic background and low-resolution images.

Rao and Shah (2001) attempted to compute a view-invariant trajectory based representation for action recognition. The method aims to recognize action including: opening and closing overhead cabinets, picking up and putting down a book, picking up and putting down a phone, erasing a whiteboard, etc. The proposed representation is based on spatio-temporal curvature analysis of hands-only trajectories where the trajectories are generated using a skin detector tracker. Fig. 2.9(a) shows some examples of detected trajectories. The action trajectory is represented by a sequence of dynamic instants and intervals. A dynamic instant is defined as an instantaneous entity, which occurs for only one frame, and represents an important change in motion characteristic: speed, direction, acceleration, and curvature. While, an instant is detected by identifying maxima (a zero crossing in a first derivative) in the spatio-temporal curvature, an interval represents the time period between any two dynamic instants, during which, the motion characteristics remain pretty much constant. Instants and intervals have physical meanings in terms of action continuity and discontinuity, therefore it is possible to explain an action as a sequence of meaningful instants and intervals. The view-invariant representation is derived by describing action with the number of instants, instants signs (positive or negative according to the direction) and

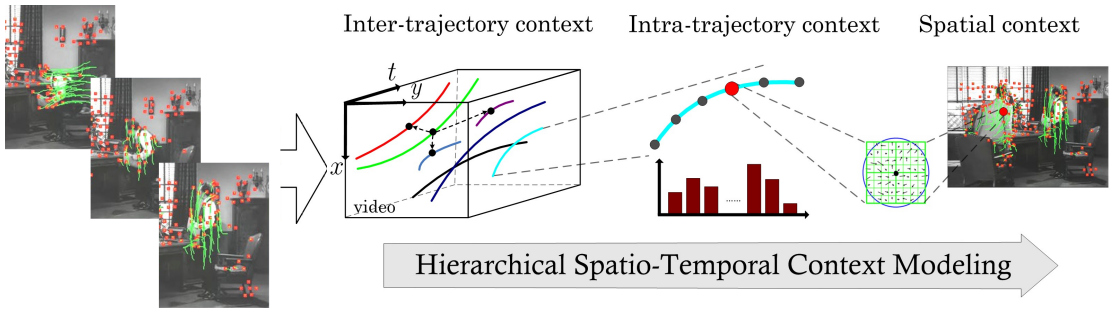


Figure 2.10: Schematic diagram on hierarchical spatio-temporal context modelling: 1) the point-level context with SIFT average descriptor, 2) intra-trajectory context, and 3) inter-trajectory context (trajectory proximity descriptor) (Sun et al. 2009a).

a matching vote (measuring the similarity with other examples). Although the authors properly classify their testing dataset, the method is not designed to be applied in a realistic environment. Hand tracking is very sensitive to initialization and light changes, and the representation is limited to hand movements. Furthermore, the method fails in the presence of camera movements and dynamic background.

Sheikh et al. (2005) proposed an alternative approach able to describe full body actions observing multiple trajectories generated by landmarks on the body as presented in Fig. 2.9(b). The proposed representation seeks to describe the landmark movements over a 4D space (X,Y,Z,t) by using a matrix formulation. Basically, the measurements of the imaged position of the anatomical landmarks of an actor are collected in a single matrix. The classification of an unknown sequence is then computed estimating the likelihood between a set of known actions. Long trajectories tracking and robust landmarks selection clearly represent a shortcoming of this method. Additionally, if any landmark is lost during the tracking, the recognition process becomes unstable. The full-body tracking appears not always practicable because of the generally unpredictable and complex nature of the human movements and self-occlusions. Seeking to address this problem, John et al. (2010) propose a framework able to track full-body articulated human motions recorded under different viewpoints. Promising results are obtained in a constrained set up.

An alternative method, able to recognize actions in realistic scenarios without relying on accurate tracking, has been presented by Sun et al. (2009a). In contrast with earlier researches, here a dense set of trajectory segments is used (segments of 5 to 25 frames maximum). This relaxes the constraints on tracking performance and key point selection. Practically, if a tracked key point is lost, in the next frame a new trajectory is generated from a new key point. The

proposed representation aims to model the spatio-temporal context information encoded in unconstrained videos using a three-level hierarchical structure: at the lower level the SIFT average description is computed along the trajectory; at the intermediate level, the transition and dynamics of the trajectory is evaluated; and at the coarse level, the spatio-temporal co-occurrence and distribution of the trajectories is observed. The schematic block diagram of this method is reported in Fig. 2.10. This representation produces a compact and efficient representation of trajectories context, dynamics and distribution, which is stable over different real world distortions such as camera movements, camera zooming, dynamic background, illumination changes and shadow. The authors validate the method with a challenging dataset reporting the best state-of-art result. Although trajectory segments have been proven efficient in unconstrained scenarios, in circumstances such as: low textured, fast moving, unstable frame rate, or small size subject the generated trajectories may be sparse. This shortcoming can be observed in scenarios such as amateur videos or clips collected on YouTube, where small size subjects are moving fast and few key points can be properly tracked over 10 to15 frames. The trajectories based method is then insufficient to provide discriminative information to perform action recognition. To address this issue, Chapter 4 merges a trajectory based method with an interest points based method. Doing this, sparse trajectories scenarios are handled by relying on interest points, which are easily extracted even in extreme conditions.

A shared problem of both trajectories based and interest point based representations is represented by the difficulty of isolating noisy features, which negatively influence the action classification task. This problem is more severe in unconstrained scenarios where noisy and redundant features are accumulated due to the complex sequence analysed. A well known strategy to address this problem is feature selection, which seeks to identify and filter out redundant and noisy components. More details on feature selection are presented in the following section.

2.2 Feature Selection

Once low-level features are extracted and represented in a descriptor, the feature selection step plays a crucial role in isolating redundant and noisy components as well as in reducing the feature space dimension. The main objectives of feature selection are: (a) to avoid over-fitting and redundancy, (b) to provide faster and more cost-effective models, and (c) to best represent the underlying structure of the data. In other words, the goal of feature selection is to find a subset of

features as small as possible, while simultaneously optimising the action labelling. However, it has been recognised that, in feature selection, combinations of individually good features do not necessarily lead to good classification performance (Cover 1974; Jain et al. 2000). Guyon and Elisseeff (2003) showed that the issue of feature redundancy and its relationship to usefulness is further complicated by the realisation that highly-correlated variables may in principle still be complementary to each other. It is then important to opt for a relevant feature selection approach.

For classification problems, feature selection techniques can be organised into two main categories: filter methods and wrapper methods. Filter methods can be seen as a pre-processing step. They select features on the basis of their relevance or discriminators power with regard to the targeted classes. Whereas wrapper methods select feature subsets by evaluating the performance of a learning algorithm (e.g. classifier). Basically, the main difference is that wrapper methods make use of the classifier, while filter methods do not. Hence, filter methods are computationally much more efficient, but usually perform worse than wrapper methods.

Filter methods employ feature ranking criteria for selection and operate independently from the classification algorithm. As a result, they look only at the intrinsic properties of the features and therefore ignore the effects of the selected feature subset on the performance. However they filter the (1) irrelevant and /or (2) redundant features to obtain a better and generic representation of the data according to its class membership, hence the term filter defined by John et al. (1994). To achieve the first task, i.e. filter irrelevant features, one of the simplest schemes is to evaluate each feature individually based on its correlation with the target class and then to select the k features with highest value. To that end, univariate feature ranking criteria such as the Relief algorithm (Kira and Rendell 1992) can be used to rank each feature independently from the others. This technique measures the linear dependency between features. An alternative method also able to measure non-linear dependency is based on the mutual information criterion (Zaffalon and Hutter 2002). Specifically, mutual information measures how a single feature is discriminative for a specific class. For instance, if a feature is shared with all the classes the associated rank is low. Otherwise, if the feature is a perfect indicator of the specific class the associated rank is maximum.

A more complex feature selection framework has been proposed by Peng et al. (2005) named minimal redundancy- maximal relevance (mRMR). This technique has been proven to be beneficial for selecting features for classification. It maximises the mutual information between the

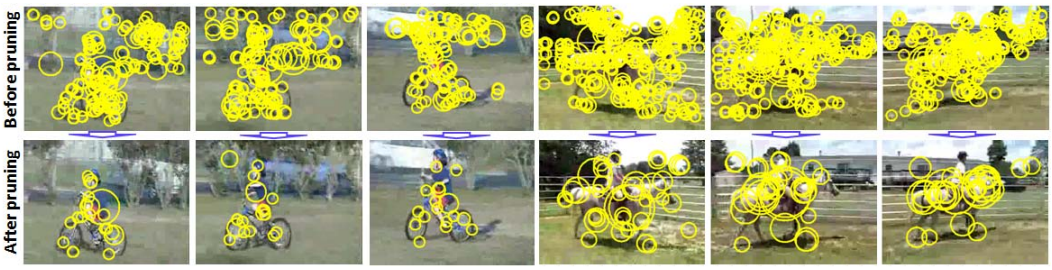


Figure 2.11: Feature selection using PageRank (Liu et al. 2009a). The first row shows the original static features, and the second row shows the selected features.

selected features and classes, whilst minimising the interdependence among the selected features.

In the wrapper approach feature selection is carried out using any machine learning algorithm. In its most general formulation, wrapper methods consist of using the test performance of a given machine learning algorithm to assess the relative usefulness of subsets of features. However the prediction performance is computed for each candidate feature at every stage (e.g. adding or removing a feature), which provokes a computationally expensive process. To prevent overfitting. In wrapper based feature selection, the more states that are visited during the search phase of the algorithm the greater the likelihood of finding a feature subset that has a high internal accuracy while generalising poorly. When this occurs, the algorithm overiftted the model to the training data. Greedy search strategies, such as sequential feature selection seem to alleviate the problem. The latter consists of two variants (Whitney 1971; Kittler 1978): (1) Sequential Forward Selection (SFS) in which features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion; and (2) Sequential Backward Elimination (SBE), in which features are sequentially removed from a full candidate set until the removal of further features increases the criterion. However both of these two techniques suffer from the so-called nesting effect [1]. To prevent the nesting of feature subsets, advanced methods have been developed, such as Sequential Forward Floating Search (SFFS), Sequential Backward Floating Search (SBFS) (Pudil et al. 1994) and the Plus q take-away r strategy (Ferri et al. 1994). All three methods backtrack as long as they find improvements compared to previous feature sets of the same size.

With regard to the specific task of action recognition, the feature selection problem is further complicated by real-world conditions and dynamics thus more sophisticated techniques are

[1]The nesting effect refers to the consequence of finding a local extreme (suboptimal solution) rather than a global (optimal) solution

needed. In particular, it can be seen that different action classes are often visually similar due to the shared atomic action components. For instance, *running* and *jogging* would involve mostly the same body parts moving in a very similar way. *Hugging* and *kissing* may look identical at the beginning of the action sequences. It is therefore critical to perform feature selection in order to identify the most discriminative features per class before classification. However, different feature sets are useful for separating different groups of actions, and there will rarely be features that are universally informative for separating all classes simultaneously. More generally, in an unconstrained environment with by high intra-class variance and high inter-class similarity traditional feature selection methods appear inefficient. For this reason, existing works prefer ad-hoc methods specifically designed for action recognition. For instance, in the presence of constantly changing background, a general requirement consists of removing the background components and selects a robust set of features associated with the subject only.

To reduce the influence of background components Liu et al. (2009a) propose an unconventional PageRank (PR) technique to select the important features. This ides is based on the assumption that if the background changes throughout the video, a consistent feature is a foreground feature. To this end, they build a large directed graph of features and evaluate if a feature is consistently matched with many other features. Thus, in case of high matching the feature is considered more significant than others. PR is employed to analyse the interaction between the features by assigning a ranking score to each feature as its relative significance in the feature network. This approach contains two major steps: visual similarity graph constructed by image matching and visual feature ranking by PR. Finally, the top informative features are selected. Some examples of feature selection using this method are reported in Fig. 2.11. Experimental results show the efficacy of PR technique in single subject scenarios where the majority of the background components, generated by camera movements, are removed. However, in the presence of multiple subjects the feature selection loses power.

Recently, to address feature selection in multiple subject scenarios Gilbert et al. (2009) exploit a complete by different idea. They are especially interested in learning discriminative low-level feature configurations, which appear frequently in the specific action sequence, and rarely on other actions or the background. In their paper feature configurations represent atomic motions or action segments without any link to human silhouette or subject location. This allows to detect simultaneously feature configurations belonging to different actions located in different

areas or to detect different feature configurations belonging to the same action but representing different motions. To this end, they propose to use a low-level feature hierarchical neighbourhood grouping exploring appearance and location concurrently. Specifically, they initially extract a dense 2D interest points set and represent each point with a 3 digit code (scale, channel, and orientation). Then, each point-code is extended incorporating the neighbouring points relationships. Finally, in order to identify the frequently reoccurring patterns data mining is used. All the compound features are combined into a single database and fed into a data mining algorithm called APriori (Agrawal and Srikant 1994). APriori finds the feature configurations which are frequently occurring in the same action class. The resulting frequent configurations are then used to group the features over a larger grid in the next hierarchical stage. The principal advantages of this method are that it can select reliable features directly at low-level and it can be used to spatio-temporally localize actions. Moreover, the reported results are comparable with the state of art. The computational cost is probably the main limitation; in fact the method has been tested only on relatively small datasets.

To perform feature selection in unconstrained videos with tractable computational cost, two alternative solutions are formulated in Chapter 4 and 5 respectively. 1) The first solution addresses the feature selection problem delineated by an ambiguous multi-class task with a sparse training set available. To this end a feature selection cascade is proposed, which simplifies the simultaneous multi-class feature selection with an iterative binary selection. Specifically, instead of separating multiple classes simultaneously, the overall task is decomposed automatically into easier sub-tasks of separating two groups of the most separable classes at a time. Then, for each classification sub-task the optimal features are selected using mutual information as a measure of feature relevance.

2) The second solution addresses feature selection in a large and ambiguous multi-class task problem, where the selection process requires analysing feature relationships. In these circumstances, conventional methods are computationally intractable, so a novel Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA) topic model for collaborative feature selection is introduced. MC-ΔLDA is designed to retain any correlation among features and select them collaboratively. MC-ΔLDA is an extension of ΔLDA proposed by (Andrzejewski et al. 2007b), which was used for understanding code bugs in computer programs with binary document classes (with and without bugs). Here, the formulated MC-ΔLDA is aimed at discovering action feature groups

(topics), some of them corresponding to features shared across different actions categories and others corresponding to unique features from specific action classes. By grouping all features jointly and collaboratively, MC-ΔLDA provides more effective feature selection for action discrimination.

To summarise, it can be said that the main ideas behind these two methods differ in terms of which aspect of the features is observed. The first method aims to identify unique features capable to strongly characterize a binary classification independently without relying on accurate training. In other words there is no interest in observing shared features and multi-class feature influences. In contrast, the last method takes into account the shared feature behaviours across different classes. Both shared and unique components are evaluated and used to jointly evaluate the feature relevance. This is achieved through an accurate learning phase.

2.3 Feature Fusion

Representation fusion is largely used to merge information coming from different sources, where the final aim is to enhance the recognition performance. Within the context of action recognition, fusion is mainly used to merge complementary representations (which explore alternative action patterns or aspects) to generate a richer action representation. It has to be mentioned that fusion brings improvement only if the fused representations mutually benefit each other. However, a straight foreword fusion is frequently impracticable because different representations may differ in dimensionality, scale or availability. Thus, the choice of the correct fusion strategy is challenging. To address feature fusion there are two main strategies that have been adopted by existing techniques: feature level fusion where feature spaces are merged at low-level and used as a single representation in the final classification (e.g. concatenation). And decision level fusion where separate recognisers are trained for each feature space before a joint decision is designed to make the final classification.

In the action recognition field a popular choice is feature level fusion. Along this line, Tran and Sorokin (2008) present an action representation obtained by concatenating histograms of optical flow and silhouette shape. Similarly, Lin et al. (2009) fuse motion and shape descriptors employing a weighted concatenation which generates a joint motion-shape descriptor. Here the optimal weights are estimated by cross-validation. Another work reporting performance improvements by fusion features is presented in (Schindler et al. 2008). Here the authors con-

catenate histograms of interest point descriptors (space-time gradient, optical flow and SIFT) increasing the recognition performance by an average of 4.5%. Despite its efficacy and simplicity, feature concatenation requires that all the features are properly normalized. Moreover, if they have a notable dimensionality difference the one with higher dimensionality needs to be reshaped otherwise the joined feature space becomes unbalanced.

More generally, it is possible to say that in a multi-class action classification problem, feature fusion should ideally weight features (which may have different scales and dimensionality) according to their relevance to the classification task. Furthermore, different weightings should be used for classifying different actions and ideally these weightings should be learned automatically from a training dataset. To this end, in Chapter 3 a multiple kernel learning (MKL) method is formulated to fuse interest point descriptors. MKL was first introduced in (Bach et al. 2004) to address the problem of selecting the optimal combination of kernel functions for a specific feature for Support Vector Machine (SVM) classification. Recently it has been used in computer vision for addressing a closely related problem, that is, given a specific kernel function but different features capturing different aspects of a visual object, how to best combine them together to achieve the optimal classification performance (Gehler and Nowozin 2009; Sun et al. 2009a). In this work, MKL is adopted to learn the optimal combination of different features without requiring any prior knowledge of them.

2.4 Action Classification

Given a training set, supervised multi-class classification algorithms aim to assign a class label for each testing sequence. The simplest case involves a two classes problem (binary classification) where the unknown sequence can be labelled as +1 or -1. Several algorithms, which have been proposed to solve binary problems, can be naturally extended to the multi-class case; others need a special formulation. Among the multi-class techniques used in action recognition, k-Nearest Neighbour (k-NN) is a well known and competitive classifier (Lin et al. 2009; Tran and Sorokin 2008; Wang and Suter 2006; Blank et al. 2005; Ali and Aggarwal 2001). k-NN performs classification measuring the distance (e.g. Euclidean) from the given unknown sequence to every other training data. The k smallest distances are identified, and the most common label among the k identified is chosen as a class label. The value of k is normally determined using cross-validation. However, k-NN is very simple and for a large training dataset the testing phase

involves a large number of comparisons that turns out to be computationally expensive. To reduce the computational cost of the nearest neighbour searches, the k-dimensional tree idea has been proposed (Moore 1991). Basically the feature space is partitioned in the training phase and the classification involves a quick tree search. A valid alternative is the Support Vector Machine (SVM) (Vapnik 1995). SVM builds a model during the training phase to represent data, then the classification is carried out matching the unknown sequence with the model. So it turns out to be computationally independent from the training data size. By its nature SVM is essentially a binary classifier. However, it can be extended to multi-class by using two common methods: (i) one-versus-all or (ii) one-versus-one. In the former method a binary classifier per action is built then the one with the highest output assigns the class label. In the last method, for each pair of classes a binary classifier is built, while discarding the rest of the classes. Then, through a voting among the classifiers output the label is assigned according to the class with the maximum number of votes. A number of papers employ the multi-class SVM classifier (Dollar et al. 2005; Danafar and Gheissari 2007; Wong and Cipolla 2007; Schüldt et al. 2004; Zhang et al. 2008; Ikizler et al. 2008), and observing the presented comparisons it emerges that generally SVM outperforms k-NN. In any case, the classifier performance depends greatly on the characteristics of the data to be classified. In other words there is no single classifier that works best on all given problems. Hence, empirical tests on classification performance are commonly employed to select the best classifier.

A different classification approach is used by (Fathi and Mori 2008; Liu et al. 2009a) where a boosting framework is applied to improve the accuracy of any given classifier. Specifically, boosting algorithms such as Adaboost (Adaptative boosting) (Freund and Schapire 1997) employ an iterative procedure to increase the performance of weak classifiers by reinforcing training on misclassified samples. Differently from the previous methods, the decision boundaries optimization is obtained gradually, generating more complex classifiers in each iteration. Advantage of Adaboost is to improve the generalisation properties without overfit the data, however sparse and noisy training set particularly reduce the classification performance.

In the context of realistic human action recognition, multi-class classification is further complicated by the fact that actions may be performed by subjects of different sizes, appearance and poses. Moreover, in an unconstrained environment the problem is compounded by the inevitable occlusion, illumination change, shadow, and camera movement. In this circumstance, character-

ize by strong presence of intra-class variation and inter-class similarity, standard classifiers such as multi-class SVM or k-NN appear inadequate. This because they use the same feature set to classify different action, furthermore the classification is done in a single step.
The idea to employ different feature-subsets to separate different classes is more suitable in this circumstance.

To address these problems, Chapter 5 proposes a novel action classification approach which utilises a cascade of binary classifiers. Instead of separating multiple action classes simultaneously, the overall task is decomposed automatically into easier sub-tasks of separating two groups of the most separable action classes at a time with different features selected for different binary classification sub-tasks. More specifically, our classifier iteratively splits a group of action classes into two sub-groups until each sub-group only contains a single action class. Compared with the standard multi-class classifiers, a binary classifier in the cascade only needs to draw a single decision boundary between two groups of data that are most separable at a time. In addition, it allows for the selection of different sets of optimal features for separating different classes of actions.

The idea of using cascaded classifiers for solving difficult vision problems has been exploited before by Viola and Jones (2002) and Athitsos et al. (2005). Specifically, Viola and Jones (2002) propose a cascade of AdaBoost classifiers to address the face detection problem. Athitsos et al. (2005) employ a cascade of approximate k-NN classifiers for recognising handwritten digits. Similar to our approach, their classifiers utilise different sets of features at different stages in a cascade with the later stage facing harder classification problems. Our approach is also closely related to the classification trees used in the machine learning community (Safavian and Landgrebe 1991). Similarly to classification tree, the cascade structure is automatically built and different feature-subsets are exploited in different separation tasks.

2.5 Summary

From this chapter one can appreciate how much attention action recognition has recently received from the computer vision community. A large number of alternative approaches have been proposed but still it remains an open area of research. In addition, very recently action recognition moved from the laboratory environment to realistic scenarios introducing a large variety of new challenges.

The first problem concerns **action representation**, which is the process of describing low-level observations in a multidimensional space suitable for machine learning techniques. Current representations differ in the explored features such as: shape, contours, motion flow, key points, or motion trajectories. According to the literature, there are four different ways to perform action representation: 1) spatio-temporal shape template based, 2) optical flow based, 3) interest points based, and 4) trajectories based. Both spatio-temporal shape template and optical flow based methods require a reliable subject localization or region of interest estimation, otherwise the representation is unable to capture meaningful information. Alternatively, spatio-temporal interest points and trajectory based methods sample the video sequence and use key points to describe the action. Specifically, interest point based methods construct action descriptors observing the surrounding area extracted around the detected points. Trajectory based methods track key points over time and generate trajectories; actions are represented using descriptors representing the spatio-temporal appearance of the trajectories themselves and the trajectories' distribution. However these two methods are more competitive and reliable in unconstrained environments. They require a dense sampling of the action that is not always possible, especially with low resolution frames, and small size, fast moving subjects. It has been observed that interest point based methods do not explicitly describe the property of the point's distribution, but they solely rely on the discriminative power of individual points. Moreover the recognition framework used (Bag of Words) involves tedious parameter tuning. In order to address these issues, Chapter 3 presents a novel action representation method which differs significantly from the existing interest point based representation in that only the global distribution information of interest points is exploited.

Chapter 4 tackles a different aspect linked to robust action representation obtained by exploiting **feature selection** and **fusion** as well as an alternative action representation. These steps are crucial when the recognition is performed in unconstrained sequences which involve large degrees of occlusions from multiple objects, illumination changes, shadows, cluttered backgrounds, and scale variations. In more detail, a novel trajectories based representation is formulated, which is able to retain local motion information (trajectory orientation and magnitude), trajectory shape and static appearance information. After a collaborative feature selection phase, this trajectories based method is fused with a standard interest point method aiming to produce a more reliable action representation.

The last problem studied is relevant to **multi-class action classification**. Among the standard methods available to solve this task the most commonly used are k-Nearest Neighbours (k-NN) and Support Vector Machine (SVM). The former performs classification measuring the distance (e.g. Euclidean), from the given unknown sequence to every other training data. The latter builds a model during the training phase learning the decision boundary, then the classification is carried out matching the unknown sequence with the model, which is computationally cheaper compared to k-NN. Alternatively, a boosting framework such as AdaBoost, may be employed to improve the accuracy of any given classifier. AdaBoost iteratively optimise the decision boundaries generating more complex classifiers at each step. Generally, in classification tasks delineated by large class overlap and intra-class variation standard classifiers appear inadequate. This is because they aim to simultaneously estimate the optimal decision boundaries that separate highly ambiguous multiple action classes.

To address this problem, Chapter 5 proposes an action classification approach which utilises a cascade of binary classifiers. More specifically, our classifier iteratively splits a group of action classes into two sub-groups until each sub-group only contains a single action class.

Chapter 3

Robust Action Representation Using Clouds of Interest Points

Most of the recent action recognition methods represent actions as bags of space-time interest points. Although these methods report promising results, they rely solely on the discriminative power of individual local space-time descriptors while ignoring the potentially useful information about the global spatio-temporal distribution of interest points. Consequently, they are unable to capture global motion components as well as smooth and fast motions. This is due to the lack of both multiple-temporal-scale and points-distribution information.

To address these limitations, this chapter puts forward an action representation method that aims to explicitly and globally exploit spatio-temporal information associated with interest point distributions. In particular, holistic features from clouds of interest points accumulated over multiple temporal scales are used. Representative frames explaining this idea are presented in Figure 3.1.

The proposed action representation, named Clouds of Points, merges the extracted features by using a Multiple Kernel Learning strategy. This allows different weights to be automatically assigned to each temporal scale, obtaining a more robust representation for each action class as a result.

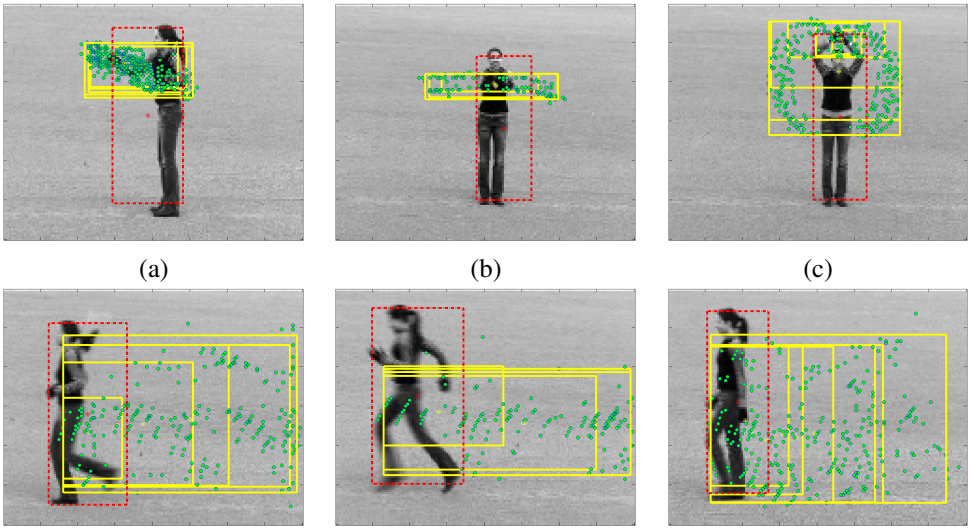


Figure 3.1: Examples of Clouds of Interest Points extracted from the KTH dataset. The clouds at different temporal scales are highlighted in yellow boxes. (a) *Boxing* (b) *Clapping* (c) *Hand waving* (d) *Jogging* (e) *Running* (f) *Walking*.

3.1 Interest Point Sampling

As presented in Chapter 1, the first step of the action recognition framework involves low-level observations extraction. With regard to the proposed Clouds of Points representation, low-level observations are extracted using an interest points approach.

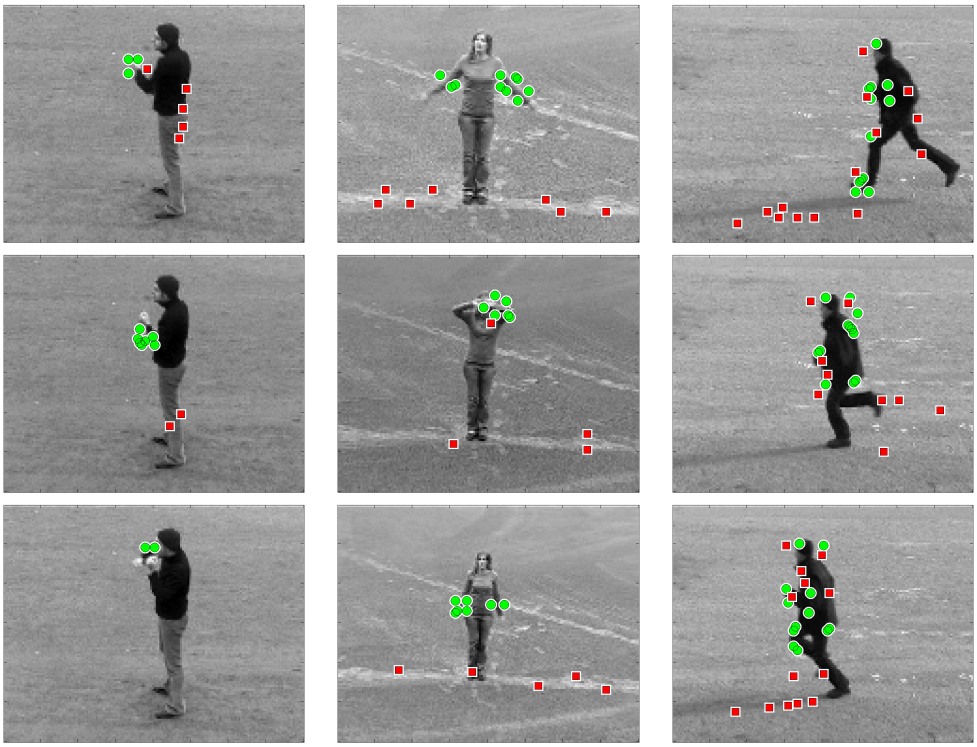
Interest points are local spatio-temporal features which are considered to be salient or descriptive of actions captured in a video. Among various interest point detection methods, the one proposed by Dollar et al. (2005) is perhaps the most widely used for action recognition. Using their detector, intensity variations in the temporal domain are detected using Gabor filtering. The detected interest points correspond to local 3D patches that undergo complex motions. Specifically, the response function of the Gabor filters has the following form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{3.1}$$

where $g(x, y : \sigma)$ is the Gaussian smoothing kernel applied in the spatial domain, while h_{ev} and h_{od} are the 1D Gabor filters applied temporally, defined as:

$$h_{ev}(t; \tau, \omega) = -cos(2\pi t \omega) e^{-t^2/\tau^2} \tag{3.2}$$

$$h_{od}(t; \tau, \omega) = -sin(2\pi t \omega) e^{-t^2/\tau^2} \tag{3.3}$$



(a) *Boxing* (b) *Hand waving* (c) *Running*

Figure 3.2: Comparison between interest points detected using our detector (green circle points) and the Dollar et al. (2005) detector (red square points). The frames present the detection process in three different conditions: (a) slow object movements (b) camera zooming (c) presence of shadow

The algorithm first applies the Gaussian smoothing on the all video sequence followed by the Gabor filtering. As reported in the original paper (Dollar et al. 2005), by setting $\omega = 4/\tau$, there are essentially two free parameters τ and σ which roughly control the spatial and temporal scales of the detector.

Despite its popularity, the Dollar detector has a number of drawbacks. As mentioned by the authors, areas undergoing pure translational motion will in general not induce a strong response. For instance a moving smoothed-edge will cause only a gradual change in intensity at a given spatial location, thus it induces a weak response.

Additionally, since it does not use any region of focus, it also tends to generate spurious detection in highly textured background areas irrelevant for the action. As shown in Figure 3.2, Dollar detector is particularly ineffective given slow object movement, small camera movement, or camera zooming.

A new interest point detector is developed here to overcome the shortcomings of the Dollar detector. In particular, most of the shortcomings of the Dollar detector are caused by its design

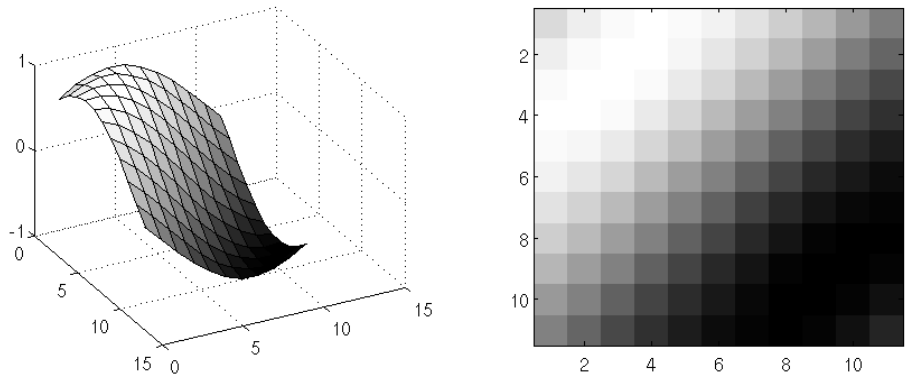


Figure 3.3: Examples of first-order derivative filter oriented along 45°.

of spatial and temporal filters and the way these filters are combined to give the final response. Especially, the 1D Gabor filter applied in the temporal domain is sensitive to background noise and highly textured background/foreground areas, which have nothing to do with the action being performed. To overcome this problem, the proposed detector adopts different and more effective filters for detecting salient space-time local areas undergoing complex motions. More specifically, our interest point detection method consists of two steps: 1) frame differencing for focus of attention and region of interest detection[1]; and 2) first-order derivative filtering on the detected regions of interest along different orientations. Via these two steps, saliency detection in both the temporal and spatial domains are combined together to give the filter response.

The first-order derivative filters are applied on the frame difference result. Specifically, the filters are composed of two parts. The first part $c(x,y;i)$ represents the real part of a complex sinusoid:

$$c(x,y;i) = cos(2\pi(\mu_0 x + \upsilon_0 y) + \theta_i) \tag{3.4}$$

where θ_i defines the orientation of the filter 8 orientations are considered:

$$\theta_{i=1,..,5} = \{0^\circ, \pm 22^\circ, \pm 45^\circ, \pm 67^\circ, 90^\circ\} \tag{3.5}$$

and μ_0 and υ_0 are the spatial frequencies of the sinusoid controlling the scale of the filter. The

[1]Although it is a very simple technique, frame differencing is found to be sufficient for our interest point detector given moderate camera motions such as those in the KTH dataset; When larger camera movements are present, a more sophisticated foreground detection method need to be adopted (e.g. one can employ an object detector such as (Felzenszwalb et al. 2008)).

second part of the filter $G(x,y)$ represents a 2D Gaussian-shaped function:

$$G(x,y) = exp\left(-\frac{\frac{x^2}{\rho^2}+\frac{y^2}{\rho^2}}{2}\right) \tag{3.6}$$

where ρ is the parameter that controls the width of $G(x,y)$. By setting $\mu_0 = \upsilon_0 = \frac{1}{2\rho}$, the only parameter controlling the scale is ρ, which is set to 11 pixels in this study[2]. The filters F_t are separately applied and 8 different responses are computed at each frame t. An example of oriented filter is shown in Fig. 3.3.

$$F_t = I_t * c_t * G_t \tag{3.7}$$

These responses are combined together to compute a bi-dimensional saliency map F_t^{map} as follows:

$$F_t^{map} = (F_t^{0^\circ})^2 + (F_t^{22^\circ})^2 + (F_t^{-22^\circ})^2 + (F_t^{45^\circ})^2 + (F_t^{-45^\circ})^2 + (F_t^{67^\circ})^2 + (F_t^{-67^\circ})^2 + (F_t^{90^\circ})^2 \tag{3.8}$$

where that image coordinates (x,y) are omitted for conciseness. Finally, interest points are detected as local maxima of the saliency map.

Figure 3.2 shows examples of our interest point detection results obtained on the KTH dataset. It is evident that the detected interest points are much more meaningful and informative compared with those detected using the Dollar et al. (2005) detector. In particular, the interest points detected by our approach tend to correspond to the main body parts contributing to the action being performed, whilst those detected by the Dollar detector often drift to static body parts or to background areas with strong edges. The experiments presented in Section 3.4.2 also suggest that a better recognition performance can be obtained when our interest point detector is used in place of the Dollar et al. (2005) detector, either with the standard Bag of Words representation or the proposed Clouds of Points representation.

3.2 Action Representation

Consider an action video sequence $\mathbf{V}$ consisting of T image frames, represented as:

$$\mathbf{V} = [\mathbf{I}_1, \ldots, \mathbf{I}_t, \ldots, \mathbf{I}_T]. \tag{3.9}$$

[2]The value of ρ is set empirically. It could be set in a more principal way via cross validation. It has been observed in our experiments that the recognition performance is not sensitive to the value of ρ.

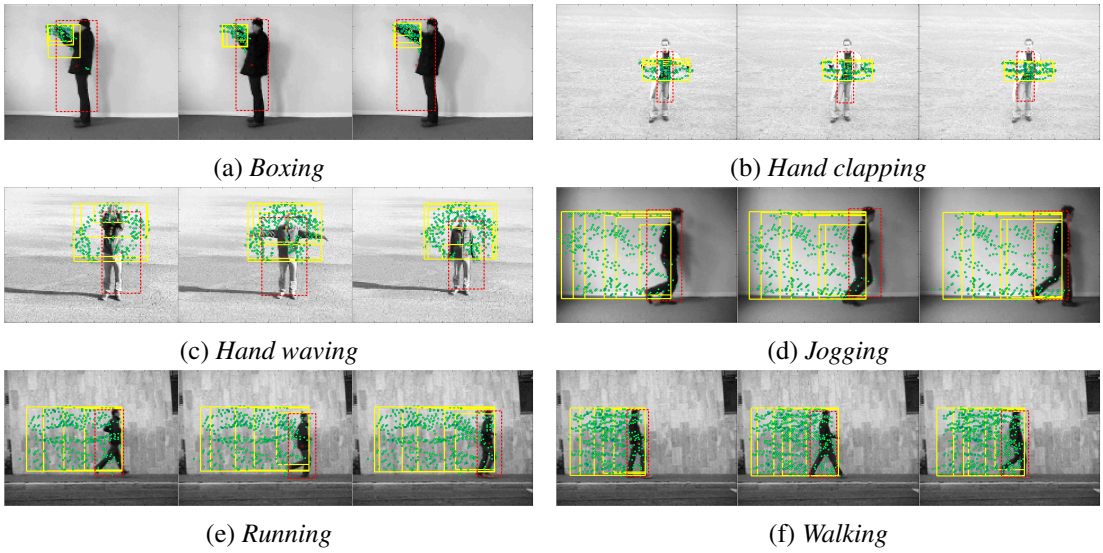
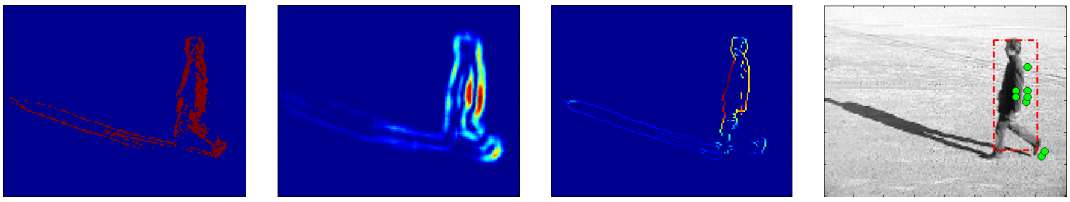

(a) *Boxing*
(b) *Hand clapping*
(c) *Hand waving*
(d) *Jogging*
(e) *Running*
(f) *Walking*

Figure 3.4: Examples of clouds of space-time interest points obtained using $S = 6$ and $N_s = 5$. In each frame the red rectangle represents the foreground area, the green points are the extracted interest points, and the yellow rectangles illustrate clouds of different scales.

Where $\mathbf{I}_t$ is the tth image frame. For the image frame $\mathbf{I}_t$, a total of S interest point clouds of different temporal scales are formed. They are denoted as: $[\mathbf{C}_t^1, \ldots, \mathbf{C}_t^s, \ldots, \mathbf{C}_t^S]$. More specifically, an interest point cloud of the s-th scale is constructed by accumulating the interest points detected over the past $s \times N_s$ frames, where N_s is the difference between two consecutive scales (in the number of frames). Examples of Clouds of Interest Points formed using the KTH and Weizmann datasets are shown in Figure 3.4. It can be seen from Figure 3.4 that different types of actions result in interest point clouds of very different shapes, relative locations (w.r.t body location), and distributions. It is also evident that interest point clouds of different temporal scales capture different aspects of human motion that potentially have different levels of discriminative power. (The cloud spatial scale is deified by the point distribution itself). This will be exploited by the feature selection method detailed later (Section 3.2.2).

3.2.1 Feature Extraction

For the S interest point clouds constructed for the t-th image frame $[\mathbf{C}_t^1, \ldots, \mathbf{C}_t^s, \ldots, \mathbf{C}_t^S]$, two sets of features are extracted. These features are significantly different from the local descriptors computed by conventional interest point based approaches. In particular, the interest point cloud features are global and holistic capturing distribution information of interest points, whilst the conventional descriptor features, computed from a cuboid centred at each interest point are local, describing appearance information of individual interest points. Advantage of the proposed



(a) *Frame difference* (b) *Filtering* (c) *Edge detection* (b) *Target localization*

Figure 3.5: Intermediate steps for target detection and localization. (a) Regions of interest are detected via frame difference. (b) 8 first-order derivative filters are applied to the image. (c) Prewitt edge detector is employed to segment the object. (d) Target localization with boundary box

method is the ability to overcome local problems such as background noise, short-term occlusions and outliers by exploiting the global representation. On the other hand, in the presence of sparse interest point the representation standard representation are more efficient.

The first set of interest point cloud features is concerned with the shape and speed of foreground objects. To reliably detect and segment a foreground object given camera movement, zooming, strong shadows, and noisy input is a non-trivial task. This is accomplished by the following procedure. Firstly, a binary mask is obtained via frame difference (this mask can be seen as a saliency map which identifies areas containing strong movements). Secondly, 8 first-order derivative filters oriented along $\{0°, \pm 22°, \pm 45°, \pm 67°, 90°\}$ are applied to the image frame. Thirdly, the responses of these filters are fused together with the frame difference mask. Finally, a Prewitt edge detector (Parker 1997) is employed to segment the object from the detected foreground area. The above mentioned four steps are presented in Figure 3.5. Once an object is segmented from the frame, two features are computed: O_t^r measuring the height to width ratio of the object, and O_t^{Sp} measuring the absolute speed of the object.

The second set of features is extracted from interest point clouds of different scales, they are thus scale dependent. Particularly, from the s-th scale cloud, 8 features are computed and denoted as:

$$[C_s^r, C_s^{Sp}, C_s^D, C_s^{Vd}, C_s^{Hd}, C_s^{Hr}, C_s^{Wr}, C_s^{Or}] \tag{3.10}$$

Note that subscript t is omitted for clarity. Specifically, C_s^r is the height to width ratio of the cloud; C_s^{Sp} is the absolute speed of the cloud; C_s^D is the density of the interest points within the cloud, which is computed as the total number of points normalised by the area of the cloud; C_s^{Vd} and C_s^{Hd} measure the spatial relationship between the cloud and the detected object area. Specifically, C_s^{Vd} is the vertical distance between the geometrical centre (centroid) of the object area and the

cloud, and C_s^{Hd} is the distance in the horizontal direction. C_s^{Hr} and C_s^{Wr} are the height ratio and width ratio between the object area and the cloud respectively. C_s^{Or} measures how much the two areas overlap. Overall, the 8 features can be put into two categories: C_s^r, C_s^{Sp}, and C_s^D measure the shape, speed, and density of the cloud itself; the 5 remaining features capture the relative shape and location information between the object and the cloud areas. Table 3.1 schematically lists all the features.

To make these features insensitive to outliers in the detected interest points, an outlier filter is deployed before the feature extraction, which evaluates the interest point distribution over 4 consecutive frames and removes those points that are too far away from the distribution centroid. Specifically, the points distribution centroid is estimated in each frame and compute the average distance from each point to the centroid. If the distance between an interest point and the centroid is 4 times or more of the average distance, it is most likely to be caused by background noise and thus removed.

Now each frame is represented using $8S+2$ features where S is the total number of scales (i.e. 8 features for each scale plus 2 scale-independent features O_t^r and O_t^{Sp}). By using a total of $(8S+2)\times T$ features to represent the whole action sequence leads to a feature space of an extremely high dimension. It is well known that a high dimensional feature space can cause over-fitting resulting in poor recognition performance. To reduce the dimensionality of the feature space, and more importantly, to make our representation less sensitive to feature noise and invariant to the length of each action sequence, a histogram of N_b bins is constructed for each of the $8S+2$ features collected over time via linear quantization. Consequently, each action sequence is represented as $8S+2$ histograms or $(8S+2)\times N_b$ scalar features with $N_b \ll T$. Instead of using fixed-width histogram binning as most existing work does, here it is adopted a histogram of non-uniform bin width with more bins being given to the high density area of the feature space (Kontkanen and Myllymaki 2007).

It has to be mentioned that the formulated representation appears local in time and global along the spatial domain. Specifically, multiple temporal scales are selected aiming to sample both local (short temporal scale) and global (larger temporal scale) action dynamics. Contrary, in the spatial domain a single region of interest is applied. This introduces some limitations in detecting a variety of scenarios such as: multiple targets, group actions and simultaneous actions. Additively, it does not allow exploring the action at atomic-motion level such as dividing the

target area in sub-regions. This is principally motivated by the fact that clean target detection is complicated, and with the available video resolution, a more sophisticate detector may introduce large amount of noise.

O_t^r	Target height to width ratio
O_t^{Sp}	Target absolute speed
C_s^r	Cloud height to width ratio
C_s^{Sp}	Cloud absolute speed
C_s^D	Interest points density within the cloud
C_s^{Vd}	Vertical distance between the target centroid and the cloud
C_s^{Hd}	Horizontal distance between the target centre and the cloud
C_s^{Hr}	Height ratio between the object area and the cloud
C_s^{Wr}	Width ratio between the object area and the cloud
C_s^{Or}	Measures the cloud and target overlap

Table 3.1: List of features used in the COP representation.

3.2.2 Feature Selection

Using the $(8S+2) \times N_b$ features as described above, the feature space dimension is still very high and needs to be further reduced. Moreover, there are uninformative and redundant features one would wish to eliminate from the feature set. To that end, a simple and intuitive yet effective feature selection method is formulated below.

Our feature selection can be defined as filters method (Yu and Liu 2004; Wang et al. 2007a) where the

$$R_{f_i} = \frac{\sqrt{\frac{1}{A}\sum_{a=1}^{A}(\mu_{f_i}^a - \hat{\mu_{f_i}})^2}}{\frac{1}{A}\sum_{a=1}^{A}\sigma_{f_i}^a} \tag{3.11}$$

where $\hat{\mu_{f_i}} = \frac{1}{A}\sum_{a=1}^{A}\mu_{f_i}^a$ is the inter class mean of the A intra class feature means. The numerator and denominator of the above equation correspond to the standard deviation of the intra class means, and the inter class mean of the intra class standard deviations respectively. The former measures how the feature value varies across different classes (the higher the value is, the more informative the feature f_i is); the latter tells how the value varies within each class (the lower the value, the more informative the feature). Overall, features with higher R_{f_i} values are preferred over those with lower ones. Finally, all features are ranked according to their R_{f_i} and a decision is made as to how many percent of the features are to be kept for recognition.

The proposed feature selection method, although intuitive, seems to have a number of drawbacks. Firstly, different features are selected separately as if they were independent of each

other. It has been widely recognised that combining good features together does not guarantee good recognition performance (Peng et al. 2005). So, ideally one would like to select the features collectively. However, this means that the feature search space is too high for an exhaustive search and even a sequential-search based approximation scheme is considerably expensive. Secondly, more sophisticated relevance measures such as mutual information (Peng et al. 2005) can be used. Nevertheless, compared with alternative feature selection approaches, one of the method advantages is that it has an extremely low computational cost. It is also shown empirically through experiments (see Sec. 3.4.2) that the proposed method is more effective than a far more complicated state-of-the-art method (Peng et al. 2005).

3.3 Combining Multi-scale Clouds of Interest Point Features

The Clouds of Interest Points (COP) features are of multiple (S) temporal scales. Features of different scales may not be equally informative in representing different actions. This is because each action can be better described by a subset of temporal scales instead of using the all. The majority of the observed actions are periodic (e.g. *running*, *walking*, *hand-clapping*), for them, the speed and lengths of their period are direct indications of their dominant temporal scales. For instance, in the KTH dataset at 25Hz, a full cycle of the *running*, *hand-clapping* and *walking* actions lasts around 20, 25, and 30 frames respectively. Intuitively, longer scale COP features are more useful in describing longer scale (slower) actions. Therefore it is necessary to weight the features of different scales according to their relevance to the classification task, and different weightings should be used for classifying different actions. Ideally these weightings should be learned automatically from a training dataset.

To this end, a multiple kernel learning (MKL) method is formulated for learning the optimal weighting of COP features of different scales for multi-class action classification. MKL was first introduced in (Bach et al. 2004) to address the problem of selecting the optimal combination of kernel functions for a specific feature for Support Vector Machine (SVM) classification. Recently it has been used in computer vision for addressing a closely related problem, that is, given a specific kernel function but different features capturing different aspects of a visual object, how to best combine them together to achieve the optimal classification performance (Gehler and Nowozin 2009; Sun et al. 2009a). In this work, the COP features of different scales capture the characteristics of an action class under multiple temporal scales and MKL is adopted to learn the

optimal combination of these features.

To formally define the multiple class action recognition problem, it is possible start to analyse the validation schema. In this work, the one-versus-rest scheme is employed, where C binary classifiers are learned to classify an action sequence into one of the C classes. By assuming that the training set is composed of N instances $(x_i, y_i)_{i=1,\dots,N}$; each training sample x_i is a video sequence containing an action with a class label y_i. To represent the action, S features are extracted as described in Section 3.2.1. Each feature is a histogram corresponding to COP features at one specific scale. s-th scale feature can be denoted as $f_s(x)$, where $f_s()$ is the feature extraction function. Using multiple kernel learning, a set of kernel functions is to be computed, each of which is essentially a distance/similarity measure. Specifically, a kernel function

$$k_s(x,x') = k(f_s(x), f_s(x')) \tag{3.12}$$

measures the similarity between a pair of action sequences represented using the s-th scale COP features. For notational convenience, given an action sequence x, its kernel response of the s-th feature to all N training samples is denoted as:

$$K_s(x) = [k_s(x,x_1), k_s(x,x_2), \dots, k_s(x,x_N)]^T \tag{3.13}$$

Now, it is studied how different kernels corresponding to different COP features are combined in an SVM framework. Using MKL, the objective is to learn an optimal weighting so that the combined kernel function has the following form:

$$k^*(x,x') = \sum_{s=1}^{S} \beta_s k_s(x,x') \tag{3.14}$$

where β_s is the weight associated to the s-th temporal scale. To learn an SVM for classifying one action class against the rest, an optimisation problem needs to be solved with the following objective function:

$$\begin{aligned} \min_{\alpha,\beta,b} \quad & \tfrac{1}{2}\sum_{s=1}^{S} \beta_s \alpha^T K_s \alpha + C\sum_{i=1}^{N} L\left(y_i, b + \sum_{s=1}^{S} \beta_s K_s(x)^T \alpha\right) \\ sb.t. \quad & \sum_{s=1}^{S} \beta_s = 1, \quad \beta_s \geq 0, \quad s = 1,\dots,S \end{aligned} \tag{3.15}$$

Where α is a N-dimensional feature vector which can be seen as the weights of each training sample, b has a scalar value, K_s is defined in Eqn. (3.13) and $L(y,z)$ denotes the Hinge Loss function (Bishop 2006). The two constraints put on β_s are to make sure that the estimated value of β_s is sparse and interpretable (i.e. as weights, they should be either zero or a positive number, and the sum of all weights should be 1). Various methods can be used to solve the above optimisation problem. In this work the semi-infinite linear program (SILP) (Sonnenburg et al. 2006) is adopted. Conventionally the multiple kernel learning problem is formulated as a convex quadratically constrained quadratic program and solved using a local descent algorithm such as Sequential Minimization Optimization (SMO). However, it is slow and only feasible for small scale problems. The method in (Sonnenburg et al. 2006) reformulates the multiple kernel learning problem as a semi-infinite linear program (SLIP), which can be efficiently solved using an off-the-shelf linear program solver and a standard SVM implementation. Two linear program solvers are formulated in (Sonnenburg et al. 2006); one is a wrapper algorithm and the other a chunking algorithm. The wrapper algorithm was used in the current implementation. Note that the regularisation constant C is determined via cross validation (only the training set is use in the validation process). Given the learned parameters β_s, α, and b, the final binary decision function of MKL is of the following form:

$$F_{MKL}(x) = sign\left(\sum_{s=1}^{S} \beta_s (K_s(x)^T \alpha + b)\right) \tag{3.16}$$

where the 'sign' function is a function that returns a value 1 if its parameter is positive and -1 if otherwise, $K_s(x$ is defined in Equation (3.13) which measures the similarity between the test data x with all N training data samples (both positive and negative). If $F_{MKL}(x)$ assumes the value 1, the test sequence x is deemed as being a member of the target action classes for which the MKL binary classifier is trained. Since it is required to solve a multiple class classification problem, multiple binary classifiers are trained and a test action sequence is classified as the action class with the highest value of $F_{MKL}(x)$.

3.4 Experiments

3.4.1 Experimental Settings

For the formulated MKL classifier, Gaussian kernels were used. All results were obtained using Leave-One-Out Cross-Validation (LOOCV); cross-validation imposes to learn the model param-

eters using the training set only and test the algorithm on the testing set (unknown data). It involved employing a group of clips from a single subject in a dataset as the testing data and the remaining clips as the training data. This was repeated so that each group of clips in the dataset is used once as the testing data. More specifically, for the KTH dataset (see Appendix A.2), the clips of 24 subjects were used for training and the clips of the remaining subject was used for validation. For the Weizmann action recognition dataset (see Appendix A.1), the training set contains 8 subjects. As for the Weizmann robustness test dataset, the whole Weizmann action recognition dataset was used as training set, and each of the 20 robustness test sequences were classified as one of the 10 action classes.

For constructing the multi-scale interest point clouds, the difference between two consecutive scales N_s was set to 5 frames and the total number of scales S was set to 6 (parameters learned empirically). This generates to 50 features ($8S+2$), each of which was represented as a 50-bin histogram (i.e. the COP features were represented in a 2500 dimensional space). 20% of these features were removed using the introduced feature selection method (See Section 3.2.2).

Additionally, the proposed Clouds of Points representation has been compared with a standard interest point based method which uses a Bag of Words (BOW) framework (Dollar et al. 2005). For extracting these BOW features, a codebook size of 300 was used for KTH and 250 for Weizmann[3]. Note that the Bag of Words method requires generating a codebook using a k-means clustering algorithm, which is sensitive to initialisation. Therefore, results are reported as an average of 20 trials. For the proposed COP features, no such initialisation issue exists, and different trials will give identical results.

3.4.2 Recognition Performance Evaluation

	BOW		COP	
	ACA	σ	ACA	σ
KTH	85.33%	1.23	**92.83**%	0
WEIZMANN	90%	0.78	**96**%	0

Table 3.2: Performance comparison between COP and BOW representations. The results are reported in terms of average classification accuracy and standard deviation over 20 trials

[3]The codebook sizes for KTH and Weizmann dataset were set to be between 200 to 500 empirically by most space-time interest point based methods reported in the literature. A few works (e.g. (Dollar et al. 2005)) investigated the effect of the size of codebook on the recognition performance and found that the performance is insensitive to the codebook size as long as it is within that range. Similar finding are reported in the proposed experiments.

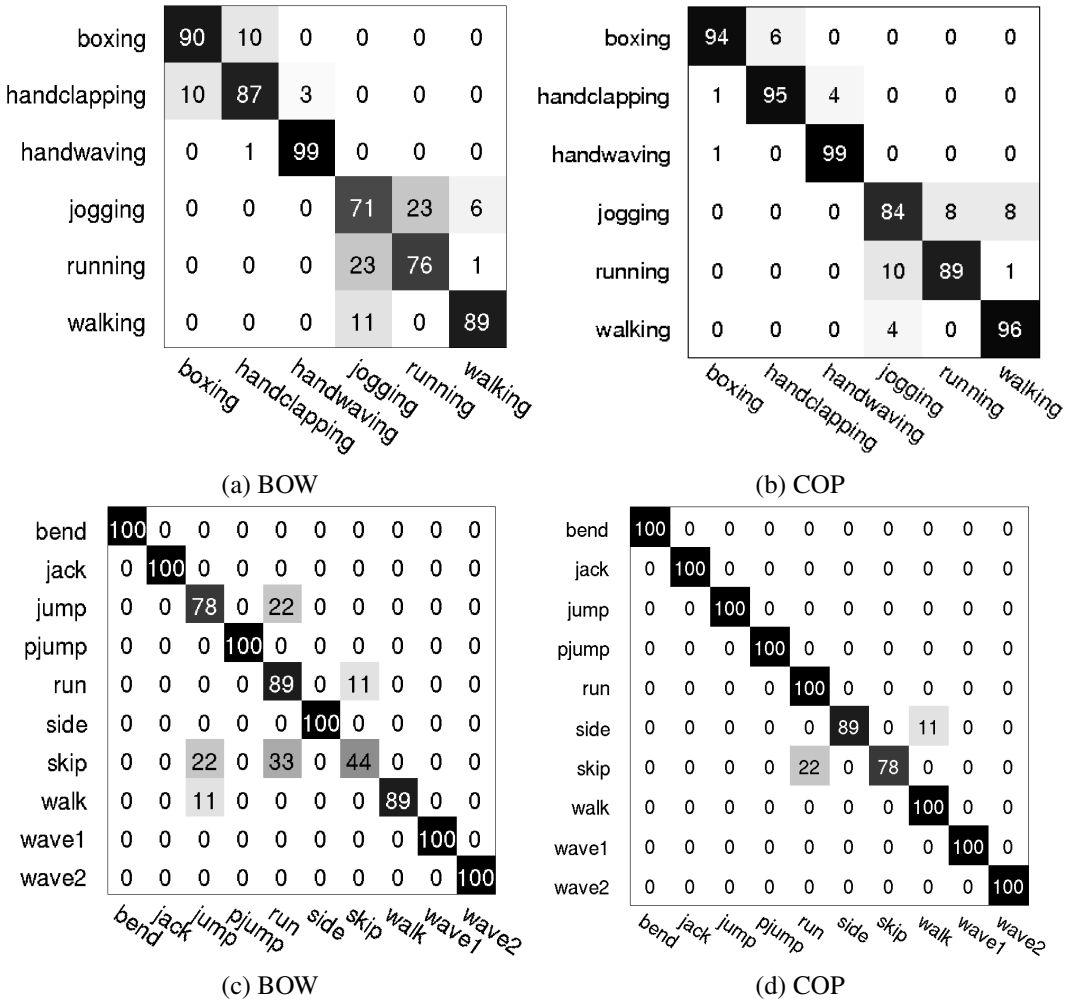


Figure 3.6: Recognition performance measured using confusion matrices: (a) KTH dataset, BOW representation, accuracy: 85.33% (b) KTH dataset, COP, accuracy: 92.83% (c) WEIZMANN dataset, BOW representation, accuracy: 90% (d) WEIZMANN dataset, COP, accuracy: 96%

Clouds of Points (COP) VS. Bag of Words (BOW) - The proposed COP is compared with BOW representation. The recognition results are presented in the form of averaged recognition rates in Table 3.2 and confusion matrices in Figure 3.6. Table 3.2 shows that the COP representation achieves higher average recognition rate on both datasets. Figure 3.6 also gives details on where the performance gain was obtained. It is noted that the COP representation is particularly strong in recognising *jogging*, *running*, and *walking* in the KTH dataset (comparing Figure 3.6(b) with (a)), and *running*, *skipping*, and *walking* in the Weizmann dataset (comparing Figure 3.6(d) with (c)). These actions are similar in terms of shape and motion appearance, but differ in terms of action speed and temporal evolution, which can only be measured globally and over different temporal scales. The BOW representation, based on interest point appearance only, is unable to

capture these differences thus its performance is inferior. On the contrary, the proposed COP representation measures explicitly and globally the spatial and temporal distribution information. Moreover, it describes actions over multiple temporal scales, which is particularly useful for distinguishing actions that differ mainly in temporal scales (e.g. *running* and *walking*). To be noticed that in the Weizmann dataset, COP outperforms BOW in all the actions except *galloping sideways*. *Galloping sideways* appears periodical and relatively slow thus well described by the BOW representation.

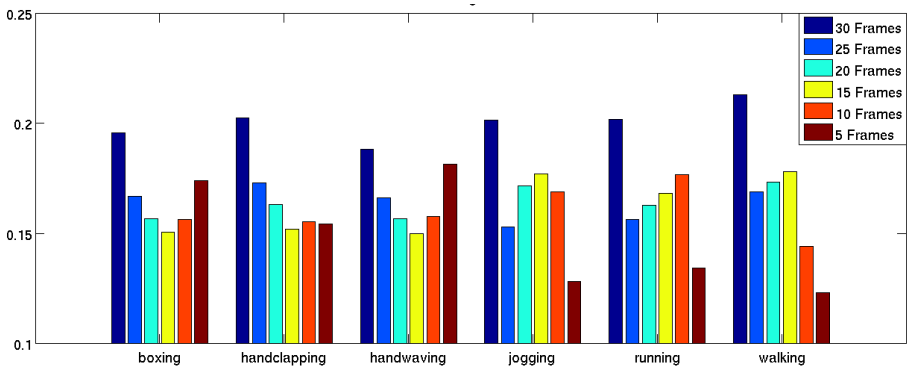
	Concatenation	MKL
KTH	92.50%	**92.83%**
WEIZMANN	95%	**96%**

Table 3.3: Performance comparison between MKL and concatenation based feature combination.

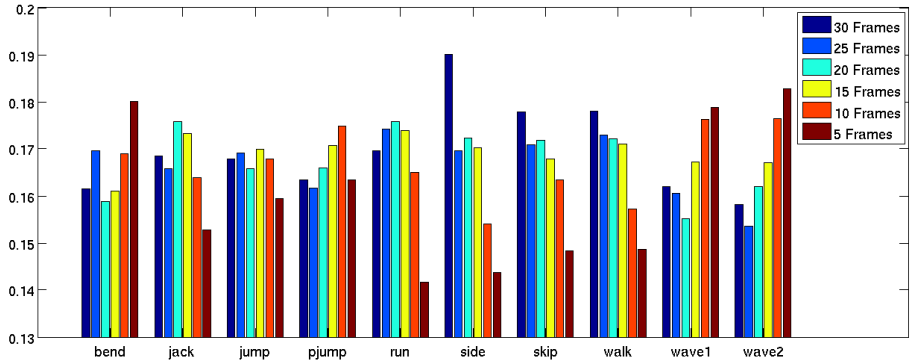
Multi-Scale Recognition: MKL vs. Concatenation - The COP representation contains features of multiple scales. Experiments were carried out to compare two ways of combining these multi-scale features: the proposed MKL method and the simple concatenation method. The former learns the optimal weighting from a training dataset, whilst the latter gives an equal weight to all the scales. (Before concatenation the features are normalized with average zero and standard deviation one).

The obtained result is shown in Table 3.3 indicates that MKL yields slight improvement compared with concatenation based feature combination.

Figure 3.7 shows the weight distributions over the multiple scale COP features learned by MKL. It is clear that different weights are assigned to COP features of different scales, and the weight distributions vary for different actions. As expected, the learned weights reflect the temporal scales of different actions. For instance, for *walking*, *jogging* and *running* in the KTH dataset, as the action is faster, more weights were assigned to shorter scale features (Figure 3.7(a)). Similarly, in the Weizmann dataset it is observed that shorter term (faster) actions such as *waving* received significantly more weights for the shorter scale features. In the meantime, longer term actions such as *hand clapping* and *galloping sideways* received more weights for the longer scale features (Figure 3.7(b)). By exploiting the different discriminative power of different feature scales, the MKL based feature combination is able to produce slight better performance than simple feature concatenation.



(a) KTH dataset



(b) Weizmann dataset

Figure 3.7: Weight distribution of 6 multi-scale COP features learned using MKL.

Table 3.4 also compares the obtained results with the existing approaches proposed recently, which are not restricted to interest points based methods. It shows that the obtained results are close to the results reported so far on each dataset, and outperform some of the recently proposed methods, especially those tested on both datasets.

Interest Point Detector Evaluation - The proposed interest point detector (Section 3.1) was compared with the widely used Dollar et al. (2005) detector and the result is shown in Table 3.5. As can be seen, using the same COP representation, the proposed detector outperforms the Dollar et al. (2005) detector on both the KTH and Weizmann datasets. This is because the proposed detector is less sensitive to dynamic backgrounds and camera movements. Moreover, it tends to select more meaningful points located near the moving body parts (see Figure 3.2). The improvement is particularly significant for the KTH dataset where dynamic background and camera motions appear frequently. Note that the presented detector differs from the Dollar et al. (2005) detector in both the way the Gabor filters are designed and the use of frame differencing as a pre-processing step. To investigate the effect of each difference individually, frame differencing is also applied to the Dollar et al. (2005) detector. The result in Table 3.5 shows that an improve-

METHOD	KTH	WEIZMANN
Proposed approach	**92.83%**	**96%**
Sun et al. (2009b)	94.00%	97.80%
Ikizler et al. (2008)	94.00%	-
Lin et al. (2009)	93.43%	-
Wang and Mori (2009)	92.51%	100%
Liu et al. (2009b)	92.30%	-
Fathi and Mori (2008)	90.50%	100%
Zhang et al. (2008)	91.33%	92.89%
Kläser et al. (2008a)	91.40%	84.30%
Niebles et al. (2008)	83.30%	90.00%
Dollar et al. (2005)	81.17%	85.20%
Liu and Shah (2008)	94.16%	-
Zhao and Elgammal (2008)	91.17%	-
Gilbert et al. (2008)	89.92%	-
Savarese et al. (2008)	86.83%	-
Nowozin et al. (2007)	84.72%	-

Table 3.4: Performance comparison with state-of-the-art.

	Dollar et al. (2005)	Dollar et al. (2005) with FD	Proposed detector
KTH	90.08%	92.00%	**92.83%**
WEIZMANN	93%	94%	**96%**

Table 3.5: Performance comparison between the proposed interest point detector and the one presented by Dollar et al. (2005) with and without frame differencing (FD).

ment can be obtained. However the result is still worse than that of our detector. This suggests that the advantage of our detector is due to both the use of frame differencing and the way the Gabor filters are designed.

	No feature selection	mRMR (Peng et al. 2005)	Proposed method
KTH	89.03 %	91.32 %	**92.83%**
WEIZMANN	93%	94%	**96%**

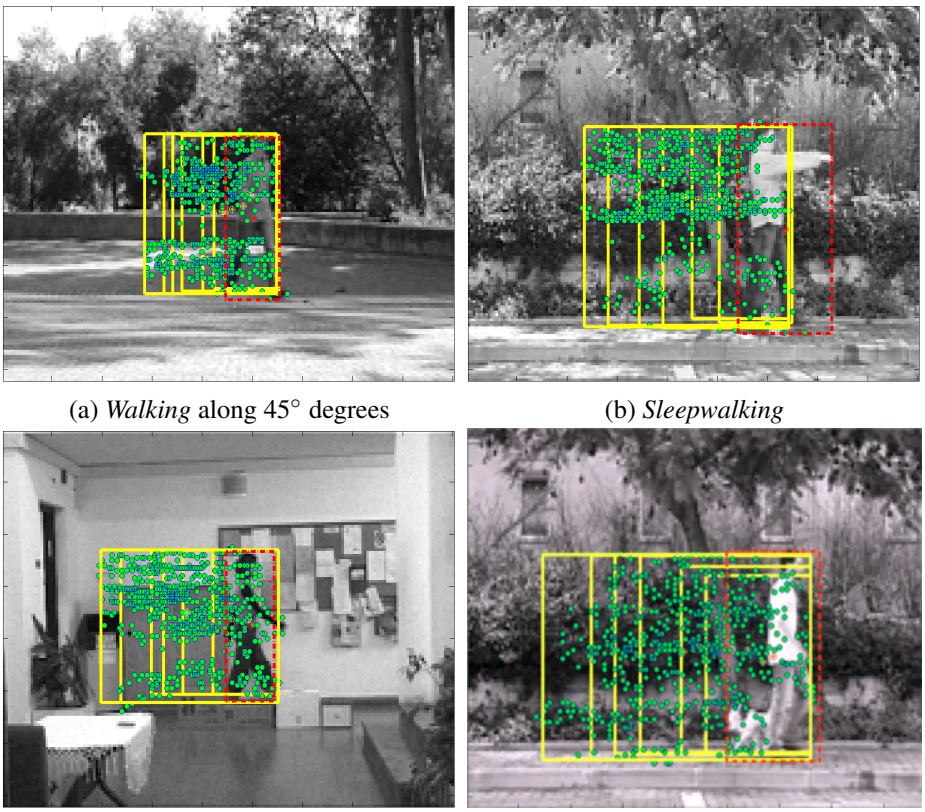
Table 3.6: Performance comparison between different feature selection approaches.

Effects of Feature Selection - The proposed COP representation was evaluated in three scenarios: without feature selection, with the proposed feature selection approach (Section 3.2.2), and with a more complex minimal-redundancy-maximal-relevance (mRMR) algorithm proposed in (Peng et al. 2005). Table 3.6 shows that feature selection improves the recognition performance and the best performance is obtained when the proposed feature selection method is employed. Note that a major attraction of the mRMR method, as compared with other existing feature selection methods, is its low computational cost. The proposed feature selection method has an even lower computational cost. Specifically, this method took less than one twelfth of the time used by the mRMR method for selecting the same amount of features (7.1 seconds using the Section

3.2.2 method to measure and rank 2500 features, as compared with 90 seconds using mRMR on a 2.1G PC platform with 4G RAM).

Processing Time - During training, most of the computation time was spent on feature extraction. Specifically for each leave-one-out run, on average the amount of time required for feature extraction was 403.40 seconds for Weizmann and 1365.60 seconds for KTH. (There are much more training clips in KTH than Weizmann). After feature extraction and selection, the training of the multiple kernel SVM classifier was much faster, needing on average 0.32 seconds for Weizmann and 3.28 seconds for KTH. During testing, the average processing times for each test clip on the Weizmann dataset were: 4.60 seconds on feature extraction and 0.0017 seconds for classification. For the KTH dataset those numbers became 23.50 and 0.0019 respectively. All implementations were in Matlab on a 2.1G PC platform with 4G RAM.

3.4.3 Robustness Evaluation



(a) *Walking* along 45° degrees (b) *Sleepwalking*

(c) *Walking* with occluded legs (d) *Walking with a dog*

Figure 3.8: Example of Clouds of Points detected in the sequences used in the robustness test experiments.

The robustness of the presented method is demonstrated using the Weizmann robustness test

	Correct recognition
COP	**19 out of 20**
BOW	10 out of 20
Blank et al. (2005)	19 out 0f 20
Wang and Suter (2007)	18 out of 20

Table 3.7: Robustness test result.

sequences. Examples of the detected Clouds of Interest Points are shown in Figure 3.8. The result is reported in Table 3.7. It can be seen that BOW based representation is very sensitive to view angle, variations in action, and occlusions, with only half of the test sequences being recognised correctly. In contrast, the proposed COP representation is much more robust, with only a single misclassified sequence (a person *walking with a dog* was recognised as *skipping*). In the sequence, the most informative human body part for the action (i.e. the legs) overlapped with another object (the dog), which was also *walking* but in a very different way (see Figure 3.8(d)). Table 3.7 shows that the proposed method outperforms as well as other existing action recognition approaches that have reported results on this robustness test dataset.

3.5 Discussions

Existing interest points based methods describe actions by employing local descriptors only, ignoring potentially valuable information associated with the global point distribution. Moreover, these methods observe actions with a fixed temporal scale, limiting their representation's discriminative power. In contrast, this chapter formulates a novel action representation method, which differs significantly from the mentioned methods in that only the global distribution information of interest points is exploited. In particular, the proposed method initially extracts holistic features from clouds of interest points which have been accumulated over multiple temporal scales. Then, it merges the features in a robust representation by using a Multiple Kernel Learning strategy. This allows the optimal weight for each temporal scale to be automatically defined in accordance with the observed action class.

Compared to existing methods, the proposed COP representation is less sensitive to background noise and occlusion. Since the action is observed globally over a window of frames, local noise from the background and short temporal occlusions are overcome. It is also robust to view changes and able to capture smooth motions. Furthermore, COP avoids the significant problems of selecting the optimal local descriptor, clustering algorithm for constructing a codebook, and

codebook size faced by previous interest points based methods. The reported performance in terms of execution time is also encouraging, being notably faster than those seen with traditional interest point based methods.

Experiments using the KTH and Weizmann datasets demonstrate that the proposed approach is comparable to the state-of-the-art. Additionally, the performed robustness test reports the best performance compared to other methods.

For the KTH dataset, the errors made by our approach come mainly from three classes: *jogging*, *running*, and *walking* all of which are visually very similar. With the global features extracted using Clouds of Points representation, fewer errors were made when compared to a conventional interest points based method (see Figure 3.6). However, there are still misclassifications between *jogging* and *running*, as there is no clear separation between these two action classes; identifying when *running* becomes slow enough to be labelled as *jogging* is a subjective human process. As for the Weizmann dataset, COP tends to mistake *skipping* for either *jumping* or *running*, and also *galloping sideways* as *walking*. Again, *skipping* is a combination of *jumping* and *running*, and *galloping sideways* is visually very similar to *walking*. In order to avoid these mistakes, it is possible to build a human body model and separate different body parts in the representation. Alternatively, features should be extracted from 3D human body shapes. However, as stated in Section 2.1.3, both model tracking based approaches and spatio-temporal shape template based approaches require highly detailed silhouettes to be extracted. They thus stand no chance on noisy data such as the KTH dataset, where silhouette extraction is very complex.

In light of these experiments, it can be noted how Bag of Words and Cloud of Points representations exploit alternative but complementary cues to describe actions (local descriptors and global distribution respectively). Consequently, a more robust representation can be achieved if the two representations are merged, leading to a more accurate action classification. Motivated by this intuition, the next Chapter will show how the Multi Kernel Learning algorithm can be extended to fuse the BOW features with the actual COP features. The advantages of feature fusion will be also discussed.

From the obtained results, the importance of feature selection also emerges. In practice, not all of the extracted features are significant, making it important to remove redundant and noisy components. As shown in Table 3.6, the proposed feature selection method, even if simple, improves performance in both the KTH and Weizmann datasets. Although efficient, one of the

limitations of this method is the evaluation of features independently rather than collaboratively. In other words, the features are singularly analysed, ignoring potential information associated with multiple-feature patterns. To address this issue, the next chapter presents an innovative Multi-Class Delta Latent Dirichlet Allocation model for feature selection in which, features are collaboratively analysed.

3.6 Summary

This chapter introduces a robust interest points based representation named Clouds of Points (COP), which is able to exploit information about the global spatio-temporal distribution of points. COP aims to observe clouds of interest points generated by human actions, and also to collect information at different temporal scales.

The first step of the proposed method consists of a **new space-time interest point detection method**, which extracts denser and more informative points when compared to existing methods. In particular, our model avoids spurious detection in both background areas and highly textured static foreground areas unrepresentative of the dynamic parts of concerned actions. The extracted interest points are accumulated over time at different temporal scales to form point clouds. Holistic features are then computed from these point clouds for action representation, and these capture *explicitly* and *globally* the spatial and temporal points distributions. The sets of holistic features collected over different temporal scales are then automatically weighted accordingly to their relevance in the classification task. To this end, a **Multiple Kernel Learning strategy** is employed. Specifically, to learn a multi-class classifier for action recognition, support vector machine with multiple kernels is employed, each kernel being associated with a certain temporal scale. Multiple Kernel Learning is then performed to learn the best linear combination of the kernels for yielding optimal classification accuracy.

The proposed approach is evaluated using two widely used public datasets, namely the KTH dataset and the Weizmann dataset. The experimental results demonstrate that our approach is comparable with the state-of-the-art. Furthermore, it is more robust against occlusion and changes in view angle condition when compared to existing methods. Crucially, the results also highlight the importance of both the formulated interest point detector and the feature selection step.

Despite the proposed representation appears robust to moderate distortions and dynamic

backgrounds, it has been observed that a more solid representation can be achieved improving the feature selection approach and exploiting more comprehensive features. Following this idea, in Chapter 4 a more sophisticated feature selection approach that aims to rank features collaboratively is formulated. Additionally, a richer feature space is obtained fusing features originated by complementary methods.

Chapter 4

Feature Fusion and Selection

To improve the action representation, it is important to 1) explore different sources of information and 2) filter out noisy and redundant features. This chapter studies these specific problems and suggests innovative approaches designed to fuse different representations and collaboratively select features.

Initially, the fact that the proposed Clouds of Points (COP) and conventional Bag of Words (BOW) representations exploit different but complementary sources of information is taken into account. COP provides a global action description by relying on interest points distribution, while the BOW representation exploits local descriptor information. The two representations are then fused with the aim of achieving a more solid action representation. Specifically, a Multiple Kernel Learning fusion strategy is formulated.

A different scenario is drawn by processing more challenging sequences such as realistic videos. Due to the presence of a constantly moving camera and a crowded background, the proposed Clouds of Points representation appears inadequate and unable to produce significant results. As such, an alternative representation based on key-point trajectories is formulated. The proposed approach removes the trajectories associated with the background and subsequently focuses attention on a specific region of interest. To further remove misleading features, a novel Multi-Class Delta Latent Dirichlet Allocation model for feature selection is formulated. The most informative features are selected collaboratively rather than independently. Collaboratively imply that the features are selected observing their group-behaviour across all the classes. Finally, to enrich the action representation, an adaptive feature fusion method is then developed to combine

the proposed trajectory based representation with a conventional BOW representation. The fusion method, in accordance with the camera movements detected, selects the optimal fusion strategy in order to cope with drastic changes in motion.

4.1 Fusion of Interest Points Based Representations

Feature fusion techniques have already been successfully employed to improve classification tasks, as presented in Section 2.3. As observed in the previous chapter, the proposed Clouds of Points representation and the conventional Bag of Words representation (Dollar et al. 2005) exploit two types of features, which capture completely different yet complementary aspects of the actions. The former (COP) contains global distribution information of interest points, while the latter (BOW) represents how each interest point looks in terms of 3D texture and localised motion characteristics. As such, to increase the action representation discriminative power, a fusion method is designed to merge both COP and BOW representations.

This fusion problem can be considered as a feature combination problem and addressed using the Multiple Kernel Learning method described in Section 3.3. More specifically, after interest points are extracted using the method described in Section 3.1 and represented as a histogram of the BOW features, a kernel function denoted as k_B can be computed and used to form a linear combination with the S COP features. Now the combined kernel function in Equation (3.14) is rewritten as:

$$k^*(x,x') = \sum_{s=1}^{S} \beta_s k_s(x,x') + \beta_B k_B(x,x') \tag{4.1}$$

Similarly, the objective function to be optimised using the SLIP algorithm (Sonnenburg et al. 2006) becomes:

$$\begin{aligned} \min_{\alpha,\beta,b} \quad & \frac{1}{2}\left(\sum_{s=1}^{S} \beta_s \alpha^T K_s \alpha + \beta_B \alpha^T K_B \alpha\right) \\ & + C\sum_{i=1}^{N} L\left(y_i, b + \sum_{s=1}^{S} \beta_s K_s(x)^T \alpha + \beta_B K_B(x)^T \alpha\right) \\ sb.t. \quad & \sum_{s=1}^{S} \beta_s + \beta_B = 1, \quad \beta_s \geq 0, \quad s = 1,\ldots,S, \quad \beta_B \geq 0 \end{aligned} \tag{4.2}$$

where β_B is the weight of the BOW features. After parameter estimation, the final binary decision function is:

$$F_{MKL}(x) = sign\left(\sum_{s=1}^{S} \beta_s (K_s(x)^T \alpha + b) + \beta_B (K_B(x)^T \alpha + b)\right) \tag{4.3}$$

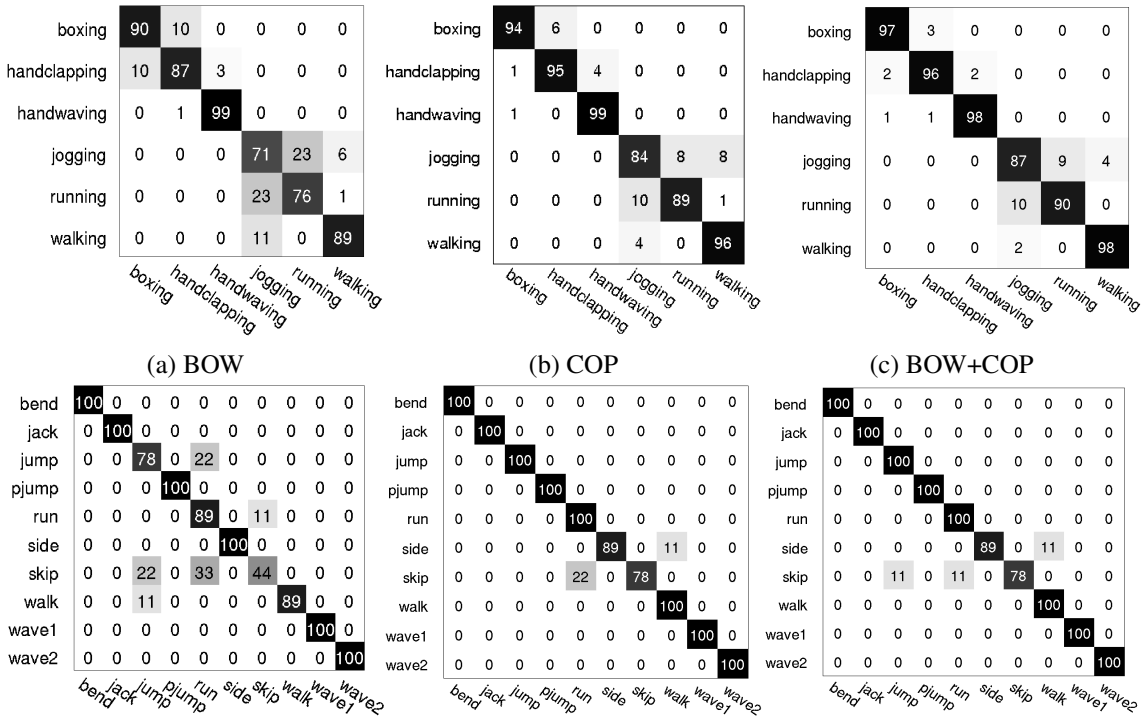
Figure 4.1: From top to bottom: Example frames from the Weizmann, Weizamnn robustness and KTH datasets. More details are reported in the Appendix A.

Although interest points based methods are robust to action distortions and noisy background, they appear inadequate in the presence of constant camera movements and crowded background. This limitation can be clearly observed processing videos recorded in unconstrained scenarios, such as the sequence contained in the YouTube Dataset (Liu et al. 2009a) (See Appendix A.6). The strong presence of interest points associated with the background cause that COP captures incorrect point distributions; consequently poor action description is experienced. A similar problem affects also conventional BOW representations, even if in a softer manner. To address this issue, Section 4.2 formulates an alternative action representation, which instead of relying on interest points, exploits key-point trajectories. As will be demonstrated, the principal advantage of the key-point trajectory representation is to be able to properly discharge background and noisy components, fundamental issue in performing action recognition in realistic scenarios.

4.1.1 Validation of Multiple Kernel Learning Fusion

Experiments have been conducted to validate the effectiveness of the proposed Multiple Kernel Learning fusion, where two complementary interest points based representations are merged (COP and BOW). In the validation process both the KTH (Schüldt et al. 2004) and Weizmman (Blank et al. 2005) datasets are analysed. To present a fair comparison, the same settings reported in Section 3.4.1 are used. More specifically, the formulated MKL classifier utilises Gaussian kernels, while the validation follows a Leave-One-Out Cross-Validation (LOOCV) schema.

For constructing the multi-scale interest point clouds and the BOW representation, the same setting reported in Section 3.4.1 is used. Note that the Bag of Words method requires generating



(a) BOW (b) COP (c) BOW+COP

(d) BOW (e) COP (f) BOW+COP

Figure 4.2: Recognition performance measured using confusion matrices: (a) KTH dataset, BOW representation, accuracy: 85.33% (b) KTH dataset, COP, accuracy: 92.83% (c) KTH dataset, BOW+COP, accuracy: 94.33% (d) WEIZMANN dataset, BOW representation, accuracy: 90% (e) WEIZMANN dataset, COP, accuracy: 96% (f) WEIZMANN dataset, BOW+COP, accuracy: 96%

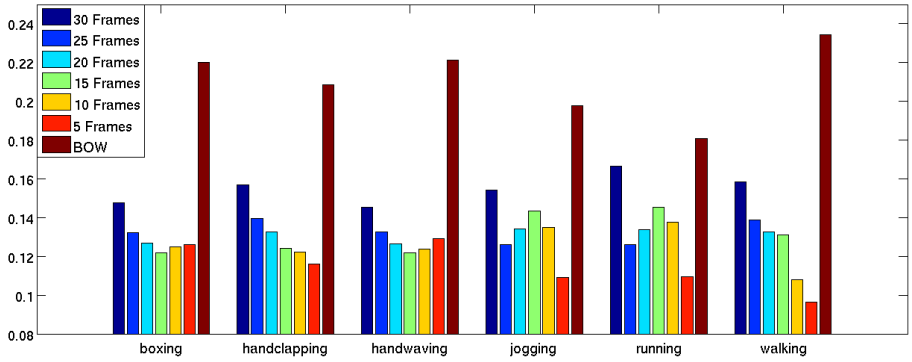
a codebook using a K-means clustering algorithm, which is sensitive to initialisation. Therefore, results are reported as an average of 20 trials. For the proposed COP features, no such initialisation issue exists, and different trials will give identical results.

Table 4.1 and Figure 4.2 present a performance comparison between using a single type of features, either BOW or COP, and the fusion of them using MKL. Table 4.1 shows that an improvement is obtained by fusing the two complementary features together on the KTH dataset. Specifically, Figure 4.2 shows that with feature fusion, the recognition rates for all 6 classes except *hand waving* were increased. It can also be seen in Table 4.1 that a simple concatenation based fusion has a negative effect on the recognition performance on both datasets. In this experiment the concatenation is performed normalizing the features with mean zero and standard deviation one. Table 4.1 also reports that for Weizmann dataset no enhancement is reported using MKL, but concatenation decreases the performance.

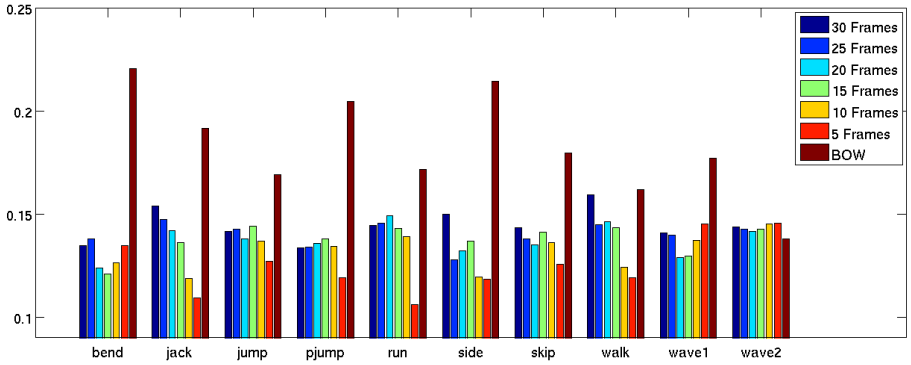
Figure 4.3 shows that different weight distributions were learned using MKL for different action classes. In general, the weights given to the BOW features are higher than those given

	BOW		COP		Concatenation		MKL Fusion	
	ACA	σ	ACA	σ	ACA	σ	ACA	σ
KTH	85.33%	1.32	92.83%	0	92.66%	0.98	**94.33%**	1.05
WEIZMANN	90%	0.78	96%	0	94 %	0.62	**96%**	0.73

Table 4.1: Effect of MKL feature fusion. Results presented in terms of average classification accuracy and standard deviation.



(a) KTH dataset



(b) Weizmann dataset

Figure 4.3: Weight distribution between Bag of Word and Clouds of Points features.

to each single scale COP feature, although overall more weight was given to COP features (It has to be mentioned that the two representations have comparable dimensionality). It is interesting to note that the weighting distribution is again largely determined by the temporal scale of different actions. In particular, BOW features are given more weight for actions with longer temporal scales (slower). For instance, in the KTH dataset, the weights of BOW features for *running*, *jogging* and *walking* are ascending in that order as the action scale gets longer. This is because when an action is performed with high motion intensity, both the detection of interest points and computation of local appearance descriptors become unreliable, which decreases the discriminative power of the BOW features.

Table 4.2 also compares the MKL results with the existing approaches proposed recently,

METHOD	KTH	WEIZMANN
MKL approach	**94.33%**	**96%**
Sun et al. (2009b)	94.00%	97.80%
Ikizler et al. (2008)	94.00%	-
Lin et al. (2009)	93.43%	-
Wang and Mori (2009)	92.51%	100%
Liu et al. (2009b)	92.30%	-
Fathi and Mori (2008)	90.50%	100%
Zhang et al. (2008)	91.33%	92.89%
Kläser et al. (2008a)	91.40%	84.30%
Niebles et al. (2008)	83.30%	90.00%
Dollar et al. (2005)	81.17%	85.20%
Liu and Shah (2008)	94.16%	-
Zhao and Elgammal (2008)	91.17%	-
Gilbert et al. (2008)	89.92%	-
Savarese et al. (2008)	86.83%	-
Nowozin et al. (2007)	84.72%	-

Table 4.2: MKL performance comparison with state-of-the-art.

which are not restricted to interest points based methods. It shows that MKL fusion performances are close to the best results reported so far on each dataset, and outperform most of the recently proposed methods, especially those tested on both KTH and Weizmann dataset.

	Correct recognition
COP+BOW	**20 out of 20**
COP	19 out of 20
BOW	10 out of 20
Blank et al. (2005)	19 out 0f 20
Wang and Suter (2007)	18 out of 20

Table 4.3: Robustness test result using MKL fusion.

Robustness Evaluation for Multiple Kernel Learning Fusion - The superior discriminative power and robustness of the formulated MKL fusion strategy is demonstrated processing the Weizmann robustness test sequences. As also reported in Chapter 3, it can be seen that BOW based representation is very sensitive to view angle, variations in action, and occlusions, with only half of the test sequences being recognised correctly. In contrast, COP only is more robust, with only a single misclassified sequence. By applying MKL to fuse COP and BOW, all the sequences are correctly classified. This result suggests that BOW and COP feature fusion does improve the robustness of action recognition. In particular, by merging two complementary representations such as COP and BOW, it is possible to overcome partial occlusions and action distortions caused by other objects (for example, those caused by the dog in the *walking with the dog* sequence). Table 4.3 shows that the formulated MKL fusion method also outperforms

(a) (b) (c)

Figure 4.4: (a) Orientation-Magnitude Descriptor: An action trajectory is quantized and converted to a histogram. (b) All the detected trajectories. (c) Trajectories of interest and ROI.

existing action recognition approaches that have reported results on this robustness test dataset.

4.2 Fusion of Trajectory and Interest Points Based Representation

Representing actions in realistic scenarios using trajectorie's segments has been shown to be successful by Sun et al. (2009a). Trajectory segments (5 to 25 frames) are easy to extract, very efficient in capturing human dynamics and at the same time no specific constraints are imposed. Inspired by this idea, this section formulates a novel method to extract and represent dense trajectory segments. It should be mentioned that a main problem of these realistic sequences consists of the strong presence of noisy trajectories. To this end, the background trajectory behaviour is estimated and used to filter out these components. Next, trajectories located outside a region of interest are discharged. After the first feature selection attempt, a more sophisticated method is formulated to collaboratively select relevant features. Finally, since the trajectory-based representation is complementary to Bag of Words interest points-based representation, the action representation is then enriched by fusing the two representations.

4.2.1 Trajectory Based Features

Trajectory Generation - The trajectories of key-points are computed using two techniques: the Pyramid Lucas-Kanade-Tomasi (KLT) tracker (Messing et al. 2009; Matikainen et al. 2009) and SIFT matching (Sun et al. 2009a). These two trackers are applied independently to a video footage so that as dense as possible trajectories can be obtained even in low textured videos. During tracking, a trajectory is considered "reliable" if it lasts more than 5 frames. Any shorter trajectory is automatically removed. When a trajectory reaches a pre-defined maximum length (25 frames), it is auto-segmented and a new trajectory is created. (The segmentation is automatically performed, practically as soon as a trajectory reach 25 frames, a new trajectory is

initialized.)

Trajectory Pruning - The extracted trajectories in a video may not always be useful for action recognition. For example, in Fig. 4.4 (b), lots of trajectories are extracted from the background area and thus need to be removed in order to retain the most relevant trajectories for describing the body actions of the person (Fig. 4.4 (c)). To that end, a trajectory pruning process is considered. The proposed approach detects a region of interest (ROI) from each video frame by measuring trajectory similarity within a temporal window. Suppose there exist N trajectories that pass through a frame f: $\mathbf{T} = \{\mathbf{t}_i\}$, $i = 1, \cdots, N$. For each trajectory, it is defined a trajectory segment $\mathbf{ts}_i$ within a temporal window of 4 frames centred at frame f and each framewise displacement vector $\mathbf{ds}_i$ as: $\mathbf{ts}_i = \{(x^i_{f-1}, y^i_{f-1}), (x^i_f, y^i_f), (x^i_{f+1}, y^i_{f+1}), (x^i_{f+2}, y^i_{f+2})\}$, and $\mathbf{ds}_i = \{d^i_1, d^i_2, d^i_3\}$, in which $d^i_k = \left(x^i_{f+(k-1)} - x^i_{f+(k-2)}, y^i_{f+(k-1)} - y^i_{f+(k-2)}\right)$. By aiming to measure similarity between any pair of trajectory displacement vectors $\mathbf{ds}_i$ and $\mathbf{ds}_j$. This results in an $N \times N$ dimensional similarity matrix $\mathbf{C}$ with entries:

$$\mathbf{C}_{i,j} = \sum_{k=1}^{3} \|ds^i_k - ds^j_k\|, \tag{4.4}$$

from which a single similarity score is computed for trajectory $\mathbf{t}_i$ as $m_i = \sum_{j=1}^{N} \mathbf{C}_{i,j}$. This score measures the similarity of this trajectory to all the other trajectories within a 4-frame temporal window centred at the frame f. It is considered that, if a trajectory is very similar to the others, it is likely that it is extracted from background clutter thus should be removed. To this end, in each frame f, an adaptive threshold $M^f_{TH} = \frac{\gamma}{N} \sum_{i=1}^{N} m_i$ is computed, where γ is empirically set to 1.3. (γ is a similarity coefficient, as small it is as selective is the filtering.)

Next, any trajectories whose similarity score is larger than M^f_{TH} is removed . After thresholding, assume N_f trajectories are left in the frame f. The centroid of the region of interest (ROI) can be computed by averaging spatial coordinates all key points on the remaining tracks:

$$\hat{x} = \frac{1}{N_f} \sum_{i=1}^{N_f} x^i_f, \qquad \hat{y} = \frac{1}{N_f} \sum_{i=1}^{N_f} y^i_f \tag{4.5}$$

and its dimensions are given by $D_x = 2\sqrt{2c_{xx}}$ and $D_y = 2\sqrt{2c_{yy}}$ where c_{xx} and c_{yy} are the second central moments of reliable key points. Any trajectory of interest which is located outside the ROI will then be removed. An example of remaining trajectories after pruning is shown in Fig. 4.4 (c). It shows clearly that the region of interest corresponds accurately to where the action takes place.

Region of interest detection has been exploited elsewhere (Liu et al. 2009a) based on 2D interest point detection. In contrast, the formulated approach is based on statistical analysis of key point trajectory distributions and is robust for videos captured by both static and moving cameras. The assumption behind this method is that the large majority of the trajectories are associated with the background and few with the target. This allows modelling and subsequently removing the background trajectories.

Advantages of the formulated ROI are: 1) the ability to filter out meaningless components and background object; 2) to be able to automatically adjust the action size and location; and to do not require any tracking method. On the other hand, the ROI is not designed to handle multiple targets. For instance, if a car and a person are moving closely a single ROI is selected including both car and person. Additionally, if the target does not generate enough trajectories the ROI cannot be computed.

Orientation-Magnitude Descriptor - The basic idea is to describe how trajectories move in terms of direction (orientation) and displacement (magnitude). Different actions generate sets of trajectories having a unique dynamic, which it can be used to recognize themselves. Along this idea, this descriptor aims to express these dynamics in a compact way. Given two consecutive points: $P = (x_l, y_l)$, $P' = (x_{l+1}, y_{l+1})$, along a trajectory as illustrated in Fig. 4.4(a), the displacement vector is computed between the two points as $d = (x_{l+1} - x_l, y_{l+1} - y_l)$. For a single trajectory $\mathbf{t}$ of length L a series of displacement vectors $\mathbf{d} = \{d_1, d_2, ..., d_{L-1}\}$ can be obtained. The quantization on $\mathbf{d}$ is performed by considering both magnitude and orientation of the displacement vectors. For magnitude quantization, each displacement vector is normalised by the largest displacement magnitude within the same trajectory and 4 uniform quantization levels are used. For orientation quantization, the top and bottom half circles are divided into 8 equal sectors , each subtending 22.5°, as shown in Fig. 4.4(a). The formulated quantization results in a track descriptor that is both scale-invariant (to be scale-invariant across different videos the same frame rate is required) and direction-invariant (direction-invariant means that an action performed left to right generates the same descriptors of the same action performed right to left). Combining magnitude and orientation quantization, each trajectory is then described by a 32-bin histogram O.

Trajectory Shape Descriptor - Complex Fourier descriptors are commonly used to represent

and compare shapes extracted from object silhouettes for object recognition (Zhang and Lu 2003). Here the Fourier descriptor to describe the shape signature of a trajectory is formulated. Assuming a trajectory consists of L key points $\{(x_1,y_1),(x_2,y_2),...,(x_{L-1},y_{L-1},)\}$ it can be then transformed in a closed contour connecting the starting and ending point; doing this the Fourier descriptor can be computed. The aim is to describe this trajectory as a 2D shape of N vertices $\{z(i) : i = 1,\ldots,N\}$. These N vertices can be computed using the N coefficients of the Fourier transform of $\{z(i)\}$:

$$z_i = \sum_{k=\frac{-N}{2}+1}^{\frac{N}{2}} c_k \exp\left(2\pi j \frac{ki}{N}\right). \tag{4.6}$$

The Fourier coefficients c_k present the frequency contents of the trajectory in which lower frequency components describe coarse shape while higher frequency components retain more trajectory details. They provide useful descriptor for trajectory global characteristics. Within the N Fourier coefficients, c_0 is omitted because it represents centre of gravity of a trajectory and by removing this term, the descriptor is invariant to translation. Moreover, c_1 is used to normalise all other Fourier coefficients, making them be invariant to homothety transformations (invariant to rotation scale and translation, as well invariant to location and trajectory phase). As a result, each trajectory is represented by an $N-1$ dimension vector F. Note that the Fourier descriptor is very different from the orientation-magnitude descriptor in that the former is a global shape descriptor concerning global motion information along a trajectory whilst the latter is a bag-of-words (BOW) descriptor of local motion information of track segments. Both types of descriptors contain complementary information.

Appearance Descriptor - Given a trajectory with a length L, first the SIFT features S_i are extracted at all the L key points $i = 1,...,L$. An appearance description of the tracked key point with this trajectory is computed as the average of L SIFT features: $S = \frac{1}{L}\sum_{i=1}^{L} S_i$. Similar approach has been used by Sun et al. (2009a).

Holistic Trajectory Representation - In order to fuse all three descriptors for action representation, the BOW method is employed. For each trajectory, its associated descriptors O,F and S are normalized and concatenated to form a global descriptor $G = [O,F,S]$. (The features normalization obtains mean zero and standard deviation equal to one.) The creation of BOW representation of a video consists of two steps. First the global descriptors G for all trajectories are quantised

using K-means to obtain a codebook with 500 words and each trajectory is assigned a codeword. Second, to retain spatio-temporal information about trajectories, the spatio-temporal volume of a ROI in a video is divided into 8 blocks including 4 non-overlapping spatial blocks and 2 temporal overlapping blocks (with a length of 2/3 of the temporal window of the ROI volume). Trajectories in each spatio-temporal blocks are then labelled separately. This results in a codebook V_1 with $500 \times 8 = 4000$ codewords to describe trajectories in videos.

4.2.2 Spatio-Temporal Interest Points Features

Local features extracted based on spatio-temporal interest points are also considered as they contain complementary information compared to trajectory features. In this model, interest points are detected using the method presented in Section 3.1. Compared with alternative methods such as Dollar et al. (2005), this approach is more reliable under realistic conditions: small camera movement, camera zooming, and shadows. The interest points are selected at the local maximal of detector response, and 3D cuboids are extracted around them. Similar to Dollar et al. (2005) and Liu et al. (2009a), the gradient vector is used to describe these cuboids and PCA to reduce the descriptor's dimensionality. To reduce the effect of spurious detection, an outlier removal method, which deletes points far from the mass centre of the points cloud, is employed. Bag-of-words is deployed again to represent each video clip. Specifically, a codebook V_2 with 300 visual words is initially built by performing K-means on a random subset of local features from the training data. Then, each clip is represented with a visual-words histogram.

4.2.3 Adaptive Feature Fusion

In this section the goal is to fuse adaptively trajectory based descriptors with Bag of Words interest points based descriptors according the presence of camera movement. The presence of a moving camera is detected by computing the global optical flow over all frames in a clip. If the majority of the frames contain global motion, the clip is regarded as being recorded by a moving camera. (In this case global motion means that the majority of the pixels are subjected to an optical flow grater the 4 pixels). This simple method can accurately and consistently separate videos with and without camera movements. For clips without camera movement, both interest points and trajectory based descriptors can be computed reliably and thus both types of descriptors are used for recognition, resulting in a final codebook $V = [V1, V2]$ with 4300 visual words. In contrast, when camera motion can be detected, interest point based descriptors are less meaningful

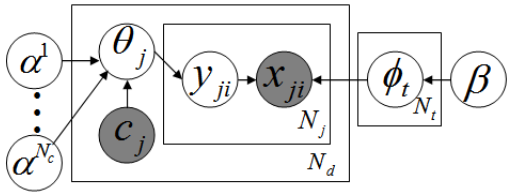


Figure 4.5: MC-ΔLDA model where α^c is the hyperparameter for the corresponding action class c, θ_j is the constrained topic distribution, ϕ_t is the Dirichlet word-topic distribution and x is the observed video clip. i and j represent the word and document index.

so only trajectory descriptors are employed, resulting the final codebook $V = V1$ with 4000 visual words. Now, a set of N_d video clips is represented as $\mathbf{X} = \{\mathbf{x}_j\}$, $j = 1, \cdots, N_d$, each of which is represented as a set of labelled features using V.

4.3 Collaborative Feature Selection

4.3.1 Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA)

For MC-ΔLDA modelling (Andrzejewski et al. 2007a), each video clip $\mathbf{x}_j$ is considered as a mixture of N_t topics $\Phi = \{\phi_t\}_{t=1}^{N_t}$ (to be discovered), each of which ϕ_t is a multi-nominal distribution over N_w words (visual features). However, different from existing topic models such as LDA model (Blei et al. 2003) which assumes uniform proportion of topic mixture for each video clip, a MC-ΔLDA model aims to constrain topic proportion *non-uniformly* and on a per-clip basis. More precisely, for each video clip belonging to action category c, it is modelled as a mixture of: (1) N_t^s topics which are shared by all N_c category of actions, and (2) $N_{t,c}$ topics which are uniquely associated with action category c. Where c represent one of the N_c available. Standard topic models such as LDA model each instance (clip) as being derived from a bag of topics drawn from a fixed (and usually uniform) set of proportions. In MC-ΔLDA, we wish to constrain the topic proportions non-uniformly and on a per-clip basis. This will enable some topics to be shared among all categories of actions and some to be uniquely associated with particular action category, thereby representing the unique aspects of that action. The structure of the proposed MC-ΔLDA model is shown in Fig. 4.5. In MC-ΔLDA, the non-uniform proportion of topic mixture for a single clip $\mathbf{x}_j$ (belonging to document j) is enforced by its action class label c_j and the hyperparameter α^c for the corresponding action class c. Given the total number of topics $N_t = N_t^s + \sum_{c=1}^{N_c} N_{t,c}$ and let T_0 be the first N_t^s elements list of shared topics and T_c be the $N_{t,c}$ element list of topics for action c_j, each hyperparameter α^c is a vector with N_t components $\alpha^c = \{\alpha_t^c\}_{t=1}^{N_t}$ in which components $t \in T_0 \bigcup T_c$ are constrained to be non-zero. To enforce

non-uniform proportion of topic mixture, a generative process of sampling video clips is given as follows:

1. Draw a Dirichlet word-topic distribution $\phi_t \sim \text{Dir}(\beta)$ for every topic t;

2. For each document j:

   (a) Draw a class label $c_j \sim \text{Multi}(\varepsilon)$;

   (b) Given label c_j, draw a constrained topic distribution $\theta_j \sim \text{Dir}(\alpha^{c_j})$;

   (c) Draw a topic $y_{j,i}$ for each word i from multinomial $y_{j,i} \sim \text{Multi}(\theta_j)$;

   (d) Sample a word $x_{j,i}$ according to $x_{j,i} \sim \text{Multi}(\phi_{y_{j,i}})$.

Given the structure of the MC-ΔLDA model and observable variables (clips $\mathbf{x}_j$ and action labels c_j), the objective is to learn the N_t^s shared topics as well as all $\sum_{c=1}^{N_c} N_{t,c}$ unique topics for all N_c classes of actions. The full joint probability of a document j in MC-ΔLDA is

$$p(\mathbf{x}_j, \mathbf{y}_j, \theta_j, \Phi, c|\alpha, \beta, \varepsilon) = \prod_i p(x_{j,i}|y_{j,i}, \Phi) p(y_{j,i}|\theta_j) p(\theta_j|\alpha, c) p(c|\varepsilon) p(\Phi|\beta). \tag{4.7}$$

Similar to the standard LDA, exact learning in this model is intractable. However a collapsed Gibbs sampler can be derived to sample the topic posterior $p(\mathbf{y}|\mathbf{x}, c, \alpha, \beta)$ (now additionally conditioned on the current class c) leading to the update

$$p(y_{j,i}| \quad \mathbf{y}_{j,-i}, \mathbf{x}, c, \alpha, \beta) \quad \propto \frac{n_{x,y}^{-i} + \beta}{\sum_x n_{x,y} + \beta} \frac{n_{y,d}^{-i} + \alpha_y^{c_j}}{\sum_y n_{y,d}^{-i} + \alpha_y^{c_j}}. \tag{4.8}$$

Here $\mathbf{y}_{j,-i}$ indicates all topics except the token i; $n_{x,y}^{-i}$ indicates the counts of topic y being assigned to word x, excluding the current item i; and $n_{y,d}^{-i}$ indicates the count of topic y occurring in the current document d. These counts are also used to point estimate Dirichlet parameters θ and Φ by their mean, e.g.

$$\hat{\Phi}_{x,y} \quad = \quad \frac{n_{x,y} + \beta}{\sum_x n_{x,y} + \beta}. \tag{4.9}$$

As reported in (Andrzejewski et al. 2007a), the presented Gibson sampling appears identical to the standard LDA in both the update step and topic constrains. The presented model has been developed in collaboration with Jian Li, extending his previous work (Li et al. 2009). The contribution of this work involves apply MC-ΔLDA in new application such as action recognition. The used MC-ΔLDA model employs $N_t^s = 5$ shared topics and at each action category assigns a single

unique topic $N_{t,c} = 1$. Moreover, the non-zero elements of α^c for shared topics are set to 0.1 and for unique topics to 10. The hyperparameter β is learned with Gibbs-expectation maximization (Minka 2003).

4.3.2 Feature Selection using MC-ΔLDA

Using MC-ΔLDA enables us to learn N_t topics $\hat{\Phi}$ to represent natural grouping of visual features either shared by all classes of actions or uniquely associated with one particular action category. This idea con be exploited to collaboratively select features, in this context collaborative means that the feature are selected observing the global behaviour of the feature space, not at single-feature level. This grouping process effectively ranks all the visual features according to their importance of being used for representing either general aspects of all action classes or unique aspects of a specific action class. Now, a sorted index of all the visual features $r(V)$ is needed to be estimated from the learned topics to represent distinctive visual features for action classification. Ideally, the ranked features $r(V)$ can be learned from the $\sum_{c=1}^{N_c} N_{t,c}$ action specific topics in $\hat{\Phi}$. However, as visual features extracted from action videos can be very noisy and not well structured, those topics can be easily corrupted by noise and are thus not suited for feature selection. Therefore, the discriminative features are learned from the N_t^s topics shared by all actions. Given the N_t^s shared topics which are represented as an $N_w \times N_t^s$ dimension matrix $\hat{\Phi}^s$, the feature selection can be summarised into two steps:

1. For each feature v_k, $k = 1, \cdots, N_w$, compute its maximum probability across all N_t^s topics according to $p(v_k) = \max(\hat{\Phi}^s_{k,1:N_t^s})$;

2. Rank the value of $p(v_k), k = 1, \cdots, N_w$ in ascending order to obtain a vector of feature index $r(V)$ in which higher ranked features correspond to more discriminative/relevant features.

As this model selects different types of features to represent videos with and without camera movements, different MC-ΔLDA models are trained separately for the two type of videos. For each model, the number of features selected for final classification is determined by cross validation.

It has to be mentioned that MC-ΔLDA differs from standard LDA because it uses shared topics (common to all the classes) as well as unique topics which appear only in special classes. Moreover, the documents are constituted of a mixture of shared and unique topics. Since the proposed feature selection exploits the shared topics only, part of the extracted information (unique

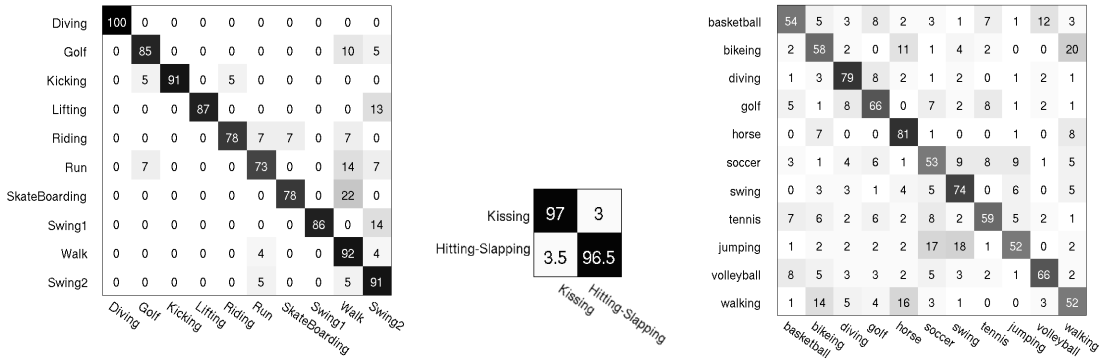
Figure 4.6: From top to bottom: Example frames from the (a) UCF Feature Films, (b) the UCF Sport Actions and (c) the YouTube datasets. More details are reported in Appendix A.

topics) is discharged. Anyhow, in the proposed feature selection, MC-ΔLDA is applied for shared topics disclosure which cannot be achieved by using standard LDA.

4.4 Experiments

The formulated trajectory based representation is validated and the results obtained from the fusion with the Bag of Words representation are discussed.

Datasets - The experiments are carried out over realistic scenarios, to this end three datasets have been explored. The UCF Feature Films Dataset (Rodriguez et al. 2008) provides a representative pool of natural samples of two action classes: *Kissing* and *Hitting/slapping*. The UCF Sport Actions Dataset (Rodriguez et al. 2008) contains 10 different types of human actions in sport broadcasting videos: *diving, kicking , weight-lifting, horse-riding, running, skateboarding, golf swinging, swinging 1* (gymnastics, on the pommel horse and floor), swinging 2 (gymnastics, on the high and uneven bars) and *walking*. The YouTube Dataset (Liu et al. 2009a) is the most extensive realistic action dataset available to the public and it is composed of 1168 videos collected from YouTube. These videos contain a representative collection of real world challenges such as: shaky cameras, cluttered background, variation in object scale, variable and changing view-point and illumination, and low resolution. Since these videos are mostly home-videos captured by hand-held cameras, the camera movements are much more unpredictable compared to the other two datasets. The YouTube dataset contains 11 action categories: *basketball shooting, volleyball*



(a) UCF Sports
ACA = 86.90%
$\sigma = 0.65$

(b) UCF Films
ACA = 96.75%
$\sigma = 1.3$

(c) YouTube
64.00%
$\sigma = 0.45$

Figure 4.7: Confusion matrices of the proposed approach on three datasets. Results presented in terms of average classification accuracy and standard deviation over 20 trials.

spiking, trampoline jumping, soccer juggling, horse-back riding, cycling, diving, swinging, golf swinging, tennis swinging, and *walking*. Some representative frames are shown in Figure 4.6. More details about these datasets are reported in Appendix A.

4.4.1 Experimental Settings

Recognition was performed using a Support Vector Machine with a Gaussian kernel and Leave-One-Out Cross-Validation (LOOCV) was used. More specifically, for the YouTube dataset the settings given in (Liu et al. 2009a) are adopted . The dataset was divided into 25 subsets, out of which 24 subsets were used for training and the remaining subset was used for testing (importantly, in the LOOCV the testing set is not used for training) .For the UCF Sport Actions and Feature Films datasets the setting in (Rodriguez et al. 2008) was followed : one clip was used for testing and the renaming for training. In this experiment, 29 Fourier coefficients were used in the trajectory shape descriptor, and for the appearance descriptor, 128-bin SIFT histogram are used. Those values were empirically selected.

Since the Bag-of-Words representation requires a clustering process to initialize the codebook, different iterations may report slightly different performance. For this reason, the results are displayed as average and standard deviation over 20 trials.

4.4.2 Trajectory Based Representation Validation

The average recognition rates obtained using the proposed representation approach are presented in Table 4.4 and compared with the results obtained by existing approaches. The classification

	UCF Films	UCF Sports	YouTube
Proposed model	**96.75%**	86.90%	64.00%
Wang et al. (2009)	86.60%	85.60%	-
Yeffet and Wolf (2009)	80.75%	79.20%	-
Rodriguez et al. (2008)	66.30%	69.20%	-
Kovashka and Grauman (2010)	-	**87.20%**	-
Yao et al. (2010)	-	86.60%	-
Liu et al. (2009a)	-	-	**71.20%**

Table 4.4: Comparison on the UCF Feature Films, UCF sport actions and YouTube datasets.

confusion matrices are also presented in Fig. 4.7. An excellent result is obtained on the UCF Feature Film dataset with 96.75% average recognition rate for the two action classes. This result is significantly better than those obtained by existing approaches, the best of which (Wang et al. 2009) achieved 86.60%. This approach also outperforms all existing methods which report results on the UCF Sport dataset except (Kovashka and Grauman 2010). As for the YouTube Dataset, an average recognition rate of 64.00% was obtained. This is a much harder dataset compared to the other two due to its home video nature. Apart from the work which first introduced this dataset (Liu et al. 2009a), no other work has presented results on this dataset. This result is slightly lower than that in (Liu et al. 2009a) which used quite different features and different classifiers, and included a number of heuristic steps that are hard to reproduce.

4.4.3 Effectiveness of Adaptive Feature Fusion

In this section we explore advantages of fusing different descriptors in an adaptive manner as presented in Section 4.2.3. Basically, if the video clip contains global motion, the interest point contribution is not used otherwise all of them are concatenated. The used descriptors are **OM** Orientation-Magnitude, **F** Fourier descriptor, **S** SIFT and **IP** spatio-temporal interest points.

Table. 4.5 clearly shows the effectiveness of each single descriptor and the fusion of them. It is evident that when the three trajectory based descriptors are fused together, action recognition performance is improved for all three datasets. The improvement is particularly significant for the YouTube dataset (around 15% increase compared to any single descriptor alone), which contains a large variety of action categories and vastly different lighting conditions, camera angles, and camera movements. It is thus more important for different complimentary information to be utilised simultaneously. Table 4.5 also shows that fusion of trajectory based features with Bag

	UCF Films	UCF Sport	YouTube
OM	80.50%	58.04%	44.7%
F	82.00%	75.02%	45.7%
S	91.60%	82.50%	44.5%
OM+F+S	94.40%	84.45%	59.90%
OM+F+S+IP without adaptive fusion	92.20%	77.33%	49.03%
OM+F+S+IP with adaptive fusion	96.75%	86.90%	64.00%

Table 4.5: Evaluation of descriptor performances and effectiveness of feature fusion. **OM**: Orientation-Magnitude Descriptor, **F**: Fourier Descriptor, **S**: SIFT Descriptor, and **IP**: spatio-temporal interest points.

of Words interest points based features can lead to better performance, provided that they are fused in an adaptive manner as proposed in this chapter. In particular, if they are fused unconditionally without considering the reliability of each type of feature given the camera movements, performance degradation is observed. This result validates the effectiveness of the novel action representation and feature fusion method formulated in this work.

4.4.4 Effectiveness of Collaborative Feature Selection

The optimal number of features used in the recognition is selected by cross-validation; for the UCF Feature Film, UCF Sport Action and YouTube datasets the following figures are used respectively: 70%, 80% and 70% of the total available feature are used. Table 4.6 compares the effectiveness of the collaborative feature selection method with a mutual information based sequential feature selection method proposed in Zaffalon and Hutter (2002). It is evident that our feature selection method indeed improves action recognition performance when compared to using all features without selection. The effect from our feature selection is in particular more significant for the difficult YouTube dataset. Here, the unpredictable camera movements introduced large number of irrelevant features that cannot be removed completely even with our trajectory pruning and region of interest detection. In comparison, the sequential feature selection process is less effective. This suggests the importance of performing feature selection jointly and collaboratively given highly correlated features extracted both instantaneously from interest points and globally over time from trajectories.

	UCF Films	UCF Sports	YouTube
MC-ΔLDA	**96.75%**	**86.90%**	**64.00%**
Mutual Information	96.10%	85.33%	62.20%
No Feature Selection	96.10%	84.00%	59.90%

Table 4.6: Comparing the effectiveness of different feature selection methods. From top to bottom: the proposed MC-ΔLDA, Mutual Information (Zaffalon and Hutter 2002) and no feature selection.

4.5 Discussion and Summary

This chapter extensively studies the problem of feature fusion and feature selection. Interestingly, the obtained results show that action recognition can be notably improved by fusing different sources of information and removing noisy and redundant components.

Initially, a novel framework for fusing interest points based representations is presented: COP is merged with a conventional Bag of Words representation. The obtained results confirm that **Multiple Kernel Learning (MKL) fusion** improves the performance in both of the tested datasets. Specifically, in terms of recognition rate, an improvement of 1.7% and 2% is observed for the KTH and the Weizmann datasets compared to concatenation. Moreover, a robustness test has been also performed and the outcome highlights the superior discriminative power of the formulated MKL fusion strategy.

Action recognition from video recorded in realistic scenarios introduces new challenges and difficulties that cannot be handled using interest points based methods only. In other words, the proposed COP and BOW representation appears unable to extract meaningful information in the presence of constant camera movement and crowded background. To address this issue an alternative **key-point trajectories representation** is formulated. The proposed trajectory descriptors (Orientation-Magnitude, Trajectory Shape and Appearance Descriptor) use alternative principles to efficiently capture action appearance and dynamics. As shown in the experiment Section 4.4.2, the best performance is obtained by merging the three descriptors. By aiming to further enrich the action representation, an **adaptive feature fusion method** is used to combine trajectory descriptors with conventional Bag of Words interest point-descriptors. The reported results show again that feature fusion notably improves the recognition rate by: 2.35%, 2.45% and 4,1% respectively for the UCF Films, UCF Sports and YouTube datasets.

Crucially, a novel feature selection approach is also introduced by formulating a discriminative **Multi-Class Delta Latent Dirichlet Allocation (MC-ΔLDA) topic model for collabo-**

rative feature selection. Interestingly, it is underlined that the unique topics appear noisy and unstable for recognition purposes, and as such, shared topics are exploited. The basic idea is to identify the less-shared topics and to use these to rank the features. Less-shared topics are distributed across a small number of classes, conserving a significant discriminative power. Additionally, they are more stable because they are less likely to represent background components.

The entire framework has been tested on three well known datasets (UCF Feature Films, UCF Sport Actions and YouTube datasets), reporting comparable results with the state-of-the-art. Although the proposed framework outperforms existing approaches, it appears inadequate for handling different challenges such as classification performed on highly overlapped action classes featured with noisy training set, short and very ambiguous sequences, strong camera movements and zooming. These difficulties are for instance included in the Hollywood dataset (Laptev et al. 2008). To address these issues a specific approach is required which simultaneously addresses both feature selection and classification. In the next chapter a cascade feature selection and classification strategy is designed to decompose the classification problem into multiple subtasks where classification and feature selection are both mutually optimised.

Chapter 5

Cascaded Feature Selection and Action Classification

As stated in the previous chapters, realistic action recognition is a very complex and still unsolved problem. Additionally, it can be further complicated if the observed video sequences are highly ambiguous and the available training set is noisy and sparse. These extreme conditions present new challenges when performing action recognition, in both feature selection and action classification phases. More specifically, in the presence of high intra-class variation and high inter-class similarity between action classes, standard feature selection methods as (Peng et al. 2005; Zaffalon and Hutter 2002) become less efficient. Similarly, a noisy and sparse training set is inappropriate for training standard multi-class classifiers. Some of these problems are highlighted in Fig. 5.1 where it can be clearly seen that, even when the image quality is relatively good, to recognise actions outside a well-controlled laboratory environment is far from simple.

In other words, it can be affirmed that:

1) It is difficult to simultaneously estimate the optimal decision boundaries that separate highly ambiguous multiple action classes. Whatever action representation approach is taken, a multi-class action dataset is featured with large intra-class variations and inter-class similarities due to the aforementioned challenges.

2) Different action classes are often visually similar due to the shared primitive components. For instance, "*running*" and "*jogging*" would involve mostly the same body parts moving in a very similar way. "*Hugging*" and "*kissing*" may look identical at the beginning of the action sequences. It is therefore critical to perform feature selection in order to identify the most discriminative features per inter-class before classification. However, different feature sets are

Figure 5.1: Examples from the Hollywood dataset (Laptev et al. 2008). (a) *Hand shaking* from an unusual view angle. (b) *Kissing* in a crowd. (c) *Getting out of a car* under challenging lighting. (d) *Hug a person* with occlusion.

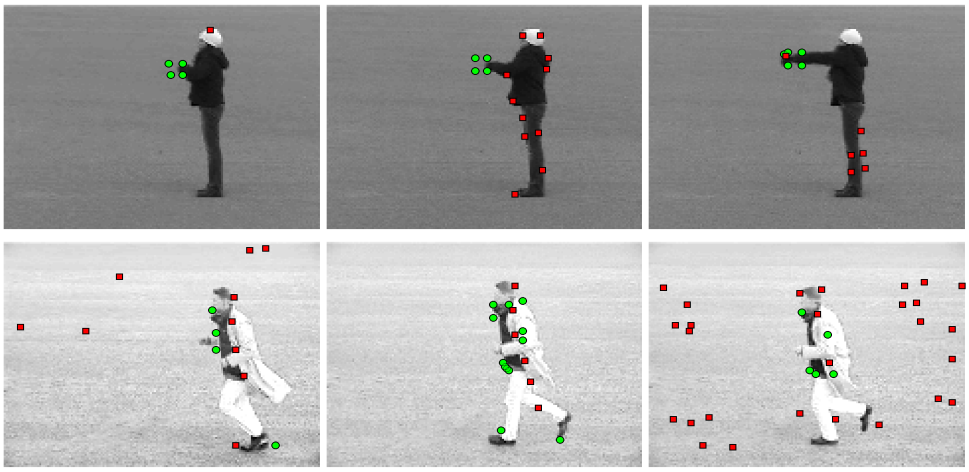
useful for separating different groups of actions, and there will rarely be features that are universally informative for separating all classes simultaneously.

To deal with these extreme circumstances, this chapter proposes a cascaded feature selection and classification that aims to classify actions by exploiting as much information as possible and at the same time trying to simplify the multi-class classification in a cascade of binary separations. The idea behind this approach consists of dynamically decomposing the classification problem, and iteratively redesigning the optimal classification context in each step.

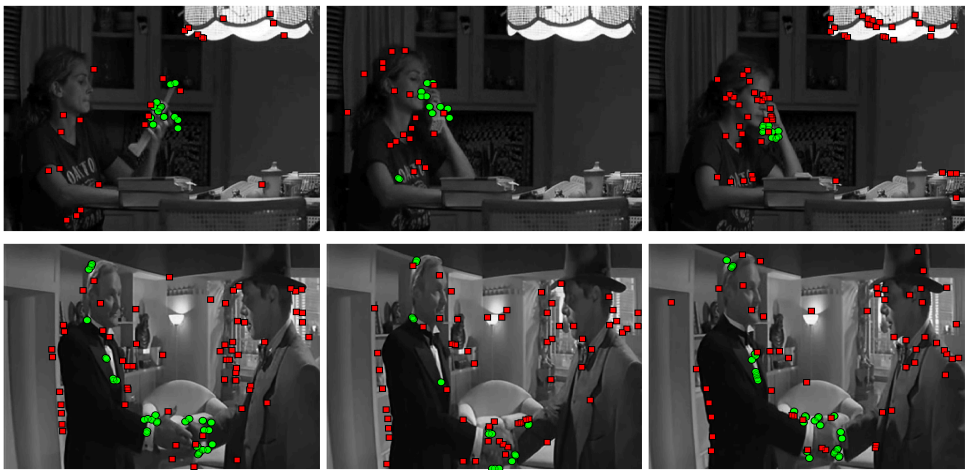
5.1 Action Representation

This chapter uses a representation based on interest point detection and Bag of Words (BOW) descriptors. That is, salient information is extracted through interest points sampling from a video sequence before clustered and represented as histograms of visual words.

Among various interest point detection methods, the one proposed by Dollar et al. (2005) is perhaps the most widely used however has limited performance in unconstrained scenario. Hence, in this study the interest point detector presented in Section 3.1 is adopted. Fig. 5.2



(a) *Boxing* and *Jogging*



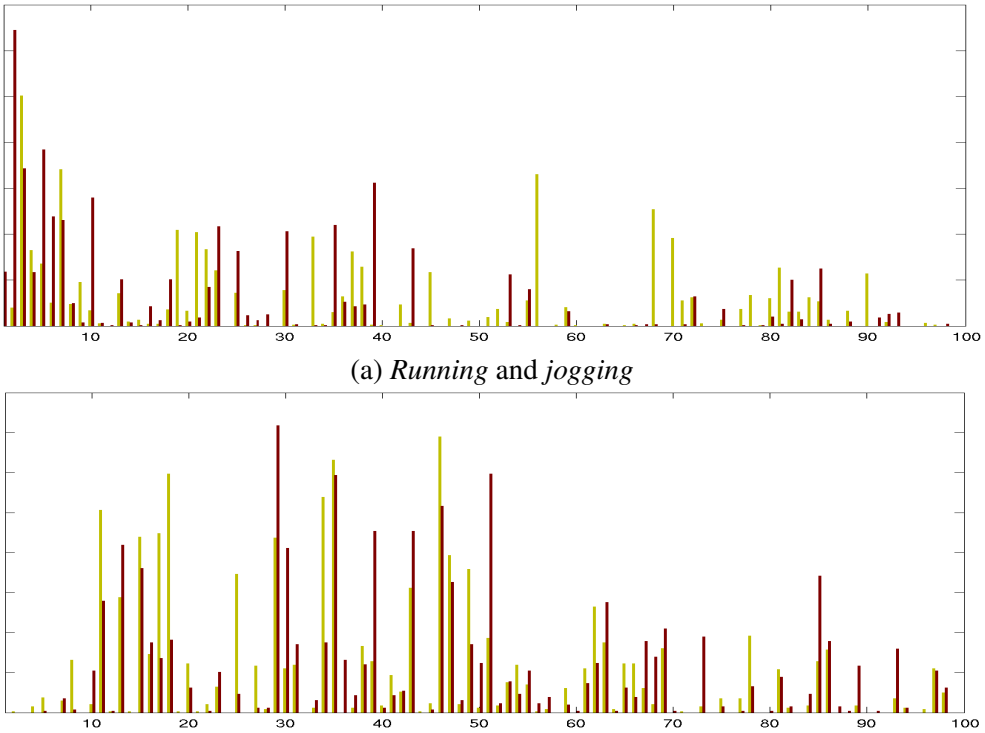
(b) *Answering the phone* and *Hand shaking*

Figure 5.2: Examples of interest point detection. Red points are extracted using (Dollar et al. 2005) while the green points using the proposed approach. (a) From the KTH dataset Schüldt et al. (2004), (b) from the Hollywood dataset (Laptev et al. 2008).

shows a comparison between these two methods. It is evident that the adopted method (green) is more meaningful and descriptive compared to the Dollar et al. (2005) detector (red).

After interest point detection, 3D cuboids are extracted from each interest point and gradient based descriptors are then employed to map these cuboids into a high dimensional feature space, generating a large collection of features. The size of the cuboid is set to 8×8 pixels in the image coordinates and 12 frames along the time axis. Next, using K-means (Spath 1985) these features are clustered and a visual word codebook is built. In the last step, each video sequence is represented as a K bin histogram of visual words, where K is the size of the codebook.

Examples of visual word histograms are shown in Fig. 5.3. It is evident, by comparing Fig. 5.3(a) and (b), that "*running*" and "*jogging*" have a very similar distribution of visual words,



(a) *Running* and *jogging*

(b) *Boxing* and *hand-clapping*

Figure 5.3: Examples of actions from the KTH dataset represented as histograms of visual words. The histograms have 100 bins (i.e. the size of the codebook is 100). (a) The light colour corresponds to *running* action while the dark one is for *jogging*. (b) The light colour is for *boxing* while the dark colour corresponds to *hand-clapping*.

which turns out to be quite different from that of "*boxing*" and "*hand-clapping*". In accordance with this observation, these four actions can be easily divided in two groups. This suggests that separating the two action classes within each group is much harder than distinguishing the two groups. Fig. 5.3 also highlights that 1) feature selection is necessary for action recognition; and 2) different sets of features should be selected for distinguishing different classes of actions. For instance, histogram bins numbered 1-20 should be selected if the task is to separate the two groups. However, they are not very helpful in separating "*running*" and "*jogging*". Similarly, bins numbered 50-70 are more discriminative for classifying "*running*" and "*jogging*", but not for "*boxing*" and "*hand-clapping*". In order to explore these characteristics more effectively in assisting action recognition, a cascaded feature selection and classification model is formulated in the following sections.

5.2 Cascaded Feature Selection and Action Classification

Given a training set containing sequences of M action classes $A = [a_1, \ldots, a_m, \ldots, a_M]$, a classification cascade is built, in each stage of which one or more binary classifiers are deployed to separate a group of action classes into the two most separable sub-groups. Any sub-group that is composed of more than one action class, will be further divided into two in the next cascade stage. Critically, for each classifier in each stage, feature selection is performed allowing different features to been selected. The total number of stages S in the cascade will depend on in which stage all sub-groups contain only a single action class. The value of S thus ranges from $\log_2 M$ to $M-1$ while the total number of binary classifiers is always $M-1$.

The structure of the cascade is determined automatically using spectral clustering (Shi and Malik 2000). Specifically, in order to group the M action classes into two sub-groups in the first stage of the cascade, the similarity between each pair of classes is computed which gives rise to a $M \times M$ similarity matrix $\mathbf{S} = \{S_{i,j}\}$ with $S_{i,j}$ measuring the similarity between the i-th and the j-th action classes. To compute $S_{i,j}$, the averaged descriptors $\mathbf{p}_i$ and $\mathbf{p}_j$ are obtained for the i-th and the j-th action classes respectively (for a single class, the average descriptor is computed averaging all the features belonging to that class). Then, it is obtained the following equation:

$$S_{i,j} = 1 - \|\mathbf{p}_i - \mathbf{p}_j\| \tag{5.1}$$

where $\|\mathbf{p}_i - \mathbf{p}_j\|$ is the Euclidean distance between the two averaged descriptors. The elements of $\mathbf{S}$ are then normalised to be in the range of $[0,1]$. Using the similarity matrix $\mathbf{S}$ as input, the normalized cut algorithm (Shi and Malik 2000) is employed to cluster the M action classes into two sub-groups. The same process is repeated for each sub-group that has more than two members.

The cascade is dynamically determined by training binary classifiers for each classification step. Within each step, feature selection is initially performed then the classifier is trained and the optimal decision boundary is defined. To that end, a mutual information based feature selection method (Zaffalon and Hutter 2002) is utilised, ranking the features (histogram bins in this case) according to their relevance. The optimal number of features to be kept is determined by cross-validation. An example of the obtained cascade structure is illustrated in Fig. 5.4.

After the cascade classifier training process is terminated, it is straightforward to classify an

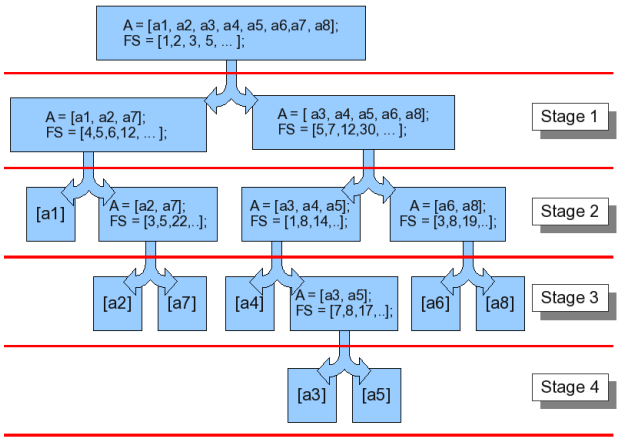


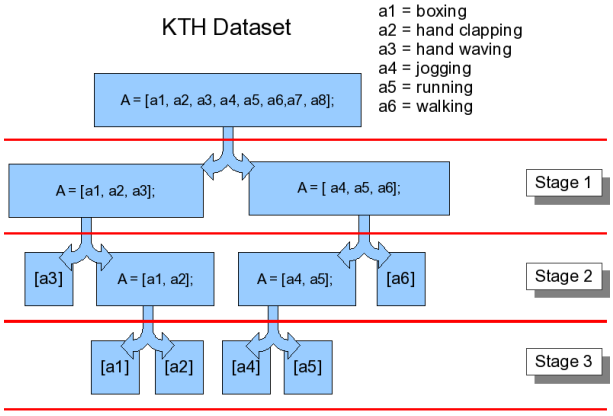
Figure 5.4: Cascaded action classifier example. Starting with a group of $A = 8$ action classes in stage 1, all action classes are separated by stage 4. Note that for dividing each sub-group into two, a different set of features FS is used. In this example, the cascade consists of 4 stages and 7 binary classifiers.

unknown action into one of the M classes. Specifically, the unknown action is firstly processed by the initial stage, falling into one of the two sub-groups in stage 1. If the sub-group into which it is classified contains more than one action class, it is further classified following the binary classifier cascade in the remaining stages. This process is repeated until the action falls into a sub-group with only one action class (not necessarily in the last cascade stage). Therefore a minimum of one and a maximum of $M - 1$ binary classifications are needed to assign a label to an action sequence.

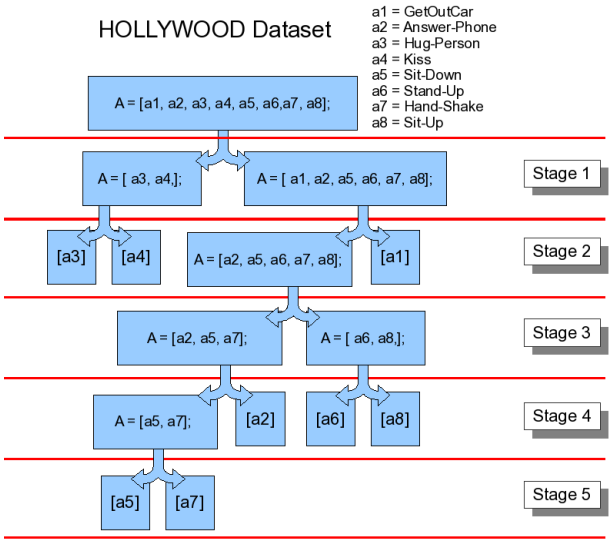
5.3 Experiments

5.3.1 Experimental Settings

Any type of binary classifier can be used to form the classification cascade. In this experiment absolute distance based k-NN and an SVM with polynomial kernel are used. The k value and polynomial degree were determined by cross validation performed for each separation step (the cross-validation is done on training data only). The proposed framework is validated over the KTH and Hollywood datasets. Additional information about these dataset is available in Appendix A. Different codebook sizes were tested for the BOW representation. The results are reported using a visual word codebook size of 200 for KTH and 300 for Hollywood when not stated otherwise. In order to present a fair comparison with the state-of-the-art results, the most widely used validation procedure is followed. This means that for the KTH dataset the average



(a)



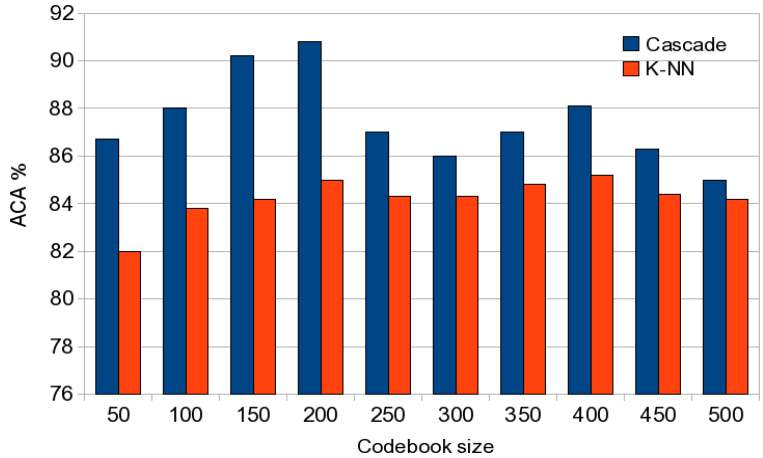
(b)

Figure 5.5: Cascade classifier structure. (a) KTH dataset and (b) Hollywood dataset.

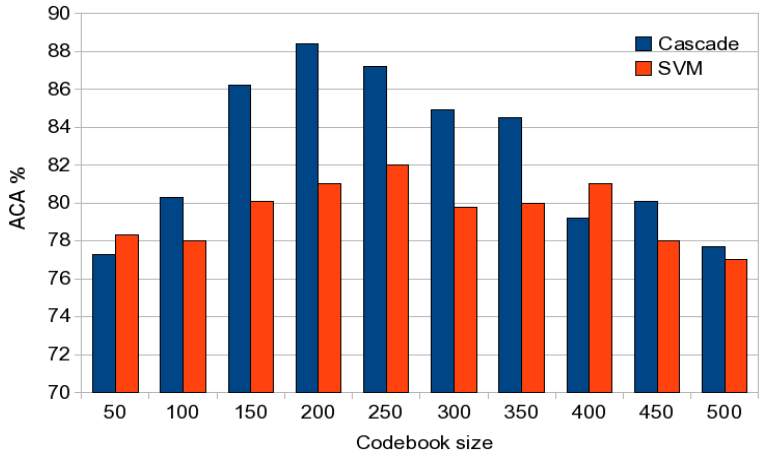
class accuracy (ACA) was used (Dollar et al. 2005), and the average precision (AP) of the precision/recall curve was used for the Hollywood dataset (Laptev et al. 2008). Since the codebook was generated using K-means which is sensitive to initialisation, the results are reported based on the average of 20 trials.

5.3.2 Learning Classifier Structures

Fig. 5.5 shows the learned cascade structures for both datasets, which were automatically determined. It can be seen that both learned structures reflect accurately the natural grouping of the different action classes. For instance, Fig. 5.5 (a) shows that the 6 action classes in the KTH dataset were divided into two sub-groups in the first cascade stage: "*jogging*", "*running*" and "*walking*" in one sub-group which all involve movement from legs, and "*boxing*",



(a) k-NN vs. cascaded k-NN



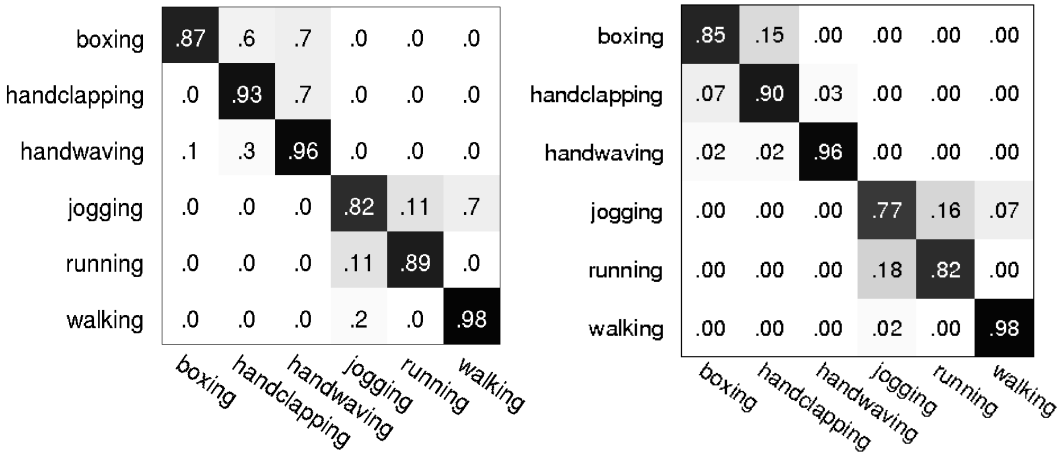
(b) SVM vs. cascaded SVM

Figure 5.6: Comparing cascaded and standard classifiers on the KTH dataset given different codebook sizes.

"*hand-clapping*" and "*waving*" in the other which are featured mainly with movements from the upper body. For the Hollywood dataset in the first stage of the cascade two action classes "*hug a person*" and "*kissing*" are grouped together and separated from the other 6 classes. These two classes are visually very similar whilst being distinctive from other action classes (see Fig. 5.1). It can also be seen from Fig. 5.5 that similar action classes such as "*stand-up*" and "*sit-up*", "*jogging*" and "*running*" stay grouped until a binary separation between them is carried out in the last cascade stage.

5.3.3 Cascaded Classifiers VS. Standard Classifiers

This experiment compares the performance of the proposed cascaded classifier with that of standard multi-class classifiers including k-NN and SVM using identical action representation. Fig. 5.6 shows the performance of cascaded k-NN, cascaded SVM, standard k-NN and SVM on the



(a) Cascaded k-NN classifiers,
ACA = 90.8%
$\sigma = 1.58$

(b) Cascaded SVM classifiers,
ACA = 88.4%
$\sigma = 1.12$

Figure 5.7: Confusion matrix computed on KTH dataset; average classification accuracy and standard deviation are reported.

KTH dataset. The results obtained with different codebook sizes are also shown to examine its effect on different classifiers. The confusion matrices obtained using the proposed cascaded classifiers are shown in Fig. 5.7. It is evident from Fig. 5.6 that cascade classifiers significantly outperform the standard k-NN and SVM (more visible when using codebook size of 200). The results obtained on the Hollywood dataset are shown in Table 5.2. Again, this is a large improvement in performance regardless of the type of classifier used. With the same action representation and the same type of classifier, this improvement can only be contributed by the proposed cascaded feature selection and classification method.

Due to cluster initialization, different runs of the proposed approach return slightly different results. The final values are presented in Table 5.1 as average of 20 trials, moreover the standard deviation is also shown.

	Average Recognition Rate	Std. Deviation
KTH	90.8%	1.58
Hollywood	31.31%	2.42

Table 5.1: Average recognition rate and standard deviation for KTH and Hollywood dataset obtained using the proposed cascade classifier. The results are observed over 20 trials.

	Cascaded k-NN	k-NN	Cascaded SVM	SVM
GetOutCar	19.3 %	17.4 %	22.3 %	20.2 %
AnswerPhone	22.5 %	18.2 %	38.4 %	32.4 %
HugPerson	21.6 %	12.7 %	33.5 %	28.8 %
Kiss	47.6 %	41.8 %	46.7 %	32.1 %
SitDown	31.5 %	30.1 %	37.9 %	17.3 %
StandUp	41.3 %	33.4 %	42.3 %	32.0 %
HandShake	13.5 %	12.1 %	21.0 %	19.5 %
SitUp	4.2 %	4.2 %	8.4 %	6.6 %
Average	25.19 %	21.24 %	**31.31 %**	23.61 %

Table 5.2: Comparing cascaded and standard classifiers on the Hollywood dataset.

	KTH	Hollywood
Proposed Method	**90.8%**	**31.31%**
Laptev et al. (2008)	91.8%	38.39%
Kläser et al. (2008b)	91.4%	24.7%
Bregonzio et al. (2009a)	93.17%	-
Niebles et al. (2008)	83.33%	-
Dollar et al. (2005)	81.17%	-

Table 5.3: Comparative results on the KTH and Hollywood datasets.

5.3.4 Comparison with the State-Of-The-Art

Table 5.3 shows a performance comparison on both KTH and Hollywood datasets between the proposed method and the state-of-the-art. It emerges that for the KTH dataset the proposed approach outperform existing methods that are based on a similar action representation (Niebles et al. 2008; Dollar et al. 2005). For those that give a slightly better result (Kläser et al. 2008b; Bregonzio et al. 2009a; Laptev et al. 2008), much more sophisticated action representation methods are employed. In particular, they all explored spatial distribution information of the visual words. In contrast, this information has not been taken into account by the current model. As for the Hollywood dataset, so far only two previous studies have reported results. Among them (Kläser et al. 2008b) does not exploit the spatio-temporal distribution of interest points but uses a more sophisticated interest point descriptor than us. However, this result is still clearly superior to that in (Kläser et al. 2008b). The result obtained in (Laptev et al. 2008) is better than ours. However, different spatio-temporal Bag-of-Words representations were exhaustively examined in their work and only the best one for each action class was used to produce their result. The same idea employed in the proposed method will possibly enhance the performance.

It should be noted that in this work the goal is to improve the performance of action recognition via the novel cascaded feature selection and classification method regardless of the adopted action representation and binary classifier. The results in Fig. 5.6 and Table 5.2 have clearly

demonstrated that this goal has been achieved. It is expected that these results will be further improved when more descriptive action representation methods such as those in (Laptev et al. 2008; Liu and Shah 2008) are employed.

5.4 Discussion

Much of the previous action recognition work focuses on action representation whilst using standard multi-class classifiers such as SVM and k-NN for action classification. It has been shown that these standard classifiers are inadequate in addressing more challenging action recognition problems encountered in an unconstrained environment where the training set available is noisy and sparse. To overcome these problems a novel action classification approach based on cascaded feature selection and classification is proposed. Specifically, instead of separating multiple action classes simultaneously, the difficult multi class task is decomposed automatically into easier subtasks. Practically, in each step the two easier-separable subgroups are identified and the optimal features for the specific task are selected. The algorithm is iterated until all the action classes are separated. Experiments are carried out using challenging public datasets to demonstrate that, with identical action representation, the formulated cascaded classifier significantly outperforms standard multi-class classifiers.

The obtained results also reveal that the learned cascade structure reflects the natural grouping of the actions, for instance very similar actions such as "*running*" and "*jogging*" are separated only in the last stage. Similarly to decision tree approaches, the formulated classification structure is simple to interpret and additionally it is robust to noise and sparse training set (such as Hollywood dataset). In principle there are no differences between the formulated cascade and automatically generated decision trees; the name cascade classifier has been used explicitly to emphasize the aim of the classifier. While decision trees frequently are employed for object detection or recognition speed-up, our formulation aims mainly to improve the classification performances.

5.5 Summary

Action recognition in realistic environments is a very complex and still unsolved problem. Additionally, it can be further complicated if the observed video sequences are highly ambiguous and the available training set is noisy and sparse. To perform action recognition in these extreme

conditions a novel approach based on cascaded feature selection and classification is proposed. Specifically, instead of separating multiple action classes simultaneously, the difficult task is decomposed automatically into easier subtasks of separating two groups of the most separable action classes at a time with different features selected for different subtasks.

A **cascade classifier** is then formulated, whereby one or more binary classifiers are deployed at each stage to separate a group of action classes into the two most separable sub-groups. Any sub-group that is composed of more than one action class will be further divided into two in the next cascade stage. Critically, for each classifier in each stage, an optimal **feature selection** is performed via cross-validation. This allows to specifically select features in accordance with a determined classification task.

Experiments are carried out using challenging public datasets: the KTH Schüldt et al. (2004) and Hollywood Laptev et al. (2008) datasets. It has been demonstrated that with identical action representation, the formulated cascaded classifier can significantly outperforms standard some multi-class classifiers.

Chapter 6

Conclusion and Future Work

This thesis has set out to explore human action recognition from video sequences by studying in detail the problem of action classification in unconstrained scenarios. Here the term *action* refer to a sequence of primitive body movements that may involve part or the whole body such as *walking*, *running*, *clapping* or *jumping*. Automatically recognising actions, in the context of artificial intelligence, plays a crucial role because it permits machines to interact and understand human requests through a video camera, without the need of a physical interface. Action recognition has recently received a large amount of attention from the computer vision community owing to the innumerable applications, including, but not limited to: medical surgery, security, education, media, and the military sector.

As presented in Chapter 2, the available literature suggests that action recognition is moving from a well constrained laboratory environment to unconstrained real world scenarios. In light of this, new problems and challenges are emerging while recent methods appear to be limited. The thesis contributions are summarized below, addressing action recognition in unconstrained scenarios, and future developments are discussed.

6.1 Robust Action Representation Using Clouds of Interest Points

Recent action recognition methods (Schüldt et al. 2004; Laptev and Lindeberg 2003; Dollar et al. 2005; Niebles et al. 2008) represent actions as bags of space-time interest points. These methods rely solely on the discriminative power of individual local space-time descriptors, while ignoring the potentially useful information about the global spatio-temporal distribution of interest points.

Consequently, they are unable to capture global motion components as well as smooth and fast motions. This is due to the lack of both multiple-temporal-scale and points-distribution information. Chapter 3 develops the Clouds of Spatio-Temporal Interest Points method (COP), which aims to *explicitly* and *globally* exploit spatio-temporal information associated with the interest points distribution. In particular, holistic features from clouds of interest points accumulated over multiple temporal scales are used. The formulated COP representation is robust to noise and outliers. Additionally, when compared with conventional interest point based methods, it appears more discriminative and invariant to changes in the recording set up and action distortions. Furthermore, a novel interest point detector is formulated to select more stable and meaningful points then standard available approaches. This detector, compared with existing methods, appears more robust against shadow, noise and dynamic background.

The proposed approach has been evaluated using two widely used public datasets, namely the KTH dataset (Schüldt et al. 2004) and the Weizmann dataset (Blank et al. 2005). The obtained results demonstrate that our approach is comparable with most of the existing methods. Furthermore, the proposed approach is more robust against occlusion and changes in viewing angle, clothing, and carrying condition compared to existing methods.

6.1.1 Future work

The proposed interest point based method can be divided in two major steps: interest point sampling and action representation, as presented in Chapter 3. With regard to interest point sampling, the proposed approach has different limitations. For instance, it extracts points with fixed spatial-temporal scale, and it is sensitive to fast camera movements and crowded backgrounds. Additionally, the estimated region of interest relies on background subtraction which may be impractical in realistic scenarios. Similarly, the action representation step appears inadequate in the presence of a crowded background, multiple subjects and fast camera movements. These circumstances generate points associated with background and surrounding objects, which are captured by the Clouds of Points (COP) representation, leading to an incorrect representation.

The possible extensions identified to enhance the proposed COP method are:

1. Extend the proposed interest point detector to multi-scale detection, where the interest point scale is automatically selected by observing the surrounding area. As discussed in section 3.3, different actions have different temporal scales and speeds. Moreover, they

may be recorded at different camera distances. An interest point detector that takes into account these issues will provide a more consistent and robust action representation such as: (Oikonomopoulos et al. 2006) which use entropy to select the point scale or (Liu et al. 2009a) where interest points are extracted at different scales and all used in the representation.

2. The proposed method requires a region of interest detection to compute different holistic features. This region of interest should be located around the subject and the quality of the extraction directly influences the performance. The present implementation relies on a frame difference approach, which is sensitive to dynamic background. Furthermore, in the presence of multiple targets the region of interest extraction fails. In light of this, it is important to improve the region extraction and increase its robustness. To this end, it would be possible to introduce an object detection approach to better initialise the subject and then to employ a tracking algorithm to maintain the subject localization.

6.2 Feature Fusion and Selection

Real world environments introduce new challenges still not addressed by existing action recognition approaches. Video sequences recorded in these circumstances are characterized by large degrees of occlusions from multiple objects, illumination change, shadow, cluttered background, scale variation, and constant camera movements. To overcome these difficulties, in Chapter 4 innovative feature selection and feature fusion strategies are deployed. Initially, in order to enhance the proposed Clouds of Points representation, this representation is fused with a conventional Bag of Words representation. This fusion is motivated by the fact that these representations contain different but complementary information, leading to a more robust and informative action description. Despite encouraging results, this combined representation still suffers from camera movements and crowded background. In light of this, a novel action representation based on key point trajectories analysis is formulated. A robust set of trajectory descriptors are computed and used in a Bag of Words paradigm. To remove redundant and noisy components, a novel collaborative feature selection method is formulated (Multi-Class Delta Latent Dirichlet Allocation model). Finally, an adaptive feature fusion method is employed to combine the proposed trajectory-based representation with a conventional interest points-based representation.

Extensive experiments have been conducted to evaluate the effectiveness of the proposed

method using realistic action datasets: YouTube (Liu et al. 2009a), UCF Sport Actions and Feature Films datasets (Rodriguez et al. 2008). The obtained results demonstrate that the proposed methods significantly outperform existing techniques. Furthermore, the proposed Multi-Class Delta Latent Dirichlet Allocation feature selection model and the adaptive fusion strategy notably contribute to the action classification process.

6.2.1 Future work

In this section are summarized some ideas to further extend the proposed method, with an aim to build a more robust and stable action recognition framework.

1. The presented feature fusion between the two interest points representations is achieved by using Multiple Kernel Learning. The Bag-of-Words (BOW) representation is treated as a single feature block while six kernels are associated with the COP features. Each kernel is associated with a different temporal scale. It will be interesting to explore the advantage of dividing the BOW features in sub-blocks and associating them to different kernels. For instance, the BOW features can be divided in two: visual words generated by the upper body and by the legs. Eventually, if multi-scale interest points are available, it will be possible to associate different kernels to each interest point scale.

2. In the presented key point trajectories representation, three different descriptors are extracted from each single trajectory, ignoring the potential information associated with the analysis of the surrounding area. It has been shown in (Sun et al. 2009a) that grouping neighbouring trajectories allows one to capture interesting patterns, as well, to filter out possible outliers. Along a similar idea, the global behaviour of the trajectories can be explicitly represented using a cloud of trajectories approach similarly to the work presented in Chapter 3.

3. Although the proposed collaborative feature selection method outperforms the existing methods, it is still unable to identify meaningful unique topics associated with each action class. For this reason, only shared topics are involved in the feature ranking. This suggests that instead of using single action classes as topics it may be more meaningful to use groups of similar action such as: lower body action (*running*, *jogging*, *walking*), riding action(*bike riding* and *hours riding*), jumping action (*jumping*, *basketball*, *volleyball*). When

doing this, the topics will be more easily separable and more robust in performing feature selection.

4. The region of interest (ROI) estimation in unconstrained video sequences is a complicated issue. The actual implementation employs a simple algorithm based on trajectory centroid estimation. Since the ROI estimation partially influences the recognition performance, it is interesting to explore more sophisticated approaches. Moreover, it can be observed that in realistic scenarios a number of actions may involve groups of people, for instance *playing volleyball*, *basketball* or *running*. In these circumstances, the ROI should be able to handle and identify action performed by an individual or a group, improving the generalization of the problem. This issue may also be solved with an alternative solution as presented in the work presented by (Gilbert et al. 2009), where ROI estimation is replaced by a feature mining approach that is able to localize the action.

5. The current work considers fusion of features extracted from interest points and trajectories. Recently it has been demonstrated that alternative features such as optical flow and scene context descriptors give strong performance on benchmarking datasets. These features explore different action aspects and contain information that is highly complementary to both interest points and trajectories-based representations. Motivated by the observed performance offered by feature fusion, it will be interesting to formulate a framework which combines three or more alternative representations.

6.3 Cascade Feature Selection and Action Classification

Automated action recognition in unconstrained environments is a complex problem and still remains partially unsolved. As discussed in the previous chapters, the final aim is to formulate a solid action recognition method capable of handling real-world challenges. In this context, it has been observed that in the presence of a very sparse training set and video sequences featuring by high intra-class variation and high inter-class similarity, standard feature selection methods and multi-class classifiers appear extremely inefficient. Thus, the multi-class classification and feature selection problem needs to be reformulated in a more efficient way. The proposed solution involves simplifying the multi-class classification task into easier subtasks, where in each subtask feature selection and classification should be simultaneously formulated according to the specific context.

This idea is developed in Chapter 5, where a cascaded feature selection and classification approach is presented. Specifically, the multi-class classification task is decomposed in a cascade of binary separations. Where in each separation step, as much information as possible is exploited to optimize the classification.

Experiments are carried out using challenging datasets such as KTH (Schüldt et al. 2004) and Hollywood (Laptev et al. 2008). It has been demonstrated that with an identical action representation, the presented cascaded feature selection and classification approach significantly outperforms standard multi-class classifiers.

6.3.1 Future work

The proposed study underlines that in the presence of ambiguous visual information and a sparse training set (such as the Hollywood dataset), alternative cues should be employed to capture relevant information as presented in (Laptev et al. 2008). Meaningful information can be extracted from the context, as well as from the movie scripts and subtitles when available. Similarly, key object detection and posture shapes may enrich the action description. By observing the cascade framework performance, the training process can becomes computationally expensive principally due to an exhaustive cross-validation. Moreover, the cascade structure is not updated during the testing process, missing the chance to further consolidate its discriminative power.

Below is a summary of some further investigations to improve the proposed cascade framework:

1. The study of context information may provide a rich source of extra information, crucially important in the presence of motion blur, serious occlusions and low resolution. Under such challenges, the cues associated with the scene and/or moving objects can be used to complement features extracted from the subject. The intuition behind this is straightforward: the presence (or absence) of particular objects or scene properties can often be used to infer the possible subset of actions that can take place. For example, if there is a swimming pool within the scene, then *diving* becomes a possible action. On the contrary, if there is no swimming pool, but a basketball court, then the probability of the *diving* action is reduced. Exploring the relationships between objects, scenes and actions will be compelling.

2. The used action representation is based on conventional interest points, which has been

shown in Chapter 3 to be unreliable in the presence of camera movements. This suggests that a trajectory-based representation should be able to outperform the achieved results. Additionally, as presented in Section 4.2.1 as well as in (Laptev et al. 2008), the usage of a spatio-temporal grid to contribute to capture low-level features should make the representation more robust.

3. Chapter 5 also raises the problem of classification training with sparse and ambiguous data. Although the decision boundaries are optimally estimated with the available data, the recognition process does not ensure a high recognition rate. To partially overcome this limitation, an additional supervision from an external user can be requested. By doing this, the learning algorithm is able to interactively query the user obtaining the needed information to overcome ambiguities. Although this solution will notably improve the performance, on the other hand it will be expensive in terms of human supervision.

4. Database: The available datasets cover small groups of actions within a similar context such as sport, films or primitive actions. Additionally, these datasets contain between 6 to 11 action classes. The action recognition community needs a larger and more comprehensive dataset containing 20 to 30 different actions covering different contexts and environments. Moreover, the frame size, quality, and rate should meet the level of modern video cameras. The collection of a new dataset will strongly contribute to improve future action recognition approaches.

Appendix A

Human Action Datasets

A.1 Weizmann Dataset

The Weizmann dataset was introduced by Blank et al. (2005). It contains 90 video clips from 9 different subjects. Each video clip contains one subject performing a single action. There are 10 different action categories: *walking, running, jumping, galloping sideways, bending, one-hand-waving, two-hands-waving, jumping in place, jumping jack, and skipping*. Each clip lasts about 2 seconds at 25Hz. The image size is 180 by 144 pixels, some examples are reported in Fig. A.1(a).

The same Weizmann group also provides a robustness test dataset, some example frames are shown in Fig. A.1(b). It includes 11 walking sequences with partial occlusions and non-rigid deformations: *walking in skirt, walking with a briefcase, knees up walking, limping man, occluded legs, walking swinging a bag, sleepwalking, and walking with a dog*. The dataset also includes 9 walking sequences captured from different viewpoints from 0° to 81° with 9° increments from the horizontal plane. This dataset is ideal for testing the robustness of an action recognition approach under occlusions, different views, and non-rigid deformations.

The sequences are recorded in a very constrained environment with static camera and clear background, this allows to for the extraction of a detailed silhouette and easily segments out the human body. Moreover, the actions are performed specifically for the dataset minimizing the intra-class variation. This makes the Weizmann dataset one of simplest benchmarks available, where current methods already achieve 100% recognition rate.

(a)

(b)

Figure A.1: Examples frames form a) Weizmann Dataset, from top left to right: "*bending*", "*jumping-jack*", "*jump-in-place*", "*jumping-forward*", "*gallop sideways*"and "*wave-one-hand*". b) Robustness Test Dataset, from top left to right: "*Walking in 45 degree*", "*Walking in 81 degree*", "*Walking with a dog* ", "*Sleepwalking* ", " *Walking occluded by a pole*"and " *Walking with occluded Legs* ".

A.2 KTH Dataset

The KTH dataset was provided by Schüldt et al. (2004) and still represents a benchmark for the action recognition community. It contains 6 types of actions: *boxing, hand clapping, hand waving, jogging, running and walking* performed by 25 subjects in 4 different scenarios including indoor, outdoor, changes in clothing and variations in scale. Each video clip contains one subject performing a single action. Each subject is captured in a total of 23 or 24 clips, giving a total of 599 video clips. Each clip has a frame rate of 25Hz and lasts between 10 to 15 seconds. The size of each image frame is 160 by 120 pixels. Examples of the KTH dataset are shown in Fig. A.2.

Figure A.2: Examples frames form the KTH Dataset, from top left to right: "*boxing*", "*hand waving*", "*clapping*", "*jogging*", "*running* and "*walking*

Despite the fact that some of the sequences contain some real-world difficulties, such as strong shadow, camera motion and zooming, changes in view angle and low grey-scale resolution, current methods already achieve more then 90% recognition rate. This is due to the fact that the intra-class variation is low, actions are performed by actors in clear and static background without any occlusion or distortion. Anyhow, KTH still remains the most popular action recognition benchmark.

A.3 UCF Feature Films Dataset

Figure A.3: Examples frames form the UCF Feature Films Dataset, top line: "*Kissing*", bottom line: "*Hitting/Slapping*"

The UCF Feature Films Dataset (Rodriguez et al. 2008) provides a representative pool of natural samples of two action classes including *Kissing* and *Hitting/Slapping* as reported in Fig. A.3. It contains 92 samples of *Kissing* and 112 samples of *Hitting/Slapping*, extracted from a range of classic movies. The actions were captured in a wide range of scenes and viewpoints with different camera movement patterns. The clips have different frame rates and different image sizes, lasting between 5 to 15 seconds.

Differently from KTH and Weizmann dataset, these UCF clip are recorded in complex realistic scenarios featured by multiple subjects, people interacting, occlusion and dynamic background. Since the clips are short, it is difficult to properly identify the subjects as well as to extract clear silhouettes. The two action classes differ on motion components and visual appearance. *Hitting/Slapping* involves fast movements, while *Kissing* involves a specific posture and slow movements.

A.4 UCF Sport Actions Dataset

Figure A.4: Examples frames form the UCF Sport Actions Dataset, from top left to right: "*golf*", "*kicking*", "*weight-lifting*", "*running*", "*skateboarding*"and "*swinging 2*"

The UCF Sport Actions Dataset (Rodriguez et al. 2008) contains 10 different types of human actions in sport broadcasting videos: *diving, kicking , weight-lifting, horse-riding, running, skateboarding, golf, swinging, swinging 1* (gymnastics, on the pommel horse and floor), *swinging 2* (gymnastics, on the high and uneven bars) and *walking*. Some examples are reported in Fig. A.4. The dataset consists of 150 video samples, which show a large intra-class variability. The videos have different frame rates and image sizes and they last an average of 5 seconds.

This dataset offers a good collection of realistic clips recorded in unconstrained scenarios. These clips contain a large range of real-world difficulties such as occlusion, camera movements and zooming, shadow, multiple subjects, dynamic and crowded background. Additionally, the dataset highlights a number of challenging aspects typical in the context of sport, for instance *Golf* and *Diving* involve fast movements that are very difficult to visually detect. Actions such as *horse riding* and *swinging* involve human-item interaction, while *soccer-kicking* and *running* are usually preformed in a very crowded scenario. Due to the mentioned issues, current methods still are unable to report high recognition performances.

A.5 Hollywood Dataset

Figure A.5: Examples frames form the Hollywood Dataset, from top left to right: "*hand shaking*", "*kissing*", "*getting out of the car*", "*answering the phone*", "*hugging*"and "*standing up*"

The Hollywood Dataset (Laptev et al. 2008) contains 8 different action classes: *answering the phone, getting out of the car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up*. These actions were collected from 32 different Hollywood movies. The full dataset contains 663 video clips sampled at 25 Hz and each of them has a different frame size and duration. The dataset is divided into manually and automatically labelled clips (Laptev et al. 2008). In this work experiments, the manually labelled set only is used where the training set contains 219 clips, while the testing set contains 211 clips. As shown in Fig. A.5, the dataset is composed of realistic sequences, and actions are performed by more than one person in a crowded and dynamic background. The variations in lighting, view angle and drastic camera movement make the dataset challenging.

Laptev et al. (2008) introduced the dataset with the idea of recognising actions exploiting simultaneously visual and text information (subtitles and scripts). In this work experiments, the visual cue only are used. The principal limitation of this dataset consists of the low visual correlation between training and testing set. Moreover, the intra-class variation is notably high, for instance, the class *sitting-up* contains videos where only the actor's face is recorded and the camera follows the face movement, losing all the motion and contextual information. Due to these problems, Marszałek et al. (2009) released a new dataset version more suitable for visual only processing.

A.6 YouTube Dataset

Figure A.6: Examples frames form the YouTube Dataset, from top left to right: "*basketball*", "*cycling*", "*diving*", "*horse-riding*", "*soccer juggling*"and "*volleyball*"

The YouTube Dataset (Liu et al. 2009a) is the most extensive realistic action dataset available to public. it is composed of 1168 videos collected from YouTube. These videos contain a representative collection of real world challenges such as: shaky cameras, cluttered background, variation in object scale, variable and changing viewpoint and illumination, and low resolution. Particularly, since these videos are mostly home videos captured by hand-held cameras, the camera movements are much more unpredictable compared to other datasets. The YouTube dataset contains 11 action categories: *basketball shooting, volleyball spiking, trampoline jumping, soccer juggling , horse-riding, cycling, diving, swinging, golf swinging, tennis, swinging, and walking*. Clips have different frame rates but constant frame size of 320 by 240 pixels. The clips last between 3 and 15 seconds. Examples are shown in Fig. A.6.

The main difficulties of this dataset are the inconsistent frame rate, fast camera movements and low image quality. In these circumstances, standard feature extraction methods appear inadequate, thus more sophisticated strategies and additional pre-processing steps are required. Furthermore, dynamic backgrounds and occlusions make target identification complicated. The dataset offers a good collection of realistic challenges, and for this reason current methods are unable to perform very high recognition rate.

Bibliography

R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *International Conference on Very Large Data Bases*, pages 487–499, 1994.

A. Ali and J.K. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, page 28, 2001.

D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu. Statistical debugging using latent topic models. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 6–17, 2007a.

D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu. Statistical debugging using latent topic models. In *European Conference on Machine Learning*, pages 6–17, 2007b.

V. Athitsos, J. Alon, and S. Sclaroff. Efficient nearest neighbor classification using a cascade of approximate similarity measures. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *International Conference on Machine Learning*, 2004.

J. Barresi and C. Moore. Intentional relations and social understanding. In *Behavioral and Brain Sciences*, volume 19, pages 107–122, 1996.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision*, 2005.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022, 2003.

A. F. Bobick, S. S. Intille, J. W. Davis, F. Baird, C. S. Pinhanez, L. W. Campbell, Y. A. Ivanov, A. Schutte, A. Schu Tte, A. Wilson, and Immersive Story Environment. The kidsroom:

Perceptually-based interactive and immersive story environment. In *PRESENCE*, pages 367–391, 1999.

A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 23, pages 257–267, 2001.

K. L. Boyer and S. Sarkar. Perceptual organization in computer vision: status, challenges, and potential. In *Computer Vision and Image Understanding*, volume 76, pages 1–5, 1999.

A. A. Branzan, B. Trevor, V. B. Naznin, and B. Cheryl. Analysis of irregularities in human actions with volumetric motion history images. In *IEEE Workshop on Motion and Video Computing*, page 16, 2007.

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*, 2009a.

M. Bregonzio, S. Gong, and T. Xiang. Action recognition with cascaded feature selection and classification. In *International Conference on Imaging for Crime Detection and Prevention*, 2009b.

M. Bregonzio, J. Li, S. Gong, and T. Xiang. Discriminative topics modelling for action feature selection and recognition. In *British Machine Vision Conference*, 2010.

T.M. Cover. The best two independent measurements are not the two best. In *IEEE Transaction on Systems, Man, and Cybernetics*, volume 4, pages 116–117, 1974.

S. Danafar and N. Gheissari. Action recognition for surveillance applications using optic flow and svm. *Asian Conference on Computer Vision*, 2007.

J. Decety, J. Grèzes, N. Costes, D. Perani, M. Jeannerod, E. Procyk, F. Grassi, and F. Fazio. Brain activity during observation of actions. influence of action content and subject's strategy. In *Brain: a journal of neurology*, volume 120 (Pt 10), pages 1763–1777, 1997.

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

A. A. Efros, A. C. Berg, Er C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *International Conference on Computer Vision*, 2003.

A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

F. Ferri, P. Pudil, M. Hatef, and J. Kittler. Comparative study of techniques for large-scale feature selection. *Pattern Recognition in Practice*, pages 403–413, 1994.

Y. Freund and R. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences*, volume 55, pages 119–139, 1997.

P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *International Conference on Computer Vision*, 2009.

A. Gilbert, J. Illingworth, and R. Bowden. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In *European Conference on Computer Vision*, 2008.

A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *International Conference on Computer Vision*, 2009.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. In *Journal of Machine Learning Research*, volume 3, pages 1157–1182, 2003.

I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. In *Pattern Analysis and Machine Intelligence*, volume 22, pages 809–830, 2000.

C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

N. Ikizler, R. Gokberk Cinbis, and P. Duygulu. Human action recognition with line and flow histograms. In *International Conference on Pattern Recognition*, 2008.

A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: a review. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, pages 4–37, 2000.

G. Johansson. Visual perception of biological motion and a model for its analysis. In *Perception And Psychophysics*, volume 14, pages 201–211, 1973.

G.H. John, R. Kohavi, and K. Peger. Irrelevant features and the subset selection problem. *International Conference on Machine Learning*, 1994.

V. John, E. Trucco, and S. J. McKenna. Markerless human motion capture using charting and manifold constrained particle swarm optimisation. In *British Machine Vision Conference (Workshops)*, pages 4.1 – 4.11, 2010.

G. Kim, C. Faloutsos, and M. Hebert. Unsupervised modeling and recognition of object categories with combination of visual contents and geometric similarity links. In *International conference on Multimedia information retrieval*, pages 419–426, 2008.

K. Kira and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. *Association for the Advancement of Artificial Intelligence*, pages 129–134, 1992.

J. Kittler. Feature set search algorithms. *Pattern Recognition and Signal Processing*, pages 41–60, 1978.

A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference*, 2008a.

A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008b.

C. Koch and B. Mathur. Neuromorphic vision chips. *IEEE Spectrum*, 1996.

P. Kontkanen and P. Myllymaki. MDL histogram density estimation. In *International Conference on Artificial Intelligence and Statistics*, 2007.

G. Kosta and M. Benoit. Group behaviour recognition for gesture analysis. In *Circuits and Systems for Video Technology*, volume 18, pages 211–222, 2008.

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

V. Kruger, D. Kragic, A. Ude, and C. Geib. The meaning of action: A review on action recognition and mapping. In *Advanced Robotics*, volume 21, pages 1473–1501, 2007.

I. Laptev and T. Lindeberg. Space-time interest points. *International Conference on Computer Vision*, 2003.

I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

J. Li, S. Gong, and T. Xiang. Discovering multi-camera behaviour correlations for on-the-fly global activity prediction and anomaly detection. In *IEEE International Workshop on Visual Surveillance*, 2009.

Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *International Conference on Computer Vision*, 2009.

J. Liu and M. Shah. Learning human actions via information maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009a.

J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009b.

A. P. Brandão Lopes, E. Alves do Valle Jr., J. Marques de Almeida, and A. de Albuquerque Araújo. Action recognition in videos: from motion capture labs to the web. *Computing Research Repository*, abs/1006.3506, 2010.

D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60, pages 91–110, 2004.

M. Marszałek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *International Conference on Computer Vision*, 2009.

S. J. McKenna and H. Nait-Charif. Summarising contextual activity and detecting unusual inactivity in a supportive home environment. In *Pattern Analysis and Applications*, volume 7, pages 386–401, 2004.

R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *International Conference on Computer Vision*, 2009.

T. P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft Research, 2003.

T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. In *Computer Vision and Image Understanding*, volume 104, pages 90–126. Elsevier Science Inc., 2006.

A. W. Moore. An intoductory tutorial on kd-trees. In *Thesis Technical Report, University of Cambridge*, volume 209, 1991.

J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *International Journal of Computer Vision*, volume 79, pages 299–318, 2008.

S. Nowozin, G. H. Bakir, and K. Tsuda. Discriminative subsequence mining for action classification. In *International Conference on Computer Vision*, pages 1–8, 2007.

A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. In *IEEE Transactions of Systems, Man, and Cybernetics*, volume 36, pages 710–719, 2006.

A. Oikonomopoulos, I. Patras, and M. Pantic. An implicit spatiotemporal shape model for human activity localization and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshops)*, page 2733, 2009.

J. Parker. *Algorithms for Image Processing and Computer Vision*. Wiley Computer, 1997.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 2, pages 1226–1238, 2005.

R. Poppe. A survey on vision-based human action recognition. In *Image and Vision Computing*, volume 28, pages 976–990, 2010.

D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? In *Behavioral and Brain Sciences*, volume 1, pages 515–526, 1978.

P. Pudil, J. Novovicov, and J. Kittler. Floating search methods in feature selection. In *Pattern Recognition Letters*, volume 15, pages 1119–1125, 1994.

C. Rao and M. Shah. View-invariance in action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 316–322, 2001.

H. Ren and G. Xu. Human action recognition in smart classroom. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 417–422, 2002.

H. Riemenschneider, M. Donoser, and H. Bischof. Bag of optical flow volumes for image sequence recognition. *British Machine Vision Conference*, 2009.

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*, 2008.

S.R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. In *IEEE Transactions on Systems, Man and Cybernetics*, volume 21, pages 660–674, 1991.

S. Savarese, A. Del Pozo, J. Niebles, and L. Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, 2008.

G. Schindler, L. Zitnick, and M. Brown. Internet video category recognition. *Internet Vision*, pages 1–7, 2008.

C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *International Conference on Computer Vision*, 2005.

J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, pages 888–905, 2000.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. In *Journal of Machine Learning Research*, volume 7, pages 1531–1565, 2006.

H. Spath. Cluster dissection and analysis: Theory, fortran programs, examples. In *Halsted Press*, volume 1, 1985.

J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009a.

X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009b.

D. Tran and A. Sorokin. Human activity recognition with metric learning. *European Conference on Computer Vision*, 2008.

V. N. Vapnik. *The nature of statistical learning theory*. Springer, 1995.

P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, volume 57, pages 137–154, 2002.

H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. In *IEEE Transactions on Image Processing*, volume 16, pages 1646–1661, 2007.

L. Wang and D. Suter. Informative shape representations for human action recognition. *International Conference on Pattern Recognition*, 2006.

S. Wang, C. Liu, and L. Zheng. Feature selection by combining fisher criterion and principal feature analysis. In *International Conference on Machine Learning and Cybernetics*, pages 1149–1154, 2007a.

Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Y. Wang, K. Huang, and T. Tan. Human activity recognition based on $\Re$ transform. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007b.

A.W. Whitney. A direct method of nonparametric measurement selection. In *IEEE Transactions on Computers*, volume C-20, pages 1100–1103, 1971.

S. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. *International Conference on Computer Vision*, 2007.

L. Xie, P. Xu, S. Chang A, A. Divakaran, and H. Sun B. Structure analysis of soccer video with hidden markov models. In *Pattern Recognition Letters*, pages 767–775, 2002.

A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *International Conference on Computer Vision*, 2009.

A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 984–989, 2005.

L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. In *Journal of Machine Learning*, page 12051224, 2004.

L. Yu-Ming, S. Sheng-Wen, A.C Shih, H.-Y.M. Liao, and L. Cheng-Chung. A language modeling approach to atomic human action recognition. In *IEEE Workshop on Multimedia Signal Processing*, pages 288–291, 2007.

M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In *International Conference on Uncertainty in Artificial Intelligence*, pages 577–584, 2002.

D. Zhang and G. Lu. A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *Journal of Visual Communication and Image Representation*, number 14, pages 41–60, 2003.

Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: A new representation for human action recognition. In *European Conference on Computer Vision*, volume 4, pages 817–829, 2008.

Z.P. Zhao and A.M. Elgammal. Information theoretic key frame selection for action recognition. In *British Machine Vision Conference*, 2008.