Human action capturing and classification

Feng, Zhou

2006

Feng, Z. (2006). Human action capturing and classification. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/13599>

<https://doi.org/10.32657/10356/13599>

NANYANG TECHNOLOGICAL UNIVERSITY

Human Action Capturing And Classification

A thesis submitted for the degree of Master
of Engineering at Nanyang Technological University

by **Feng Zhou**

Supervisor: Dr. Cham Tat Jen

March 2006

Abstract

Recognizing humans and activities in video is a key enabling technology for ultra-intelligent environments. These environments, which we call sentient spaces, are envisioned to be fully capable of sensing the motives and desires of occupants, and proactively reacting to them in ways to dramatically improve work productivity and quality of life.

In this report, a vision-based framework is proposed for learning and inferring occupant activities at different levels. These levels range from short temporal interval movements, to intermediate level events and long temporal term activities. Our research is focused on using a combined tracking-classification framework for the unsupervised classification of human action.

An initial comparative study was done to evaluate several existing foreground segmentation methods that employ background modeling. Our own probabilistic foreground-background segmentation method is proposed to extract human-centric reference frames.

Based on the human-centric reference frames, a principled analysis of the correspondence problem leads to a novel probabilistic action representation called the correspondence-ambiguous feature histogram array (CAFHA) that is robust to variations across similar actions. CAFHA is shown to be effective in unsupervised action classification and quasi real-time action inference.

A novel feature selection method is proposed to select the optimal features to improve the CAFHA representation, such that the best discrimination between different action clusters may be found via unsupervised spectral clustering.

Finally, a number of potential future directions are proposed that are targeted at further improvements to our framework and creating new research methods required to recognize human activity at longer temporal scales.

Acknowledgments

This thesis summarizes my postgraduate research work in School of Computer Engineering, Nanyang Technological University. Over the past two years, I've received precious help from lots of people. At the top of my list is my supervisor Dr. Cham Tat Jen. His passion to science, constant encouragement, heuristic instruction, kindly criticism and forgiveness help me come through these years. Without him, I could never have done the job.

I want to thank my colleagues: Chen Xi, Song Peng, Pong Hon Keat, and Pham Minh Tri. We are in the same boat, and share a good time together. Many thanks to Li Jiang, who gave me a lot of inspiration and suggestions in my research.

Friend in need is a friend indeed. I'd like to express my heartfelt thanks to my friends: Han wei, Hu Yiqun, Qiu Jie, Zhan Wenning, Wu Jun, Bai Jie, Yu Renjun, Qiu Yuanfu, Shi Xiaomeng, Hao Yanyan, Liu Bing, Yin Zongyou, Loo Rui, Shen Zhuoqian, Huang Huamei, Cao Jia, Li Da *et al.* Your true friendship encourages me whenever I am facing difficulties. Lots of thanks to my Cement colleagues: Wang huiqing, Zhou Zhi, Guo Jing, Ye Minghua, Li Jiaming *et al.* You are part of everyday life.

A very deep thanks to Madam Mary Loo, who cares and encourages me like her own son. Sincere thanks to my cousin Lin Mei and Xiang Lin for the great help and advice you gave me, especially during my conference trip in US. Also thank Tracy Lim for motivating me a lot last year.

I'd like to thank SCE and NTU for providing me the chance to study in this beautiful campus. I broaden my eyesight and find my way to the future here. Thanks to Dr. Chia Liang Tien, Clement and Dr. Leung Karhang Maylor, who helped me when I first started my research in NTU.

Finally I wish to express my deepest gratitude to my parents and sister. It is their constant love, support and trust that driving me ahead.

Contents

Abstract	i
Acknowledgments	ii
List of Figures	vi
List of Tables	x
1 Introduction	1
1.1 Motivation	1
1.2 Research Issues	3
1.2.1 Problems to Solve	3
1.2.2 Framework	4
1.3 Outline of Report	6
2 Literature Review	8
2.1 Human Tracking	9
2.1.1 Foreground Segmentation	9
2.1.2 Temporal Tracking	11
2.2 Human Movements Recognition	15
2.2.1 Approaches with Explicit Shape Models	15
2.2.2 Approaches without Shape Models	18
2.2.3 Event Based Video Analysis	21
2.3 Human Activity Understanding	25
3 Tracking and Human-centric Reference Frames	29
3.1 Overview	29
3.2 Experiments and Results	29

3.2.1	Temporal Differencing	30
3.2.2	Background Modeling (Averaging)	31
3.2.3	Background Modeling (Median Value)	33
3.2.4	Background Modeling (Gaussian)	34
3.2.5	Background Modeling (Mixture of Gaussians)	36
3.2.6	Comparison	39
3.3	Probabilistic Tracking and Human-centric Reference Frame	40
3.3.1	Probabilistic Foreground-Background Segmentation	40
3.3.2	Estimating the Human-Centric Spatial Reference Frame	42
3.3.3	Experiments	43
3.4	Multiple People Tracking	45
4	Human Action Classification with Ambiguous Correspondences	47
4.1	Definition of Action	47
4.2	Discussion of Zelnik-Manor and Irani's Work	48
4.2.1	Spatial and Temporal Information	48
4.2.2	Effect of Background	49
4.3	Spatial Correspondence in Action Classification	51
4.4	Overview of Action Classification Framework	53
4.5	Basics	55
4.6	Action Representation	55
4.6.1	Feature-Set Action Representation	56
4.6.2	Action Variation and Observational Errors	57
4.6.3	Correspondence Ambiguity	58
4.7	Probabilistic Analysis of Features	58
4.7.1	Discussion of the Prior Correspondence Ambiguity Term	60
4.7.2	Construction of Correspondence-Ambiguous Feature Histogram Array (CAFHA)	61
4.7.3	CAFHA Implementation	62
4.8	Action Classification	63
4.8.1	CAFHA Distance Measure	63

4.8.2	Computing the Affinity Matrix	64
4.8.3	Spectral Clustering	64
4.9	Human Action Learning and Online Inference	66
4.10	Experiments	67
4.10.1	Probabilistic Segmentation and Tracking	67
4.10.2	Action Classification	67
4.10.3	Action Learning and Inference	71
4.11	Comparison with Zelnik-Manor and Irani's Method	72
4.12	Conclusion	72
5	Feature Selection and Optimization for Unsupervised Action Classification	74
5.1	Feature Selection and Action Classification Framework	75
5.2	Feature Selection and Optimization	75
5.2.1	Intuition	76
5.2.2	A Metric based on Affinity Matrix for Classification	77
5.2.3	Optimization	78
5.3	Experiments	79
5.3.1	Optimization	79
5.3.2	Action Classification	80
5.3.3	Real-time Online Action Inference	83
5.4	Conclusion	85
6	Conclusion and Future Work	86
6.1	Conclusion	86
6.2	Future work	87
	References	93
	Publications	102

List of Figures

1.1	(a) Sample picture of sentient space [2] (b) Sample picture of visual surveillance of monitoring pedestrians in public place [3]	1
1.2	(a) Sample picture of user interfaces [30] (b) Sample picture of virtual reality [1]	2
1.3	Research framework proposed in this thesis	4
2.1	Samples of pedestrian detection results [37]	11
2.2	(a) Block diagram of color object tracking by CAMSHIFT (b) Tracking Samples [13]	13
2.3	Flow diagram of the Porikli's [63] decision process	14
2.4	A 2-D stick-figure model (By Leung and Yang [51])	16
2.5	Samples of Pavlovic <i>et al.</i> [60]'s figure model	16
2.6	3-D human model "ELLEN" and "DARIU" using tapered super quadrics [36] .	17
2.7	Sample motion magnitude feature for walking (top) and running (bottom) [62] .	18
2.8	Comparison of MEI and MHI	19
2.9	Overview of the framework	21
2.10	Example of temporal pyramid	22
2.11	The empirical distribution of N_t^1 for 6 different clips [82]	23
2.12	(a)-(d) Sample frames of a basketball video. (e) Example of distance measure computed by 2.2 [82]	24
2.13	Four level decomposition of human dynamics [14]	26
2.14	System Diagram of Trajectory guided tracking and understanding	27
2.15	(Upper left) An image of the scene, (Center) the classification hierarchy (Upper right) Co-occurrence matrix and normalized pmfs. [72]	28
3.1	Tracking by temporal differencing	30

3.2	Background modeling by averaging	32
3.3	Background modeling by median value	33
3.4	Gaussian distribution	35
3.5	Background modeling by mean and variance	36
3.6	Background modeling by mixture of Gaussians	37
3.7	Examples of manually labeled foreground	39
3.8	ROC curve for five tracking algorithms	40
3.9	(a) Sample probability density frame and color map (b) $\phi(x)$ (c) $\phi(y)$	42
3.10	(a) Sample frames of walking, and images representing the probability density value of $P(L_u=1 \mathbf{u})$ (b) Sample frames of waving, and images representing the probability density value of $P(L_u=1 \mathbf{u})$ (c) Cropped images of human centered in the box	44
3.11	Sample tracked images	44
3.12	Tracking module	46
3.13	An example of tracked foreground in bounding boxes	46
4.1	Top three figures are sample frames of three different events: running, waving and walking [82]. Middle three are empirical distributions N_t^1 of the whole frames (a) Running (b) Waving (c) Walking. Bottom three are empirical distribution of N_t^1 for the foreground only: (d) Running (e) Waving (f) Walking . . .	50
4.2	Sample result of a Correspondence-Exact method [60]	51
4.3	Sample result of a Correspondence-Free method [11]	51
4.4	Correspondences, even if ambiguous, can be exploited to distinguish the two waving actions.	53
4.5	Action Classification Framework	54
4.6	Action Reference Frame	55
4.7	Construction of the Correspondence-Ambiguous Feature Histogram Array (CAFHA)	62
4.8	Framework of action learning and inferring	66
4.9	Sample frames of walking, and images representing the probability density value of $P(L_u=1 \mathbf{u})$	67
4.10	Sample frame of (a) walking fast (b) running (c) walking at medium speed . . .	67

4.11	Clustering results by (a) our method (b) Zelnik-Manor and Irani's method and (c) ground-truth (Blue represents walking fast, red represents running, and cyan represents walking at medium speed)	68
4.12	Sample frame of (a) walking fast, (b) walking slowly, (c) waving over head, (d) waving in the front of body.	69
4.13	Clustering results by (a) our method, (b) Zelnik-Manor and Irani's method, (c) ground-truth (Red represents walking fast, yellow represents walking slowly, blue represents waving over head and cyan represents waving in front of body)	69
4.14	Sample frame of walking fast, waving over head and waving in front of body by three different people	70
4.15	Unsupervised clustering results by (a) our method (b) Zelnik-Manor and Irani's method (c) ground-truth (yellow represents walking , red represents waving over head and blue represents waving in front of body)	70
4.16	(a) Inference result of multiple person actions based on learned CAFHA representation, (b) Ground-truth information	70
5.1	Action classification framework with optimization	76
5.2	Example of different feature points	76
5.3	D_1, D_2, D_3, D_4, D_5 imply clusters of same action. D_6, D_7 are the distance values between subsequences of two different action sequences	77
5.4	(a) The initial Gaussian kernels in human-centered frame (b) The optimized Gaussian kernels	80
5.5	(a) Affinity matrix computed by Gaussian kernel in figure 5.4(a) (b) Affinity matrix by the optimal Gaussian in figure 5.4(b)	81
5.6	Unsupervised clustering results by (a) Optimized CAFHA (b) Naive CAFHA with 4 Gaussian Kernels (c) Zelnik-Manor and Irani's method (d) ground-truth (yellow represents walking , red represents waving over head and blue represents waving in front of body)	82
5.7	Sample frames video dataset for sequence 1 to sequence 3	82
5.8	(a) Sample frame of walking fast, (b) Sample frame of waving over head, (c) Sample frame of waving in the front of body	84

5.9	(a) Inference result of multiple person actions based on optimized CAFHA representation of selecting only 3 points (b) Inference result of multiple person actions by naive CAFHA of selecting 9 points (c) Ground-truth information . .	84
6.1	Visualization of affinity matrix	88
6.2	Multiple people activity understanding	89

List of Tables

4.1	Results for inference of different actions	71
5.1	Comparison of Zelnik-Manor and Irani's method, Naive CAFHA and Optimized CAFHA on action classification	83

Chapter 1

Introduction

1.1 Motivation

Capturing and recognizing humans and their activities in video is receiving increasing attention from computer vision researchers. This interest is motivated by a wide spectrum of applications that involve designing machines capable of interacting intelligently and effortlessly within a human-inhabited environment.

Figure 1.1: (a) Sample picture of sentient space [2] (b) Sample picture of visual surveillance of monitoring pedestrians in public place [3]

An important application domain is sentient spaces. Future homes and offices can be equipped with computing devices, which can intelligently learn and understand the occupants activities and attentively respond to meet the need of the occupants, as shown in figure 1.1(a). For example, when a person wants to watch TV, he does not need bother to find the remote

Figure 1.2: (a) Sample picture of user interfaces [30] (b) Sample picture of virtual reality [1]

control. Instead, he can sit easily in the sofa, and give a certain gesture. The television will turn on automatically and adjust the screen to a location which is a best distance and angle to his face. Sentient spaces can greatly improve the way we live and work.

Another application domain is visual surveillance. There is a enormous need for automated surveillance systems especially in military applications and security applications. Video surveillance technology has also been proposed to measure traffic flow, detect accidents on highways, monitor pedestrians in public spaces (as shown in figure 1.1(b)), compile consumer demographics in shopping malls and amusement parks, log routine maintenance, and robotic applications. For example, in 1997, the Defense Advanced Research Projects Agency (DARPA) started a program to develop Visual Surveillance and Monitoring (VSAM) [20] technology, which is intended for keeping track of people, vehicles, and their interactions in urban and battlefield settings.

Activity understanding can be a powerful cue for analysis of video, including event-based video indexing, browsing, clustering and segmentation. It is widely researched as a tool for natural user-interfaces (as picture shown in figure 1.2(a)). The desired goals include recognizing the user, observing facial expressions and body gestures. The capability for understanding occupant activities also support other applications, such as virtual reality (as picture shown in figure 1.2(b)), model-based coding and motion analysis.

1.2 Research Issues

1.2.1 Problems to Solve

The goal of our research is to endow computers with the ability of human vision to see and understand the dynamic scene in a video sequence, which is a very difficult task. Generally, tracking techniques are first used to extract the relevant motion information, and the information is then processed by pattern recognition techniques to understand the activities occurring in the video.

The key challenges in the tracking part are robustness and reliability in the complex environments, as it is difficult to build models that accurately describe images of real scenes. Motion is an effective cue for visual perception, and the challenge is to reproduce this capability in computer vision systems. The purpose of our tracking module is to extract the relevant foreground motion information for further understanding.

In the activity understanding stage, defining actions and activities is the first problem. Creating a satisfactory representation of the motion attributes for action and activity is a big challenge because of the complexity and variety of human motion. In scenes of multiple interacting persons, the difficulties increase many folds.

Many research efforts have focused on tracking detailed human body parts, such as heads, faces, hands, legs etc, combined with features like shape, texture and edge, and building sophisticated models [51, 64, 59, 18, 5] for understanding specific actions in constrained environment. These methods do a good job in classifying pre-defined actions; however they are usually application dependent, sensitive to noise, and require constrained environments, because the models used are often very specific to the data observed. These systems cannot be generalized to broader situations.

Unlike these works, we are interested in building a robust, generic and multiple stage system which is able to handle broad range of actions. We want our system to automatically

capture and understand the occupant activities robustly at different levels of detail without the necessity of sophisticated tracking and modeling. The tracking results may not be accurate, but this level of information is sufficient to be exploited in a subsequent action understanding stage. With more captured information and prior knowledge about the activities and environments, the action understanding at different levels can be refined recursively.

1.2.2 Framework

Figure 1.3: Research framework proposed in this thesis

This project is based in Centre for Multimedia and Network Technology (CeMNet) of Nanyang Technological University as part of the activities of the pervasive computing group. It is aimed at researching a critical component in the SmartSpace project of creating a sentient space.

Figure. 1.3 shows the proposed research framework. We build an automatic understanding system, which understands occupant activities at different levels from low level short temporal interval movements, to intermediate level events and high level long term actions. Bobick [10] proposed a three-level motion understanding taxonomy: movements, activities and actions. Our taxonomy is similar to Bobick's, since it is based on the temporal scale. We consider low level movements to be the atomic motion elements of short temporal interval. Examples are an isolated walking step, the kicking of a ball, sitting down *etc.* Intermediate events refer to motions of larger temporal scale which is a certain grouping of low level movements, *e.g.* extended people walking and running sequence. High level action is a temporal sequence of even greater temporal scale. The taxonomy of activities has strong temporal characteristics.

A detailed description of our framework is given below:

- (i) Tracking module. Tracking humans in video is the first step. A simple but robust tracking system is used to extract the foreground occupants in human-centric reference frames, and record the locations of the occupants as trajectories. A human-centric reference frame can be considered as a projected canonical reference frame with a person's mass center located in the center of this frame. The horizontal and vertical spatial scales of the frames are proportionally linked to the girth and height of the person. Human-centric reference frames are used as the basic elements for subsequent research on activity understanding, since they capture the foreground motion of the original frames, which is the key information for activity understanding.
- (ii) Low level movement understanding module. Based on the tracked motion and trajectories, motion histories of occupants are analyzed to learn lower level individual movements with short temporal intervals. Some atom movements such as walking, running, turning around and squatting can be recognized. This is the stage of low level understanding.

- (iii) Intermediate action understanding module. The atomic movements we recognized from the low level stage may be then automatically grouped to form intermediate events with longer temporal intervals. Intermediate level understanding is to understand this kind of longer temporal action.
- (iv) Actions classification module. This module includes our main research focus. The interesting issue in intermediate action understanding stage is to classify different actions for each person by segmenting and clustering different intermediate actions. A non-parametric unsupervised clustering method is applied for the video segmentation based on a novel action representation will be studied. An optimized method through feature selection for action classification will be discussed.
- (v) High level action understanding module. Eventually, these intermediate events are automatically grouped to high-level, multi-person activities with extensive temporal range. People interaction and different activity patterns can be recognized, and the system will be able to find abnormal activities. These groupings will form a taxonomy of activities that allows sentient spaces to react intelligently in a context-aware manner to user input.

With more prior information about the actions that take place in the video, the video segmentation in the intermediate stage can be refined. Meanwhile, the high level understanding result can be used as feedback to refine the intermediate level segmentation, and even the low-level recognition. It is a recursive processing (figure 1.3) that uses higher level feedback to refine the low level process.

1.3 Outline of Report

In chapter 2, a broad range of research of human tracking, human motion analysis and human activity understanding techniques are reviewed.

1.3 Outline of Report

7

Five different established tracking techniques are discussed in Chapter 3. Experiments are tested and compared, and the advantages and disadvantages of each technique are studied in this chapter. Adaptive background modeling by mixture of Gaussians are analyzed in some detail. Our own probabilistic foreground-background segmentation is proposed, which is used to extract human-centric reference frames for further research on human activity understanding. Finally, the implementation of a multiple people tracking module based on mixture of Gaussians is presented, which is proven to be the best method among the five techniques from our experiments.

Based on the human-centric reference frames, a combined tracking-classification framework for the unsupervised classification of human action is proposed in Chapter 4. A principled analysis of correspondence problem leads to a novel probabilistic action representation called the correspondence-ambiguous feature histogram array (CAFHA) that is robust to variations across similar actions. Our results show that the new framework outperforms the recent Zelnik-Manor and Irani method [82] for unsupervised event classification. Additionally, the framework is extended to quasi real-time action inference, with good recognition accuracy despite changes in person identity and variations in the actions.

Chapter 5 describes a principled feature selection method for unsupervised action classification based on the work described in chapter 4. An optimization process is applied to selecting the most informative features, such that the representation allows the best discrimination between different action clusters found via unsupervised spectral clustering.

Conclusion and future work is presented in Chapter 6.

Chapter 2

Literature Review

The number of publications in the area of human capture and understanding has grown exponentially over the last few years. We first classify some of the important works into three different categories:

- Human tracking (discussed in section 2.1), which is aimed at tracking foreground motion regions.
- Low level human motion recognition (discussed in section 2.2), which is aimed at recognizing low level short temporal movement.
- High level activity understanding (discussed in section 2.3), which is aimed at understanding long temporal scale human actions.

This classification fits nicely into our research framework (section 1.2.2).

Because our intermediate to high level recognition modules are based on temporal scales, video segmentation along the temporal axis is a very important topic. Various elements of the recent Zelnik-Manor and Irani's method [82] for unsupervised event classification are used in our research. Their work is viewed in detail in section 2.2.3.

2.1 Human Tracking

The goal of human tracking is to segment the foreground (The term: *foreground* in this report refers to the regions of human motion.) and establish correspondence of foreground regions between frames. We further classify human tracking into two categories:

- Human region segmentation
- Temporal tracking

which are discussed in the following two sections.

2.1.1 Foreground Segmentation

Human region segmentation is aimed at segmenting foreground, which are regions of human motion, from the background. It is the fundamental step for most tracking systems. Dealing with changing illumination, shadows and background clutter are the major challenges.

One of the popular approaches is based on determining which pixels in the current image frame are different from a reference background. To cope with noise, the probability distributions for the grey levels of each pixel may be separately modeled. A pixel is classified as foreground if the intensity deviates substantially from the high probability gray levels of the corresponding pixel in the reference background.

Friedman and Russell [34] modeled each pixel by an adaptive parametric mixture model of three Gaussian distributions. Their method classifies the pixel values into three separate, predetermined distributions corresponding to the road color, the shadow color, and vehicle color. However it is insufficient in dealing with problems of repetitive motions, shadows and reflectance. Stauffer and Grimson [71] improved Friedman's method by modeling the values of a particular pixel as a mixture of Gaussians and updated its parameters online. This method deals robustly with changing illumination, repetitive motions, tracking through cluttered regions, slow-moving objects, *etc.* Moreover, it is computationally efficient and thus can

be implemented in real time. However it may be undesirable to detect moving shadows as foreground. KaewTraKulPong *et al.*'s [47] solved this problem by incorporating a shadow detection scheme. Their system is able to learn fast and accurately, and adapts effectively to changing environments. In the W4 system by Haritaoglu *et al.* [41], an interesting statistical background model is built by representing each pixel with three values: its minimum, maximum intensity values, and the maximum intensity difference between consecutive frames observed during the training period.

Temporal differencing [48] uses the information of pixel-wise difference between consecutive frames of an video sequence to extract moving regions. It is adaptive in dynamic environments, but it generates holes inside moving objects and therefore does a poor job in extracting the entire foreground regions. It can be combined with other motion detection techniques to achieve better results. One example is part of the video surveillance and monitoring (VSAM) program developed by Carnegie Mellon University (CMU) [20]. A hybrid algorithm for motion segmentation is proposed by combining an adaptive background subtraction algorithm with a three-frame differencing technique. It is very fast and effective.

Optical flow [8] is used for extracting measurements such as spatial and temporal derivatives to produce a 2-D flow field. Meyer *et al.* [55] proposed an active rays method to use the optical flow to initialize a contour-based tracking algorithm. Generally, the optical flow method can be used to detect independent moving objects in the presence of camera motion. However most optical flow methods are computationally complex and very sensitive to noise.

Another approach is to use active contours [35, 33] to extract the entire outline of the subject [9] or individual body parts. Wren *et al.* [79] proposed Pfinder, an approach in which the objects are modeled by blobs with individual color and shape statistics. Each pixel in the current image is then classified as one of the blobs according to its color and spatial properties.

Gavrila and Philomin [37] presented a shape-based object detection method based on Distance Transforms (DT). The Distance Transform [40] is a morphological operator for binary

Figure 2.1: Samples of pedestrian detection results [37]

image, which is used for template matching in their work. A template hierarchy is generated to capture the shape of objects, which can be generated off-line using stochastic optimization techniques. Figure 2.1 shows the examples of the detection result.

2.1.2 Temporal Tracking

Temporal tracking is the problem of inferring the motion of an object given a temporally-ordered sequence of images [32]. It involves finding corresponding objects in consecutive frames. The difficulty of the task is related to the complexity of the scene and target objects. For example, the detected region may be split apart and then merged into a new region due to occlusion or noises, and the appearance of an object may change due to shadows and illumina-

tion changing.

Compared with non-temporal object detection, temporal tracking utilizes the dynamic information between frames to establish foreground correspondence. The correspondence analysis is often supported by prediction, which is based on a dynamic model. Some representative methods for prediction are discussed in the following.

Kalman filter [39] is widely used in tracking as a linear dynamic model. It is based on the assumption that all random variables have Gaussian distributions and they are linear related. Unfortunately, Kalman filter is only effective in the situations where the probability distribution of the state parameters is unimodal. It does not handle simultaneous multi-modal distributions which may occur in the presence of occlusion, cluttered background, *etc.* The Kalman filter can be extended to various models, such as: extended Kalman filter, unscented Kalman filter [76], switching Kalman filter [57], mixture Kalman filter [17] and particle filter [45].

The CONDENSATION algorithm [45] is an early application of particle filtering in tracking. It uses factored sampling, in which the probability distribution of possible interpretations is represented by a set of samples in state space. Particle filtering approximates the posterior distribution by a set of weighted particles. It weighs the particles based on a likelihood score and then propagate these particles according to a motion model. By combining a dynamic model with visual observation, it can accomplish highly robust tracking of objects. Thus it is a powerful alternative to the Kalman filter. However, it is highly dependent on the availability of an unbiased predictor, which is difficult to obtain and also requires an unwieldy number of samples in high dimensional state space.

Cham and Rehg [15] proposed a novel multiple hypothesis tracker. The distribution of the state is formed by a set of modes with piecewise Gaussians characterizing the neighborhood around these modes. The modes of the distribution are given by the propagated hypotheses, the covariances describe the shape of the distribution in the neighborhood of the hypotheses, and the temporal evolution of the posterior distribution is achieved by sampling the prior distribution. Compared with particle filtering, it does not need a huge number of samples, because

2.1 Human Tracking

13

the representation uses the modes of the distributions. It is more effective in high dimensional problems and much less dependent on the unbiasedness of the predictor.

Figure 2.2: (a) Block diagram of color object tracking by CAMSHIFT (b) Tracking Samples [13]

Mean shift is an important algorithm which computes the offset of current location to the new location by the mean shift vector. Recently it is used for tracking by several researchers. Bradski [13] improved this algorithm to deal with dynamically color distribution changing. The new algorithm is called CAMSHIFT, and applied in human face tracking. Figure 2.2(a) illustrates the procedure of CAMSHIFT and (b) shows the tracking face with background distractors. Comaniciu *et al.* [23] analyzed the mean shift algorithm theoretically, and applied it in real-time non-rigid object tracking. In their framework, color distribution are used in the mean shift iteration to find the most probable target position. It is formulated as a maximum

Bayes error estimating problem. Porikli *et al.* [63] suggested an automatic real time human tracking system by integrating a mean-shift update technique with an adaptive change detection method, as shown in figure 2.3. In order to automatically initialize the mean shift tracking, adaptive background subtraction is used to find the initial regions. Collins [22] studied the problem of mean-shift kernel scale variation when tracking blobs are changing in size. Lindeberg's theory of feature scale selection is adopted to choose and update the scale.

Figure 2.3: Flow diagram of the Porikli's [63] decision process

Toyama and Blake [75] proposed an exemplar-based, probabilistic visual tracking in metric space. The probabilistic representation has advantage in handling uncertainties, but it requires complex modeling. In order to preserve the advantages of probabilistic representation and avoid building explicit models, exemplars are used. Exemplars are selected representatives of raw training data, and used to represent probabilistic distributions of object configurations. By using exemplars, hand-construction of object models and problems with changes of topology are avoided.

2.2 Human Movements Recognition

Human movement here refers to low level movement, which is the atomic motion element, for example, an isolated walking, the kicking of a ball. It has short temporal scale.

A lot of works have been done in recognizing the low level movement. Spatial and temporal models are largely used to encode the unique characteristics of how humans move. They may provide direct correspondence between labeled body parts and regions in the image, enforce registration constraints and deliver predictive power that allows us to make inferences about human motion robustly in the presence of noise and uncertainty. There are also some works that achieved good recognition without using explicit models.

We classify the works by whether information of the object shape is employed in the motion analysis and review them in the following two sections: approaches with explicit shape models and approaches without a priori shape model. Some specific topics such as gait analysis, gesture recognition are also considered into this section.

2.2.1 Approaches with Explicit Shape Models

Many methods for the motion analysis of human body use models to fit to the given image in order to focus on high-level tasks at a later stage. The models may be presented as stick, figures, 2D contours or volumetric models.

Leung and Yang [51] proposed a human body outline labeling system which labels body parts from the silhouette of a person. Their basic body model consists of five U-shaped ribbons and a body trunk, various joints and mid points, plus a number of structural constraints, as shown in figure 2.4.

Elliptical cylinders are one of the commonly used volumetric models for modeling the human body parts. Hogg [43] and Rohr [66] modeled the human body using 14 elliptical cylinders in order to recover the 3D pose of human body. Each cylinder is represented by three parameters: the length of the axis, the major and minor axis of the ellipse cross section.

Figure 2.4: A 2-D stick-figure model (By Leung and Yang [51])

Chen *et al.* [18] recovered the 3D configuration of a moving subject according to its 2D image projection by modeling human with 17 line segments and 14 joints to represent the head, torso, hip, arms and legs.

Figure 2.5: Samples of Pavlovic *et al.* [60]'s figure model

2.2 Human Movements Recognition

17

Pavlovic *et al.* [60] used a Dynamic Bayesian Networks (DBNs) based switching linear dynamic system (SLDS) model for figure motion analysis and tracking. Their kinematic model had 8 degrees of freedom (DOF), corresponding to rotations at the knees, hip, and neck, as shown in figure 2.5. They proposed a Viterbi approximation technique which overcame the exponential complexity of exact inference in mixed-state DBNs.

Figure 2.6: 3-D human model "ELLEN" and "DARIU" using tapered super quadrics [36]

Gavrila and Davis [36] used a 3-D model to track unconstrained human movements. Figure 2.6 shows the 3-D models, which has 22-DOF (3 DOF for the positioning of the root of the articulated structure, 3 DOF for the torso and 4 DOF for each arm and each leg).

Figure 2.7: Sample motion magnitude feature for walking (top) and running (bottom) [62]

2.2.2 Approaches without Shape Models

Approaches without explicit shape models use simpler representations that do not directly capture the articulated constraints of the human body, but nevertheless provide sufficient uniqueness to distinguish human from non-human motion.

Polana and Nelson [62] proposed a method of recognizing repetitive human motion without finding his body parts. For one cycle of human action, the human in the image is segmented, and spatially scaled to form a new image sequence in a uniform size, in which the person is at the center of the frame. Each frame is divided by a spatial grid of $X \times Y$ dimension. Motion magnitudes in each spatial tile are summed separately and together they form a feature vector which represents the human action. Figure 2.7 shows one cycle of activity divided into six time divisions shown horizontally, each frame shows the spatial distribution of motion in a 4×4 spatial grid with size of each square proportional to the amount of motion in the neighborhood. There are similar techniques that involve superimposing grids on regions of interest, and computing a variety of different features. Takahashi *et al.* [73] used an average edge vector for each tile, while Yamato *et al.* [80] used the number of foreground pixels in each tile.

Cutler and Davis [25] proposed techniques to detect and analyze periodic motion by tracking objects of interest and computing an object's self-similarity over different frames. Based

2.2 Human Movements Recognition

19

on the observation that the self-similarity measure of periodic motion is also periodic, they applied time-frequency analysis to detect and characterize the periodic motion. Because different objects in the scene normally display different periodicities, detecting the periodicity can also be used to classify objects.

Chomat and Crowley [19] generated motion templates by using a set of temporal-spatial filters computed by principle component analysis (PCA). The appearance of human motion is statistically described by the appearance of small spatio-temporal neighborhoods over a set of sequences. A Bayes classifier is finally used to perform action selection.

Figure 2.8: Comparison of MEI and MHI

Bobick *et al.* [11] proposed a view-based approach to the representation and recognition of human actions using temporal templates. They used the binary motion energy image (MEI) which represents where motion has occurred in an image sequence, and gray-scale motion history image (MHI) which represents how recently did the pixel experience motion, as examples

shown in figure 2.8. Motion images in a sequence were extracted via temporal differencing, and these motion images in a sequence were accumulated in time to form a binary MEI. The MEI can be enhanced into the MHI by the way that each pixel is given a value proportional to the duration of motion at that position. The positions where motion recently happened are given higher values, and positions where motion happened longer ago are given lower values. Finally, these view-specific templates were matched against the stored models of views of known actions during the recognition process.

Bobick *et al.* [12] presented a state-based technique for the representation and recognition of gestures. A gesture is first defined as an order sequence of states restricted by motion constraints in measurement space. The prototype trajectory of the gesture is then computed and states that capture the relevant variances and constraints are defined. A technique based on dynamic programming is applied at last for matching and classification.

There is early work by Harris [42] on corner detection. The idea is based on the observation that shifting a window in any direction of a corner should give a large change in intensity. It is rotation invariant, and further improved to be scale invariant. Laptev and Lindeberg extended this idea from spatial domain to spatio-temporal domain [49], and proposed a scale-invariant spatio-temporal descriptors for human action representation. This representation is used in detecting human walking in moving background. However the robustness of the method is questionable. Some other spatial scale invariant detectors includes: Harris-Laplacian [56], SIFT [53]. Another local descriptor Histogram of Oriented Gradient (HOG) [26] is recently used in detecting human.

Laptev and Lindeberg [50] later investigated local space-time descriptors for representing motion patterns. They have shown that in the combination of feature and measurement, position dependent histograms of optical flow is the best for spatio-temporal action recognition, and the second best is position dependent histogram of spatial-temporal gradients.

There are also a large amount of work to segment the images sequence into regions of consistent motions. The term *consistent motion* however has inconsistent definitions. An important

Figure 2.9: Overview of the framework

work of Costeira and Kandade [24] presented a method for separating and recovering motion and shape of multiple independently moving objects without prior knowledge of the number of object. However their method assumes that the features are extracted and tracked. Irani *et al.* have done a series of works [44, 54, 83, 84] on motion factorization. In Irani and Anandan's work in 2000 [44], they gave a deep analysis of factorization with uncertainty, which dealt with the correlated positional noises. While the theory is interesting, it has only been tested on synthetic data. In 2002 Machline *et al.* [54] suggested a new definition of motion consistency, which is based on temporal consistency of behaviors across multiple frames, and proposed a multi-body segmentation algorithm without requiring prior correspondence estimation nor feature tracking. Zelnik-Manor and Irani [84] further extended the method of spatial factorization to temporal factorization.

2.2.3 Event Based Video Analysis

Zelnik-Manor and Irani [82] proposed a framework, shown in figure 2.9, for event segmentation. Based on the perception of event, an event-based distance measure is proposed to measure

the difference between different events. Isolating and clustering events within long video sequence is achieved by applying data clustering methods.

Two events are considered similar if their empirical distributions at different temporal scales are similar. The distance between the events distributions provides a quantitative measure of the difference between two video sequences. It is constructed by the following three steps:

Step1: Construct temporal pyramid

Figure 2.10: Example of temporal pyramid

To get features at multiple temporal scales, a temporal pyramid $S^1, S^2, S^3, \dots, S^L$ of the video sequence is built by blurring and sub-sampling the sequence along the temporal direction as shown in figure 2.10. In this notation S is a sequence of images, while S^l has half the number of frames of the higher resolution sequence S^{l-1} .

Step2: Define local space-time feature measurement

Local intensity gradients (S_x^l, S_y^l, S_t^l) at each space-time point (x, y, t) of each sequence S^l is estimated. Because the gradients direction contain the information of local surface orientation which depends largely on the local behavioral properties of the moving objects, the gradient vector is normalized to conserve this information and to be more sensitive to spatial properties. In order to be invariant to the direction of the motion, absolute value is further taken to get the magnitude, leading to the space-time feature measurement equation of:

$$(N_x^l, N_y^l, N_t^l) = \frac{(S_x^l, S_y^l, S_t^l)}{\sqrt{((S_x^l)^2 + (S_y^l)^2 + (S_t^l)^2)}} \quad (2.1)$$

2.2 Human Movements Recognition

23

where N_k^l is the histogram of each feature component ($k = x, y, t$) at each temporal scale ($l = 1, 2, 3 \dots L$), which contains the spatial-temporal statistic information of the event. Each event is associate with a set of empirical distributions N_k^l .

Figure 2.11: The empirical distribution of N_t^1 for 6 different clips [82]

Figure 2.11 illustrates the empirical distributions of 6 different clips. Three clips figure 2.11(a, e, f) show the walking event of different people wearing different clothes, walking in different directions, and in front of different background viewed from different angles. The four clips figure 2.11(a, b, c, d) show the same person performing different activities wearing the same clothes in front of the same background. The results show that distributions for all the walking clips are much closer to each other, even though appearance-wise they have different visual content. Different activities have quite different distributions.

Step3: Measure the distance between two sequences

In step 2, each event is associated with $3L$ distributions, *e.g.*: if $L = 4$, *i.e.* 4 temporal scales, then each event will be associated with 12 histograms. To measure the difference

between two events which are represented by histograms, the χ^2 test is used:

$$D^2 = \frac{1}{3L} \sum_{k,l,i} \frac{[h_{1k}^l(i) - h_{2k}^l(i)]^2}{h_{1k}^l(i) + h_{2k}^l(i)} \quad (2.2)$$

where h_k^l is the discrete, smoothed and normalized histogram of N_k^l . $h_{1k}^l(i)$ represents the i -th bin value of h_k^l for the first event, and h_{2k}^l the i -th bin value for the second event. D is the distance measure of the two events.

Figure 2.12: (a)-(d) Sample frames of a basketball video. (e) Example of distance measure computed by 2.2 [82]

Figure 2.12 illustrates an example of the distance measure. A short sample clip which shows a player throwing the ball at the basket is compared against a sliding window of the same length as the sample clip continuously shifted across a long basketball video sequence. The sequence consists of actions like dribbling, dunking *etc.* Figure 2.12(a)-(d) are sample frames of a basketball video. Figure 2.12(e) is the result of distance measure computed by (4.2). The low values indicate the event happened in the sliding window at that time has high similarity with the sample clip.

2.3 Human Activity Understanding

Human activity understanding here can be considered as classification problems involving temporal sequence of feature data, which is derived from the earlier stages. It is about recognizing actions of long temporal intervals (including intermediate events and high level actions as defined in our framework), inferring semantic and logic descriptions, and deducing cognitive meaning.

Two kinds of approaches are commonly used: Template matching approaches and state-space approaches as discussed in Wang *et al.* [77]. Temporal matching approaches compare the feature extracted from the given image sequence to the pre-stored patterns during the recognition process. State space approaches define each atom action or posture as a state, which is connected by probabilities of transition to other states or to its own. The Hidden Markov Model (HMM) [65] is a representative state space model used to study discrete time series. HMM has been very popular in speech recognition, and has recently been adopted for recognition of human motion sequences.

Bregler [14] proposed a probabilistic decomposition of human dynamics into 4 levels, and showed how to propagate hypotheses across these levels. Low-level primitives are regions of coherent motion found by Expectation Maximization (EM) clustering, which are tracked over time. In middle level processing, those regions are fitted into simple movements represented by dynamical systems. High level complex gestures are represented by HMM as successive phases of the middle movements. Figure 2.13 shows the 4-level framework.

Gong and Tao [38] proposed Dynamically Multi-Linked Hidden Markov Model (DML-HMM) to recognize group activities of multiple object events in outdoor noisy situations. Relationships of multiple temporal events are modeled using Dynamic Probabilistic Networks (DPNs). The temporal order is inferred automatically by factorizing the state transition using Schwarz's Bayesian Information Criterion [68]. Airport cargo loading and unloading activities are modeled as a specific example.

Figure 2.13: Four level decomposition of human dynamics [14]

Rosales *et al.* [67] proposed a trajectory guided tracking and recognition system. Their framework consisted of five levels: low-level image processing, tracking, 3D trajectory recovery, occlusion reasoning and motion recognition in a recursive manner, as shown in figure 2.14. Low level information, such as the foregrounds, trajectories and occlusions, are the input for high level recognition, and the higher level modules provide feedback to maintain the tracking of non-rigid objects. A trajectory-guided recognition is carried out at last to classify actions.

Haritaoglu *et al.* [41] proposed the W4 system, which is a real-time visual surveillance

Figure 2.14: System Diagram of Trajectory guided tracking and understanding

system for tracking multiple people and their body parts, and understanding their activities. W4 does not use color information; instead it integrates the techniques of shape analysis and tracking techniques to locate people and their body parts, and build their silhouette models. W4 is able to understand the events of people interacting with each other and people interacting with the objects.

Stauffer and Grimson [72] developed a codebook of prototype representations using on-line Vector Quantization (VQ) to classify activities. A codebook is a set of prototype representations that represent the input distributions. A co-occurrence matrix is then defined over the prototypes and used to break apart the prototypes into a binary tree representation. A classification hierarchy tree is finally built to classify activities. Figure 2.15 shows an example of process for classifying activities.

Figure 2.15: (Upper left) An image of the scene, (Center) the classification hierarchy (Upper right) Co-occurrence matrix and normalized pmfs. [72]

Chapter 3

Tracking and Human-centric Reference Frames

3.1 Overview

Our human tracking module is aimed at acquiring human-centric reference frames of video sequences from a static camera. The human-centric reference frames serve as the basis for further higher level understanding. In order to find the best tracking method that serves our purpose, five different tracking techniques are tested and compared, and the advantages and disadvantages of each technique are studied in this chapter. A probabilistic foreground background segmentation method is proposed to extract human-centric reference frames. An implementation of multiple people tracking is finally presented.

3.2 Experiments and Results

In this section, five established tracking method are studied and experimented for use with a static camera. They are

- temporal differencing
- background modeling by averaging
- background modeling by median value

- background modeling by Gaussian
- background modeling by mixture of Gaussians

Comparisons are made qualitatively and quantitatively. The dataset from PETS2002 (*Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*) [4] is used to train models and test our algorithm. All the experiments and results are based on gray scale images with static background.

3.2.1 Temporal Differencing

Figure 3.1: Tracking by temporal differencing

Temporal differencing is a method of using the information of pixel-wise differences between consecutive frames of a video sequence to extract moving regions.

$$L(x, y, t) = \begin{cases} 1 & \text{if } |I(x, y, t + 1) - I(x, y, t)| \geq H \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

is a typical two-frame differencing equation, where $I(x, y, t)$ is the gray value at position (x, y) and time t , while $L(x, y, t)$ is a binary image with 1 representing foreground and 0 background at time t , and H is the threshold.

Figure 3.1 shows the result of the temporal differencing algorithm. 256 gray scale pixel value is used here, and a heuristical value 20 is used as threshold. The outline of human body is clearly observed, but the incorrect holes inside the foreground region are also obvious. Based on the binary image, we observe that while the noise is very small, the foreground is not fully detected.

Temporal differencing does not require the building up of a background model. It is reasonably adaptive in dynamic environments, but it does not maintain a static appearance model.

Temporal differencing only detects the regions of obvious change between consecutive frames. Hence for homogenous regions, even should the whole region move, the overlap parts of the two frames cannot be detected. However, highly textual regions or edges and corners, can be easily detected as long as the regions exhibit obvious changes.

3.2.2 Background Modeling (Averaging)

Background modeling is a method of estimating a statistical appearance model for a static background. The simplest method is averaging the images over time, creating a background approximation which is similar to the current static scene except where motion occurs:

$$M(x, y) = \frac{1}{K} \sum_{t=1}^K I(x, y, t) \quad (3.2)$$

$$L(x, y, t) = \begin{cases} 1 & \text{if } |I(x, y, t) - M(x, y)| \geq H \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

Figure 3.2: Background modeling by averaging

$M(x, y)$ is the pixel value at position (x, y) of the background reference image, $I(x, y, t)$ the pixel value at (x, y) position in the t^{th} frame. K is the number of frames used to obtain background and $L(x, y, t)$ is a binary image with 1 representing foreground and 0 background at time t .

Figure 3.2 illustrates several frames of the experimental results. The background reference image was obtained by averaging 40 clear background images, and foreground was detected by subtracting the current frame from the reference background and thresholding. The intensity values larger than 20 were considered as foreground, otherwise they were considered background.

3.2 Experiments and Results

33

The method can be extended in an online adaptive manner, where new frames are used to train the background reference image. The ideal situation of using this method is when clear static background images are acquired. As it is difficult to obtain pure background images online, it is only effective in the situation when small objects are moving continuously and the background is visible for a significant portion of time. It breaks down in the scenes with many moving objects, particularly if they move slowly, or with the illumination changing significantly.

3.2.3 Background Modeling (Median Value)

Figure 3.3: Background modeling by median value

Background modeling by median value uses the median value of a sequence of frames for every pixel to construct background reference image. The foreground detection is exactly the same as the background modeling by averaging.

$$M(x, y) = \text{median}\{I(x, y, t)\}, t = 1, 2 \dots K \quad (3.4)$$

$$L(x, y, t) = \begin{cases} 1 & \text{if } |I(x, y, t) - M(x, y)| \geq H \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

Figure 3.3 illustrates several frames of the experimental results. The background reference image was obtained by taking the median value of 80 frames, and foreground was detected by subtracting the current frame from the reference background and thresholding. The intensity values larger than 20 were considered as foreground, otherwise they were considered background.

It is quite similar to background modeling by averaging, but it has more advantages. Firstly, it does not require clean background images, which may not be easy to acquire. Secondly, in most situations each pixel will contain the background for a much larger portion of time than the foreground during the entire image sequence. When modeling the background, the median value falls in the range of background values, and is robust against outlier foreground pixels. In contrast, background modeling by averaging will be distorted by the outlier foreground pixels when modeling the background.

3.2.4 Background Modeling (Gaussian)

Background modeling by a Gaussian distribution is a method whereby the means and variances of each background pixel are estimated and used in the thresholding process.

For each pixel at (x, y) position, the probability of observing the current pixel value $I(x, y, t)$ is:

$$P(I(x, y, t)) = \eta(I(x, y, t), \mu(x, y), \sigma(x, y)) \quad (3.6)$$

$$\mu(x, y) = \frac{1}{K} \sum_{t=1}^K I(x, y, t) \quad (3.7)$$

Figure 3.4: Gaussian distribution

$$\sigma(x, y)^2 = \frac{1}{K-1} \sum_{t=1}^K (I(x, y, t) - \mu(x, y))^2 \quad (3.8)$$

$$L(x, y, t) = \begin{cases} 1 & \text{if } |I(x, y, t) - \mu(x, y)| \geq N\sigma(x, y) \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$

where $\mu(x, y)$ is the mean and $\sigma(x, y)$ is the variance of the pixel value at position (x, y) , and $\eta()$ is the Gaussian probability density function. Different N determines different statistical thresholds. For example in figure 3.4, $N = 2$ is used.

Figure 3.5 shows the tracking results of background modeling in pixel level by mean and variance. The background reference image is trained by 40 clear background images. The mean and variance of each pixel along the 40 frames are computed. For a new frame, every pixel is compared with the reference image. If its value is within the range of $\mu \pm 2\sigma$, it is considered as background, otherwise it is foreground.

Unlike the previous methods, every pixel has a different threshold value based on the statistical attribute at that position. Thus it is more robust in dealing with small noises. If the background illumination changes slowly over time, the mean and variance of the Gaussian may be updated dynamically over time.

Figure 3.5: Background modeling by mean and variance

3.2.5 Background Modeling (Mixture of Gaussians)

In the presence of repetitive motion and repetitive illumination variation, such as leaves swinging and water fluctuating, the statistical distribution will be multi-modal. A single Gaussian will not be sufficient to model the background. Thus, multiple adaptive Gaussians are needed. One of the representative work is done by Stauffer [71].

Rather than explicitly modeling the values of all the pixels as one particular type of distribution, it models the values of a particular pixel as a mixture of Gaussians. Based on the mean and the variance of each of the Gaussians, it determines which Gaussians correspond to background colors. Pixel values that do not fit the background distributions are considered

Figure 3.6: Background modeling by mixture of Gaussians

foreground until there is a Gaussian that includes them with sufficient evidence supporting to convert it to a new background mixture.

For each pixel at (x, y) position, the pixel history value is $X_1, X_2 \dots X_t$. The probability of observing the current pixel value X_t is:

$$P(X_t) = \sum_{i=1}^K \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (3.10)$$

Where K is the number of Gaussian, $\omega_{i,t}$ is the estimated weight of the i^{th} Gaussian, $\mu_{i,t}$ is the mean and $\Sigma_{i,t}$ is the covariance matrix of the i^{th} Gaussian of the mixture at time t , and $\eta()$ is a Gaussian probability density function. The Gaussians in the mixture model are sorted

in descending order by the value of ω/σ , where ω and σ are the weight and variance of each Gaussian. The Gaussian with larger ω/σ value stands a bigger portion of the mixture model. Thus by sorting, the most likely background remains on the top and the less probable transient background gravitates to the bottom. Decision is made by the following equation:

$$L(X_t) = \begin{cases} 0 & \text{if } \exists i(i = 1, 2..K), |I(X_t) - \mu_{i,t}| \leq N * \sigma_{i,t} \\ 1 & \text{otherwise} \end{cases} \quad (3.11)$$

where $L(X_t)$ is a binary image with 1 representing foreground and 0 background at time t for the current pixel. Every new pixel value is checked against the existing K Gaussian distributions. If none of the K Gaussian matches the pixel value, it is considered as foreground, and the least probable Gaussian is replaced with a Gaussian with the current value as its mean, an initially a high variance, and low prior weight. If a match is found, the pixel fits the model and is considered as background. The μ and σ for unmatched distribution remains the same, but the parameters for the matched Gaussian is updated as follows:

$$\omega_{k,t} = (1 - \alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \quad (3.12)$$

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \quad (3.13)$$

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \quad (3.14)$$

where

$$\rho = \alpha\eta(X_t|\mu_k, \sigma_k) \quad (3.15)$$

is the learning factor for updating current distribution, and α is the learning rate. $M_{k,t}$ is 1 for the model which matched and 0 for the remaining models.

Figure 3.6 shows the tracking results of adaptive background modeling by three Gaussians. In this experiment, Gaussian number $N = 2.5$, $K = 3$, learning rate $\alpha = 0.5$, and ρ is given a constant value: $\rho = 0.05$ to reduce the computation. Compared with the previous algorithms,

3.2 Experiments and Results

39

the noises are significantly reduced. When implemented in Visual C++ 6.0 on a Pentium 4 PC, the algorithm runs smoothly in real time.

Stauffer and Grimson's tracker [71] is robust in dealing with lighting changes, repetitive motion of scene elements, tracking through cluttered regions, slow-moving objects, and introducing or removing objects from the scene.

3.2.6 Comparison

Figure 3.7: Examples of manually labeled foreground

We use foreground detection rate and foreground false positive rate to evaluate the performance of the five algorithm discussed in previous sections.

Foreground detection rate is the percentage of foreground that successfully detected.

$$\text{Detection Rate} = \frac{\text{Number of Pixels Classified as Foreground}}{\text{Number of Foreground Pixels}} \quad (3.16)$$

Foreground false positive rate is ratio of false positive pixels number to the background pixel number.

$$\text{False Positive Rate} = \frac{\text{Number of Pixels Wrongly Classified as Foreground}}{\text{Number of Background Pixels}} \quad (3.17)$$

Figure 3.7 are sample frames of manually labeled ground truth foreground, which is painted in red.

Figure 3.8 show the ROC curve of the five tracking algorithms. Except for temporal differencing, the other four background modeling methods have quite similar ROC curves. Among

Figure 3.8: ROC curve for five tracking algorithms

them, background modeling by mixture of Gaussians slightly outperforms the others. Because the dataset we used here has no obvious illumination variation nor any repetitive background motion, the result by background modeling by mixture of Gaussians does not show obvious advantages. However in noisier conditions, we expect background modeling by mixture Gaussians to show greater advantages. Hence we choose background modeling by mixture of Gaussians in our tracking module.

3.3 Probabilistic Tracking and Human-centric Reference Frame

3.3.1 Probabilistic Foreground-Background Segmentation

Instead of segmenting the foreground by thresholding, we propose a probabilistic framework for soft segmentation.

3.3 Probabilistic Tracking and Human-centric Reference Frame

41

We want to infer the probability that the pixel belongs to foreground $p(L_u=1 | \mathbf{u})$, given the pixel value $\mathbf{u}$.

$$\begin{aligned}
 p(L_u=1 | \mathbf{u}) &= 1 - p(L_u=0 | \mathbf{u}) \\
 &= 1 - \frac{p(\mathbf{u}|L_u=0)p(L_u=0)}{p(\mathbf{u}|L_u=0)p(L_u=0) + p(\mathbf{u}|L_u=1)p(L_u=1)} \\
 &= 1 - \frac{p(\mathbf{u}|L_u=0)p(L_u=0)}{p(\mathbf{u}|L_u=0)p(L_u=0) + p(\mathbf{u}|L_u=1)(1 - p(L_u=0))} \quad (3.18)
 \end{aligned}$$

where $p(\mathbf{u}|L_u=0)$ is the probabilistic distribution of the background pixels, which can be modeled by Gaussian or mixture Gaussians. $p(\mathbf{u}|L_u=1)$ is the foreground distribution. $p(L_u=0)$ is the prior probability that a pixel belongs to background.

If we model the $p(\mathbf{u}|L_u=0)$ by a Gaussian:

$$p(\mathbf{u}|L_u=0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|\mathbf{x}-\mu\|^2}{2\sigma^2}} \quad (3.19)$$

where μ and σ are mean and standard deviation of the Gaussian.

To label if the pixel is background or foreground, the decision is made by:

$$\begin{cases} \text{Foreground} & \text{if } |x - \mu| \geq k\sigma \\ \text{Background} & \text{otherwise} \end{cases} \quad (3.20)$$

When

$$|x - \mu| = k\sigma \quad (3.21)$$

we have:

$$\frac{p(\mathbf{u}|L_u=0)p(L_u=0)}{p(\mathbf{u}|L_u=0)p(L_u=0) + p(\mathbf{u}|L_u=1)(1 - p(L_u=0))} = 1/2 \quad (3.22)$$

With equation 3.22 and 3.21, we get:

$$p(L_u=0) = \frac{p(\mathbf{u}|L_u=1)\sqrt{2\pi}\sigma}{p(\mathbf{u}|L_u=1)\sqrt{2\pi}\sigma + e^{-k^2/2}} \quad (3.23)$$

$p(\mathbf{u}|L_u=1)$ is the distribution of foreground. It could simply be approximated by

$$p(\mathbf{u}|L_u=1) = 1/256 \quad (3.24)$$

when using gray scale images and assuming the distribution of foreground is uniform.

From equations 3.18, 3.19, 3.23, 3.24, $p(L_u=1 | \mathbf{u})$ is finally obtained:

$$p(L_u=1 | \mathbf{u}) = \frac{e^{-k^2/2}}{256e^{-\frac{(x-\mu)^2}{2\sigma^2}} + e^{-k^2/2}} \quad (3.25)$$

3.3.2 Estimating the Human-Centric Spatial Reference Frame

The task here is to estimate the human-centric reference frames from the $p(L_u=1 | \mathbf{u})$ that got from the previous section. The following parameters are to be estimated: the location of the human given by $(\bar{x}, \bar{y})$, and the width and height of the human-centric reference frames given by W and H respectively.

Figure 3.9: (a) Sample probability density frame and color map (b) $\phi(x)$ (c) $\phi(y)$

The $p(L_u=1 | \mathbf{u})$ values of all the pixels in a $m \times n$ frame are projected by the following equation to the X and Y directions respectively:

$$\phi(x) = \sum_{y=1}^m P(L_u=1 | V(x, y)) \quad (3.26)$$

$$\phi(y) = \sum_{x=1}^n P(L_u=1 | V(x, y)) \quad (3.27)$$

as examples shown in figure 3.9(b) (c). $\phi(x)$ and $\phi(y)$ are histograms. To remove the background and out layer noise that effects the center of the box, the first α percent and last α

3.3 Probabilistic Tracking and Human-centric Reference Frame

43

percent of the bins are removed for both $\phi(x)$ and $\phi(y)$. The trimmed histograms are:

$$\tilde{\phi}(x) = \sum_{y=Y_1}^{Y_2} P(L_u=1 | V(x, y)) \quad (3.28)$$

$$\tilde{\phi}(y) = \sum_{x=X_1}^{X_2} P(L_u=1 | V(x, y)) \quad (3.29)$$

The parameters of the human-centric reference frames are thus computed:

$$\bar{x} = \sum_{x=X_1}^{X_2} x \tilde{\phi}(x) \quad (3.30)$$

$$\bar{y} = \sum_{y=Y_1}^{Y_2} y \tilde{\phi}(Y) \quad (3.31)$$

$$\sigma_x = \sum_{x=X_1}^{X_2} \tilde{\phi}(x)(x - \bar{x})^2 \quad (3.32)$$

$$\sigma_y = \sum_{y=Y_1}^{Y_2} \tilde{\phi}(y)(y - \bar{y})^2 \quad (3.33)$$

$$W = \beta \sigma_x \quad (3.34)$$

$$H = \beta \sigma_y \quad (3.35)$$

where β is a constant scale factor to decide the size of the human-centric reference frame.

3.3.3 Experiments

The probability of every pixel belongs to foreground $P(L_u=1 | \mathbf{u})$ is computed by (3.25). Sample frames and their corresponding probability density images are shown in figure 4.9(a),(b).

The σ of Gaussian kernel used in (3.19) is $2\sqrt{\sigma_x \sigma_y}$. Figure 4.9(c) shows the foreground segmentation result of the sample images. More examples are shown in figure 3.11. The person is well-centered in the box, which shows that the method is robust to outliers, such as hands and legs.

Figure 3.10: (a) Sample frames of walking, and images representing the probability density value of $P(L_u=1 | \mathbf{u})$ (b) Sample frames of waving, and images representing the probability density value of $P(L_u=1 | \mathbf{u})$ (c) Cropped images of human centered in the box

Figure 3.11: Sample tracked images

3.4 Multiple People Tracking

In order to deal with the video of multiple persons in the scene, a multiple people tracking module is implemented. Each person is ideally placed in one bounding box, and trajectory of every person is recorded. Additionally, while we are concerned with the pixel data within the bounding box, we do not need to recover explicit motion of detailed human body parts, such as heads and hands. Areas in the boxes are our regions of interest. Based on these boxes, the trajectories of multiple occupants can be directly computed, and the motion pattern of each person can be analyzed. This analysis is very important for our subsequent work in activity understanding.

Figure 3.12 illustrates our tracking module framework. From our previous discussion, we found that adaptive background modeling of mixture of Gaussians has advantages in detecting moving foreground over other methods. So we adopt it to detect the foreground regions. It is implemented by Visual C++ 6.0. The parameters used are: $N = 2.5$, number of Gaussians $K = 3$, learning rate $\alpha = 0.5$, and learning factor ρ is given a constant value: 0.05.

Based on the detected regions, connected components analysis is carried out. One set of connected pixels is considered as one region. All the foreground regions are firstly detected and each of them is put into a bounding box separately. The boxes smaller than 2000 pixels are discarded, because they are likely to be noise. The remaining bounding boxes represents the foreground areas. Figure 3.13 is a frame of tracked foreground in bounding boxes. Four separate bounding boxes are drawn, which capture the four persons.

Each person should be identified by one labeled bounding box. In noisy situations, foreground regions may not be properly connected, so one person may be covered by 2 or more bounding boxes. When people interact with each other, two bounding boxes will merge to form one and subsequently split. These are difficult research issues. In our future work, we will investigate robust frameworks for identifying occupants by analyzing the trajectories and area histories of the bounding boxes.

Figure 3.12: Tracking module

Figure 3.13: An example of tracked foreground in bounding boxes

Chapter 4

Human Action Classification with Ambiguous Correspondences

In the computer vision community, analysis of actions has mostly focused on the recognition of predefined actions, *e.g.* the works of Cutler and Davis [25] and Polana and Nelson [61] are designed to detect periodic activities. These methods are capable of capturing the characteristics of these events by specialized parametric models with a small number of parameters and achieve good recognition of the studied actions. However, real-world applications are unlikely to be restricted only to recognition of pre-studied actions. When dealing with general video data like movies, often there is no prior knowledge about the types of events, their temporal and spatial extent, and their nature. So unsupervised and adaptive methods for action classification are required, based on a less-specialized and less restrictive approach to action modeling.

In this chapter, we first study the different definitions of action and event. The limitation of Zelnik-Manor and Irani [82] is then discussed. After a principle analysis of correspondence problem, we proposed a combined tracking-classification framework for the unsupervised classification that outperforms Zelnik-Manor and Irani's work.

4.1 Definition of Action

One of the key challenges in this problem is in defining actions and activities. One possibility is to manually assign action labels to short video sequences that capture the actions. The prob-

lem then becomes that of recognizing future observations based on similarity to these labeled action exemplars. This approach is good for applications such as interpreting the American Sign Language [81], where actions are deliberate and well-defined (even though there may be substantial variation in different instances of the same action). However other more casual forms of actions are often ill-defined, and manual labeling is highly subjective.

A more recent approach is to assume that the interesting human actions are those that *recur*. The challenge is then to be able to effectively *cluster* subsequences of video into categories of similar actions. The advantage of such an approach is that the availability of a data-labeled training set is only optional, as we can resort to unsupervised learning of these clusters.

In Zelnik-Manor and Irani [82], an event is regarded as a temporal object. It is defined as a stochastic temporal process, which is characterized by spatio-temporal features at multiple temporal scales. The term *event* is the same as *action* as used in this thesis. They are both equivalent to *intermediate actions* in our three level understanding research framework.

4.2 Discussion of Zelnik-Manor and Irani's Work

Zelnik Manor and Irani [82] presented an event-based distance measure that can be used to cluster and isolate events within long continuous video sequences, as was reviewed in section 2.2.3. This method is non-parametric and can handle a wide range of dynamic events without prior knowledge. However, it is inferior in accuracy to the more specialized parametric models and cannot analyze more complicated events with multiple dynamic objects. It also uses the spatial information in an ineffective way, so it cannot discriminate spatially different actions with similar temporal patterns.

4.2.1 Spatial and Temporal Information

Motion is a strong cue to discriminate different actions. To analyze motion, there are two keys: spatio-temporal derivatives which are related to the magnitude of motion, and relative spatial

displacements which indicate where the motion happens. Together they provide integrated information to represent actions.

Different actions that have the similar patterns of motion magnitude can often be distinguished by different spatial displacement. For example, a person waving his hands over his head and waving his hands in front of his chest may have similar motion magnitudes, and the difference lies only in the relative locations of where the motion happens.

Conversely different actions may exhibit the same spatial qualities, but different motion magnitudes, *e.g.* people walking at different speeds.

The space-time feature measurement equation in Zelnik Manor and Irani [82]'s work is constructed as follows:

$$(N_x^l, N_y^l, N_t^l) = \frac{(S_x^l, S_y^l, S_t^l)}{\sqrt{((S_x^l)^2 + (S_y^l)^2 + (S_t^l)^2)}} \quad (4.1)$$

Equation 4.1 essentially build distributions of gradient value at x, y, t directions. N_t^l captures the characteristics of the motion magnitude, but ignores the information of where the motion happens. Additionally, distributions of N_x^l, N_y^l capture local spatial texture properties, but do not contain information of where the motion happens. The measurements will not discriminate the actions of a person waving hands over its head and waving hands in front of his chest at the same speed.

By searching new measurement that combines the spatial and temporal information in a more effective way, better discrimination can be achieved.

4.2.2 Effect of Background

The key of event based video segmentation is to discriminate different actions or events by the statistical distributions of their motion patterns. We consider the impact that the background has on these distributions that affect the discriminability.

Figure 4.1: Top three figures are sample frames of three different events: running, waving and walking [82]. Middle three are empirical distributions N_t^1 of the whole frames (a) Running (b) Waving (c) Walking. Bottom three are empirical distribution of N_t^1 for the foreground only: (d) Running (e) Waving (f) Walking

In cases when the foreground motion areas are too small compared with the static background, the distributions of N_x^l , N_y^l , N_t^l are dominated by contribution from the background. Figure 4.1 (a),(b),(c) shows the running, walking and waving distribution of N_t^1 of the whole frame computed by (4.1). Despite the different actions, the distributions appear very similar.

Since we are only concerned with foreground motion, the unnecessary inclusion of background regions simply results in unwanted noise. The approach that we take is to apply the

foreground segmentation of tracking module first, before applying the event clustering algorithm on the detected foreground regions. This results in much better discrimination. Figure 4.1(d),(e),(f) shows the running, walking and waving distributions of N_t^1 based on only foreground regions computed by (4.1), and the differences are significant.

4.3 Spatial Correspondence in Action Classification

Many existing methods in action classification may be divided into two categories:

Figure 4.2: Sample result of a Correspondence-Exact method [60]

Figure 4.3: Sample result of a Correspondence-Free method [11]

- *Correspondence-Exact*. Methods in this approach assume or initially acquire correspondences between a spatial (often kinematic) model and human body parts in the video through tracking. Motion analysis and classification is applied to the high-level tracked

states rather than image features. While the methods may assume that there are errors in estimating correspondences, the use of model state space implies that estimates of the correspondences are always assumed to be available. Pavlovic *et al.* [60] proposed an approximate Viterbi inference technique in dynamic Bayesian networks using human kinematic models, as picture shown in figure 4.2. Other related methods were proposed by Alon *et al.* [6], Bregler [14], Stauffer and Grimson [72], and Haritaoglu *et al.* [41].

- *Correspondence-Free.* Methods in this approach directly analyze lower-level features extracted from video. Laptev and Lindeberg [50] investigated several spatio-temporal image descriptors for action representation. Generally, these methods adopt a pattern recognition approach to evaluate these features. No direct correspondences to human body parts are assumed or computed. Methods falling into this category are proposed by Zelnik Manor and Irani [82], Davis and Bobick [28], Cutler and Davis [25], Chomat and Crowley [19], Bobick and Davis [11], Bobick and Wilson [12]. An example image is shown in figure 4.3.

Spatial correspondences provide a much richer and relevant representation (the relevant works are reviewed in section 2.5). However, these correspondences may be estimated inaccurately, with errors in localizing image features leading to errors in computing the model state. Furthermore, the correspondences may be *ambiguous*, *e.g.* there may be a multitude of alternative image features for correspondence, each of which is associated with a different probability. Nevertheless, even having this weak correspondence information can help distinguish between the two hand waving actions as seen in figure 4.4.

In order to exploit weak correspondences, Davis [27] proposed a probabilistic action inference framework making use of motion-history images. Other methods involve dividing the foreground region into a number of spatial partitions to provide some form of approximate region-wise correspondence in order to better analyze human motion. Representative works

Figure 4.4: Correspondences, even if ambiguous, can be exploited to distinguish the two waving actions.

are proposed by Polano and Nelson [62], Takahashi and Seki [73], Yamato and Ohya [80]. The partitioning is done without deeper analysis.

Generally, these methods proposed intuitive, but nevertheless *empirically constructed* methods to use weak correspondences. In comparison, we have a principled framework that formally uses the probability of correct correspondence in defining action representation.

4.4 Overview of Action Classification Framework

A principled framework is proposed for action classification as shown in figure 4.5. In the first stage, input video is processed by a simple but fairly inaccurate tracking module combined with simple probabilistic foreground-background segmentation based on background modeling and subtraction, and then our novel probabilistic representation called CAFHA (described later) is constructed based on the cropped, human-centric image where each pixel additionally includes a probability that it is part of the foreground (*i.e.* it is a point on a person). In the second stage, action classification is done in both steps of unsupervised learning and action inference. The output is an action-labeled video sequence.

In this chapter, we take a mixed approach that exploits the benefits of correspondences, but also assume that they may be ambiguous. Various elements in the Zelnik-Manor and

Figure 4.5: Action Classification Framework

Irani's [82] framework, which is correspondence-free, are adopted in our framework conjunction but enhanced with these ambiguous correspondences.

A novel action representation: Correspondence-Ambiguous Feature Histogram Array (CAFHA) is proposed and applied in unsupervised learning of video sequences, and online human action inference.

Our work is focused on classifying the actions of a single person. While this framework is able to be extended to cope with multiple persons in a video, it does so by tracking each person

individually and does not handle interactions between people.

4.5 Basics

We consider an *action* to be characterized by a sequence of image appearance changes that are relative to a human-centric spatial reference frame in image space, and also relative to an action-centric temporal reference frame.

The human-centric spatial reference frame has an origin at $(\bar{x}, \bar{y})$ located at the projected center of mass of the person, and includes horizontal and vertical scaling factors s_x and s_y that normalize the projected image height and maximum projected shoulder width of the person to unity. The scaling factors are intended to make the event classification framework invariant to both the shape profile of the human subject as well as to the distance of the camera. The action-centric reference frame has a temporal origin at $\bar{t}$ which is the mid-point of the action timespan, and a temporal scale s_t which is used for normalizing any minor difference in the speeds of the actions. See figure 4.6 for an illustration.

Figure 4.6: Action Reference Frame

4.6 Action Representation

In order to address the issue of representing human action in a video sequence, the approach taken in this chapter involves the following steps:

- (i) Propose a feature-based representation that is ideal under contrived assumptions of actions that are invariant and formulize this ideal action A in canonical reference coordinate $\mathbf{x}'$.
- (ii) Describe how this representation is affected by more realistic conditions resulting in action variation.
- (iii) Propose an alternative probabilistic action representation resulting from the above analysis. Detailed analyse of different terms that effects the representation is conducted after mathematically decompose the probabilistic action representation.

4.6.1 Feature-Set Action Representation

Consider an ideal world where all actions (and associated motion) in the same category are *exactly* the same, and there are no observation errors. In such a world, an action A can be represented by a set of image feature measurements at all foreground pixels in a video volume of timespan T

$$A = \left\{ (\mathbf{x}', \mathbf{f}) \mid L_{\mathbf{x}'} = 1, |t'| < \frac{T}{2} \right\} \quad (4.2)$$

where

$$\mathbf{x}' = [x' \ y' \ t']^T$$

are transformed versions of spatial-temporal coordinates $[x \ y \ t]^T$ with respect to the above-mentioned reference frames

$$x' = s_x(x - \bar{x}), \quad y' = s_y(y - \bar{y}), \quad t' = s_t(t - \bar{t}); \quad (4.3)$$

Here we define $\mathbf{f}$ as an *appearance-invariant* local feature measurement at $\mathbf{x}'$ and $L_{\mathbf{x}'}$ as a foreground-background pixel label in canonical human-centric reference frame:

$$L_{\mathbf{x}'} = \begin{cases} 1, & \text{if } \mathbf{x}' \text{ is a foreground pixel} \\ 0, & \text{if } \mathbf{x}' \text{ is a background pixel} \end{cases}$$

The timespan T is a subjectively assigned value that ultimately defines the period of the action. $|t'| < \frac{T}{2}$ means that the current action starts from $-\frac{T}{2}$ till $\frac{T}{2}$ in the t direction of the canonical coordinate. The local features are not specified at this stage as our framework is independent and able to make use of different types of features. The actual features used in our implementation are described in section 4.7.3.

Once actions are defined using such image feature sets, observed actions can be compared in this ideal world simply by verifying whether the feature sets are identical.

4.6.2 Action Variation and Observational Errors

Moving towards more realistic conditions, there are three sources of variation that affect this action representation:

- natural human variation in the action that affect the distribution of $(x', y', t', \mathbf{f})$;
- errors in estimating the reference frames¹, affecting the measurement of $(\bar{x}, \bar{y}, \bar{t})$; and
- errors in foreground-background segmentation that affect the labels $L_{\mathbf{x}'}$.

We will however make a *key assumption* in this chapter: that the human action variation is predominantly due to physical motion and displacement rather than other forms of variation. This form of variation results in spatial-temporal *displacements* in the feature measurements, rather than direct changes in the feature values $\mathbf{f}$ themselves. Additionally, errors in estimating the reference frames also results only in displacement changes.

Based on the above assumption, the problem of action comparison is a *correspondence problem*. If we are able to establish exact pixel-wise correspondences between the foreground regions of two video sequences of the same action, we will find that the feature measurements of corresponding pixels are close.

¹In our work, we will ignore errors in estimating s_x and s_y ; additionally, we use an arbitrarily fixed s_t .

4.6.3 Correspondence Ambiguity

In realistic conditions, we are unlikely to be able to establish exact *pixel-wise* correspondences between video sequences of similar actions as inferring these correspondences is generally an ill-posed process. On the other hand, we are likely to be able to establish approximate *region-wise* correspondences. As will be shown in subsequent analysis, these partially ambiguous region correspondences are nevertheless useful in recognizing similar actions and discriminating different actions.

Suppose we observe an action $\hat{A}$ as

$$\hat{A} = \left\{ (\mathbf{u}, \mathbf{f}) \mid L_{\mathbf{u}} = 1, |w| < \frac{T}{2} \right\} \quad (4.4)$$

where

$$\mathbf{u} = \begin{bmatrix} u & v & w \end{bmatrix}^T$$

are the observed spatial-temporal coordinates in the human-centric reference frames. Based on the correspondence problem assumption, we can specify whether (u, v, w) corresponds to (x', y', t') through the use of a correspondence label $C_{\mathbf{u}, \mathbf{x}'}$ given by

$$C_{\mathbf{u}, \mathbf{x}'} = \begin{cases} 1, & \text{if } \mathbf{u} \text{ corresponds to } \mathbf{x}'; \\ 0, & \text{otherwise} \end{cases} \quad (4.5)$$

For cases where the correspondence is true and $C_{\mathbf{u}, \mathbf{x}'}=1$, we can treat $\mathbf{u}$ as a perturbation of the canonical coordinates $\mathbf{x}'$, i.e. $\mathbf{u} = \mathbf{x}' + \Delta \mathbf{x}'$. However as we are not able to evaluate exact correspondences, $C_{\mathbf{u}, \mathbf{x}'}$'s are true but *hidden* labels.

4.7 Probabilistic Analysis of Features

Recall the assumption made in section 4.6.2 that human action variation is predominantly due to physical motion and displacement rather than other forms of variation. An immediate result of the assumption is that the *distribution* of the foreground feature measurements expressed

in the canonical coordinates $\mathbf{x}'$ changes little, *i.e.* $p(\mathbf{x}', \mathbf{f}, L_{\mathbf{x}'}=1, C_{\cdot, \mathbf{x}'}=1)$ remains relatively constant. However, since image observations are in perturbed $\mathbf{u}$ coordinates and the correspondences are ambiguous, this distribution needs to be estimated from the observations.

Expanding the above term as a marginal over $\mathbf{u}$:

$$\begin{aligned}
 & p(\mathbf{x}', \mathbf{f}_{\mathbf{x}'}, L_{\mathbf{x}'}=1, C_{\cdot, \mathbf{x}'}=1) \\
 &= \int p(\mathbf{f}_{\mathbf{x}'}, L_{\mathbf{x}'}=1, C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, L_{\mathbf{u}}, \mathbf{x}') d\mathbf{u} \\
 &= \int \left\{ p(\mathbf{f}_{\mathbf{x}'} \mid L_{\mathbf{x}'}=1, C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \cdot \right. \\
 &\quad \left. p(L_{\mathbf{x}'}=1 \mid C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \cdot \right. \\
 &\quad \left. p(C_{\mathbf{u}, \mathbf{x}'}=1 \mid \mathbf{u}, \mathbf{x}') p(\mathbf{u}, \mathbf{x}') \right\} d\mathbf{u} \\
 &= \int \left\{ p(\mathbf{f}_{\mathbf{u}} \mid L_{\mathbf{x}'}=1, C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \cdot \right. \\
 &\quad \left. p(L_{\mathbf{u}}=1 \mid C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \cdot \right. \\
 &\quad \left. p(C_{\mathbf{u}, \mathbf{x}'}=1 \mid \mathbf{u}, \mathbf{x}') p(\mathbf{u}, \mathbf{x}') \right\} d\mathbf{u} \tag{4.6}
 \end{aligned}$$

Some of the probability terms in (4.6) have strong relevance to various modules in our proposed video representation framework, and are discussed below:

- **The histogram array term:** $p(\mathbf{x}', \mathbf{f}, L_{\mathbf{x}'}=1, C_{\cdot, \mathbf{x}'}=1)$. This term is the feature probability distribution in canonical coordinates. It can be visualized as a 3D array aligned with the canonical $\mathbf{x}'$ coordinates, where each bin contains a *histogram* of foreground feature values. It is taken as approximately invariant within each category of action, this forms the basis for our new action representation under realistic conditions.
- **The feature observation term in observed coordinate:** $p(\mathbf{f}_{\mathbf{u}} \mid L_{\mathbf{x}'}=1, C_{\mathbf{u}, \mathbf{x}'}=1, \mathbf{u}, \mathbf{x}')$. For a single pixel $\mathbf{u}$ in an observed image $\hat{A}$, this term is simply a delta function histogram corresponding to observed feature value at $\mathbf{u}$.

- **The feature observation term in canonical coordinate:** $p(\mathbf{f}_{\mathbf{x}'} \mid L_{\mathbf{x}'}=1, C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}')$.

Given the condition:

$$C_{\mathbf{u},\mathbf{x}'}=1 \quad (4.7)$$

which means that: when the point in $\mathbf{u}$ is corresponded to the point in $\mathbf{x}'$, the feature observation term in observed coordinate and canonical coordinate are equivalent for the pixel:

$$p(\mathbf{f}_{\mathbf{x}'} \mid L_{\mathbf{x}'}=1, C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') = p(\mathbf{f}_{\mathbf{u}} \mid L_{\mathbf{u}}=1, C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \quad (4.8)$$

- **Prior foreground observation term:** $p(L_{\mathbf{u}}=1 \mid C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}')$ It is the prior probability that the points in $\mathbf{u}$ being foreground. Under the condition of (4.7), the probability of the point being foreground in observed coordinate and the corresponding point in canonical coordinate are the same:

$$p(L_{\mathbf{u}}=1 \mid C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') = p(L_{\mathbf{x}'}=1 \mid C_{\mathbf{u},\mathbf{x}'}=1, \mathbf{u}, \mathbf{x}') \quad (4.9)$$

- **The prior correspondence ambiguity term:** $p(C_{\mathbf{u},\mathbf{x}'}=1 \mid \mathbf{u}, \mathbf{x}')$. This term indicates the prior probability of correspondence based only on relative spatial positions of $\mathbf{u}$ and $\mathbf{x}'$. It is further elaborated in the following section.

4.7.1 Discussion of the Prior Correspondence Ambiguity Term

The relative poses of limbs and other body parts are important factors of variation for human actions. Unlike most of the correspondence-exact and correspondence-free methods, we propose an ambiguous correspondence method, which empirically constructs the correspondence information without the need to accurately locate body parts in the images. In particular, we address the question: what is the probability that a point in the observed $\mathbf{u}$ coordinates and a point in the hidden reference $\mathbf{x}'$ coordinates are corresponding points?

In order to circumvent the need to find accurate correspondence, the approach here is to adopt the simple but somewhat crude Gestalt principle of proximity: the larger the relative

distance between the two points, the lower the probability of correspondence. A simple way to model the prior correspondence ambiguity term $p(C_{\mathbf{u},\mathbf{x}'}=1 \mid \mathbf{u}, \mathbf{x}')$ is to use a Gaussian kernel function:

$$p(C_{\mathbf{u},\mathbf{x}'}=1 \mid \mathbf{u}, \mathbf{x}') = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|\mathbf{x}'-\mathbf{u}\|^2}{2\sigma^2}} \quad (4.10)$$

It quantifies our prior assumption that proximity is an indicator of correspondence.

In this chapter, the Gaussian function parameters are empirically obtained. In chapter 5, a more principled approach is taken to determine the optimal parameters.

4.7.2 Construction of Correspondence-Ambiguous Feature Histogram Array (CAFHA)

The probability distribution $p(\mathbf{x}', \mathbf{f}, L_{\mathbf{x}'}=1, C_{\cdot,\mathbf{x}'}=1)$ is however extremely expensive to store and compute as it involves maintaining separate feature histograms in each bin of a 3D array in $\mathbf{x}'$ coordinates. For practical purposes, only a small number of feature histograms are maintained as an approximation to this distribution. These histograms may be specified by heavily subsampling the 3D array. We call this set of histograms the *Correspondence-Ambiguous Feature Histogram Array (CAFHA)*.

An direct intuition is to uniformly sample a few points. We usually sample 9 points to construct a naive CAFHA.

Figure 4.7 shows an example where the CAFHA involves estimating feature histograms at nine locations in the human-centric reference frame. A Gaussian kernel is used at each location to represent the correspondence ambiguity. The final weight of the feature pixel that will contribute to the histogram in the relevant bin is

$$W(\mathbf{u}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|\mathbf{x}'-\mathbf{u}\|^2}{2\sigma^2}} p(L_{\mathbf{u}}=1 \mid \mathbf{u}) \quad (4.11)$$

The CAFHA is used as our feature representation as it has the desirable properties of being fairly invariant under realistic conditions, and also able to exploit ambiguous correspondences.

Figure 4.7: Construction of the Correspondence-Ambiguous Feature Histogram Array (CAFHA)

Interesting questions that are not addressed here include: how many points are needed in subsampling? Which are the best points to sample? How to best model the ambiguity? These questions will be discussed in chapter 5.

4.7.3 CAFHA Implementation

The CAFHA is practically computed in the following way:

- Assume that we have a 2D array of feature vector histogram indices F and a 2D array S of foreground probabilities. These are image-sized arrays obtained from segmentation and feature measurements (discussed in later sections).
- Define 2D array A containing prior correspondence ambiguity values for relative dis-

placements of $\mathbf{u}$ and $\mathbf{x}'$. In our case, this simply contains values from a Gaussian kernel.

- Create CAFHA array $\mathbf{C}$ of size W_x by W_y by W_t (size values are number of samples per spatio-temporal dimension). Each array element contains a 1D histogram array of length M – the number of histogram bins for values of $\mathbf{f}$.
- Run the following algorithm:

begin

for all tuples (x', y', t') index to array $\mathbf{C}$ **do**

for all each tuple (u, v, w) **do**

$\mathbf{h}$ = histogram at $\mathbf{C}[x'][y'][t']$

i = index obtained from $\mathbf{F}[u][v][w]$

$h[i] = h[i] + S[u][v][w] * A[x'-u][y'-v][t'-w]$

end for

end for

end

In our implementation, our CAFHA array was $3 \times 3 \times 1$ sized, in which each array element contains a feature value histogram of 18 bins.

4.8 Action Classification

Given the foreground probabilities as well as estimates for the human-centric reference frame in all images, the second stage uses the CAFHA as an action representation for different subsequences, and attempts to cluster these subsequences.

4.8.1 CAFHA Distance Measure

Motion estimated on the original sequence is the primary information we use. To get motion at multiple temporal scales, a temporal pyramid $S^1, S^2, S^3, \dots, S^L$ of the sequence is built by

blurring and sub-sampling the sequence along the temporal direction. It is the same as the work of Zelnik-Manor and Irani [82], which is described in section 2.2.3.

The measurement we used is normalized temporal gradient at different temporal scales:

$$N_t^l = \frac{|(S_t^l)|}{\sqrt{((S_x^l)^2 + (S_y^l)^2 + (S_t^l)^2)}} \quad (4.12)$$

We do not use spatial gradient, which is slight different from Zelnik-Manor and Irani [82]. This is because the spatial information has already been contained in the correspondence ambiguity term, and experiments has shown that temporal gradient alone is enough in discriminating different actions. Using the CAFHA representation, each action is associated with $M = (W_x)(W_y)(W_t)$ histograms. To measure the difference between two actions a and b , which are represented by the M CAFHA histograms respectively, the χ^2 distance is used:

$$D^2 = \frac{1}{M} \sum_m \sum_i \frac{[H_a^m(i) - H_b^m(i)]^2}{H_a^m(i) + H_b^m(i)} \quad (4.13)$$

where D is the distance, $H^m(a)$ is the m -th histogram in the CAFHA of action a , and likewise for action b .

4.8.2 Computing the Affinity Matrix

A temporal sliding window demarcates a subsequence of length $T1$ each step and traverses $T2$ frames between steps within the whole video sequence. It generates a lot of subsequences. The affinity matrix $\mathbf{D}$ is constructed, whose entry $D_{i,j}$ is the distance between sub-sequence i and sub-sequence j computed by 4.13, and τ is a scale factor. We adopt a spectral clustering [58] method to cluster the affinity matrix.

4.8.3 Spectral Clustering

CAFHA distance measure is further used to clustering the actions in the video sequence by first computing the affinity matrix and then carrying out clustering [58] on the affinity matrix.

The cluster of actions can be directly obtained from the clustered affinity matrix by clustering methods. The spectral clustering method [78, 70, 58], which uses the eigenvectors of an affinity matrix to obtain the clustering of the data, is applied here. Several algorithms have been proposed. Zelnik-Manor and Irani [82] used Shi *et al.*'s normalized-cut method [70] to cluster the data. We adopted Ng *et al.*'s spectral clustering [58] method, because it is effective and easier to implement in Matlab. The two methods are actually equivalent, and the only difference between the two is that they use eigenvectors in slightly different ways.

Following the distance measure of previous section, the event clustering is done via the following steps.

- Step 1. Construct the affinity matrix $\mathbf{A}$.

A sliding temporal window is used to compare every sub-sequence of length T to all other sub-sequences of the same length within the long video sequence. The affinity matrix $\mathbf{A}$ is constructed whose entries

$$A_{i,j} = \exp[-D_{i,j}^2/\sigma] \quad (4.14)$$

where $D_{i,j}$ is the distance between sub-sequence i and sub-sequence j computed by 2.2, and σ is a constant scale factor.

- Step 2. Define $\mathbf{D}$ to be the diagonal matrix whose (i, i) -element is the sum of $\mathbf{A}$'s i^{th} row, and construct the matrix $\mathbf{L}$ by

$$\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2} \quad (4.15)$$

Find $x_1, x_2, \dots, x_k$, the k largest eigenvectors of $\mathbf{L}$, and form the matrix $\mathbf{X} = [x_1, x_2, \dots, x_k] \in R^{n \times k}$ by stacking the eigenvectors in the columns.

- Step 3. Form the matrix $\mathbf{Y}$ from $\mathbf{X}$ by re-normalizing each of $\mathbf{X}$'s rows to have unit length:

$$Y_{i,j} = \frac{X_{i,j}}{\sum_j (X_{i,j}^2)^{1/2}} \quad (4.16)$$

- Step 4. Treating each row of $\mathbf{Y}$ as a point in R^k , cluster them into k clusters via K-means algorithm [7]. Assign the i -th sub-sequence to cluster j , if and only if row i of the matrix $\mathbf{Y}$ is assigned to cluster j .

4.9 Human Action Learning and Online Inference

In this section, the CAFHA representation and unsupervised action classification is extended to online action inference. The following challenge is addressed: Given two video sequences A and B in which similar actions are carried by different people, unsupervised learning is to be carried out in an offline manner in A , followed by inferring the learned actions in B online.

Figure 4.8: Framework of action learning and inferring

Figure 4.8 shows the framework of action learning and inference: the short sample video A is first clustered and labeled, and the standard CAFHA representation for each clusters is extracted by computing the mean representation of each cluster. For every subsequence in long video sequence B , the corresponding CAFHA representation is constructed, and the CAFHA distances to all cluster means are computed. The class label is assigned via nearest mean classification. The online inference portion is carried out in near real time.

4.10 Experiments

4.10.1 Probabilistic Segmentation and Tracking

Figure 4.9: Sample frames of walking, and images representing the probability density value of $P(L_u=1 \mid \mathbf{u})$

Sample human-centric reference frames and their corresponding probability density images are shown in figure 4.9. It is clear that both the tracking and the segmentation are subject to substantial errors. However, this level of information is sufficient to be exploited in our action classification framework.

4.10.2 Action Classification

4.10.2.1 Experiment 1

Figure 4.10: Sample frame of (a) walking fast (b) running (c) walking at medium speed

A sequence of 395 frames is used in this experiment. Three similar actions: running, walking fast and walking at medium speed occurred in the sequence, as shown in figure 4.10.

Figure 4.11: Clustering results by (a) our method (b) Zelnik-Manor and Irani's method and (c) ground-truth (Blue represents walking fast, red represents running, and cyan represents walking at medium speed)

A sliding window of 48 frames is applied to the entire sequence. The sliding window demarcates a subsequence each step and traverses 5 frames between steps. Figure 4.11 shows the comparison of final clustering results. Zelnik-Manor and Irani's [82] method² is unable to discriminate between walking fast and running, while our method is able to do so. The likely reason for this is that walking and running at similar speeds generates similar distributions of spatio-temporal features. However the motions result in different spatial displacements – *e.g.* arms are bent when running, but straightened when walking. By incorporating probabilistic correspondence measures within our CAFHA representation, these two similar actions can be discriminated.

4.10.2.2 Experiment 2

A sequence of 370 frames is used in this experiment. Four different actions occurred in the sequences, including two similar actions: a person walking fast and walking slowly, and two other similar actions: hand waving over head and hand waving in front of body, as shown in figure 4.12. A sliding window of 24 frames is applied and 5 frames are traversed between

²Experiments on the dataset from [82] are not presented here, because results of our method and Zelnik-Manor and Irani's method on those datasets are the same. Experiments are done on new datasets that highlight the difference in performance of the methods.

4.10 Experiments

69

Figure 4.12: Sample frame of (a) walking fast, (b) walking slowly, (c) waving over head, (d) waving in the front of body.

Figure 4.13: Clustering results by (a) our method, (b) Zelnik-Manor and Irani's method, (c) ground-truth (Red represents walking fast, yellow represents walking slowly, blue represents waving over head and cyan represents waving in front of body)

steps. Experimental results are compared with Zelnik-Manor and Irani's method. Figure 4.13 shows the final clustering results. Hand waving over head and waving in front of body generate very similar spatiotemporal features. As before, Zelnik-Manor and Irani's method is unable to discriminate these actions, while our method is able to.

Figure 4.14: Sample frame of walking fast, waving over head and waving in front of body by three different people

Figure 4.15: Unsupervised clustering results by (a) our method (b) Zelnik-Manor and Irani's method (c) ground-truth (yellow represents walking , red represents waving over head and blue represents waving in front of body)

Figure 4.16: (a) Inference result of multiple person actions based on learned CAFHA representation, (b) Ground-truth information

4.10.3 Action Learning and Inference

A sequence of 753 frames, including a person walking, waving hand over head and waving hand in front of body, as shown in figure 4.14(a). A sliding window of length 24 frames is applied and traverses 5 frames between steps. The clustering results by unsupervised learning are shown in figure 4.15. Post-clustering labels are manually assigned to the clusters.

In the inference stage, a long sequence of 1240 frames is used, including 2 different people doing three different actions, as shown in figure 4.14(b)(c). The standard CAFHA representation of each action is built by the mean of randomly picked 5 sample subsequences of each action and compared with the long sequence in the way that is shown in figure 4.8. The inference results are compared with the ground-truth shown in figure 4.16.

The frame size of the sample video is 200×100 and the sliding window of length 24 frames. When running in MatlabTM on a Pentium M 1.5 GHz personal computer, it takes on average 1.86 seconds to obtain the classification label for each subsequence.

Action	Total no. of frames	Detection rate	False +ve rate
Walking	366	100.0%	2.7%
Waving over-head	459	98.3%	11.2%
Waving in front of body	415	83.1%	0.7%

Table 4.1: Results for inference of different actions

Table 4.1 shows the statistical results of the experiment. A few frames at the boundary of two actions are not counted in because they are transition actions that do not belong to neither actions. Action of walking achieves best detection rate and false positive rate because it is much different from the other two actions, which are similar to each other.

4.11 Comparison with Zelnik-Manor and Irani's Method

Zelnik-Manor and Irani's method [82] of representing action can be considered as an instance of our principled framework, when treating the whole image as foreground, modelling the correspondence ambiguity by one gaussian of infinite large variance, and constructing the histogram by sub-sampling one point at x' coordinate.

CAFHA bears many of the attributes of Zelnik-Manor and Irani's method, such as being invariant to the appearance, size, location of the person, and the image background. Moreover it outperforms their method in the following aspects: First, CAFHA conserves spatial information, which can be utilized to discriminate similar actions such as waving above head and waving in front of the body. Second, the current action classification framework can be easily extended to the case of multiple people, because CAFHA has a concept of foreground and background. Third, the CAFHA representation is computationally more efficient, because the computation is mainly focused on the cropped human-centric reference frames and requires fewer features. It makes possible for online real time video classification, which will be discussed in the next section.

Currently there is no universal powerful tracker can extract perfect foreground information. Due to background and any kinds of errors, the method of using probabilistic foreground segmentation to get human centric frames is not deemed to be accurate. However, the human-centric reference frames acquired by the coarse method are favorably sufficient for ambiguous correspondence analysis. More sophisticated methods like background modeling by mixture Gaussians are not necessary.

4.12 Conclusion

We described a multi-stage framework for the unsupervised classification of human actions, involving separate modules for foreground segmentation, tracking and classification. A principled analysis of the problem leads to a novel probabilistic action representation called the

correspondence-ambiguous feature histogram array (CAFHA) that is robust to variations across similar actions. This representation allows for ambiguous correspondences as well as probabilistic segmentation to be exploited. Results were reported showing that the new framework outperforms the Zelnik-Manor and Irani's method [82] for unsupervised action classification. Additionally, the clustered means were used for quasi real-time action inference with good accuracy despite changes in person identity and variations in the actions.

Chapter 5

Feature Selection and Optimization for Unsupervised Action Classification

Feature selection is concerned with choosing good subsets of measurement of a given classification task. Selecting and combining discriminative measures to maximize the overall classification performance [21] is a timely topic in computer vision.

John *et al.* [46] defined the notions of relevance and irrelevance for understanding of the behavior of subset selection algorithms. Levi *et al.* [52] showed that performance of their face detection system largely depends on the features that are used, and the local edge orientation histograms (EOH) outperform most other linear features. Zhou *et al.* [85] proposed a new perspective of feature selection for active recognition, which takes into account feature interaction issues. Most of the methods are focused on supervised learning. Chang *et al.* [16] proposed a Bayesian algorithm for unsupervised feature selection and density estimation framework by minimizing the model structure complexity and estimating model parameters. Shan *et al.* [69] formulated vehicle matching as a class classification problem and proposed a novel unsupervised learning method to learn the feature measurement for discrimination without labeled data. There are also lots of works about detecting the salient features in spatial temporal domain for action recognition, such as SIFT [53] and spatial-temporal interest points [49].

In the previous chapter, the concept of the correspondence-ambiguous feature histogram array (CAFHA) that is robust to variations across similar actions is discussed. CAFHA is

constructed by heavily subsampling the 3D histogram array in $\mathbf{x}'$ coordinates. A simple and direct subsampling of uniformly choosing 9 points is used in our implementation, which we call *naive CAFHA*, as examples shown in figure 4.7. Interesting questions are:

- how many points are needed in subsampling?
- Which are the best points to sample?
- How to best model the ambiguity?

Answers to those questions can be found through principled analysis of feature selection.

In this chapter, we discuss another important issue: how to select the most informative features, such that the representation allows the best discrimination between different action clusters found via unsupervised spectral clustering.

5.1 Feature Selection and Action Classification Framework

Our new action classification and optimization framework is shown in figure 5.1. Based on the probabilistic representation of human action by CAFHA, action classification is done in both steps of unsupervised learning and online inference. A feature selection and optimization module is introduced to iteratively enhance the CAFHA representation and improve the unsupervised learning. The output is an action-labeled video sequence.

In this chapter, a principled feature selection method for unsupervised action classification is proposed based on a deeper analysis of spatial correspondence in action recognition.

5.2 Feature Selection and Optimization

In the previous CAFHA representation, 9 uniformly distributed points are selected in constructing a naive CAFHA. In this section, an optimization procedure is carried out to determine *informative points* and obtain the optimal parameters for the Gaussian kernels used in section 4.7.

Figure 5.1: Action classification framework with optimization

5.2.1 Intuition

Figure 5.2: Example of different feature points

As shown in figure 5.2, the actions to be compared are *waving over head* and *waving in front of body*, in which most of the motion are within areas $R1$ and $R2$. Three points $P1$, $P2$ and $P3$ are sampled: $P3$ is least effective since it is far way from the motion area and provides little information on the type of actions taking place. $P1$ is an informative point, since it is located in the area $R1$, where the two different actions exhibit different motion patterns. It provides sufficient discrimination for the two actions.

5.2.2 A Metric based on Affinity Matrix for Classification

Figure 5.3: D_1, D_2, D_3, D_4, D_5 imply clusters of same action. D_6, D_7 are the distance values between subsequences of two different action sequences

In order to achieve better classification, the key is to select good features so that the feature representations maximize the difference between classes of actions and minimize the variance within the same action. However this is a very difficult problem in pattern classification, since it is hard to find a proper metric.

The affinity matrix of distances, which is described in section 4.8.2 provides rich and interesting information for evaluating classification. As shown in figure 5.3, D_1, D_2, D_3, D_4 and D_5 are clusters of the same action, suggesting that the original sequence has 5 segments, as verified by the spectral clustering result of this affinity matrix (shown in figure 5.6(a)). The regions D_6, D_7 in figure 5.3 show higher distances between two actions.

The problem of maximizing the difference between different classes and minimizing the

within class variance is equivalent to maximizing the following ratio J :

$$J = \frac{\text{Sum of Affinity Matrix Elements}}{\text{Sum of Within Cluster Variance}} \quad (5.1)$$

Since the element value of the affinity matrix represents the distance between two subsequences of actions, the *Sum of Affinity Matrix Elements* term expresses the difference between actions. By maximizing this term, the discrimination between different actions is maximized. The spectral clustering in section 4.8.3 has clustered the affinity matrix, with each cluster implying one action. The *Sum of Within Cluster Variance* term expresses the within class variance. By minimizing this term, subsequences of the same action will have small distances to each other. The ratio J is thus used as a metric for evaluating feature selection. The Min-max Cut Algorithm [29] proposed by Chris *et al.* has the similar concept as the ratio J here.

5.2.3 Optimization

Tong *et al.* [74] has investigated an optimization-based learning method for image retrieval. The feature selection for unsupervised action classification is also formulated as an optimization based on the analysis of the affinity matrix. We are not interested in going into the details of optimization method, instead we propose a easy understood metric for optimization our CAFHA representation.

Given that the foreground feature measurement $\mathbf{x}'$ changes little, *i.e.* $p(\mathbf{x}', \mathbf{f}, L_{\mathbf{x}'}=1, C_{\cdot, \mathbf{x}'}=1)$ remains relatively constant, the problem is to find the best subset of points in $\mathbf{x}'$ space, and the optimal subset of parameters for modeling $C_{\cdot, \mathbf{x}'}=1$ in its parameter space.

Our CAFHA representation can be re-expressed in the form of:

$$\mathbf{A}(\vec{X}, \vec{G}) \quad (5.2)$$

where $\vec{X}$ is a 2-D array of the spatial locations of all the informative points in $\mathbf{x}'$ space. Each informative point is associated with a 2-D Gaussian kernel. The covariance matrices of these Gaussian kernels are stored in $\vec{G}$, which is a 3-D array.

A metric for evaluating the feature selection is formed as:

$$J(\vec{X}, \vec{G}) = \frac{\sum_j^N \sum_i^N D_{i,j}(\mathbf{A}(\vec{X}, \vec{G}))}{\sum_j^M \sum_i^M \text{var}(R_{i,j})} \quad (5.3)$$

The size $N \times N$ affinity matrix $D(A)$ may be considered to be a function of the CAFHA representation A . Here we assume the spectral clustering yields $M \times M$ clusters on the affinity matrix. The numerator is the sum of the affinity matrix elements, while the denominator is the sum of within cluster variances.

The optimal $\vec{X}$ and $\vec{G}$ can be found by the following optimization procedure:

- (i) Uniformly choose K points $\vec{X}_k, k = 1, 2, \dots, K$ in $\mathbf{x}'$ space. The corresponding Gaussian kernel $\vec{G}_k, k = 1, 2, \dots, K$ for each points is manually initialized. Construct a naive CAFHA $\mathbf{A}(\vec{X}, \vec{G})$
- (ii) Optimization by minimizing $\frac{1}{J(\vec{X}, \vec{G})}$ with respect to $\vec{X}, \vec{G}$. An optimal $\vec{X}, \vec{G}$ will be found at convergence.
- (iii) Reconstruct the CAFHA with the optimal $\vec{X}$ and $\vec{G}$.

Zelnik-Manor and Irani's method of representing action can be considered as an instance of our principled framework, where the whole image is treated as the foreground, only one sample region selected (*i.e.* $K = 1$), and the correspondence ambiguity is modeled by one gaussian of infinite large variance.

When K increases, the dimension of A in (5.2) increases and the computation cost of computing A increases. Empirically $K = 3, 4$ are sufficient for our action classification experiments.

5.3 Experiments

5.3.1 Optimization

A sequence of 753 frames is used, which captures a person walking, waving hand over head and waving hand in the front of body; sample frames are shown in figure 5.8(a). A sliding

Figure 5.4: (a) The initial Gaussian kernels in human-centered frame (b) The optimized Gaussian kernels

window of length 24 frames is applied and traverses 5 frames between steps. Four points are initially sampled, and the corresponding ambiguity model are symmetric Gaussian kernels with standard deviation of $\sqrt{(HW)}$, as shown in figure 5.4(a), to construct naive CAFHA. H and W are the height and width of the canonical frame respectively. After optimization, the location and shape of the Gaussian kernels are changed to figure 5.4(b). The affinity matrix computed by the initial Gaussian kernels is compared with the affinity matrix after optimization in figure 5.5. The clustering results are compared in figure 5.6. Obviously, the optimized affinity matrix one has much better segmented clusters, which means less within-class variances, and larger distances between classes.

5.3.2 Action Classification

In this section, we test the three mentioned algorithms Zelnik-Manor and Irani's method, naive CAFHA and optimized CAFHA on three sequences and compare the results quantitatively.

Our dataset consists of 3 different sequences captured by different surveillance cameras in different conditions, as corresponding sample frames shown in figure 5.7:

- Sequence 1 is of 395 frames. It is composed of three actions: walking fast, walking slowly and jogging performed by three different people.

(a)

(b)

Figure 5.5: (a) Affinity matrix computed by Gaussian kernel in figure 5.4(a) (b) Affinity matrix by the optimal Gaussian in figure 5.4(b)

Figure 5.6: Unsupervised clustering results by (a) Optimized CAFHA (b) Naive CAFHA with 4 Gaussian Kernels (c) Zelnik-Manor and Irani's method (d) ground-truth (yellow represents walking , red represents waving over head and blue represents waving in front of body)

- Sequence 2 is of 858 frames. It is composed of three actions: kicking, punching and ducking performed by the same person. This sequence is obtained from Zelnik-Manor

Figure 5.7: Sample frames video dataset for sequence 1 to sequence 3

	Zelnik-Manor and Irani's method		Naive CAFHA		Optimized CAFHA	
	Detection rate	False +ve rate	Detection rate	False +ve rate	Detection rate	False +ve rate
Sequence 1	24.3%	60.5%	88.3%	12.5%	91.3%	11.5%
Sequence 2	92.7%	5.5%	98.7%	2.5%	98.7%	2.0%
Sequence 3	61.1%	33.7%	81.1%	17.7%	86.2%	13.9%

Table 5.1: Comparison of Zelnik-Manor and Irani's method, Naive CAFHA and Optimized CAFHA on action classification

and Irani [82].

- Sequence 3 is of 2900 frames. It is composed of five different actions: waving over head, waving in the front, walking, kicking and jumping performed by three different people in front of cluttered background.

A sliding window of length 24 frames is applied to the entire sequence. The sliding window demarcates a subsequence each step and traverses 5 frames between steps. The above mentioned three algorithms are tested separately on these sequences and a quantitative comparison is shown in table 5.1. The optimized CAFHA use only 4 Gaussian kernels. Both of the naive CAFHA and the optimized CAFHA are invariant to foreground appearance, as proved by the results on sequence 1 and sequence 3 of multiple people doing different actions. They obviously outperform Zelnik-Manor and Irani's method. The optimized CAFHA achieves the best results.

5.3.3 Real-time Online Action Inference

The same sequence in experiment 1 is used for unsupervised learning. Only 3 Gaussian kernels are used in optimization, and the optimized CAFHA representation for each action is constructed. In the inference stage, a long sequence of 1240 frames is used, including 2 different people doing three different actions, as shown in figure 5.8(b)(c). The inference results

Figure 5.8: (a) Sample frame of walking fast, (b) Sample frame of waving over head, (c) Sample frame of waving in the front of body

outperform the result computed by naive CAFHA with a 9 point selection, as shown in figure 5.9.

The frame size of the sample video is 160×70 and the sliding window is of length 24 frames. When running the algorithm in MatlabTM on a Pentium M 1.5 GHz personal computer, it takes on average 0.76 seconds to obtain the classification label for each subsequence.

Figure 5.9: (a) Inference result of multiple person actions based on optimized CAFHA representation of selecting only 3 points (b) Inference result of multiple person actions by naive CAFHA of selecting 9 points (c) Ground-truth information

5.4 Conclusion

A principled feature selection method for unsupervised action classification is proposed in this chapter which is able to selecting the most informative features to construct probabilistic action representation CAFHA [31]. Experiments shown that the optimized CAFHA outperforms manually specified CAFHA. The optimized CAFHA representation is also efficient and compact, and allows for real-time action inference.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

We proposed a framework for automatically tracking and understanding human in video at different temporal scales.

In order to find the best tracking method that fits our framework, five foreground segmentation methods were studied, implemented and compared qualitatively and quantitatively. A bounding box tracking module was eventually developed based on probabilistic foreground and background segmentation, and applied for extracting human-centric reference frames.

A multi-stage framework for the unsupervised classification of human action, involving separate modules for foreground segmentation, tracking and classification is discussed in chapter 4. A novel probabilistic action representation CAFHA is proposed for action classification. While most existing approaches assume that feature-wise correspondences on people are either available or not at all, this method explicitly formalizes how the probability of correspondences can be used in computation when the correspondences are ambiguous.

An extension of CAFHA representation is to select the most informative features, such that the representation allows the best discrimination between different action clusters. A principled feature selection method is proposed for unsupervised action classification in chapter 5. An optimization process is applied to selecting the most informative features. The learned

representation outperforms existing representations that are specified manually. In particular, good recognition accuracy is achieved across different people performing a variety of actions, and is robust to changes in the background and foreground appearance. The representation is also efficient and compact, and allows for real-time action inference.

As a summary, this thesis has made the following contributions:

- A novel probabilistic action representation CAFHA is derived from principled analysis of the correspondence problem;
- a combined tracking and action classification framework is described that exploits potentially ambiguous correspondences;
- an automatic human action learning and online inference framework is proposed, which substantially improves the efficiency of action classification and can run in near real-time;
- experimental results demonstrated substantial improvement on the state-of-the-art Zelnik-Manor and Irani's method
- a novel feature selection module by optimization was proposed, which greatly improved the effectiveness of action representation for unsupervised action classification;
- various experiments have shown the effectiveness of the action representation in unsupervised learning. The optimized action representation is able to run in real-time for online action inference.

6.2 Future work

Based on our combined tracking and action classification framework, there are several potential directions for future research. We discuss them in the following sections:

Figure 6.1: Visualization of affinity matrix

- *Temporal motion segmentation*

Our current segmentation is the intermediate level event segmentation according to our three-level taxonomy. The visualization of the affinity matrix computed in section 4.8.2 implies that the events can be well clustered. Taking a careful look at figure 6.1, each region of roughly the same value in the Z axis is a cluster of similar events. However, if we were to look more closely within a region corresponding to an event, we can also observe peaks (marked with dark circles) located in a uniform pattern. They correspond to the low-level atomic motions: one step of walking, running *etc.* It may be possible that the low-level atomic motions can also be roughly segmented from the affinity matrix

Figure 6.2: Multiple people activity understanding

by locating the peaks of specific patterns. We will research on developing an algorithm to extract this information for temporal motion segmentation.

- *Multiple people activity understanding*

Our framework is currently focused on classifying the actions of a single person in the scene. Future work may include extending the current framework to cope with multiple persons in a video by tracking each person individually.

We have shown in figure 6.2 that when multiple people are in the video, our framework can record the separate trajectories of various people and segment the activity of each person in the video sequence. Our combined tracking and understanding framework is aimed at understanding multiple activities carried out in isolation. However, in the situation when multiple people interact with each other, the understanding becomes complicated. Further research will be conducted to understand multiple people scene by using the trajectories and temporally segmented foreground.

- *Probabilistic people identification*

In noisy situations, foreground regions may not be properly connected, so one person may be covered by 2 or more bounding boxes. When people interact with each other, two bounding boxes will merge to form one and subsequently split. These are difficult research issues. In our future work, we will investigate probabilistic frameworks for identifying occupants by 2D motion trajectory estimation techniques. Trajectories and area histories of the bounding boxes will be analyzed, and a probabilistic model for the identification of multiple people will be created. The probability of each identification is estimated and passed to the next recognition stage. The feedback from the high level module is used to refine the probabilities. Instead of using only skin color, shape or other features, which are application-dependent for identification, our method of additionally using trajectory information can make identification a lot more robust.

- *High level understanding*

In the short and intermediate temporal scale, motion and event fragments are observed. For example, in a basketball match, these fragments are: a player dribbling, free throwing, dunking or running. However, in the long temporal scale, what does a sequence of these fragments mean? Is it a defence or an offence? Are they playing a good match? These questions are potentially interesting. This is the high level understanding in our research framework. After the temporal segmentation of intermediate events, it is interesting to look at how to infer longer temporal meanings using these segments combined with possible prior knowledge as a high level understanding stage.

- *Recursive understanding*

Our current approach is based on estimating lower-level states from noisy data at shorter temporal scales, and making higher level inferences through the coupling of these rough estimates and higher level constraints. Information flows one way from lower to higher levels in our framework. However, there should be a mechanism for improving the noisy

low-level estimates based on our higher-level estimates. This may be done in a recursive manner, with the goal of achieving accurate inferences not just at the uppermost level but throughout all levels of our framework. For example, when the long temporal meaning is acquired, which indicates an offence in basketball match, it increases the probabilities of ball passing and throwing. Thus the intermediate level and even low level estimation could be refined. A probabilistic recursive understanding framework will be developed in the future.

References

- [1] <http://vrlab.epfl.ch/research/>.
- [2] NSF science and technology center for graphics and visualization. In <http://www.cs.unc.edu/Research/stc/>.
- [3] <http://pets2001.visualsurveillance.org/>. In *Proc. Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii, 2001.
- [4] <http://pets2002.visualsurveillance.org/>. In *Proc. Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, Denmark, 2002.
- [5] J.K. Aggarwal, Q. Cai, and W. Liao. Articulated and elastic non-rigid motion: a review. *Annals of Statistics*, pages 2–14, 1994.
- [6] J. Alon, S. Sclaroff, G. Kollios, and V. Pavlovic. Discovering clusters in motion time-series data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–381, Madison, WI, 2003.
- [7] K. Alsabti, S. Ranka, and V. Singh. An efficient k-means clustering algorithm. In *Proceedings of the First Workshop on High Performance Data Mining*, Orlando, FL, 1998.
- [8] J. Barron and D. Fleet. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:42–77, 1994.
- [9] A.M. Baumberg and D.C. Hogg. An efficient method for contour tracking using active shape models. In *Proc. Workshop on Motion of Non-Rigid and Articulated Objects*, pages 2–14, Austin, Texas, 1994.

-
- [10] A. Bobick. Movement, activity and action: The role of knowledge of the perception of motion. *Philosophical Transactions of the Royal Society London*, B(352):1257–1265, 1997.
 - [11] A. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.
 - [12] A. Bobick and A.D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:1325–1337, 1997.
 - [13] G.R. Bradski. Computer video face tracking for use in a perceptual user interface. *Intel Technology Journal*, 1998.
 - [14] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 568–574, San Juan, Puerto Rico, 1997.
 - [15] T.J. Cham and J.M. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 239–245, Fort Collins, CO, Jun 1999.
 - [16] S. Chang, N. Dasgupta, and L. Carin. A bayesian approach to unsupervised feature selection and density estimation using expectation propagation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1050, San Diego, CA, 2005.
 - [17] R. Chen and J. S. Liu. Mixture kalman filters. *Journal of the Royal Statistical Society*, 62:493–508, 2000.
 - [18] Z. Chen and H. J. Lee. Knowledge-guided visual perception of 3d human gait from a single image sequence. *IEEE Transactions on Systems, Man and Cybernetics*, 22:336–343, 1992.
 - [19] O. Chomat and J.L. Crowley. Recognizing motion using local appearance. In *Proc. International Symposium on Intelligent Robotic Systems*, pages 271–279, Edinburgh, United Kingdom, 1998.

-
- [20] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, and O. Hasegawa. A system for video surveillance and monitoring: VSAM final report. *CMU Technical Report*.
- [21] R. Collins and Y. Liu. On-line selection of discriminative tracking features. In *Proc. International Conference on Computer Vision*, pages 1043–1050, Nice, France, Oct 2003.
- [22] R.T. Collins. Mean-shift blob tracking through scale space. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 234–240, Madison, WI, 2003.
- [23] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 142–149, Hilton Head, SC, 2000.
- [24] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *Proc. International Conference on Computer Vision*, pages 1071–1076, Cambridge, MA, June 1995.
- [25] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:781–796, 2000.
- [26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 886–893, San Diego, CA, June 2005.
- [27] J.W. Davis. Sequential reliable-inference for rapid detection of human actions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 111–118, Washington D.C., 2004.
- [28] J.W. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–934, San Juan, Puerto Rico, 1997.

- [29] Chris Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and Horst Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proc. IEEE International Conference on Data Mining*, pages 107–114, San Jose, CA, 2001.
- [30] D. Fellner, G. Albuquerque, H. Kim, and S. Havemann. <http://www.cg.v.tugraz.at/cgv/research>.
- [31] Z. Feng and T.J. Cham. Video-based human action classification with ambiguous correspondences. In *Proc. IEEE Workshop on Vision for Human-Computer Interaction*, San Diego, CA, Jun 2005.
- [32] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice-hall International, Inc, 2003.
- [33] D. Freedman and T. Zhang. Active contours for tracking distributions. *IEEE Transactions on Image Processing*, 13:518–526, 2004.
- [34] N. Friedman and S. Russell. Image segmentation in video sequences: A probabilistic approach. In *Proc. The Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 175–181, Rhode Island, USA, 1997.
- [35] M. Gastaud and M. Barlaud. Tracking video objects using active contours. In *Proc. IEEE Workshop on Motion and Video Computing*, pages 90–95, Orlando FL, 2002.
- [36] D.M. Gavrila and L.S. Davis. 3D model-based tracking of humans in action: a multi-view approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, CA, 1996.
- [37] D.M. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *Proc. International Conference on Computer Vision*, pages 87–93, Kerkyra, Greece, 1999.
- [38] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *Proc. International Conference on Computer Vision*, pages 742–749, Nice, France, 2003.
- [39] M. S. Grewal and A. P. Andrews. *Kalman Filtering: Theory and Practice*. Prentice-Hall, 1993.

-
- [40] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley Publishing Company, 1992.
 - [41] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:809–830, 2000.
 - [42] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.
 - [43] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1:5–20, 1983.
 - [44] M. Irani and P. Anandan. Factorization with uncertainty. In *Proc. European Conference on Computer Vision*, pages 539–553, Dublin, Ireland, June 2000.
 - [45] M. Isard and A. Blake. CONDENSATION—conditional density propagation for visual tracking. *International Journal of Computer Vision*, 1:5–28, 1998.
 - [46] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proc. IEEE Conference on Machine Learning*, pages 121–129, New Brunswick, NJ, 1994.
 - [47] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for realtime tracking with shadow detection. In *Proc. Second European Workshop on Advanced Video Based Surveillance Systems*, Kingston, UK, 2001.
 - [48] Y. Kameda and M. Minoh. A human motion estimation method using 3-successive video frames. In *Proc. IEEE Conference on Virtual System and Multimedia*, pages 135–140, Gifu, Japan, 1996.
 - [49] I. Laptev and D. Lindeberg. Space-time interest points. In *Proc. International Conference on Computer Vision*, pages 432–439, Nice, France, 2003.
 - [50] I. Laptev and D. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proc. W4 Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, Czech Republic, 2004.

- [51] M.K. Leung and Y.H. Yang. First sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:359–377, 1995.
- [52] K. Levi and Y. Weiss. Learning object detection from a small number of examples: The importance of good features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 53–60, Washington, D.C., 2004.
- [53] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- [54] M. Machline, L.Zelnik-Manor, and M. Irani. Multi-body segmentation: Revisiting motion consistency. In *Proc. Workshop on Vision and Modeling of Dynamic Scenes*, Copenhagen, Denmark, June 2002.
- [55] D. Meyer and J. Denzler. Model based extraction of articulated objects in image sequences for gait analysis. In *Proc. IEEE International Conference on Image Processing*, pages 78–81, Santa Barbara, CA, 1997.
- [56] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. International Conference on Computer Vision*, pages 525–531, Vancouver, Canada, 2001.
- [57] K. P. Murphy. Switching kalman filter. *Technical Report, Compaq Cambridge Research Laboratory*, 1998.
- [58] A.Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. Neural Information Processing Systems*, Vancouver, BC, 2001.
- [59] S. Niyogi and E. Adelson. Eanalyzing and recognizing walking and gestures in xyt. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–474, Seattle, WA, 1994.
- [60] V. Pavlovic, J.M. Rehg, T.J. Cham, and K. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamical models. In *Proc. International Conference on Computer Vision*, pages 94–101, Kerkyra, Greece, 1999.

- [61] R. Polana and R. Nelson. Detecting activities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 569–574, New York, NY, 1993.
- [62] R. Polana and R. Nelson. Low level recognition of human motion. In *Proc. IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 77–82, Austin, Texas, 1994.
- [63] F. Porikli. Human body tracking by adaptive background models and mean-shift analysis. In *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Graz, Austria, Mar 2003.
- [64] R.J. Qian and T.S. Huang. Estimating articulated motion by decomposition. *Time-Varying Image Processing and Moving Object Recognition*, pages 275–286, 1994.
- [65] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1993.
- [66] K. Rohr. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 59:94–115, 1994.
- [67] R. Rosales and S. Sclaroff. 3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2123, Fort Collins, CO, Jun 1999.
- [68] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [69] Y. Shan, H. Sawhney, and R. Kumar. Unsupervised learning of discriminative edge measures vehicle matching between non-overlapping cameras. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 894–901, San Diego, CA, 2005.
- [70] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [71] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, Fort Collins, CO, 1999.

- [72] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:747–757, 2000.
- [73] K. Takahashi and S. Seki. Recognition of dexterous manipulations from time-varying images. In *Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 23–28, Austin, Texas, 1994.
- [74] H. Tong, J. He, M. Li, W. Ma, C. Zhang, and H. Zhang. A unified optimization based learning method for image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 230–235, San Diego, CA, 2005.
- [75] K. Toyama and A Blake. Probabilistic tracking in a metric space. In *Proc. International Conference on Computer Vision*, volume 2, pages 50–57, Vancouver, Canada, 2001.
- [76] E. A. Wan and R. van der Merwe. The unscented kalman filter for nonlinear estimation. In *Proc. IEEE Symposium for Adaptive Systems for Signal Processing, Communication and Control*, pages 153–158, Alberta, Canada, Oct 2000.
- [77] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.
- [78] Y. Weiss. Segmentation using eigenvectors: A unifying view. In *Proc. International Conference on Computer Vision*, pages 975–982, Kerkyra, Greece, Sep 1999.
- [79] C.R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfunder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:780–785, 1997.
- [80] J. Yamato and J. Ohya. Recognizing human action in time-sequential images using hidden markov models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 376–385, Urbana-Champaign, Illinois, 1992.
- [81] M.H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074, 2002.

-
- [82] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 123–130, Kauai, Hawaii, 2001.
 - [83] L. Zelnik-Manor and M. Irani. Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–293, Madison, WI, June 2003.
 - [84] L. Zelnik-Manor and M. Irani. Temporal factorization vs. spatial factorization. In *Proc. European Conference on Computer Vision*, pages 434–445, Prague, Czech Republic, May 2004.
 - [85] X. S. Zhou, D. Comaniciu, and A. Krishnan. Conditional feature sensitivity: A unifying view on active recognition and feature selection. In *Proc. International Conference on Computer Vision*, pages 1502–1511, Nice, France, 2003.

Publications

- [1] Z. Feng and T.J. Cham. Video-based human action classification with ambiguous correspondences. In *Proc. IEEE Workshop on Vision for Human-Computer Interaction* held in Conjunction with *IEEE Conference on Computer Vision and Pattern Recognition*. San Diego, CA, Jun 2005.
- [2] Z. Feng and T.J. Cham. Feature Selection and Optimization for Unsupervised Action Classification. In preparation for journal submission 2006.