University of Southampton
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

Learning Human Behaviour in Complex Scenes

by

Ping-Lin Chang

2010-09-24

A dissertation submitted in partial fulfilment of the degree of
MSc Artificial Intelligence
by examination and dissertation

Abstract

Recognition of human actions is a challenging issue in computer vision. The automatic recognition systems are necessary due to the ubiquitous surveillance video at present. Through understanding the content of the video, an intelligent recognition system provides a more effective and efficient searching. On the other hand, a real-time surveillance system requires an on-line recognition to detect suspicious actions and perform the reaction immediately, such as for in banks and shopping mall, or for traffic control. All the techniques rely on first analyzing the human actions in the surveillance videos. In this dissertation, we propose a novel interest point detector for detecting human actions.

Specifically based on part-based approach, a general human action recognition framework includes spatio-temporal interest point detection, building the descriptor, constructing the codebook, and testing on the pre-trained classifier. We intentionally focus on the detectors for precisely detecting the human action and the classifiers for performing accurate classification. Noting that actions contain not only moving parts but also static regions, such as the static heel strike in walking, a stasis interest point detector is addressed. More specifically, we exploits two popular spatial interest point detectors, SIFT and FAST, to detect the base interest points, where are then filtered by optical flow to decide the interest points with moving information. Stasis interest points thus can be found by looking for static base interest points around the moving interest points. Empirical studies show the improvement of classification accuracy by adding the proposed stasis interest point.

Different classifiers have been evaluated with the devised interest point detectors. Experiment results show that there is still room for the researches in kernel design and the classification model choosing. Three datasets are employed in the evaluations which are KTH, Weizmann, and an industry environment surveillance video. Preliminary studies have shown the promise of the stasis interest point in human action recognition.

Acknowledgement

I would like to thank everything happening in my life and the love surrounding me.

Contents

Chapter 1

Introduction

In this dissertation, a robust human behaviour analysis algorithm has been studied. The human behaviour is regarded as actions performed by people in the sequence. The intelligent algorithm is devised to analyze the sequence to recognize various actions in complex scenes and variant circumstance. While there are several steps included in the recognition system, we especially emphasize the action detection and classification step. Action detection is the most critical process in the entire system since the noisy environment and cluttered background are managed in this step. A failed detector generates nothing useful for the rest of steps, and results in an error classification. In additions, the classifier plays an important role at the last step where distinct input data are classified. After learning from a set of training data, a robust classifier ensures a reasonable prediction with the testing data. In this work, we have devised a novel algorithm in action detection step and have investigated suitable classifiers for the proposed algorithm. The system is then evaluated by standard datasets KTH [30], Weizmann [3], and a real industry surveillance video.

The part-based approach which utilizes the spatio-temporal interest points has been targeted. Noting that the proper description for actions should contain both the moving and the static information at the same time, prior detection approaches are insufficient. Previous studies have only focused on motion detection through the significant variation of patterns in the temporal dimension. In contrast, we have developed a novel interest point detector which determines not only the motion part but also the static region efficiently. More specifically, the stasis interest points which represent meaningful static regions can be detected through incorporating the state of the art spatial interest point detector with a common optical flow detector. The final classification affects the final recognition accuracy as well. We have found that the kernels utilized in support vector machine (SVM) are still worth to be researched. Moreover, while Probabilistic Latent Semantic Analysis (*p*LSA) provides additional detailed and flexible probability distribution for the further analysis, it suffers from the increase of the diversity of human actions.

1.1 Motivation

Recognition of human actions in videos is a challenging issue in computer vision, and has attracted significant interest for some time. Via our visual system, human can easily understand human actions in a complex scene. However, it is not trivial to enable a computer to process the low-level digital signals obtained via a camera to obtain the high-level interpretation of human being. The task of human action recognition can be regarded as labelling videos containing human motion with action classes. Such a problem is referred to as pattern recognition, or more specifically, automatic categorization and localization of human actions. The interest in this topic is motivated by the emergence of demand for both off-line and on-line applications.

Off-line annotation of videos, for example, enables more efficient and effective video searching queried by motion keywords, such as querying for the segment of action movies, the sports news, or the clip containing particular actions of interest by users. On-line application allows for intelligent real-time surveillance, such as in shopping malls, banks, ATMs, or in airport, and moreover, outside traffic control and inside intelligent care for elder people all require a precise on-line surveillance. Besides, automatic human action recognition provides more intuitive media in the applications of human computer interaction (HCI) as well as augmented reality (AR).

However, constructing a robust human action recognition system for generic use is an extremely tough problem. To formalize the problem, given a set of consecutive images with one or more subjects performing various actions, can we devise a system that can automatically recognize what actions are performed? Although the question might seem naïve, the solution has tormented researchers for more than 20 years [24] [33]. Well-known issues include the variation of posture, appearance, scale, camera motion, illumination, background, and occlusions, which make the problem ill-pose even in a very easy task.

The goal of this work is to develop a robust human action recognition system that can adapt to various changes of view point, appearance, scale, illumination, and environment. In this work, we focus on a part-based approach for human action recognition in which the treatment has been shown a better performance under influence of varying conditions [8]. The part-based approach usually involves several steps which are modelling environment, interest point detection, interest point description, constructing codebook, and action classification. The work is focussed specifically on the interest point detection and the action

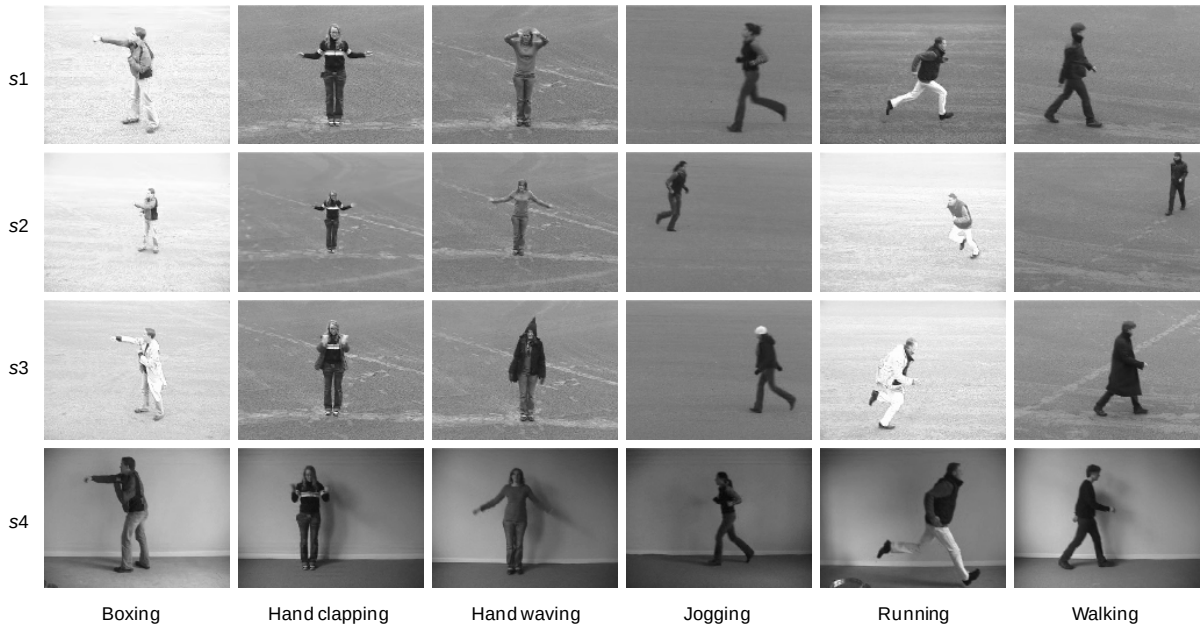


Figure 1.1: Example frames of KTH dataset [30]. The dataset contains six action categories, performed by 25 subjects, and four different scenarios: *s*1 is outdoor view; *s*2 is scale variance; *s*3 is change of appearance; *s*4 is indoor environment. Different actions are performed by the same people under different scenarios.

classification, and has been verified with several commonly-used datasets showing that successful recognition can be achieved by these new approaches.

1.2 Datasets

Three datasets are employed in our experiments. The choice is dependent on their diversity for different evaluations. The first one is KTH [30], which is known as a largest dataset currently with six different actions as well as simple backgrounds and moving cameras. The second one is Weizmann [3], which provides ten different actions but rather less subjects. The third one is an as yet unpublished industry video dataset for top view surveillance, which supports an evaluation with cluttered background and partial occlusion.

1.2.1 KTH

The KTH human motion dataset is the largest available sequence dataset of human actions [30]. Each video has the frame size 160×120 with 25 FPS, and has only one action performed. The dataset contains six categories of human actions which are boxing, hand clapping, hand

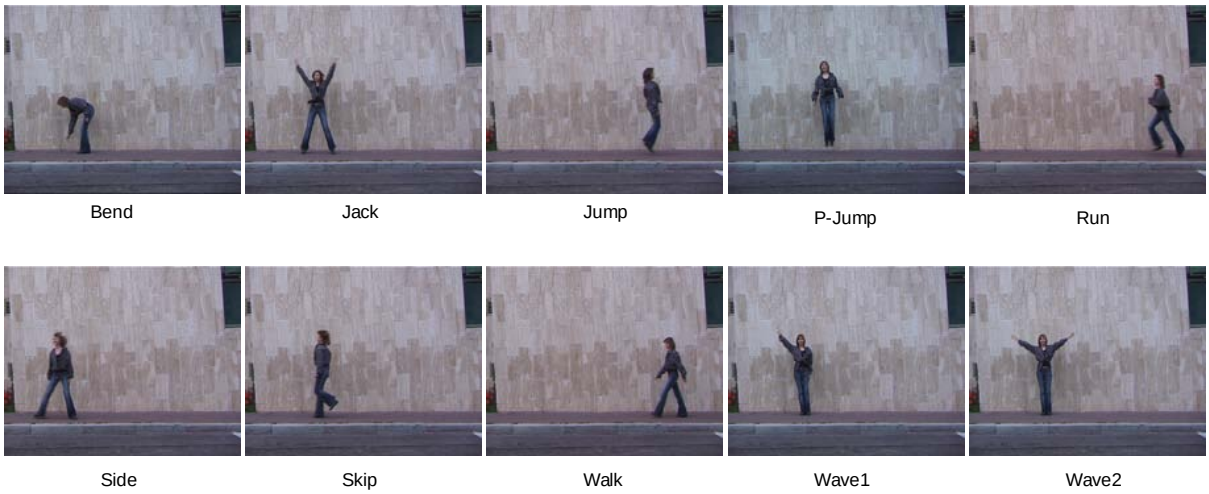


Figure 1.2: Example frames of Weizmann dataset [3]. The dataset contains ten action categories, performed by nine subjects. The sequences are taken by static camera with static background.

waving, jogging, running and walking performed iteratively by 25 subjects in four different scenarios: outdoor view, scale-variance, change of appearance, and indoor environment. In total, it contains 598 sequences. Figure 1.1 shows some frames of the KTH dataset.

1.2.2 Weizmann

The Weizmann dataset contains ten action categories performed by nine people, providing a total of 90 sequences. Each video has the frame size 180×144 with 25 FPS, and has only one action performed. Different action categories in frames are shown in Figure 1.2. This dataset contains sequences with static camera and simple background with single scenario [3]. The Weizmann dataset contains fewer sequences compared with the KTH dataset though it provides a good evaluation if the training set is limited. In addition, the provided categories are more than in the KTH dataset which gives a good testing ground to investigate the performance of the algorithm when the number of categories is increased.

1.2.3 Human actions within industrial environments

This dataset is chosen for a real application within the industry environments which provide additional conditions for evaluating the proposed algorithm, such as partial occlusion of subjects and cluttered background. The data was recorded during the manufacturing cycle of one production station at the Barcelona Nissan factory, under the SCOVIS EU research

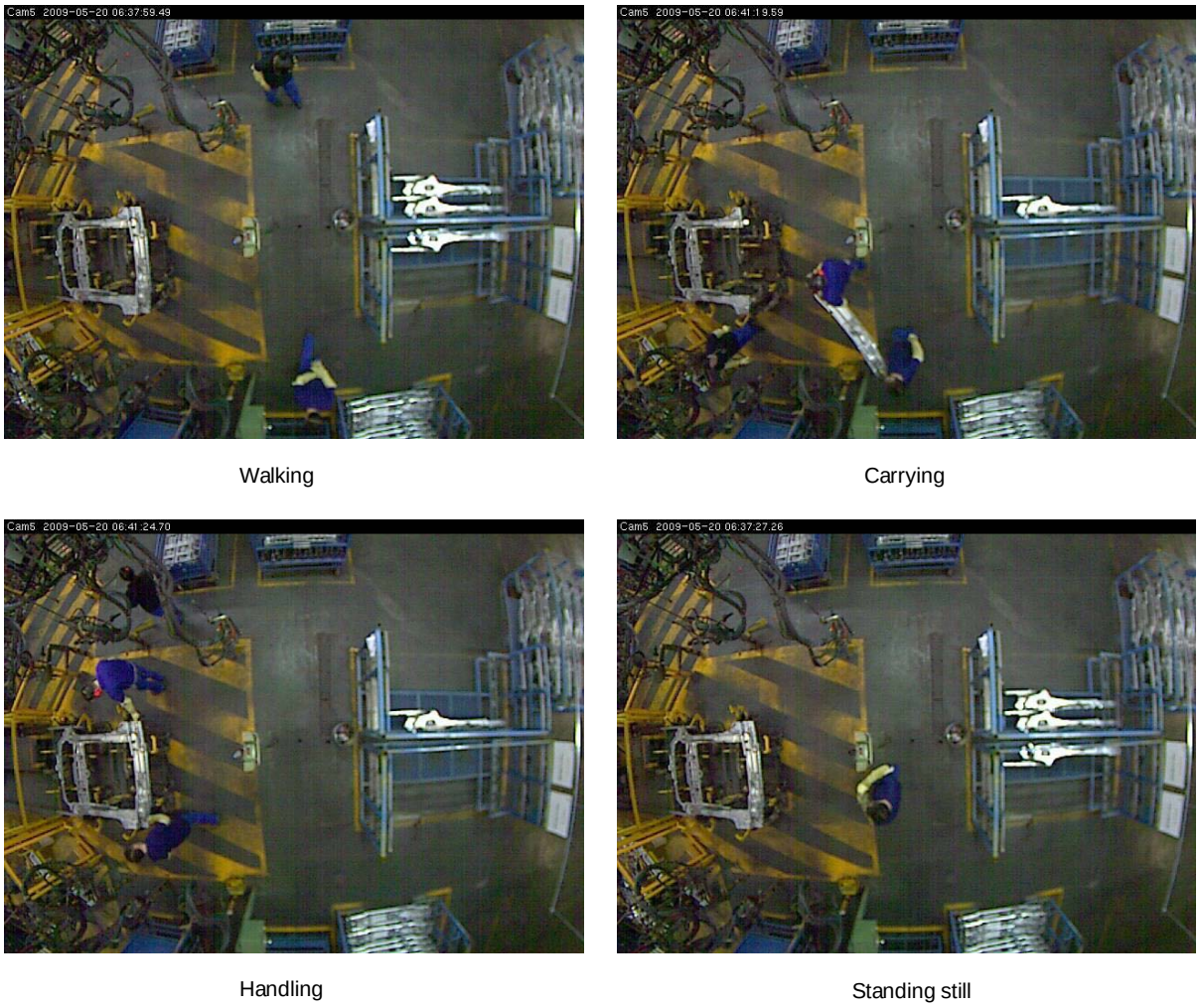


Figure 1.3: Example frames of real industry surveillance video. Note that several subjects would appear in the same frame, and the circumstance is cluttered.

program (http://www.scovis.eu/). The surveillance video is taken by a static camera from an oblique views as it does not suffered from too much occlusion, but still retains powerful cues about the activity of individuals. The dataset is extracted from a daily routine of 10,000 frames which contains four action categories: walking, carrying, handling, and standing still that form the common procedure for the industry operation. Each frame is with 640×480 size. Note that several subjects would appear in the same frame to perform the same or different actions in this dataset. The example frames are shown in Figure 1.3.

1.3 Dissertation statement and contributions

The dissertation focuses on a part-based approach for the human action recognition in complex scene. The contributions of this work are two-fold. First, we have proposed a novel

idea for interest point detection to boost the classification accuracy while being not losing the generality of various spatial interest point detectors for diverse actions in various environments. Second, two popular classification models, SVM for the discriminative model and *p*LSA for the generative model, have been investigated for their utility. The empirical studies have shown promise for the proposed detector as well as the proposed classifiers. This dissertation is organized as follows: Chapter 2 introduces the related works in human action recognition. The proposed framework as well as the novel interest point detector is described in Chapter 3 . In Chapter 4 , several experiments have been conducted for evaluating our algorithms. The dissertation concludes in Chapter 5 with further areas of study to confirm and improve on the advances already made.

Chapter 2

Related Works

For decades, human action recognition has remained a challenge issue and has attracted a considerable number of attention. In literature, the solutions which analyze actions in space-time domain can be divided into two categories: holistic approaches and part-based approaches [24]. Holistic approaches utilize global information to recognize actions, and finding a region of interest (ROI) is required at first to localize the subject. This can be reached by using background subtraction or tracking techniques. Holistic approaches generally perform quite well because all information in the sequence are analyzed at once. However, the good performance is only under particular circumstances since holistic approaches rely on a good ROI discovery and are very sensitive to noise, viewpoint, occlusion, and cluttered environment.

Part-based approaches depend on local patches detected between the frames to represent the actions. Spatio-temporal interest point detectors are usually used to discover the local patches first. The local patches are later combined to represent the sequence using with the bag of words model. Due to the local representation, such approaches are less sensitive to partial occlusion, noise, and clutter. Moreover, part-based approaches do not strictly require a background subtraction, and thus are more efficient in feature extraction. The performances of these approaches depend on a sufficient amount of relevant local interest points to represent a specific action, and thus the significant characteristics of that action should be encoded precisely. Previous works only focused on the characteristics of moving part, which is concerned insufficient in our work.

This chapter reviews the previous works for each approach. Holistic and Part-based approaches are discussed in Section 2.1 and Section 2.2 respectively.

2.1 Holistic approaches

Silhouettes, edges or optical flow are common media utilized in holistic approaches. The first work using silhouettes was by Bobick and Davis [4]. In their approach, silhouettes are extracted from a single view, and are accumulated differences between consecutive frames. The procedure results in a binary motion energy image (MEI) to represent the motion, and further forms a motion history image (MHI) in which the intensity is controlled by the time when the motion is happened on the

corresponding location. This representation can be matched using global statistics, such as Hu moment. Note that the silhouette extraction requires a well segmented foreground and background.

Efros et al. [9] proposed a correlating optical flow measurements which requires first stabilized each subject in the sequence to create stabilized spatio-temporal volumes for classification using the annotated labels. The descriptor is formed by a blurred dense optical flow from each volume. This method ignores the detailed information of appearance, and thus is suited to recognize small subjects in low resolution sequence. Blank et al. [3] proposed space-time shapes which utilizes the silhouettes as well. The Poisson equation is then utilized to extract features, such as local space-time saliency, action dynamics, shape structure and orientation. The method also requires static backgrounds for segmenting the foreground.

Liu and Yuen [18] introduced a spatio-temporal Information Saliency Map (ISM) which is calculated from a video sequence by estimating pixel density function. Human actions are segmented into a set of coarse motion cycles first, and are represented by a Salient Action Unit (SAU), which is then performed on Principle Component Analysis (PCA) to determine the EigenAction. Note that holistic approaches generally perform well under controlled environments [24], and they are computationally expensive because of the pre-processing requirements which may include background subtraction, shape extraction, optic flow calculation, or object tracking.

2.2 Part-based approaches

Part-based approaches have overcome the limitations of holistic approaches (i.e. the necessary of background subtraction and object tracking) in human action recognition and provide a more efficient way to extract action features. In additions, they do not require a global appearance analysis which is highly utilized in holistic approach. The idea was first inspired by object recognition in static two-dimensional image in which objects are represented by local salient interest points. The representation benefits overcoming occlusion and the change of postures, unlike the global appearance representation used in holistic approaches. Following the recent success of the part-based approach [13], researches have paid considerable attention on developing the spatio-temporal interest point detector and descriptor [7] [8] [12] [14] [15] [28] [30] [31] [36] [38].

Laptev [13] proposed an extension of Harris corner detector from two-dimensionality to three-dimensionality for the spatio-temporal interest point detection. Locations of sequences which show strong variations of intensity both in spatial and temporal directions interesting parts are of interest. The scale-space in temporal dimension is extended for automatic scale-selection. The detected spatio-temporal interest points are successfully utilized in the work by Schuldt et al. [30] for human action

recognition which exploits the bag of words representation for each sequence and the discriminative model for classifier.

Following Laptev's philosophy, Dollár et al. [8] have suggested to treat time differently from space and to look for periodic motion patterns, which includes space-time corners, using a separable liner Gabor filters. Compare with Laptev's detector, their approach produces a denser sampling of the spatio-temporal volume. They also proposed several descriptors for the local patches around the detected interest points. Based on the same classification model, their report has shown an improvement by the proposed spatio-temporal interest point detector compared with the one used by Schuldt et al. [30] on the same dataset. Using the detector by Dollár et al., Niebles et al. [23] have devised an unsupervised learning model using the generative classifier, inspired by latent topics analysis in text-mining field, resulting in a better classification performance. Willems et al. [38] proposed a new efficient and scale-invariant spatio-temporal detector using the determinant of a three-dimsional Hessian matrix. Wong et al. [37] has conducted an evaluation on the interest point detectors of Dollár et al. [8], Laptev [13], and Willems et al. [38], and found that a dense sampling detector outperformed them.

On the other hand, local descriptors contain the information around the detected interest points, which should be taken into account of background clutter, the variance of appearance due to occlusion, rotation and scale change. Therefore a considerable number of researches have focused on interest point descriptor as well. Scovanner et al. [31] extends the SIFT descriptor to the temporal dimension, and Willems et al. [38] extends SURF features in which each cell contains the sums of Harr-wavelets. More recently, Laptev et al. [15] have addressed a structural representation based on dense temporal and spatial scale sampling inspired by spatial pyramids [16], which encodes appearance and motion by histogram of gradients (HoG) and histogram of optical flow (HoF) respectively. Kläser et al. [12] aggregates three-dimensional gradients by binning each gradients into regular polyhedrons. According to the evaluation of [37] on local descriptors, in general a descriptor combining HoG and HoF has the best performance.

Most of previous researches assume that spatio-temporal interest points should be located at the parts with significant variance of appearances along temporal dimension, which largely ignores the explicit motion information in spatial domain. Under such an assumption, smooth gestures like rounding motions which lack sharp space-time extreme will be omitted. Chen [7] has shown that incorporating explicit motion detection in spatial domain indeed gains a much better result in human action recognitions.

Chapter 3

Motion with Spatio-temporal Interest Point

Processing all the low-level pixels to understand the content of image sequences is infeasible. This chapter elaborates effective techniques utilized in the proposed part-based approach for human action recognition. An entire system flow chart is shown in Figure 3.1. The first step is interest point detection, which is also called feature extraction. Specifically, interest point detection significantly reduces a volume of pixels, and only representative pixels will be preserved for further analysis. In the spatial domain, the representative and high-repeatable interest points among similar images should be kept. The detected interest points (with motion) can be further filtered by using optical flow in the temporal domain. In the case of analyzing sequences, detected spatio-temporal interest points describe the characteristic of different actions, and in principle, should possess high repeatability between similar poses and actions.

Constructing an effective descriptor for each detected spatio-temporal interest point plays an important role in action recognition. The information of appearance and motion should be well-encoded in the devised descriptor. To perform the bag of words model, each spatio-temporal descriptor is regarded as a spatio-temporal word and is employed to train a codebook for obtaining representative vocabulary by clustering algorithm. Each sequence is then expressed by the histogram of spatio-temporal words for the ultimate classification. Two different classification models are introduced. A discriminative model provides a straightforward approach to classify sequences but is constrained from further applications by its principle. Instead, a generative model gives a probability distribution to each spatio-temporal word, and hence enables recognition and localization of multiple actions.

3.1 Interest point detection

Depending on the different approaches mentioned in Chapter 2, several techniques have been proposed to select good features which describe pose and motion. A part-based approach,

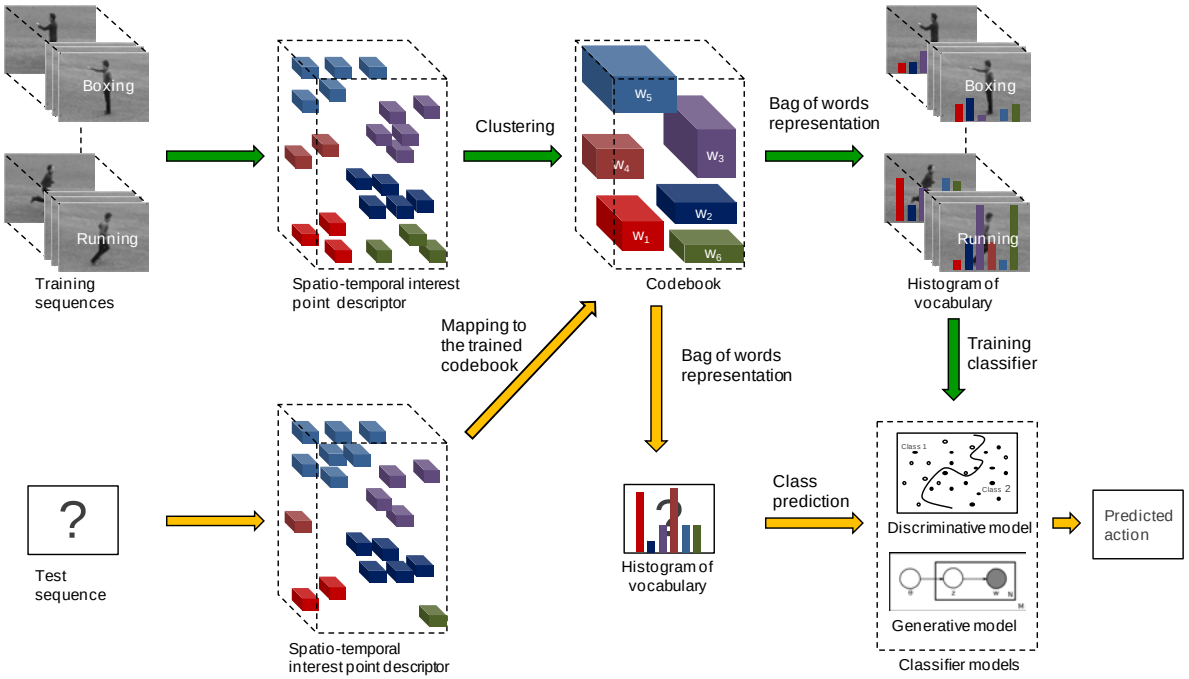


Figure 3.1: The system flow chart for action recognition. The arrow with green colour indicates the training process. The arrow with yellow colour indicates the test process. The figure is best viewed in colour mode.

which utilizes spatio-temporal interest points, is exploited in this work because of its flexibility and tolerance to clutter and partial occlusion [24]. In addition, Dollar et al. [8] have shown that compare with a holistic approaches, a part-based approach using local features or spatio-temporal interest points can provide a more effective description in action recognition. The most intuitive way to detect spatio-temporal interest points is to redesign present two-dimensional detection algorithms with an extra analysis in the temporal dimension for three-dimensional video analysis.

Laptev proposes a spatio-temporal interest point detector based on Harris corner detector [10] [13]. The responding local structure regarded as of interest has high variation in spatial and temporal dimensions. However, the detection generates sparse interest points which hardly characterize many actions in a complex scene. Such a drawback was first noted by Dollar et al [8], and instead, they suggest a detector based on a set of separable linear filters where local regions with complex motion patterns are discovered. The method produces more interest points for various motions, with a better performance in human action recognition [8] [23].

In principle, various spatial interest point detectors can be incorporated into the spatio-temporal scenario, depending on the corresponding algorithms harnessing for the detection in

the temporal dimension. A considerable number of algorithms for interest point detection in the spatial domain have been compared and discussed [29] [34]. In this work, our goal is to develop a method which produces sufficient and representative interest points to describe arbitrary human actions. We thus intentionally select two state-of-the-art spatial interest point detectors. The first one is the scale-invariant feature transform (SIFT) which utilizes difference of Gaussian (DoG) to detect local extrema as interest point [7] [19], and the other one is the features from accelerated segment test (FAST) which efficiently and effectively detects corners by machine learning [27]. Details are introduced in Section 3.1.1 and 3.1.2.

While previous researches elegantly extend interest point detectors from spatial to spatio-temporal dimension, the information of explicit motion has been largely ignored. Specifically, such detectors imply motions by the difference of appearances in temporal dimension, and hence smooth gestures like rounding motions which lack sharp space-time extreme will be omitted. Chen [7] has revealed that incorporating explicit motion detection indeed gains a much better result in human action recognitions, in contrast to the most of works which incorporate only optical flow into descriptors [1] [14] [15] [36]. In addition, Chen also suggests that appearance and motion can be analyzed independently, and address an algorithm called Motion SIFT (MoSIFT) which discover meaningful key points by SIFT and harnesses Lucas-Kanade optical flow to motion detection [7]. Following the philosophy, noting that not only moving parts but also static regions would provide the meaning of action, we propose an additional detector for stasis interest point. Lucas-Kanade optical flow for calculating the magnitude of motion is briefly introduced in Section 3.1.3, and the new concept, stasis interest point, is introduced in Section 3.1.4.

3.1.1 Difference of Gaussian

In the SIFT algorithm, Lowe suggests a detector which realizes scale-invariance through convolving the image with a difference of Gaussian (DoG) kernel in multiple scales. DoG is utilized because of its efficiency and stability approximating the scale-normalized Laplacian of Gaussian (LoG). Given an input image, we can denote the scale space of the image as $L(x, y, \sigma)$ produced by the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with the image, $I(x, y)$:

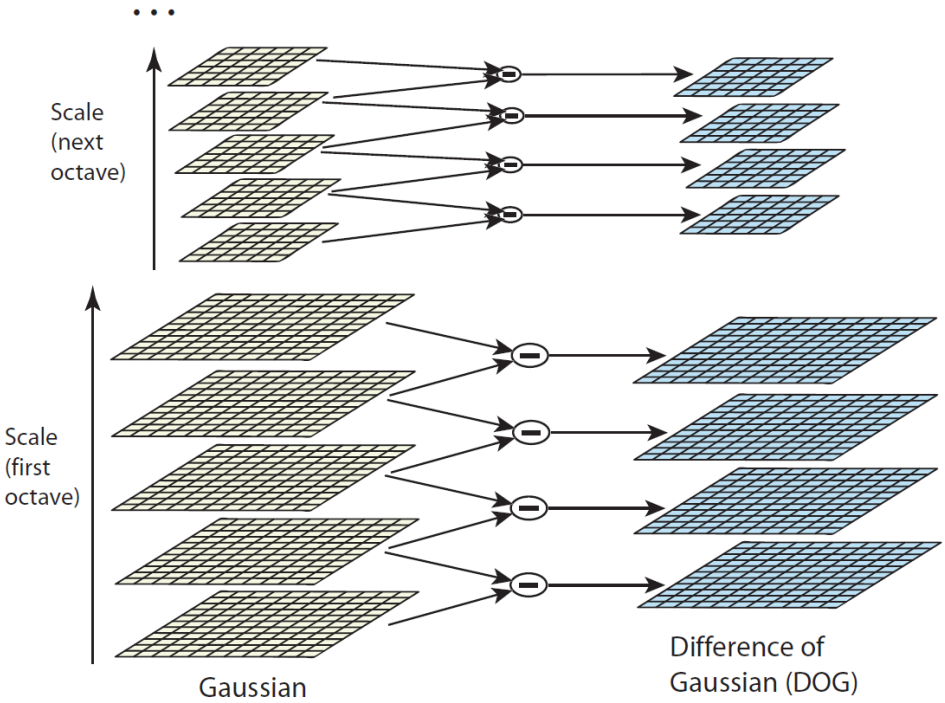


Figure 3.2: On the left side, the initial image in each octave is repeatedly convolved with Gaussians to produce blurred images with different scales. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian (DoG) images on the right side. The figure is reproduced from [19].

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

where $*$ is the convolution operation, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}.$$

With a Gaussian function with a constant factor k, we can compute a DoG function, $D(x, y, \sigma)$, by subtracting two nearby images in scale space:

$$\begin{aligned} D(x, y, \sigma) &= \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma). \end{aligned}$$

The entire scale space thus can be constructed by a sequence of octaves and each octave is divided into a sequence of blurred images convolved by $G(x, y, k\sigma)$. DoG images in the entire scale space can be obtained by iteratively subtracting each interval in each octave as shown in Figure 3.2. In order to control the first image in upper octave having exactly a doubleσ, k is chosen to be $2^{1/s}$, where s+3 is the number of blurred images for each octave.

In the SIFT operator, interest points are detected as local extrema compared with their 26 neighbours in 3x3 regions among the current and adjacent scales as shown in Figure 3.3, and are further filtered out by truncating low contrast and edge response point. Previous studies

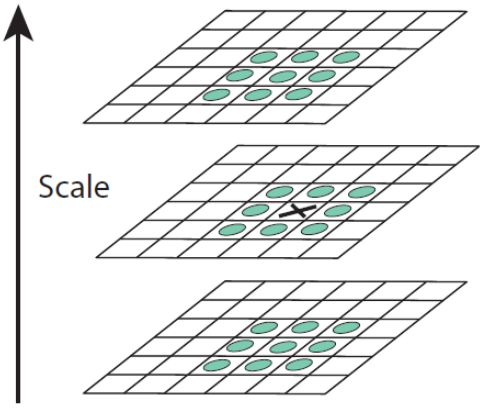


Figure 3.3: Local extrema of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 neighbours in 3x3 regions at the current and adjacent scales (marked with circles). The figure is reproduced from [19].

have shown that the SIFT with DoG interest point detector has a better performance under various transform such as with different scales, affine, shift, and rotation [19] [21] [34]. Note that in the MoSIFT, the further filtering step is omitted for obtaining more interest points.

3.1.2 Features from accelerated segment test

Rosten and Drummond devised a spatial corner detector called features from accelerated segment test (FAST), for high-speed purpose [26], and have further purposed a machine learning approach to boost the speed much faster as well as the performance [27]. The feature detector utilizes a segment test criterion operated by considering pixels in a Bresenham circle of radius three around a corner candidate point p (i.e. a circle of 16 pixels circumference). If a set of n contiguous pixels in the circle which are all brighter than the intensity of the candidate point $I_p + t$ or all darker than $I_p - t$, then p is considered to be a corner, as shown in Figure 3.4.

However, the high-speed test does not perform well for $n < 12$. To speed FAST up, a machine learning process has been included to generate a decision tree by ID3 [25]. In order to obtain the ground truth corners, with a given value for n and a suitable threshold t, the slow FAST is first performed on a set of training images which are preferably from the application domain. Second, each pixel p in the whole training images is tested with its 16 pixels on the circle to get the corresponding states. Specifically, given the set of $x \in \{1...16\}$ surrounding p denoted by $p \to x$, three possible states are:

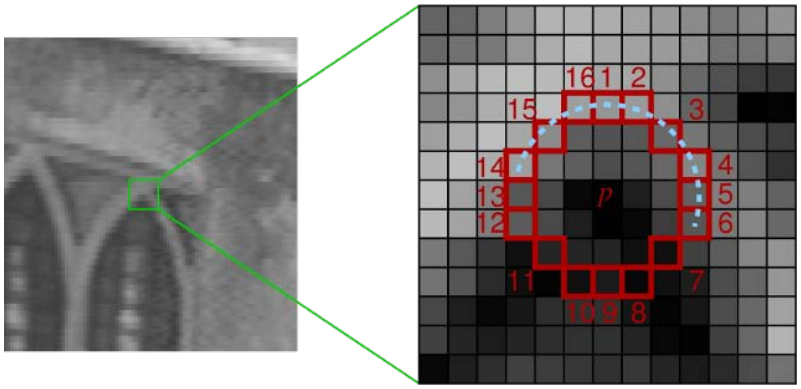


Figure 3.4: An $n = 9$ point segment test corner detection in an image patch. The pixel at p is the centre of a candidate corner judged by the pixels at red squares. The dash arc passes through 9 contiguous pixels which are brighter than p by more than the threshold. The figure is reproduced from [27].

$$S_{p\to x} = \begin{cases} d, & I_{p\to x} \le I_p - t \quad \text{(darker)} \\ s, & I_p - t < I_{p\to x} < I_p + t \quad \text{(similar)} \\ b, & I_p + t < I_{p\to x} \quad \text{(brighter)} \end{cases} .$$

Selecting an x and computing the state $S_{p\to x}$ for each pixel $p \in P$, where P is the root set, and is further partitioned into three subsets P_d, P_s, and P_b according to the $P_{S_{p\to x}}$. Utilizing the ID3 algorithm, partition is recursively processed by having selected the x which yields the most information about whether the candidate pixel is a corner. The information gain of choosing an x is measured by the entropy of K_p, a Boolean variable which is true if p is a corner and false otherwise, and the formula is

$$H(P) - H(P_d) - H(P_s) - H(P_b) ,$$

where $H(P)$ is

$$\begin{aligned} H(P) &= (c + \bar{c})log_2(c + \bar{c}) - clog_2 c - \bar{c}log_2\bar{c} \\ c &= \left|\{p|K_p \text{ is true}\}\right| \quad \text{(number of corners)} \\ \bar{c} &= \left|\{p|K_p \text{ is false}\}\right| \quad \text{(number of noncorners).} \end{aligned}$$

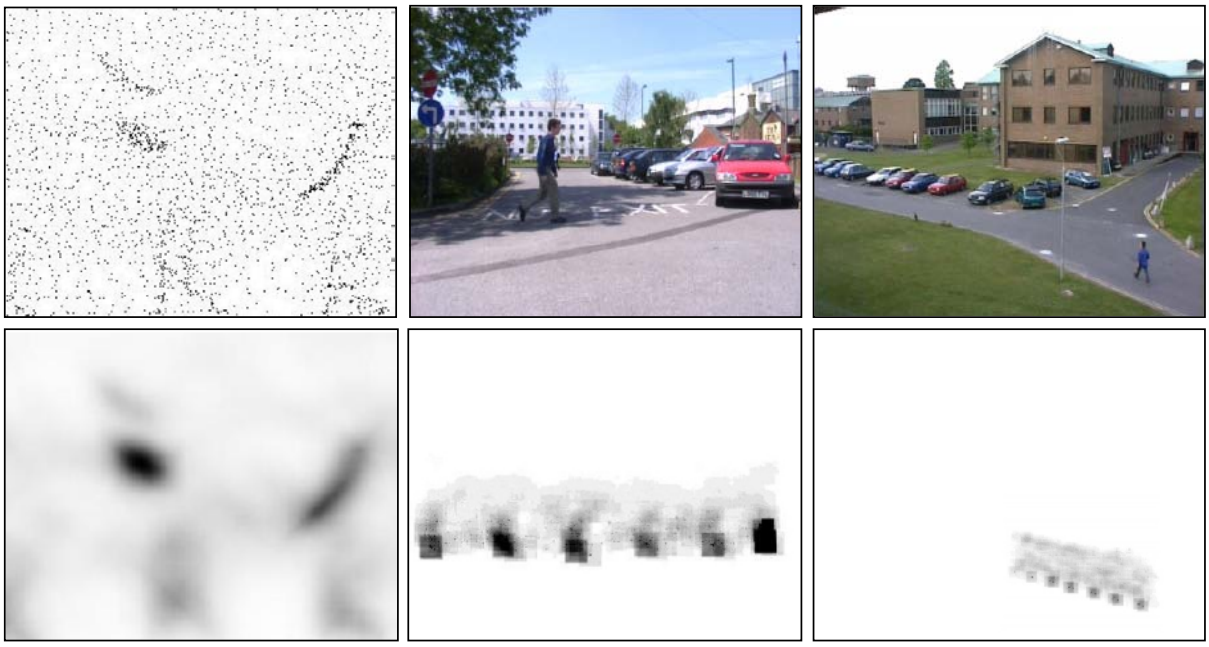
The recursive process terminates when the entropy of a subset is zero. The corner detector is created by the decision tree which can be further converted into programming codes as a long string of nested if-then-else statements to be compiled. Although the enhanced FAST is designed for speed, the performance has not been sacrificed. The reports have shown a high repeatability of FAST points, and the score is even better than DoG using in the SIFT except for the case under severe noise [27].

3.1.3 Stasis interest point

In human action recognition, most of previous works considered that spatio-temporal interest points have significant variance in temporal dimension [7] [8] [13]. Only emphasizing on moving regions, however, is insufficient to precisely describe an action. Instead, the static parts such as heel strikes of the subject, for example, can provide an accurate measure for gait periodicity as well as the gait stride and step length [5]. Specifically, during the strike phase, the foot of the striking leg stays at the same position for half a gait cycle, whilst the rest of the human body moves forward. The heel strike regions thus will produce a dense region if the interest points are detected and located at the heel through consecutive frames. To demonstrate this phenomenon, Figure 3.5 shows the dense regions enhanced through entire sequences by a corner proximity measure algorithm introduced by Bouchrika et al. [5].

Not only can the stasis interest points be applied to detect the static heel strikes for gait recognition, but to other actions are also possible. The hand clapping and waving actions, for example, have the static end-points at the hands' motion if the sequence has enough high frames per second (FPS). Besides, the stasis interest points detected in the background nearby the subject, or on the static part of the body, have the additional information for describing the circumstance where the action is performed. Specifically, human actions always have a implicit connection with the surrounding. Running, for example, needs a large space and the plane ground, but the space for picking motion is very small. Besides, while clapping makes hands usually cross the static chest, hand waving is just nearby the unmoved head. Recording the appearances at these different stasis interest points thus has significant meaning for describing the human action.

Such an idea can be employed in our part-based approach by detecting the stasis interest points through few frames. Our goal is to develop a precise spatio-temporal interest point detector which can produce a manageable number of interest points and involve the static and moving information. Any interest point detector such as DoG or FAST can be exploited here for extracting the base of interest points. The Lucas-Kanade optical flow is then employed to predict the motion of pixels [20]. Hence the final set of the interest points includes the moving interest points with sufficient optical flow and the stasis interest points nearby the action.



(a) (b) (c)

Figure 3.5: Example results for the corner proximity measure [5]. The input images are on the upside, and the corner proximity images are on the downside. (a) A test image; (b) heel strike extraction in sagittal indoor view; (c) heel strike extraction in oblique outdoor view.

3.1.4 The proposed interest point detector

Given $I(t)$ which represents the base set of the interest point detected by DoG or FAST in the frame t, we can extract two subsets which are $M(t)$, the set of interest points with sufficient optical flow m, and $S(t)$, the subset of stasis interest points s. Given the Lucas-Kanade optical flow predictor $LK(\cdot)$ [20] and the stasis function $stasis(\cdot, \cdot)$ which extracts the unmoving interest points between frames, the set of final spatio-temporal interest point $P(t)$ can be detected as follows:

$$P(t) = \{M(t) \cup S(t)\},$$

where

$$\begin{aligned} M(t) &= \{m \mid LK(I(t)) > th\}, \text{and} \\ S(t) &= \{s \mid dist(i,m) < r, i \in stasis(I(t), I(t-1)), m \in \{M(t), M(t-1)\}\}. \end{aligned}$$

The function $dist(\cdot, \cdot)$ is the Euclidean distance between two points in spatial domain. Two variables, th and r, are determined in advance according to the applied scenes. th is the threshold for the magnitude of optical flow and r is the search radius around a moving interest point.

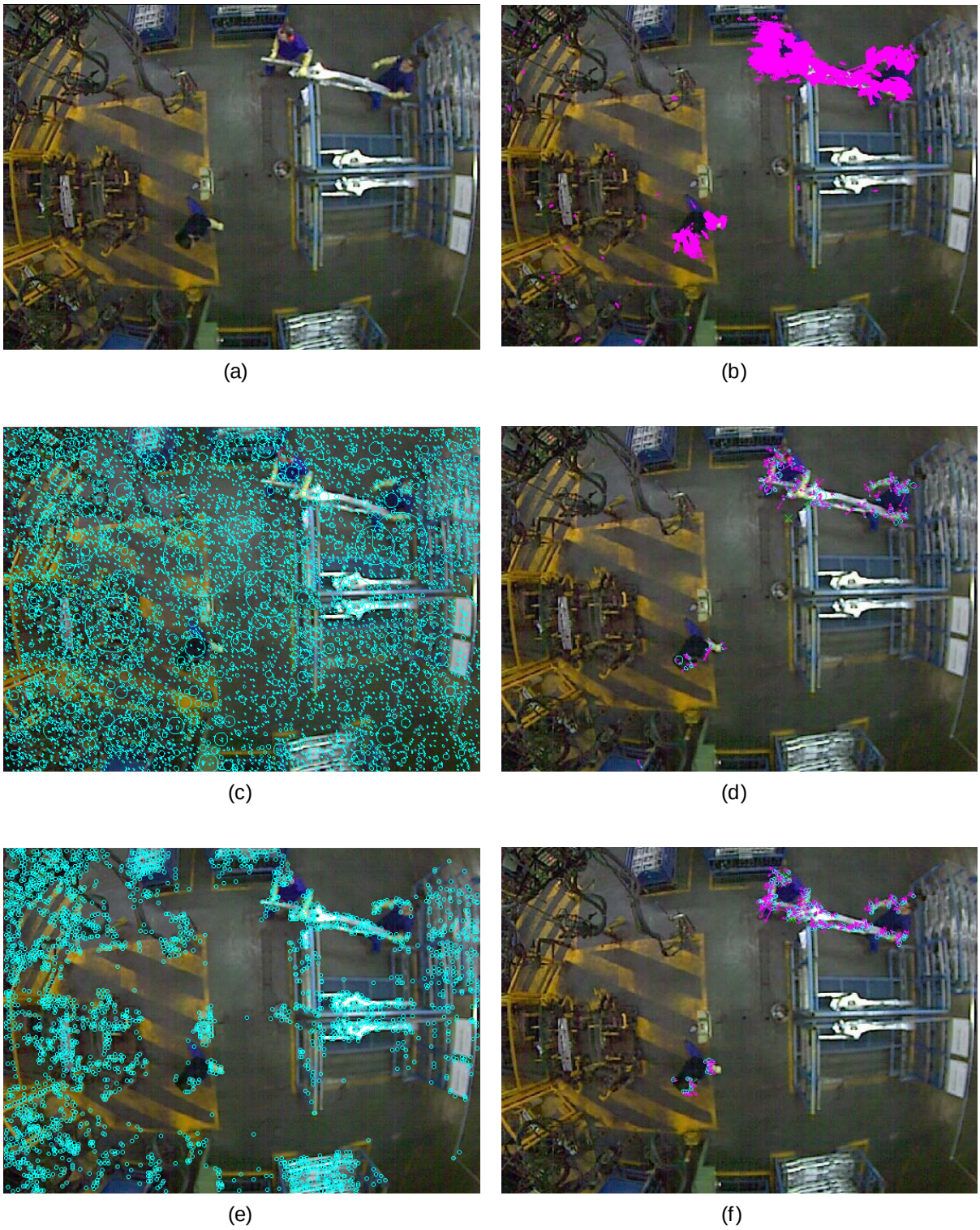
Figure 3.6 shows the examples of the proposed interest point detector using DoG and FAST. MoSIFT is simply replicated by [7], but the primary difference of our implementation is that the scale space is not expanded for motion detection. Instead, we rely on training dataset to reach the motion scale-invariance [23]. The empirical studies in Chapter 4 have shown a comparable performance between the different designs, and more importantly, computing optical flow in each expanded scale space is rather expensive. The detector using FAST, which is trained by the subset of the applied dataset in advance, is named Motion with FAST (MoFAST). Note while the results between the MoSIFT and the MoFAST show a completely different property, in both algorithms the action is precisely captured and the stasis points are located around the moving pixels.

The required computation of the proposed spatio-temporal interest point detector involves two synchronous processes: spatial interest point detection and temporal motion detection by optical flow. Such a distributed structure enables implementation of parallel computing for efficiency-demanded tasks. The minimum number of consecutive frames for the proposed detector is merely two, which provides an agile feature extraction, and would benefit potential real-time applications.

3.2 Interest point Descriptor

Various interest point descriptors have been devised to describe the appearance only in the spatial domain [8] [13]. More sophisticated descriptors have considered the change of appearance in spatio-temporal domain [12] [14] [31] [38]. Nevertheless, these researches focused on the appearance, and further augment spatio-temporal representations with histograms of optical flow (HoF) into the descriptor, the performance has shown a significant boost [7] [15] [28] [36]. To accurately encode the appearance and motion information of the detected interest points, as in [7], we simply extend the SIFT descriptor to form an additional 128-dimensional HoF.

Following briefly introduces the SIFT descriptor. Figure 3.7 shows the illustration of the SIFT descriptor. The magnitude and direction for the gradient are calculated for every pixel in a neighbouring region around the interest point in the scale space. To reach the rotation invariance, an orientation histogram with 36 bins is formed to equally cover the compass. Each sample in the neighbouring window is aggregated into the corresponding bin weighted



(a) (b)

(c) (d)

(e) (f)

Figure 3.6: The examples of the proposed spatio-temporal interest point detector with SCOVIS dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. (a) The original input frame; (b) the predicted optical flow calculated by (a) and the immediate frame; (c) the result after doing DoG. Note that the interest points detected in the different scale space are drawn as the circles with different radiuses; (d) the final interest points using MoSIFT incorporated with stasis interest points; (e) the result after doing FAST; (f) the final interest points using MoFAST incorporated with stasis points. The figure is best viewed in colour mode.

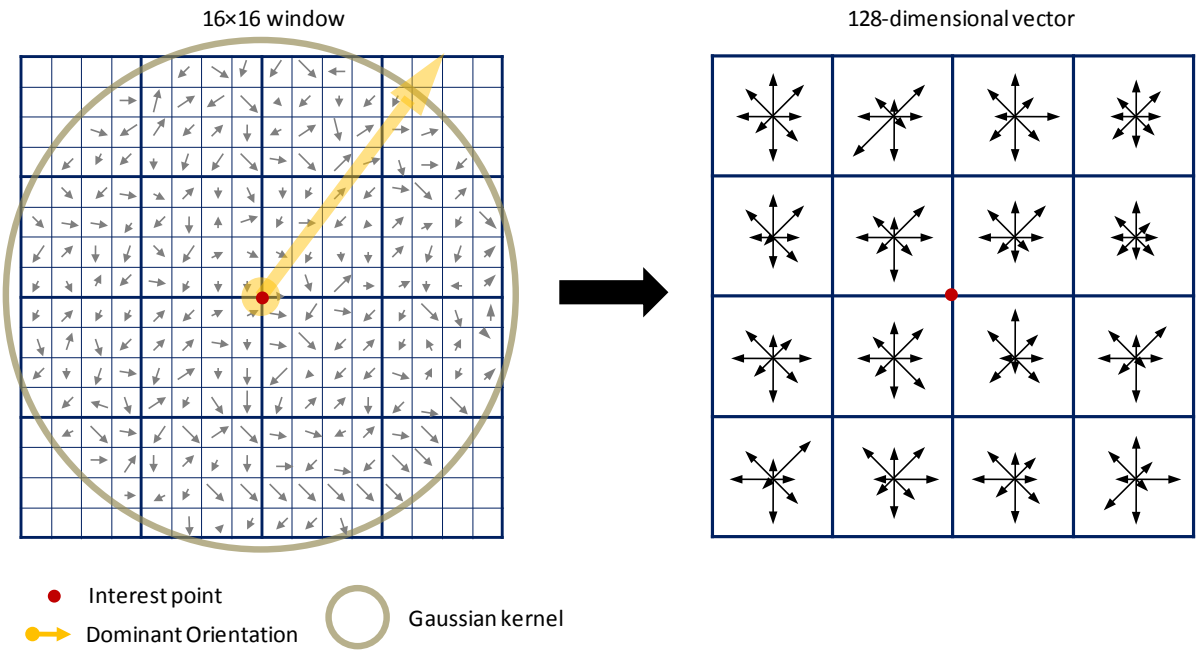


Figure 3.7: An interest point descriptor is constructed by first computing the gradient magnitude and orientation at each pixel in a region around the interest point location, as shown on the left. Before accumulating the orientation histograms, all orientations of the gradient are adjusted with the dominant orientation, and a Gaussian kernel convolves the region which is indicated by the overlaid circle. The final descriptor is formed by the 16×16 window with the contents of each 4x4 sub-region summarized by an orientation histogram with the length of each arrow corresponding to the sum of the gradient magnitude.

by its gradient magnitude that is multiplied with a Gaussian kernel with an σ which is 1.5 times the scale of the interest point. The peaks in the histogram thus represent the dominant orientations. In the case of multiple orientations being assigned, an additional interest point is generated with the same location and scale as the original interest point but with different orientation. To construct the SIFT descriptor, the direction of every pixel in a 16×16 windows around the interest point is adjusted to the dominant orientation, and the magnitude of each gradient in the region is convolved by a Gaussian kernel with an σ equal to one half the width of the descriptor window. By concatenating a local 8 bins histogram which equally divides the compass and is aggregated by each gradient within a 4×4 region, the whole 16×16 window forms a 128-dimensional vector (8×4×4=128) called histogram of gradient (HoG) which is the descriptor of the interest point. The vector is further normalized and clipped values larger than 0.2, and renormalized. For encoding the appearance of the interest point, with the normalization and clip procedures has shown a better resistant to the affine changes in illumination and the effects of non-linear illumination [19].

To encode the motion information, such aggregation, normalization, and clip schemes used in the SIFT can be exploited to aggregate HoF as well. We simply concatenate the

two128-dimensional HoG and HoF consecutively which generates a 256-dimensional descriptor. For stasis interest point, the motion vector is set with all zero. Note that the HoF is not designed for scale-invariance and rotation-invariance because the distinct magnitude and orientation of optical flow are considered representative in various actions.

3.3 Codebook formation and bag of words model

To exploit the bag of words model, the video codebook is constructed by standard *k*-means algorithm to obtain the vocabulary. The centre of each resulting cluster is regarded as a spatio-temporal word. Each detected interest point therefore can be assigned a spatio-temporal word, and is then collected to form the histogram of spatioal-temporal words for each input video. Although prior studies have shown the significance of the codebook design [1] [7] [17] [41], we reveal that using the naïve clustering algorithm with the representative visual descriptors and the effective classification model, it is possible reach or even beyond the performance of the state of the art. The size of the codebook is another critical factor for the recognition accuracy. A small codebook size is hard to capture the diversity of the spatio-temporal descriptor, while a large codebook size is suffered from the curse of dimensionality which produces a sparse distribution of the spatio-temporal words in the high-dimensional space, and decreases the efficiency and the accuracy of the final classification.

3.4 Classification

The classification problem can be divided into two stages which are inference stage and decision stage. At the inference stage, to construct the classifier, the classification model is learned from the training dataset which is formed by a subset of various sequence samples represented by the vectors as the histogram of spatio-temporal words. At the subsequent decision stage, we simply input an unknown data, and the classifier gives the predicted class. Two classic classification models, discriminative model [1] [7] [8] [12] [15] [28] [31] [38] [41] and generative model [23] [36] [40] [39], have been harassed in human action recognition with the bag of words model. Discriminative model generates a discriminative function by the training dataset, and later addresses the classified decision directly given an unknown data, in which probabilities play no role. On the other hand, generative model is based on Bayes' theorem which attempts to find the posterior probabilities of classes given training samples, and makes the classification decision for an unknown data by the learned posterior. For investigating the effect of learning model, we employ Support Vector Machine

(SVM) [35] as the discriminative model and Probabilistic Latent Semantic Analysis (*p*LSA) [11] as the generative model in our part-based approach. Following subsections give the briefly introductions.

3.4.1 Discriminative model

Here we introduce the non-linear SVM model which is utilized in our classification task. The fundamental theory of SVM can be further referred to [35]. Given the training dataset of the pairs of instance-label ($\mathbf{x}_i$, y_i), i = 1, …, n where $\mathbf{x}_i \in \mathbf{R}^p$ (i.e, the p-dimensional vectors) and $y_i \in \{1, -1\}_{i=1}^n$, the non-linear SVM with soft margin is an optimization problem as:

$$\min_{\mathrm{w},b,\varphi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i\right),$$

which is subjected to

$$y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0.$$

C is a positive penalty parameter for adjusting the soft margin, and can be decided by cross-validation. The kernel is defined by $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$, and $\Phi(\cdot)$ is the function for mapping the training vector $\mathbf{x}_i$ into a higher dimensional space for seeking a linear separable hyper-plane.

A considerable number of researches have utilized the SVM classifier with the part-based approach proposed by Schuldt et al. [30] which has shown that the SVM generally has a better performance than the naïve algorithm, k-nearest neighbour search. Following the high classification accuracy in general cases evaluated by [42], x^2 kernel is utilized in most of works for the non-linear classification such as to classify the histograms of spatial-temporal words [1] [2] [7] [8] [12] [15] [28] [31] [38] [41]. The x^2 kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{1}{A}D(\mathbf{x}_i, \mathbf{x}_j)\right),$$

where

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}\sum_{i=1}^p \frac{(u_i - v_i)^2}{u_i + v_i},$$

with $\mathbf{x}_i = (u_1, \ldots, u_p)$ and $\mathbf{x}_j = (v_1, \ldots, v_p)$. The only scale parameter A can be decided by cross-validation.

Another common kernel function is radial basis function (RBF) which is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \text{ and } \gamma > 0.$$

The parameter γ is also decided by the cross-validation. While incorporating with the MoSIFT descriptor has been shown an impressive recognition accuracy using the SVM with the x^2 kernel [7], our empirical study reveals that if the grid search algorithm [6] in both the x^2 and RBF kernel for the heuristic parameters such as *C*, *A*, and γ, RBF kernel performs even better. Regarding the multiple-classes learning, we adopt the one-versus-all strategy.

3.4.2 Generative model

According to the original definition of *p*LSA model [11], in the human action recognition the corresponding terms are mapped as 'documents' to 'sequences', 'words' to 'spatio-temporal words' (i.e. the clusters of the spatio-temporal descriptors), and 'topics' to 'actions'. To exploit the *p*LSA model, *N* video sequences containing spatio-temporal words from a vocabulary of size *M* (i.e. the codebook size) is represented by $d_j = (w_1, \ldots, w_M)$ with j = 1…*N*, and a *M*×*N* co-occurrence matrix **C** can be built by $n(w_i, d_j)$ with i = 1, …, *M* which records the number of occurrence of the spatio-temporal word w_i in the sequence d_j. An assumption is made by the model existing a latent topic variable z_k with k = 1, …, *K* which is associated with each $n(w_i, d_j)$, and each latent topic z_k corresponds to an action category in our scenario.

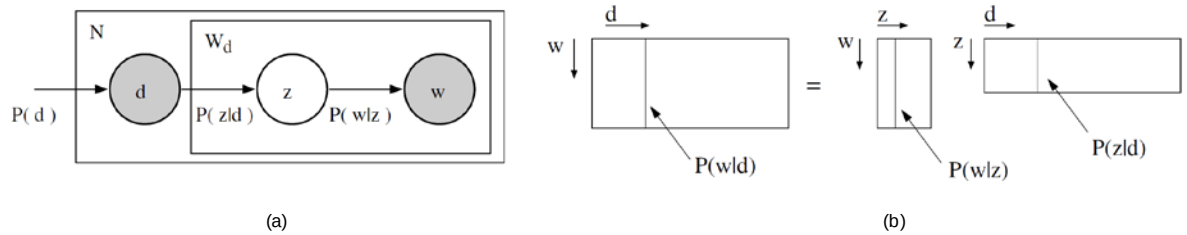
*p*LSA is based on the graphic model as shown in Figure 3.8 [32] which takes an assumption of the joint probability $P(w_i, d_j, z_k)$ as:

$$P(w_i, d_j) = P(w_i|d_j)\, P(d_j).$$

We assume that each pair of (w_i, d_j) is generated independently. Therefore the marginalized conditional probability over optics z_k can be formed as:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(w_i|z_k) P(z_k|d_j),$$

where $P(z_k|d_j)$ is the probability of topic z_k occurring in video d_j, and $P(w_i|z_k)$ is the probability of spatio-temporal word w_i occurring in a particular topic z_k. In the training stage, in order to find the highest probability $P(w_i|d_j)$ and the most explainable parameters



(a) (b)

Figure 3.8: (a) pLSA graphical model. Nodes inside a given box (plate notation) indicate that they are replicated the number of times indicated in the top left corner. Filled circles indicate observed random variables; unfilled is the latent variable. (b) In pLSA the goal is to find the topic specific word distributions P($w|z_k$) and corresponding document specific mixing proportions P($z|d_j$) which make up the document specific word distribution P($w|d_j$). The figure is reproduced from [32].

$P(w_i|z_k)$ and $P(z_k|d_j)$, an expectation-maximization (EM) algorithm is utilized with the following objective function:

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}.$$

In the test stage, using EM again, the fitted *p*LSA model then gives the most possible $P(w|d_{test})$ with the learned $P(w_i|z_k)$ to address the probability $P(z|d_{test})$ in which the maximum probability is the predicted topic.

Niebles et al. first devised an unsupervised learning model for human action recognition with part-based approach using generative model [23]. According to the report, the *p*LSA model surprisingly performs better than the more sophisticated model, Latent Dirichlet Allocation (LDA). Although the unsupervised learning does not require labels for instance data, the model still depends on ground truth testing data for deciding the predicted action. Nevertheless, using the similar interest points detector and descriptor, the *p*LSA is outperforms the discriminative classification model investigated in [8]. The other advantage of using generative model is to realize object localization straightforwardly by assigning the probability $P(z_k|w_i, d_j)$ for each spatio-temporal word according to the learned model,

$$P(z_k|w_i, d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)},$$

unlike other works rely on performing additional procedures for the localization task [22] [2]. Specifically, given the probability $P(z_k|w_i, d_j)$ for each spatial-temporal word, we can further analyze the geometric information to perform object localization and multi-objects tracking [23].

3.5 Summary

In this chapter, we have described each step shown in Figure 3.1. In the Section 3.1, two different spatio-temporal interest point detectors, MoSIFT and MoFAST, with the novel stasis interest point detector have been introduced. The effect of stasis interest point is demonstrated in the next chapter. The effective descriptor is described in Section 3.2 where a 256-dimensional vector is formed by equally encoding the appearance and the motion information around the detected spatio-temporal interest point. In Section 3.3, the naïve clustering algorithm, *k*-means, is employed in our framework to build the codebook (i.e. the spatio-temporal words) because of its simplicity and efficiency. Section 3.4 describes two different classification models which are investigated in the next chapter.

Chapter 4

Empirical Studies

In this chapter, the standard datasets, KTH and Weizmann, are separately utilized to evaluate the proposed interest point detector in server aspects including the efficiency and accuracy, using the SVM with the radial basis kernel first, and the issue of classifiers regarding SVM and *p*LSA are later investigated. The real industry surveillance video is employed to demonstrate the effectiveness of our algorithm. All experiments are run on a duo-core 2.67 GHz CPU with 4 GB physical memory, and the programs are implemented in MATLAB. We set the two parameters $th = 1$ and $r = 5$ for the interest point detector in all dataset. The FAST corner detector is fast trained by a subset of the given dataset.

4.1 Evaluation on KTH data

Since the KTH dataset is very large, we adopt the methodology as the experiment in [23]. To verify the adaptability of the proposed part-based approach in camera motion, the sequences are not pre-processed for stabilization which is used in [36], and are not divided into action sub-sequences in advance which is used in most of literatures based on [30]. In other words, there is no any pre-processing on KTH data in our experiments. To construct the codebook, ideally we should exploit all spatio-temporal feature descriptors in all training sequences. However, since the total number of descriptors is very large as shown in Table 4.1, we construct the codebook using only two sequences of each action from three subjects to fit the memory requirement and the reasonable processing time. These used sequences are later taken out of the training and testing sets. We have found that such a methodology is stricter than using random sampling among descriptors in whole dataset that has been used by [7].

The evaluation is adopted the common leave-one-out testing paradigm. Specifically, in each run, the model is learned from the sequences of 24 subjects without the sequences used to construct the codebook, and then is tested by the remaining subject. The results are

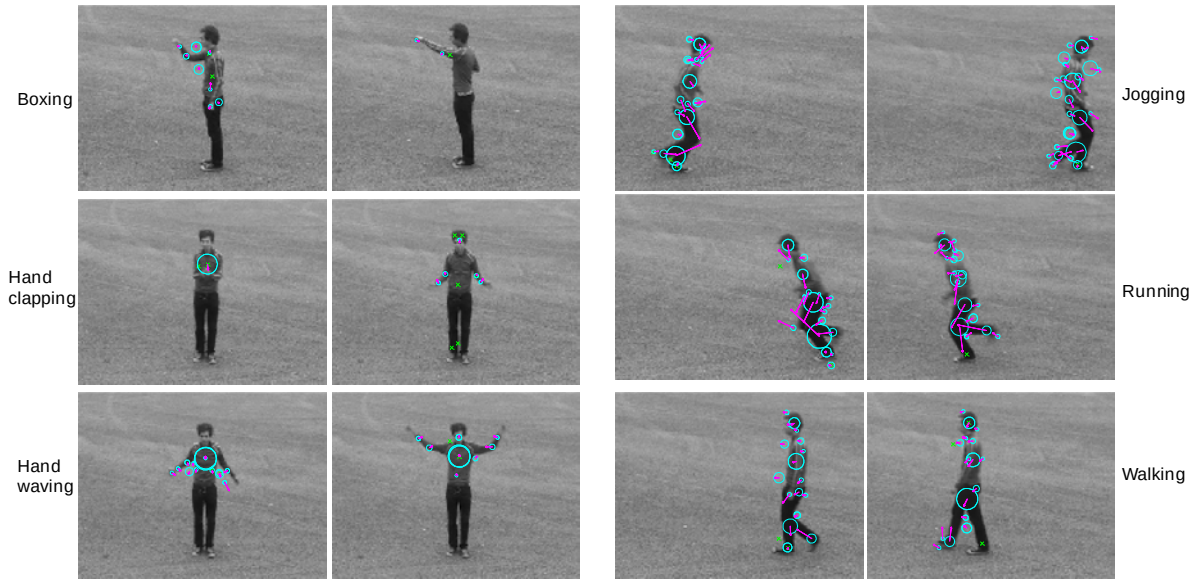


Figure 4.1: MoSIFT with stasis interest point performed in KTH dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. Note that the interest points detected in the different scale space are drawn as the circles with different radiuses. The figure is best viewed in colour mode.

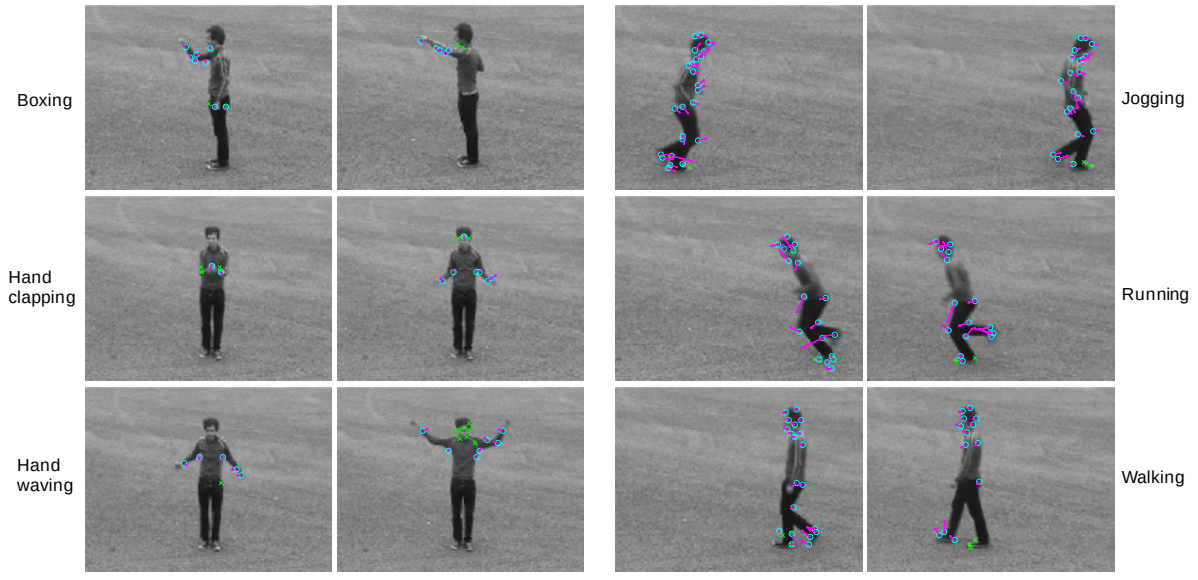


Figure 4.2: MoFAST with stasis interest point performed in KTH dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. The figure is best viewed in colour mode.

calculated on the average of 25 runs. The parameter γ and C for the SVM with radial basis kernel are decided by 24 cross-validation with the grid search algorithm [6]. Table 4.3 shows the average detecting speed and the number of detected spatio-temporal interest points. The number of descriptors produced by MoFAST is clearly much larger in KTH data, and MoFAST is also more efficient than MoSIFT.

Table 4.1: The detection speed and the number of detected interest points in KTH data.

Detector	Extracting time per frame (*ms*)	# of stasis points	# of descriptors in total
MoSIFT	78	527,244	3,262,718
MoFAST	20	1,428,633	4,840,286

Figure 4.1 and Figure 4.2 show the examples of MoSIFT and MoFAST with stasis interest point detector performed in the frames of different actions on KTH data. The interest points have been determined in the relevant places, as expected. Figure 4.3 shows the distribution of stasis interest points using in the MoSIFT and MoFAST. Note that MoFAST detects much more spatio-temporal interest points as well as the stasis points than MoSIFT. Figure 4.4(a) shows the effect of codebook size in different detectors. MoFAST in general does not perform better than MoSIFT. The best classification result occurs on the 600 codebook size by the MoSIFT with stasis interest point detector, which has 94.67% accuracy on average.

Figure 4.4(b) shows the confusion matrix of the best result. The rows are ground truth testing data, and the columns are the predicted actions. The confusion happens between ‘jogging’ and ‘running’ the most, which is expected as these two actions are very similar. The other confused part is the hand actions. Interestingly, some components are indeed similar between ‘boxing’ and ‘waving’ as well as ‘boxing’ and ‘walking’.

To demonstrate the effect of using different strategies for constructing the codebook, we reproduce the methodology utilized in [7], in which the codebook is trained by 1% random sampling from all descriptors in the whole dataset. As shown in Figure 4.5, using such a scheme, the classification accuracies are boosted in both the results using MoSIFT and MoFAST. The issue is considered a data contamination [23] since the descriptors for constructing the codebook should not include descriptors from testing set.

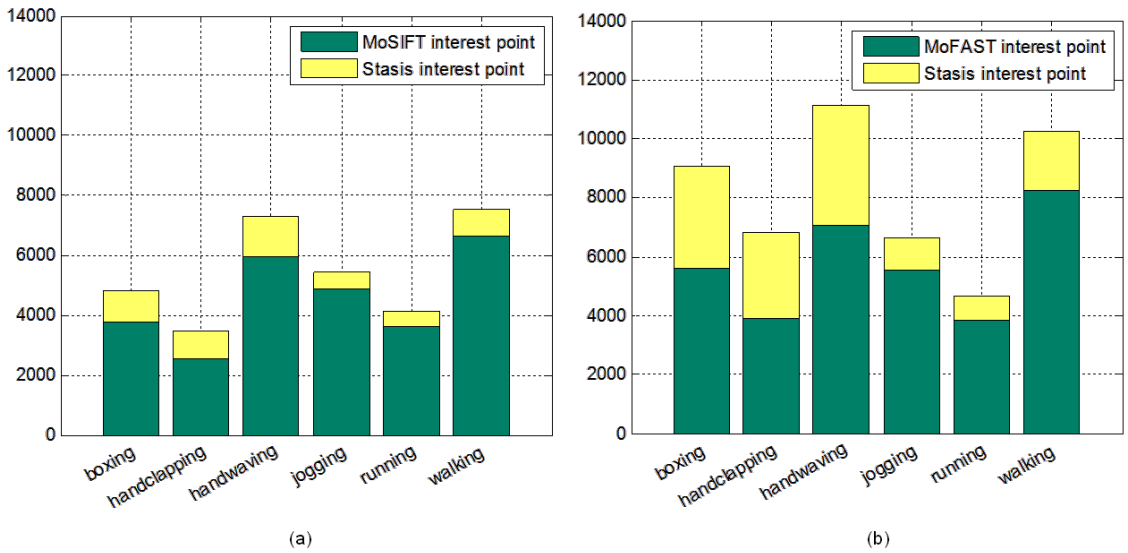


Figure 4.3: (a) The average number of stasis interest points and MoSIFT interest points detected in the KTH dataset in each sequence. (b) The average number of stasis interest points and MoFAST interest points detected in the KTH dataset in each sequence.

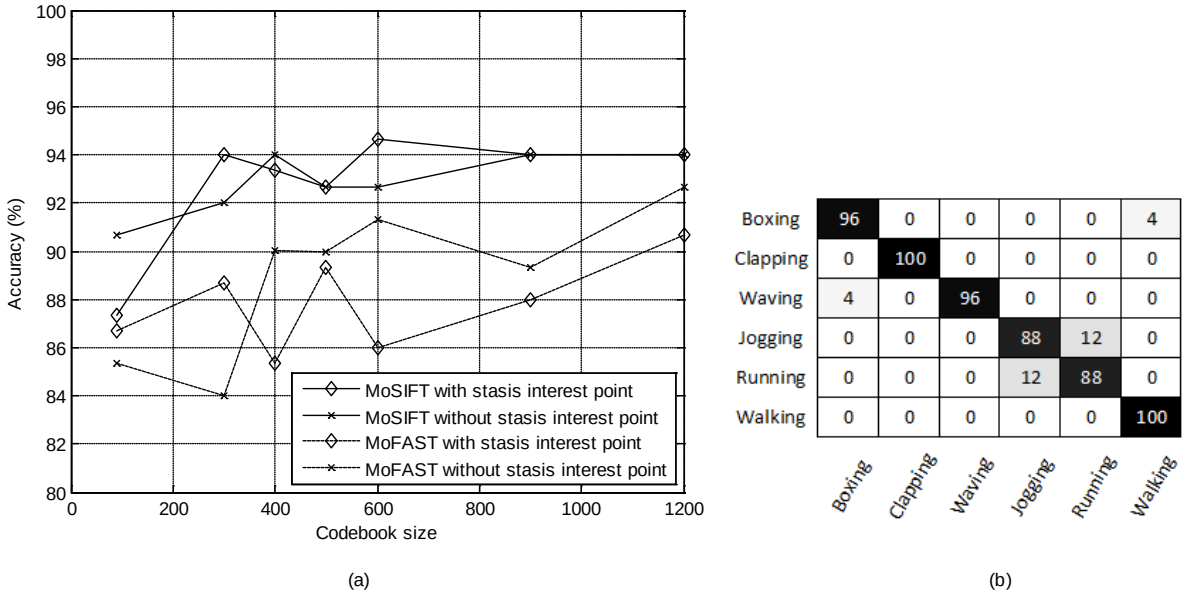


Figure 4.4: (a) The classification accuracy of different algorithms on KTH data. (b) The corresponding confusion matrix using the best algorithm which is the MoSIFT with stasis point with 600 codebook size. The best model gives a 94.67% classification accuracy.

We test different classifiers on KTH data using the best interest point detector (i.e. MoSIFT with stasis interest point) with codebook size 600. The results and the comparison with previous works based on paritial-based approaches are shown in Table 4.2. The SVM with x^2 kernel which has the parameter A determined by grid search algorithm [6] has a worse performance than the SVM with radial basis kernel. The *p*LSA model outperforms the state of the art approaches with the proposed spatio-temporal interest point detector. Note that even

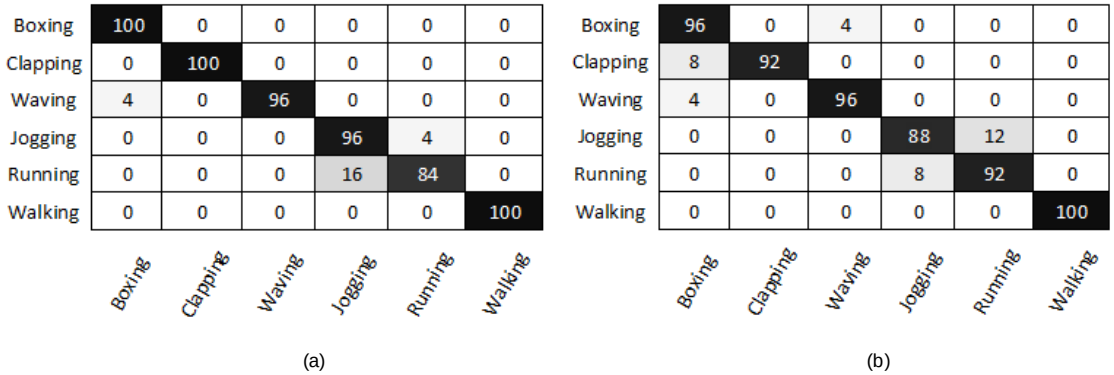


Figure 4.5: The corresponding confusion matrix produced by the model using 1% randomly sampling for constructing codebook with size 500. (a) The result of using MoSIFT detector has classification accuracy 96%. (b) The result of using MoFAST detector has classification accuracy 94%.

if our approach has better performance, we cannot directly compare the result with the approach by Ballan et al. [1] since in contrast to the common framework, they use a more sophisticated clustering algorithm to build the codebook, as well as the works by Blank et al. [3] and Liu [18] which exploit holistic representations that are not comparable to the proposed part-based approach.

Table 4.2: Comparison of different classifiers and the previous works on KTH data. SP stands for the stasis points, RB stands for the radial basis kernel, and x^2 is the x^2 kernel.

Approach	generative model	Discriminative model	Accuracy
MoSIFT + SP + SVM + RB		✔	94.67%
MoSIFT + SP + SVM + x^2		✔	92.70%
MoSIFT + SP + *p*LSA	✔		**97.33%**
Chen [7]		✔	95.00%
Dollár et al. [8]		✔	81.20%
Kläser et al. [12]		✔	91.4%
Laptev et al. [15]		✔	91.80%
Niebles et al. [23]	✔		83.33%
Schuldt et al. [30]		✔	71.70%
Wang et al. [36]	✔		92.43%
Willems et al. [38]		✔	84.26%
Wong and Cipolla [39]		✔	86.62%

4.2 Evaluation on Weizmann data

We adopt a general setting for the experiment which is a leave-one-out scheme to measure the recognition accuracy. Specifically, in each run we use eight subjects as training data, and test the learning model with the remaining subject. The result takes the average of total nine runs. The speed and the detected number of the interest point are shown in Table 4.3. Since the large number of interest points, we randomly sample 30,000 interest points for constructing the codebook for fitting the reasonable processing time. The parameter γ and C for the SVM with radial basis kernel are decided by eight cross-validation with the grid search algorithm [6]. Table 4.3 shows the average detecting speed and the number of detected spatio-temporal interest points. In Weizmann data, the number of descriptors generated by MoFAST is smaller and the approach is still much more efficient.

Table 4.3: The detection speed and the number of detected interest points in Weizmann data.

Detector	Extracting time per frame (*ms*)	# of stasis points	# of descriptors in total
MoSIFT	110	14,369	105,325
MoFAST	22	8,974	67,391

Figure 4.6 and Figure 4.7 show the examples of MoSIFT and MoFAST with stasis interest point detector performed in the frames of different actions on Weizmann data. The interest points have indeed been determined in the correct places, as expected. Figure 4.8 shows the distribution of stasis interest points using in the MoSIFT and MoFAST. Note that MoFAST detects much fewer spatio-temporal interest points as well as the stasis points than MoSIFT. Figure 4.9(a) shows the effect of codebook size in different detectors. MoFAST in general does not perform better than MoSIFT. The result is expected when we investigate the distributions of the interest points by MoSIFT and MoFAST in Figure 4.8. The variance of the interest point distribution among different actions has a relationship to recognition accuracy. This is due to the bag of words representation in which each sequence is represented by the histogram of spatio-temporal words. The number of detected spatio-temporal interest point in different action thus can affect the final classification. The dependency happens on stasis interest points as well. According to Figure 4.9(a), the best

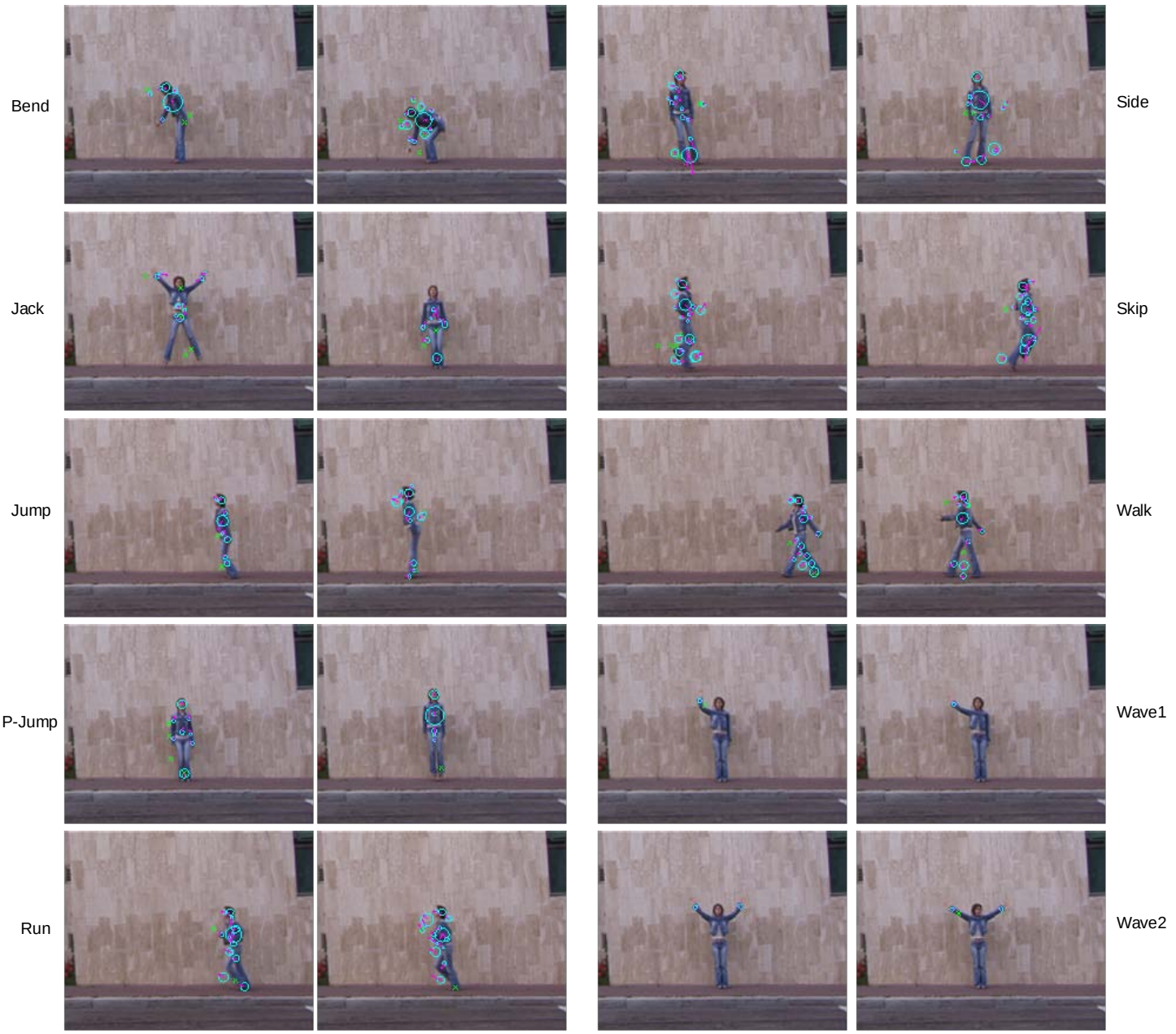


Figure 4.6: MoSIFT with stasis interest point performed in Weizmann dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. Note that the interest points detected in the different scale space are drawn as the circles with different radiuses. The figure is best viewed in colour mode.

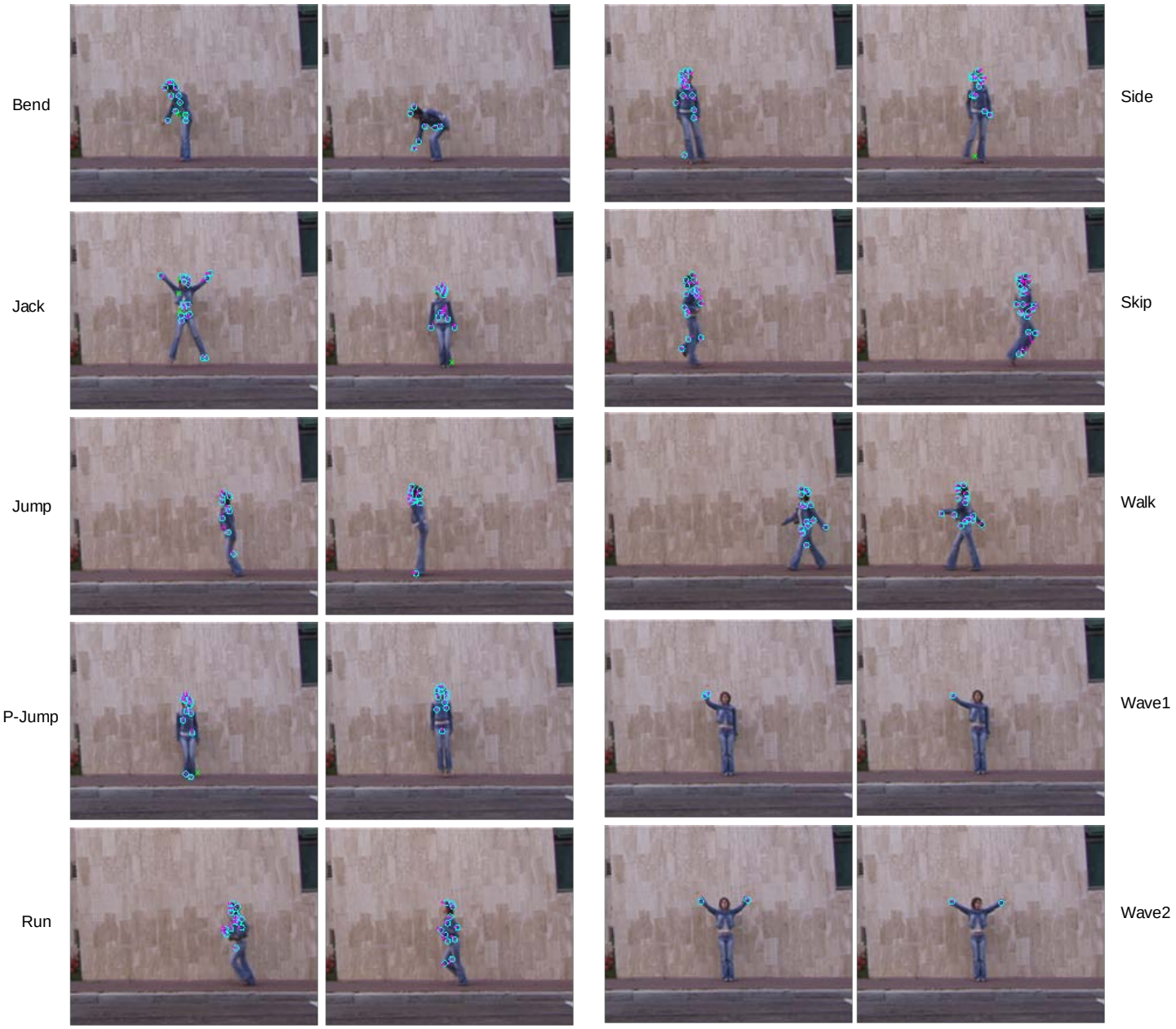


Figure 4.7: MoFAST with stasis interest point performed in Wiezmann dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. The figure is best viewed in colour mode.

classification result occurs on the 500 codebook size by the MoSIFT with stasis interest point detector, which has 92.3% accuracy on average.

The worse performance of the MoFAST probably is due to that we ignore the issue of scale-invariance in the appearance descriptor which may result in a coarse clustering to affect the ultimate classification result. On the other hand, MoSIFT has scare-invariance in the detection step through the use of DoG. Unlike on KTH dataset which has scale-variant training data (i.e. the second scenario) which can provide MoFAST additional scale-variant interest points, Weizmann has no such a scenario. This inference corresponds to that the MoFAST performs not too worse on KTH but comparatively worse on Weizmann. Thus we can infer that scale-invariance for representing appearance part is a rather important factor to design a spatio-temporal interest point detector or descriptor.

Figure 4.9(b) shows the confusion matrix of the best result. The rows are ground truth testing data, and the columns are the predicted actions. The most confused actions occur between ‘run’, ‘skip’ and ‘walk’. In Weizmann dataset, the action ‘run’ and ‘skip’ are fairly similar which results in having similar spatio-temporal interest points. This issue could be resolved by using a longer sequence to provide more frames to make them distinct in the number of spatio-temporal interest points. The other confusion surprisingly happens on ‘wave1’ and ‘wave2’. These two actions are completely distinct since one is with only one hand and the other one has two hands waving. Investigation shows that the motion threshold used for *th* is too high for the waving action. As shown in the Figure 4.8, ‘wave1’ and ‘wave2’ are detected with rather few spatio-temporal interest points whether for MoSIFT interest points or for stasis interest points. This result supports that a suitable *th* is significant to the final classification especially when the diversity of actions increases.

We test different classifiers on Weizmann data using the best interest point detector (i.e. MoSIFT with stasis interest point) with codebook size 500. The results and the comparison with previous works based on paritial-based approaches are shown in Table 4.4. The SVM with x^2 kernel which has the parameter A determined by grid search algorithm [6] has a worse performance than the SVM with radial basis kernel. The *p*LSA works better than the SVM with x^2 kernel but compared with the performance on KTH data, is undermined by the increase of action categories. The proposed approach generally outperforms the state of the art. Note that we cannot directly compare the result with the approach by Ballan et al. [1] and Liu et al. [17] since in contrast to the common framework, they use a more sophisticated

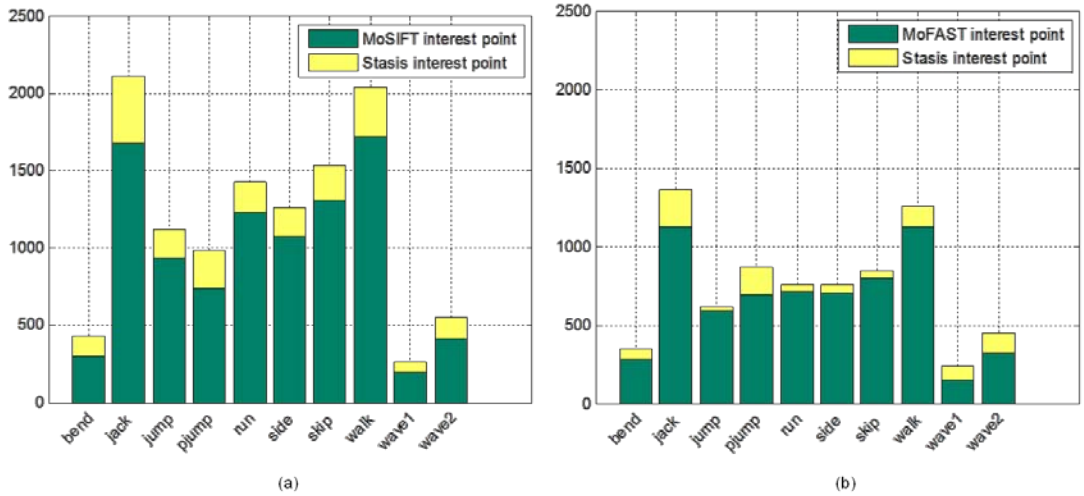


Figure 4.8: (a) The average number of stasis interest points and MoSIFT interest points detected in the Weizmann dataset in each sequence. (b) The average number of stasis interest points and MoFAST interest points detected in the Weizmann dataset in each sequence.

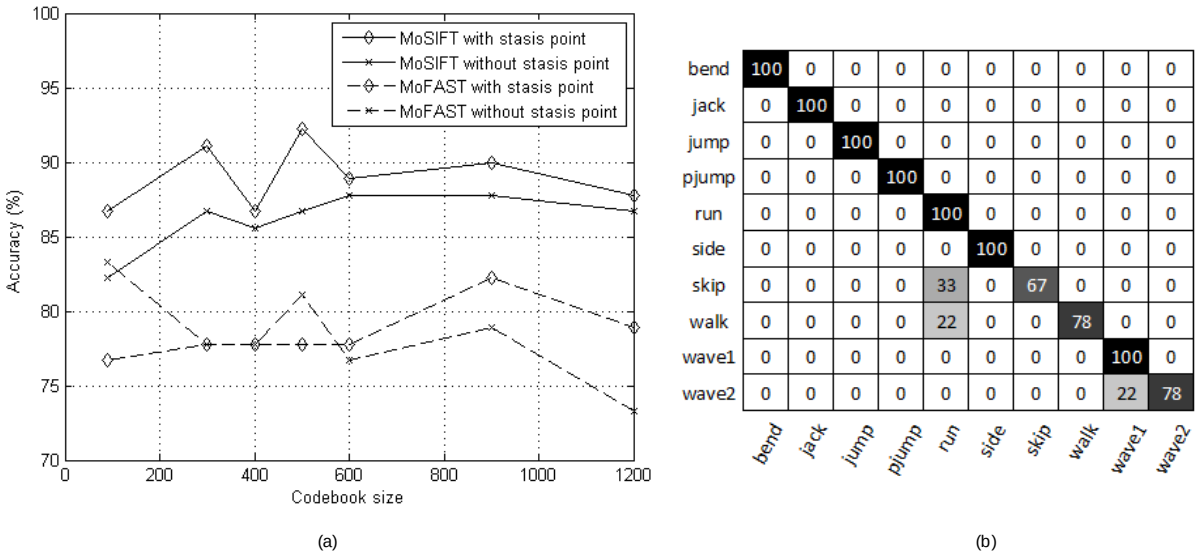


Figure 4.9: (a) The classification accuracy of different algorithms on Weizmann data. (b) The corresponding confusion matrix using the best algorithm which is the MoSIFT with stasis point with 500 codebook size. The best model gives a 92.3% classification accuracy.

Table 4.4: Comparison of different classifiers and the previous works on Weizmann data. SP stands for the stasis points, RB stands for the radial basis kernel, and x^2 is the x^2 kernel.

Approach	generative model	Discriminative model	Accuracy
MoSIFT + SP + SVM + RB		✔	**92.3%**
MoSIFT + SP + SVM + x^2		✔	82.22%
MoSIFT + SP + *p*LSA	✔		85.56%
Kläser et al. [12]		✔	84.30%
Niebles et al. [23]	✔		90.00%
Scovanner et al. [31]		✔	82.60%

clustering algorithm to build the codebook, as well as the works by Blank et al. [3] and Liu [18] which exploit holistic representations that are not comparable to the proposed part-based approach.

4.3 Evaluation on industrial environment data

In this experiment, we extract the surveillance video from the daily routine in the factory, and label the actions by hand. Four actions are of interest which are walking, carrying, handling, and stand still. We then divide the sequences into eight segments for each four action, and each sequence contains ten frames. The evaluation is again adopted leave-one-out scheme which utilizes seven sequences as the training dataset, and tests the learning model by the remaining sequence. The result takes the average of total eight runs.

The experiment is designed intentionally for recognizing actions using rather few consecutive frames. We can thus obtain a small number of interest points to enable the codebook being trained by all spatio-temporal interest point descriptors to observe the performance of different approaches. Besides, compared with the KTH and Weizmann data, the backgrounds in the industry environments are rather cluttered as shown in Figure 1.3, and partial occlusion happens frequently as well. The dataset is a very good testing bed to evaluate the proposed stasis interest point in such a complex scene. According to the previous experiments, we choose MoSIFT as the base interest point detector since it has the best performance in general case. Figure 4.10 shows the examples of MoSIFT with stasis interest point detection performed on the industrial environment data. Note that noises are also possible to be detected as the interest point and hence produce stasis interest points. We currently omitted this issue and rely on the sufficient amount of interest points detected around the main subject.

Figure 4.11(a) shows the results of different setting with MoSIFT interest point detector, and Figure 4.11(b) is the confusion matrix of the best combination in which the rows are ground truth testing data, and the columns are the predicted actions. Clearly that the detectors without stasis interest point detection generally perform worse. People actions in a complex industry environment are under a large constraint in which motion magnitudes are rather small and the range for activity is narrow. In this situation, conventional detectors consider

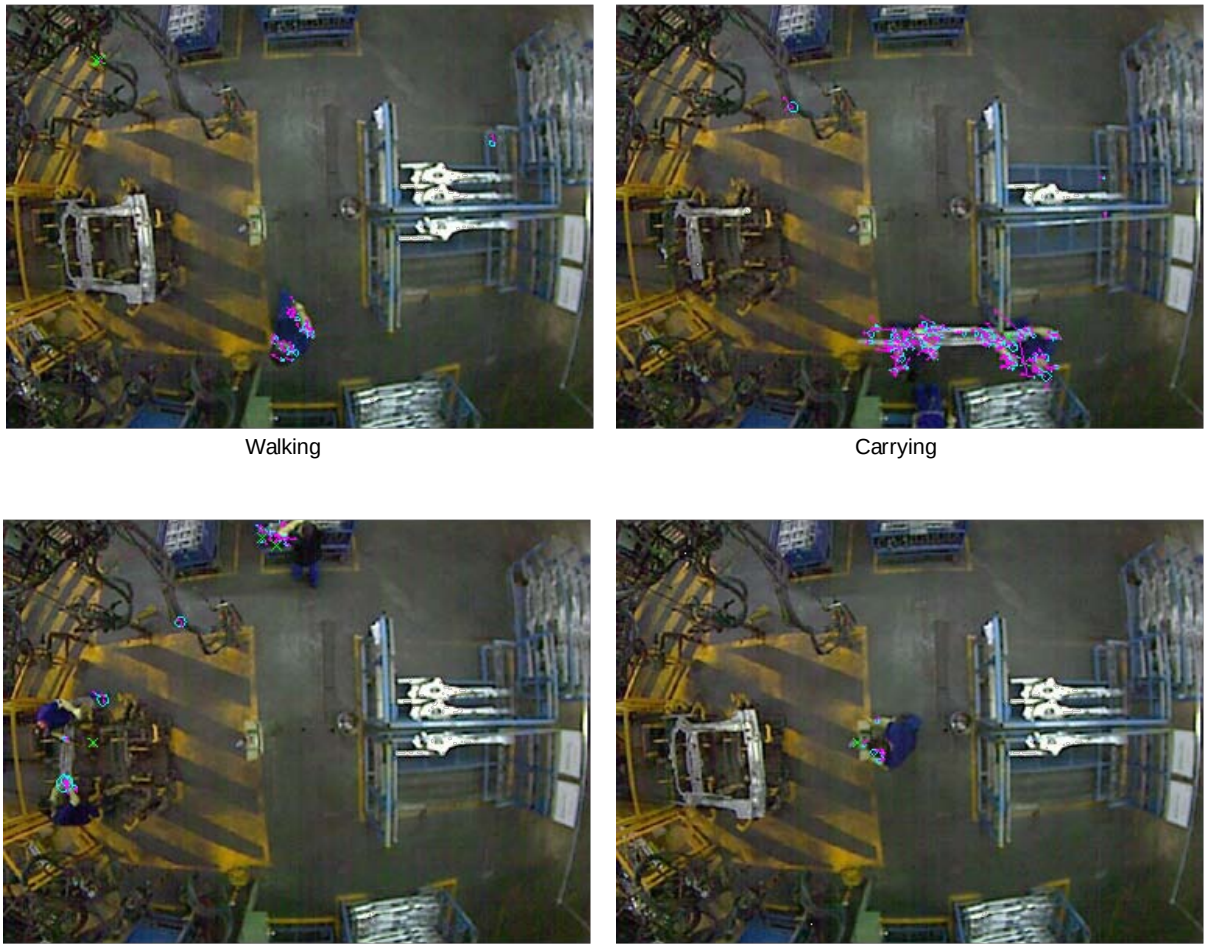
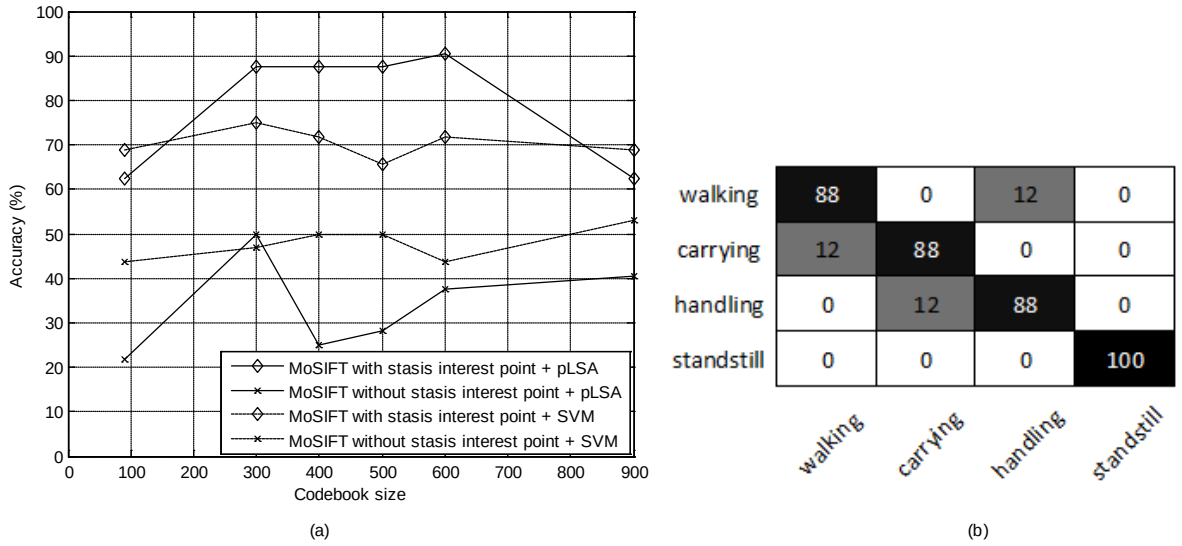


Figure 4.10: MoSIFT with stasis interest point performed on the industrial dataset. The length of the magenta arrows represents the magnitude of the optical flow. The cyan circle is the location of the interest points. The stasis interest points are drawn by green cross. Note that the interest points detected in the different scale space are drawn as the circles with different radiuses. The figure is best viewed in colour mode.

only motion information that is not enough. Stasis interest points, instead, provides additional information regarding the appearance of static surroundings for classifying actions. For example, ‘carrying’ should always accompany with some metal stuff, and ‘working’ usually happens in an open area. However, confusing motions indeed exist in these actions as shown in Figure 4.11(b) since either action sometimes happens with the other one together. Specifically, ‘walking’ is performed through a metal shelf or ‘carrying’ passes through an open area. This issue can be solved by training more sequence samples to make actions more distinct under a controlled environment.

The confused cases of handling reveals that multi-subjects appearing in the same frame will significantly confuse the classifier when the sequence contains only few frames. As shown in Figure 4.11(b), subjects ‘walking’ though the other subjects who are performing



(a) (b)

Figure 4.11: (a) The classification accuracy of different algorithms on industrial environment data. (b) The corresponding confusion matrix using the best algorithm which is the MoSIFT with stasis point with 500 codebook size. The best model gives a 91% classification accuracy.

'handling', or two subjects are performing 'carrying' together while one of them stops and puts stuff down like 'handling' first. The part-based approaches usually assume only one subject is performing actions in the sequence. How to separate multi-subjects to training the learning model requires additional researches. On the other hand, 'stand still' is detected very well even if the motion magnitude is very small. Once there is a MoSIFT interest point being detected, finding the stasis interest points around it is possible. Hence the almost static ''stand still' can have a good description by MoSIFT with stasis interest point detection.

The empirical studies in using different classifiers show that the *p*LSA performs better than the SVM when the number of action types are rather few (i.e. four types in this experiment) as shown in Figure 4.11(a). *p*LSA provides another function which is object localization as shown in Figure 4.12. This is achieved by finding the maximum marginal probability $P(z_k|w_i, d_j)$ for each spatio-temporal word, and the spatio-temporal interest points in the frames are then mapping into the corresponding codebook words to obtain the prediction. Figure 4.12 also shows that confusions are caused by the actions consisting of the different categories of spatio-temporal interest point. Nevertheless, further analyzing the probability of spatio-temporal words helps realizing the multi-subjects with multi-actions detection, if the different actions are separated far enough and there are dominant interest points with the same category for each action.

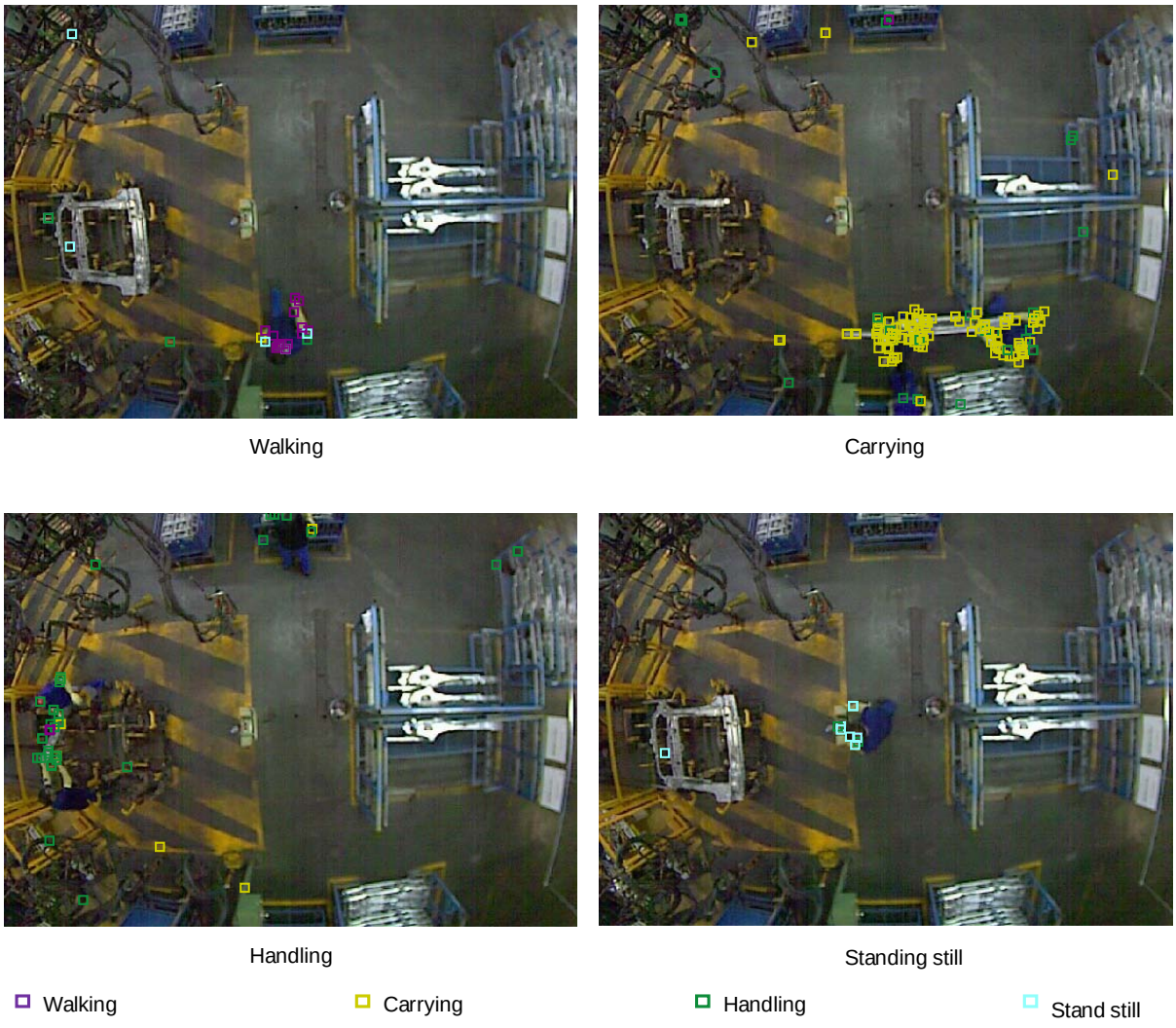


Figure 4.12: The predicted categories for each spatio-temporal interest point using *p*LSA. The figure is best viewed in colour mode.

Chapter 5

Conclusions and Further Works

In this work, robust human behaviour analysis algorithms based on part-based approach have been studied. We have particularly emphasized the action detection and classification part. A more precise spatio-temporal interest point detector has been developed by combining a spatial interest point detector, optical flow, and the proposed stasis interest point detector. We have investigated the performance of the collaborations between stasis interest points, MoSIFT, and MoFAST. The empirical studies have demonstrated that adding the stasis interest points into the action description indeed makes an improvement of classification accuracy. On the other hand, most of previous researches study part-based approaches by using SVM with x^2 kernel for classification directly. The evaluations for radial basis kernel and x^2 kernel show that the non-linear kernel classification using in human action recognition still has room to be improved. The other classifier is *p*LSA based on the generative model. Our results have shown that a generative model outperforms a discriminative model such as SVM is possible. How to select the suitable classier for particular human actions is still an open question.

The work will be extended to a more profound investigation. Although preliminary results show that stasis interest points work well on the KTH and Weizmann data, more evaluations shall be conducted, especially for cases in more cluttered background and with more diverse actions. More sophisticated classifiers will be tested with the proposed algorithms, such as the Latent Dirichlet Allocation (LDA) generative model [23] and the EMD kernel for SVM [42]. The codebook designed has also a significant influence to the final action recognition [1] [17]. Suitable algorithms for the proposed spatio-temporal interest points to construct the codebook are worth to be researched further.

References

[1] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra, "Effective Codebooks for Human Action Categorization," in *Proceedings of the 12th International Conference on Computer Vision Worksohps*, Kyoto, Japan, 2009, pp. 506-513.

[2] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra, "Human Action Localization and Recognition using Spatio-Temporal Interest Points and Tracking," *Submitted to Special Issue on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis in Expert Systems: The Journal of Knowledge Engineering*, 2010.

[3] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as Space-Time Shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 1395 - 1402.

[4] Aaron F. Bobick and James W. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257 - 267, March 2001.

[5] Imed Bouchrika and Mark S. Nixon, "Model-Based Feature Extraction for Gait Analysis and Recognition," in *Proceedings of Proceedings of the 3rd International Conference on Computer Vision/Computer Graphics Collaboration Techniques*, Rocquencourt, France, 2007, pp. 150-160.

[6] Chih-Chung Chang and Chih-Jen Lin. (2001) LIBSVM: a Library for Support Vector Machines. [Online]. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] Ming-yu Chen, "Long Term Activity Analysis in Surveillance Video Archives," Carnegie Mellon Univer, Pittsburgh, PhD Thesis 2010.

[8] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features," in *Proceedings of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, 2005, pp. 65 - 72.

[9] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik, "Recognizing Action at a Distance," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 726 - 733.

[10] Chris Harris and Mike Stephens, "A Combined Corner and Edge Detector," in *Proceedings of the 4th Alvey Vision Conference*, Manchester, UK, 1988, pp. 147-151.

[11] Thomas Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* , Berkeley, US, 1999, pp. 50-57.

[12] Alexander Kläser, Marszałek Marcin, and Cordelia Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *Proceedings of British Machine Vision Conference*, Leeds, UK, 2008, pp. 995-1004.

[13] Ivan Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107-123, June 2005.

[14] Ivan Laptev and Tony Lindeberg, "Local Descriptors for Spatio-Temporal Recognition," in *Proceedings of the 1st International Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, Czech Republic, 2006, pp. 91-103.

[15] Ivan Laptev et al., "Learning Realistic Human Actions from Movies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK , 2008, pp. 1-8.

[16] Svetlana Lazebnik, Schmid Cordelia, and Jean Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, 2006, pp. 2169 - 2178.

[17] Jingen Liu, Saad Ali, and Mubarak Shah, "Recognizing Human Actions Using Multiple Features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK , 2008, pp. 1-8.

[18] Chang Liu and Pong Chi Yuen, "Human Action Recognition Using Boosted EigenActions," *Image and Vision Computing*, vol. 28, no. 5, pp. 825-835, May 2010.

[19] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, January 2004.

[20] Bruce D. Lucas and Takeo Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver, BC, Canada , 1981, pp. 121-130.

[21] Krystian Mikolajczyk and Cordelia Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, October 2005.

[22] Krystian Mikolajczyk and Hirofumi Uemura, "Action Recognition with Motion-Appearance Vocabulary Forest," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK , 2008, pp. 1-8.

[23] Juan Carlos Niebles, Wang Hongcheng, and Li Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299-318, September 2008.

[24] Ronald Poppe, "A Survey on Vision-Based Human Action Recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

[25] John Ross Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, March 1986.

[26] Edward Rosten and Tom Drummond, "Fusing Points and Lines for High Performance Tracking," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 1508-1515.

[27] Edward Rosten and Tom Drummond, "Machine Learning for High-Speed Corner Detection," in *Proceedings of the 9th European Conference on Computer Vision*, Graz, Austria, 2006, pp. 430-443.

[28] Konrad Schindler and Luc Van Gool, "Action snippets: How many frames does human action recognition require?," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, 2008, pp. 1-8.

[29] Cordelia Schmid, Roger Mohr, and Christian Bauckhage, "Evaluation of Interest Point Detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151-172, June 2000.

[30] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing Human Actions: A Local SVM Approach," in *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 2004, pp. 32 - 36.

[31] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional Sift Descriptor and its Application to Action Recognition," in *Proceedings of the 15th International Conference on Multimedia*, Augsburg, Germany, 2007, pp. 357-360.

[32] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman, "Discovering Objects and Their Location in Images," in *Proceedings of the 10th IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 370 - 377.

[33] Pavan Turaga, Rama Chellappa, V. S. Subrahmanian, and Octavian Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473 - 1488 , 2008.

[34] Tinne Tuytelaars and Krystian Mikolajczyk, "Local Invariant Feature Detectors: A Survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177-280, January 2008.

[35] Vladimir N. Vapnik, *Statistical Learning Theory*. NY: Wiley-Interscience, 1998.

[36] Yang Wang, Payam Sabzmeydani, and Greg Mori, "Semi-Latent Dirichlet Allocation: A Hierarchical Model for Human Action Recognition," in *Proceedings of the 2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation (ICCV)*, Rio de Janeiro, Brazil, 2007.

[37] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *Proceedings of the British Machine Vision Conference*, London, UK, 2009.

[38] Geert Willems, Tinne Tuytelaars, and Luc Gool, "An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector," in *Proceedings of the 10th European Conference on Computer Vision*, Marseille, France, 2008, pp. 650-663.

[39] Shu-Fai Wong and Roberto Cipolla, "Extracting Spatiotemporal Interest Points using Global Information," in *Proceedings of the 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1-8.

[40] Shu-Fai Wong, Tae-Kyun Kim, and Roberto Cipolla, "Learning Motion Categories using both Semantic and Structural Information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, US, 2007, pp. 1-6.

[41] Jianxin Wu and M. James Rehg, "Beyond the Euclidean Distance: Creating Effective Visual Codebooks Using the Histogram Intersection Kernel," in *Proceedings of the 12th IEEE International Conference on Computer Vision*, Kyoto, Janpan, 2009, pp. 630-637.

[42] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213-238, June 2007.