การจำแนกกริยาของมนุษย์โดยใช้ลักษณะการเคลื่อนที่และสภาพปรากฏ
เพื่อการเข้าใจกิจกรรมและการตรวจหาความผิดปกติในการเฝ้าระวังจากภาพ

นางสาวกนกพรรณ เลิศนิพนธ์พันธุ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

HUMAN ACTION CLASSIFICATION USING MOTION AND APPEARANCE FEATURES FOR ACTIVITY UNDERSTANDING AND ANOMALY DETECTION IN VISUAL SURVEILLANCE

Miss Kanokphan Lertniphonphan

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2015

Thesis Title	HUMAN ACTION CLASSIFICATION USING MOTION AND APPEARANCE FEATURES FOR ACTIVITY UNDERSTANDING AND ANOMALY DETECTION IN VISUAL SURVEILLANCE
By	Miss Kanokphan Lertniphonphan
Field of Study	Electrical Engineering
Thesis Advisor	Assistant Professor Supavadee Aramvith, Ph.D.
Thesis Co-Advisor	Assistant Professor Thanarat Chalidabhongse, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

....Dean of the Faculty of Engineering
(Associate Professor Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

....Chairman
(Associate Professor Chedsada Chinrungrueng, Ph.D.)

....Thesis Advisor
(Assistant Professor Supavadee Aramvith, Ph.D.)

....Thesis Co-Advisor
(Assistant Professor Thanarat Chalidabhongse, Ph.D.)

....Examiner
(Assistant Professor Supatana Auethavekiat, Ph.D.)

....External Examiner
(Sanparith Marukatat, Ph.D.)

กนกพรรณ เลิศนิพนธ์พันธุ์ : การจำแนกกริยาของมนุษย์โดยใช้ลักษณะการเคลื่อนที่และสภาพปรากฏ เพื่อการเข้าใจกิจกรรมและการตรวจหาความผิดปกติในการเฝ้าระวังจากภาพ (HUMAN ACTION CLASSIFICATION USING MOTION AND APPEARANCE FEATURES FOR ACTIVITY UNDERSTANDING AND ANOMALY DETECTION IN VISUAL SURVEILLANCE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. สุภาวดี อร่ามวิทย์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร. ธนารัตน์ ชลิดาพงศ์, 57 หน้า.

การรู้จำกริยาอาการของมนุษย์จากวีดิทัศน์ เป็นงานวิจัยที่ได้รับความสนใจเรื่องหนึ่งในงานด้านคอมพิวเตอร์วิทัศน์ อันเนื่องมาจากความต้องการในการนำไปประยุกต์ใช้ในงานทางด้านการเฝ้าระวังจากภาพแบบอัตโนมัติ เพื่อทดแทนการดูแลที่ไม่เพียงพอและขาดประสิทธิภาพของมนุษย์ อย่างไรก็ตามโจทย์ที่ท้าทายในการรู้จำกริยาคือการสกัดคุณลักษณะจากภาพที่เหมาะสมต่อการจำแนก คุณลักษณะที่ดีควรประกอบด้วยข้อมูลที่บอกลักษณะการเคลื่อนที่และสภาพปรากฏของบุคคลในภาพ นอกจากนี้กระบวนการสกัดคุณลักษณะควรต้องมีความสามารถในการปรับตัวแบบอัตโนมัติต่อความเร็วที่หลากหลายของการแสดงกริยาของมนุษย์ เมื่อนำระบบไปใช้กับบุคคลต่างๆ กริยาต่างๆ และฐานข้อมูลวีดิทัศน์ที่ต่างกัน วิทยานิพนธ์นี้นำเสนอการสกัดคุณลักษณะของกริยาในช่วงเวลาระหว่างเฟรมหลักแบบปรับตัวได้ (AKFI) เพื่อแบ่งย่อยกริยาของมนุษย์ในชุดลำดับภาพเป็นลำดับภาพย่อยๆของกริยาพื้นฐาน โดยความกว้างของช่วงเวลาดังกล่าวจะถูกปรับเปลี่ยนโดยอัตโนมัติตามลักษณะของกิจกรรม และความเร็วที่แตกต่างกันของแต่ละบุคคล เมื่อตรวจพบเฟรมหลัก ระบบจะเข้ารหัสคุณลักษณะของกริยาภายในช่วงเวลาระหว่างเฟรมหลัก ให้อยู่ในรูปของภาพประวัติการเคลื่อนที่แบบปรับตัวได้ (AMHI) และภาพประวัติท่าทางหลัก(KPHI) ซึ่งคุณลักษณะดังกล่าวประกอบด้วยลักษณะการเคลื่อนที่และสภาพปรากฏของกริยาของมนุษย์ ผลการทดลองแสดงให้เห็นว่า ระบบสามารถแบ่งแยกกริยาไม่ปกติจากสถานการณ์ปกติได้ นอกจากนี้ AMHI และ KPHI สามารถใช้ในการจำแนกกริยาได้อย่างมีประสิทธิภาพ โดยได้ทำการเปรียบเทียบผลลัพธ์การจำแนกกริยากับขั้นตอนวิธีของงานวิจัยอื่นๆ

ภาควิชา วิศวกรรมไฟฟ้า ลายมือชื่อนิสิต ....

สาขาวิชา วิศวกรรมไฟฟ้า ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ปีการศึกษา 2558 ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

5171878921 : MAJOR ELECTRICAL ENGINEERING

KEYWORDS: HUMAN ACTIVITY CLASSIFICATION / ANOMALY DETECTION / FEATURE EXTRACTION

KANOKPHAN LERTNIPHONPHAN: HUMAN ACTION CLASSIFICATION USING MOTION AND APPEARANCE FEATURES FOR ACTIVITY UNDERSTANDING AND ANOMALY DETECTION IN VISUAL SURVEILLANCE. ADVISOR: ASST. PROF. SUPAVADEE ARAMVITH, Ph.D., CO-ADVISOR: ASST. PROF. THANARAT CHALIDABHONGSE, Ph.D., 57 pp.

Human action recognition is one of the interesting research areas in computer vision. It is an important component of the automated surveillance system which is needed to reduce the insufficiency and inefficiency of human’s role in the system. However, one of the challenges in action recognition is the extracting appropriate features for classification. Good features should contain both motion and appearance information of human. Also, the feature extraction process should automatically adapt to the speed variation of human actions when applying the system to the different performers, actions, and datasets. In this thesis, we propose the Adaptive Key Frame Interval (AKFI) feature extraction to segment human action into primitive action subsequences. The interval length is automatically changed based on the action characteristic and speed of the performer. Once key frames are detected, the features within a segmented period are encoded by Adaptive Motion History Image (AMHI) and Key Pose History Image (KPHI). The features contain both appearance and motion information of human actions. The experimental results demonstrate that the system can differentiate the unusual action from the normal situation. Also, AMHI and KPHI can effectively classify action compared to other well-known algorithms.

Department:	Electrical Engineering	Student's Signature	....
Field of Study:	Electrical Engineering	Advisor's Signature	....
Academic Year:	2015	Co-Advisor's Signature	....

ACKNOWLEDGEMENTS

First, I would like to express my sincerest gratitude to my advisor Asst. Prof. Dr. Supavadee Aramvith and my co-advisor Asst. Prof. Dr. Thanarat Chalidabhongse for her tireless efforts, encouragement, support and guidance throughout the entire course of my study.

I also would like to thank the members of my thesis committee, Asst. Prof. Dr. Supatana Auethavekiat, Dr. Sanparith Marukatat and the chairperson, Assoc. Prof. Dr. Chedsada Chinrungrueng, for giving critical reviews of this work and for their advice on my dissertation.

I would like to thank NTC Scholarship from the Telecommunications Research Industrial and Development Institute, National Telecommunication Commission Fund, for financial supporting from 2009 to 2011. This research also partly supported by the 90th Anniversary of Chulalongkorn University, Rachadapisek Sompote Fund.

Also, I appreciate all members of the Video Research Group for their support during my research work and any assistance and warm collaborations.

Finally, I would like to express the highest gratitude to my parents, everyone in my family, and all of my friends for their help, their understanding and moral encouragements through my graduate journey.

CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

Page

CHAPTER 1
INTRODUCTION

1.1 Motivation and Problem Statement

In recent years, surveillance cameras have been widely used in security and anomaly detection to detect and prevent danger which may arise. Conventional visual surveillance systems rely on human operators to monitor the activities and events occurring in the scene. However, there are limitations of human perception system. The effectiveness of the analysis process may be varied and restrained by the physical condition of each human. The computer vision technology was introduced in visual surveillance system [1] to increase the overall efficiency of the system and help the insufficiency of human's role in the system. The intelligent visual surveillance system is also considered as an important research area in computer vision for human action classification and anomaly detection.

Human action classification in video is a process which analyzes and understand human movement in a scenario. The classification can be used to define action label and detect the anomaly in video files. By using the output of action classification, the searching time for locating unusual events and suspects will be dramatically reduced. The action classification process consists of human action modeling, feature extraction, and classification. Before applying the action classification in the system, human action models are constructed by using features which are extracted from the training dataset. In the testing, feature vectors are extracted and used to constructed action representation. Then, the classifier determines action label based on the trained action models.

The several surveys of human motion and action analysis [2-5] provided the overview of various researches with the comparisons among human activity analysis approaches for various applications. As shown in those surveys, the challenges of human action classification in visual surveillance include variations of performer appearances, movement patterns, performing speed, scaling, occlusion, and camera viewpoint. The selected features for action representation should consider these constraints.

The characteristic of features is various based on the applications. In anomaly detection, many researches used low-level features such as trajectory paths [6-11], moving direction, and magnitude of the displacement [9, 12], [13] for modeling the normal events. However, these features cannot be used to classify actions. Since the detailed information of human body is not considered, the situations such as leaving object, picking up stuff, and fighting might not be detected. The action representation for action classification is required for accurate anomaly detection in visual surveillance.

In action classification, the extracted features, which contain the characteristic of human posture and motion in each action, can improve the classification accuracy. In addition, the appropriate interval effects the completeness of feature. Generally, primitive actions can be effectively classified by using very short snippets of 1-7 frames [14] based on the observation of biological vision system. However, the movement of each action and the movement of each people are not the same. The informative features, which are extracted from the appropriate interval are required in classifying human actions from image sequences.

In this work, we focus on improving feature extraction process. The automatic event boundary detection by a low-level motion variation is used to segment the meaningful features for human action recognition. By detecting the beginning and ending of an event, the system adjusts an appropriate number of frames for different actions and performers. The adaptive interval also solves a problem of speed variation and extracts informative representation of human primitive actions.

We propose the Adaptive Key Frame Interval (AKFI) for extracting features and segmenting action into small primitive movements by detecting key frames at the starting and the ending time as shown in Figure 1.1. The timestamp allows system adaptation based on the speed of performers. So, we do not have to search for a suitable length of sliding window in temporal domain for each action or dataset. Therefore, the features within the same action from several performers contain similar properties.

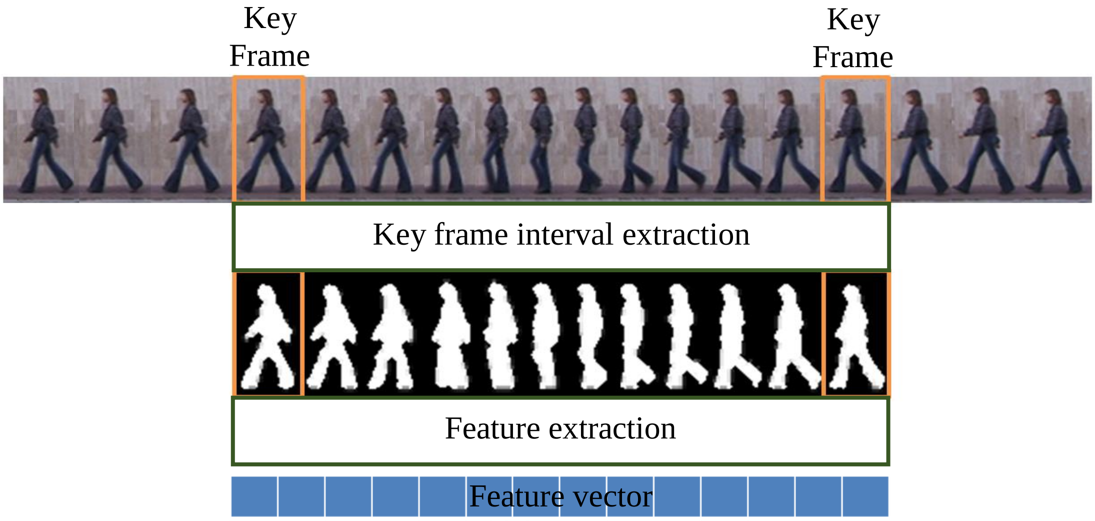


Figure 1.1 Feature extraction using Adaptive Key Fram Interval for walking sequence.

1.2 Objective

1. Investigate the motion and appearance features of human to classify actions and understand activities in the visual surveillance scenario.
2. Develop feature representation from human movements and appearance for classifying human actions in the visual surveillance scenario.
3. Detect anomaly events as the unusual actions occurring in the scene.

1.3 Scope

1. This work uses video from a stationary camera in the static environment.
2. The whole human body is clearly represented without occlusion and viewpoint variation.
3. The proposed feature extraction, which extracts of motion and appearance information, can classify human actions and can detect anomaly from the video.
4. The performance measurement of the algorithm is tested with the public datasets.

1.4 Contribution

Our main contributions consist of anomaly detection and human action classification for surveillance system.

1. Propose the histogram of angular difference between edge and motion direction to detect anomaly. The anomaly can be identified by separating the normal actions from the scene instead of detecting unusual action which we may not have enough information.
2. Propose Adaptive Key Frame Interval (AKFI) to extract key frame interval to specify action subsequence duration. By using AKFI, the system can automatically vary the number of frame, to extract features. This help solving the problem of speed variation of performer and actions.
3. Propose Adaptive Motion History Image (AMHI) and Key Pose History Image (KPHI) to extract features based on the key frame information and AKFI. AMHI and KPHI represent the motion and posture information which are accumulated over AKFI to extract the informative and compact features in each action cycle.

1.5 Outline of Thesis

This thesis is organized into five chapters including this chapter. The following paragraphs provide brief descriptions of the remaining chapters of this thesis.

Chapter 2 provides some background and structure of the human action classification in video. Literature review on various features are described.

Chapter 3 presents features that relate to the proposed method. Both appearance and motion information are used to detect anomaly situation and classify human actions. The AKFI for extracting AMHI and KPHI, is explained in details.

Chapter 4 explains the experimental setups and testing datasets. The experimental results are described and discussed in this chapter.

Chapter 5 includes conclusions and future works of the research.

CHAPTER 2
BACKGROUND AND LITERATURE REIVEW

In recent years, understanding human activities in visual analysis plays the important role in many applications including automated surveillance system, anomaly detections and alarming, crowd flux statistics and congestion analysis, human-computer interface, etc. The purpose of developing such system is to have an automatic system to track, identify persons, and understand human activities. This is very useful when there is a large number of cameras but with limited human capacity. The prerequisites of the automated surveillance system using single camera can divide into the following stages: environment modeling, motion segmentation, object classification, tracking, person identification, and human action classification as shown in Figure 2.1 [1].

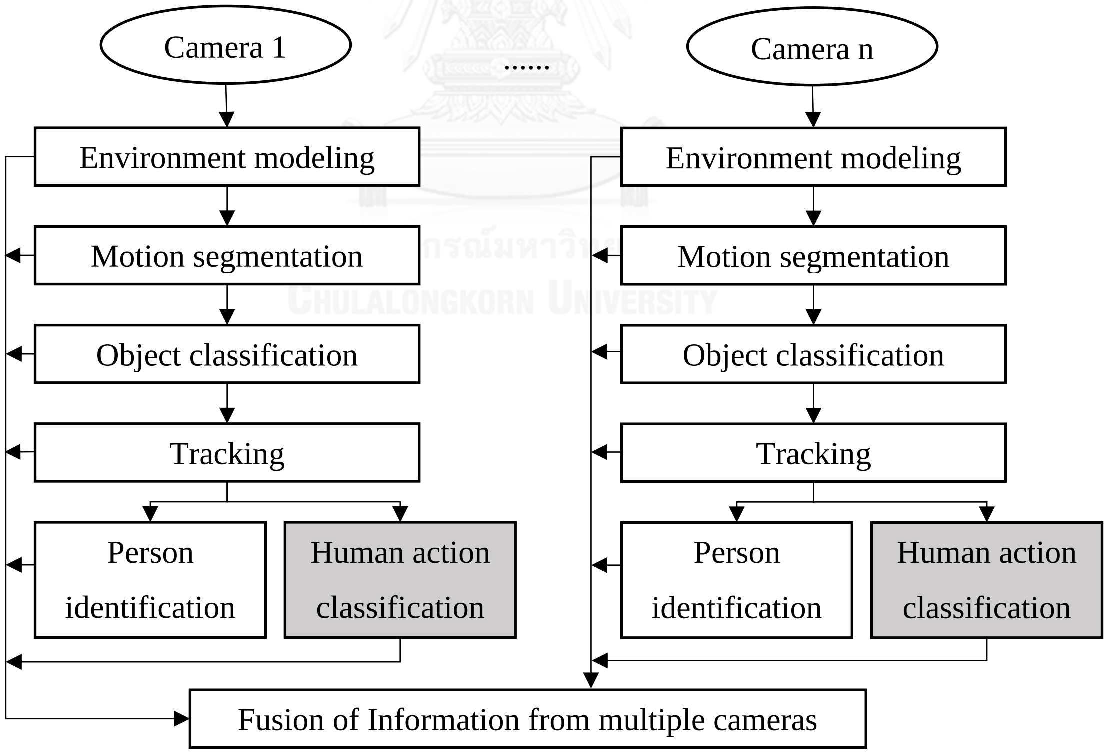


Figure 2.1 General framework of visual surveillance

The automated surveillance system begins with the environment modeling or background model process to construct the background model which does not contain moving objects. When an object moves in to the camera viewpoint, the background subtraction is used to segment the object by subtracting a current frame with the background model. The displacement of the detected object in consecutive frames can be computed by using motion segmentation. Although the segmentation results contain all of the moving objects in the scene, only some of them are used for further analysis. The object classification is used to detect and separate the interested objects such as human from the other moving objects. Then, tracking will be applied to collect the interested person information such as the current position, trajectory path, and velocity. In some specific area such as at a forbidden area entrance, the person identification is installed to prevent the outsider. Lastly, the human action classification will analyze the appearance and motion information during the tracking process to classify actions. This work will focus on feature extraction which is part of the human action classification.

In this chapter, background and literature review in the human action classification for activity understanding with a single camera are presented. The previous researches of human action representation, feature extraction methods, and key frame detection, which are parts of the human behavior understanding in surveillance system, are described in more details. In human action representation section, feature characteristics for action classification are reviewed. While feature extraction in video focuses on features based on time domain segmentation. Lastly, the key frame detection methods are also reviewed.

2.1 Human Action Representation

The human action representation consists of features which indicate the action characteristic and contain the discrimination properties. The action representation in the literatures can be grouped into two main approaches, which are appearance-based and motion-based approaches.

2.1.1 The appearance-based approach

The approach uses information of a silhouette to find the correlation in the shape of human posture. The silhouette can be obtained by using background subtraction which is based on the environment modeling. Then, the object classification identifies which object is human. Only human silhouette is further used in human action classification. The extracted human silhouettes in the different actions are shown in Figure 2.2. The appearance-based feature mainly extract feature from the shape of the extracted foreground.

In 2000, Cutler and Davis [15] analyzed the periodic motion from the set of detected silhouette as shown in Figure 2.3. Shapes of the extracted silhouette are directly used to analyze and classify human actions. For the periodic action, the similarity matrix also indicates the periodicity. The research used the periodicity to identify the moving objects in the surveillance scenarios. Ikizler et al. [16] extracts orientated histogram of the straight lines which fit to the detected human shape boundary. Chaaraoui et al. [17] uses only the contour point of the extracted silhouette to reduce the redundancy during the feature extraction. Baysal et al. [18] uses line-pairs to compute the similarity between frames. The line-pairs are extracted by fitting a line to the extracted foreground contour.

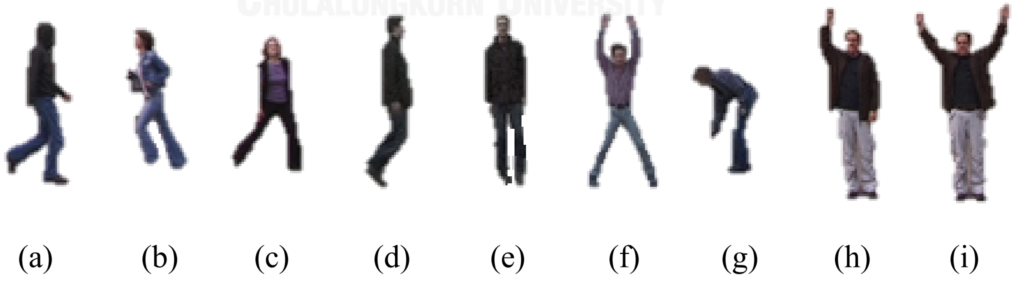


Figure 2.2 Extracted silhouettes of Weizmann dataset [19] (a) walk, (b) run, (c) gallop sideways, (d) jump- forward-on-two-legs, (e) jump-in-place-on-two-legs, (f) jumping-jack, (g) bend, (h) wave-one-hand, and (i) wave-two-hands.

The Histogram of Oriented Gradient (HOG) [20] is a descriptor which extracts local intensity gradient or direction of edge from small regions. Then, the normalized histogram of the edge direction is created. From the Figure 2.4 (a), the human model clearly indicates the human shape such as head, shoulders, and legs. After the success of HOG in human detection in an image, many researches apply HOG to extract the characteristic of human posture in each action. In Thurau and Halaváč [21], HOG is used to categorize primitive actions in still images and image sequences. Ikizler and Duygulu [22] proposed to use oriented rectangular patches over human silhouette called Histogram of Oriented Rectangles (HOR) to represent human posture. The spatial histograms represent the distribution of rectangular patches instead of the distribution of edge gradients.

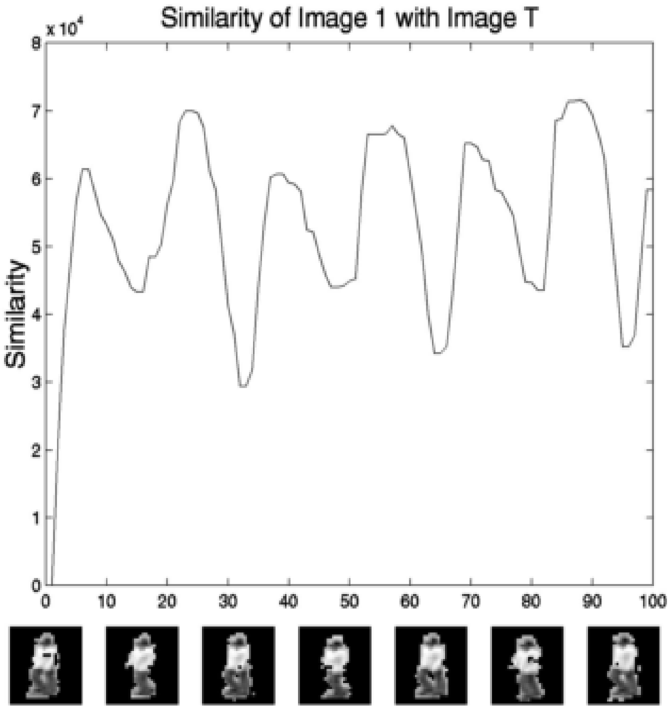


Figure 2.3 The segmented object by using the local minima of the appearance similarity [15]

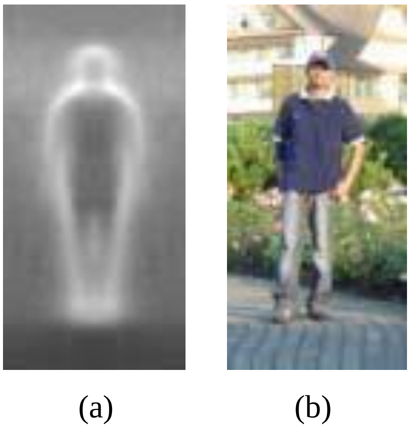


Figure 2.4 (a) The average gradien human image over the training dataset
(b) A test image of human [20].

Although the appearance-based approach can classify human actions with satisfactory results, most of them are not robust with spatial noise and imperfect of extracted human silhouette. Furthermore, the selected frame for classification is also important. As some actions, such as walking and running, may contain similar postures, misclassification can occur if the system chooses the posture that is correlated to other actions.

2.1.2 The motion-based approach

For motion-based approach, the consecutive frames are used to extract motion pattern. Many researches [23] [16] [24] [25] extract features based on optical flow computation to detect the motion area and compute direction and magnitude of the human movement. The optical flow computation in computer vision is calculated based on the assumption that the pixel intensities of an object do not change during the consecutive frames.

Given an input image $I(x, y, t)$ at time t, at time $t + dt$ an image point (x, y) is moved to $(x + dx, y + dy)$. Therefore,

$$I(x, y, t) = I(x + dx,\ y + dy,\ t + dt) \tag{1}$$

When dx and dy are small, using Taylor's expansion,

$$I\big(x + dx,\ y + dy,\ t + dt\big) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt \tag{2}$$

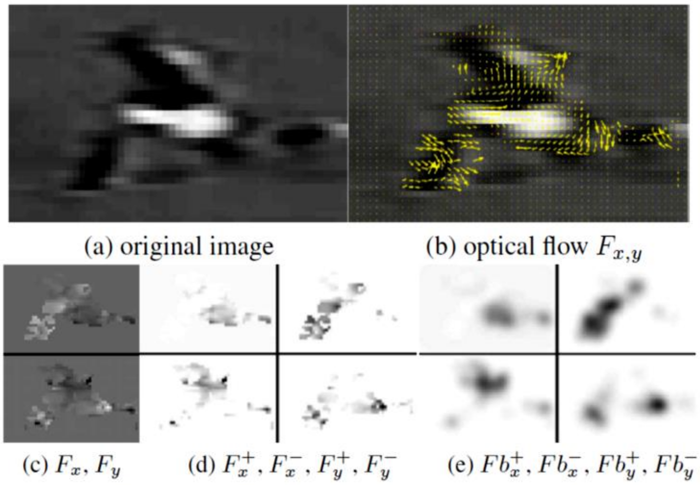
From (1) and (2),

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{3}$$

where (u, v) represents the velocity vector in spatial domain. By applying Lucas-Kanade (LK) algorithm [26] to estimate optical flow, the motion information of a moving patch is extracted.

Efros et al. [23] introduced motion descriptor based on optical flow. The noisy optical flow measurements are treated as a spatial pattern which is smoothed into four separated channels to reduce noise and preserve motion information, as shown in Figure 2.5. The other approaches deal with the noisy motion measurement of optical flow by using histograms of motion feature [24, 27-29]. The motion descriptors are presented as spatial and directional binning of optical flow [16]. Chaudhry et al. [24] proposed a histogram of oriented optical flow (HOOF) and recognize action using Binet Cauchy kernels. HOOF alleviates the effect of noise, scale and direction of motion variation. As this method uses the magnitude of the optical flow to measure its angle, the effect of scale variation is still existing. For example, the same walking person in the different scales has the same angle of motion. By using HOOF, the larger scale has a huge magnitude than a smaller one but the angle is the same. Thus, the distribution has shifted due to the scale variation.

There is another approach to extract motion information from the image sequence. Bobick et al. [30] proposed the temporal template which is a static vector image. The vector at each point is a function of the corresponding spatial location motion properties. The approach consists of motion-energy-image (MEI) and motion-history-image (MHI) as shown in Figure 2.6. MEI represents the region where motion occurs in image sequence. MHI is an intensity value of recent motion.



(a) original image

(b) optical flow $F_{x,y}$

(c) F_x, F_y

(d) $F_x^+, F_x^-, F_y^+, F_y^-$

(e) $Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

Figure 2.5 The motion descriptor (a) Original image, (b) Optical flow, (c) Separating the x and y components of optical flow vectors, (d) 4 separate channels of optical flow vectors, (e) Final blurry motion channels [23]

Given an input image $I(x, y, t)$ and a binary frame differencing image $D(x, y, t)$ at time t, the binary $MEI(x, y, t)$ is defined as,

$$MEI(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i) \tag{4}$$

while, τ is the temporal extent of a movement. The result of MEI is an accumulated motion region during τ duration. All the layered binary images are assigned the same value. So, MEI indicates location and shape of motion occurrence without the detail of magnitude and direction.

To specify the how motion image moving, MHI layer motion image with a timestamp value. The intensity value of MHI image indicates the motion image ordering as shown in Figure 2.6. The darker pixels mean the motion area was occurred before the brighter pixels. The current motion image is always the brightest. An $MHI(x, y, t)$ is defined as,

$$MHI(x, y, t) = \begin{cases} \tau & , if\ D(x, y, t) = 1 \\ max(0, MHI(x, y, t - 1) - 1) & , otherwise \end{cases} \tag{5}$$

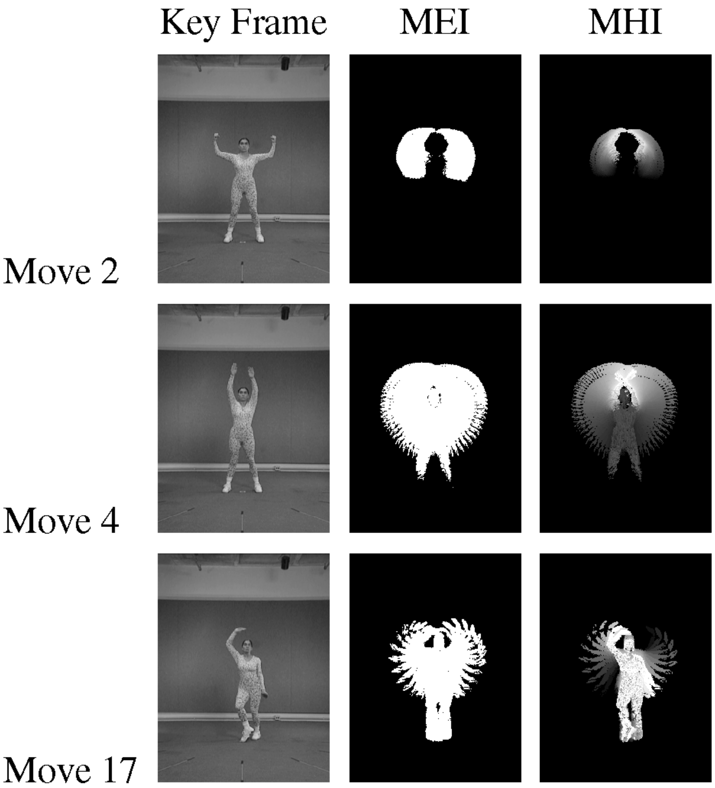


Figure 2.6 The comparison of MEI and MHI as two components of a temporal template [30]

Bradski et al. [31] extended the motion temporal template in [30] by computing gradient vector of the boundary at each step of MHI image as shown in Figure 2.7. The gradient can be obtained by using Sobel filters in x and y dimensions. Given image gradient of MHI in x and y dimensions $F_x(x,y)$ and $F_y(x,y)$, the gradient orientation is defined as,

$$\emptyset(x,y) = arctan\frac{F_y(x,y)}{F_x(x,y)} \tag{6}$$

The magnitude of the gradient defined as,

$$M(x,y) = \sqrt{F_x(x,y)^2 + F_y(x,y^2} \tag{7}$$

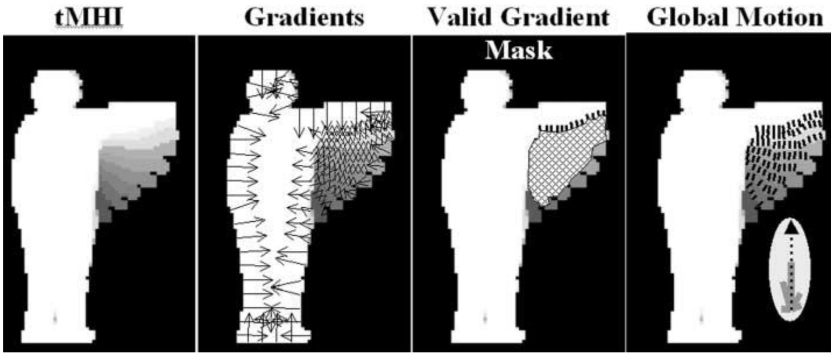


Figure 2.7 The motion history gradients process [31]

The results of gradient can be used to compute the global motion within the duration τ. Although MHI image can be used to segment motion, there is a problem of choosing the appropriate length of duration τ. If τ is too short, the layered motion images cannot be used to segment motion and recognize human movement. If τ is too large, the new motion images overwrite the informative layers which occurred long time ago. So, the parameter τ should be considered in using MHI to classify actions.

2.2 Feature Extraction in the Video

For action classification, there are several approaches which use different time spans such as an entire sequence [32], a single image [17, 22], or a subsequence of video [14, 30, 33-37] , for extracting action features. The suitable number of frames, which are required to recognize human actions, vary according to the datasets, actions and features. Schindler and Gool [14] proposed that primitive actions can be effectively classified by using very short snippets of 1-7 frames based on the observation of biological vision system. The results indicated that by applying the same setting and parameter to different dataset, the most accurate results occurred at the different time. In some cases [28], small errors in temporal alignment or few frames of missed segmentation can cause a huge effect on the recognition rate. The results of subsequence classification, which determine the primitive action, can be used to recognize more complex behaviors [38, 39].

In this work, we also consider action classification in surveillance environment which contains continuous activities and varieties of actions. The main challenge in this scenario is how to segment features that provide useful information from a video sequence which contains more than one action. Segmenting actions into subsequences, which contains primitive movement, can help feature extraction process to extract relevant information from the specified interval. The previous approaches, which recognized human activities based on a sequence of image, can be grouped into two categories: sequential approach, space-time approach, and key frame approach.

2.2.1 Sequential approach

Sequential approaches classify human activities by using a sequence of feature vectors extracted from each image. The models are constructed by using the entire sequence or fixed number of frames in a sliding window. Chaaraoui et al. [17] presented pose representation based on the contour of human silhouette at each frame to find the nearest trained key poses and used Dynamic Time Warping (DTW) to recognize the sequence of key poses. The works in references [24, 27-29] used the histogram of optical flow which is computed at each frame to construct feature vectors. Chaudhry et al. [24] used histogram of oriented optical flow time series with Binet-Cauchy kernels for nonlinear dynamical systems. Perš et al. [28] encoded histogram of optical flow descriptor to detect activities of person entering the restricted area. Ikizler and Duygulu [22] proposed histogram of oriented rectangles and classified actions by using Support Vector Machine (SVM) and DTW. The extension of Hidden Markov Model (HMM) for activity analysis from routine traces was proposed in references [6, 7]. Jiang et al. [40] proposed shape-motion prototype tree to learn and match shape and motion descriptors [23] based on the histogram of oriented gradient (HOG) [20]. Park and Arggarwal [38] estimated body postures by using a hierarchical Baysian network for representing two-person interaction. The approaches can classify complex activities due to the flexibility of feature usage. However, extracting features frame by frame in these sequential approaches requires computational time. The irrelevance is also encoded during the process.

2.2.2 Space-time approach

Space-time approaches recognize human activities by extracting spatial and temporal information from an input video. The extracted feature contains both human shape and motion information within the interested area such as whole human body, parts of the body, and key points. The main advantage of this approach is that the feature can be used to classify actions within the small period of time. By using a small interval, the action recognition can be applied to a sequence which contains more than one action.

In the key point based approaches, Laptev and Linderberg [41] extended Harris corner detection [42] to detect local structure in space-time dimension where the significant variations occurred as shown in Figure 2.8. The approach [41] is robust to noise and does not require the accurate low-level components. However, the result of space-time corner detection is sparse. Gilbert et al. [43] extracted 2D corners independently in each (*x*,*y*), (*x*,*t*), and (*y*,*t*) plane and grouped the features by using a hierarchical process which the mined compound features became more discriminative in each level.

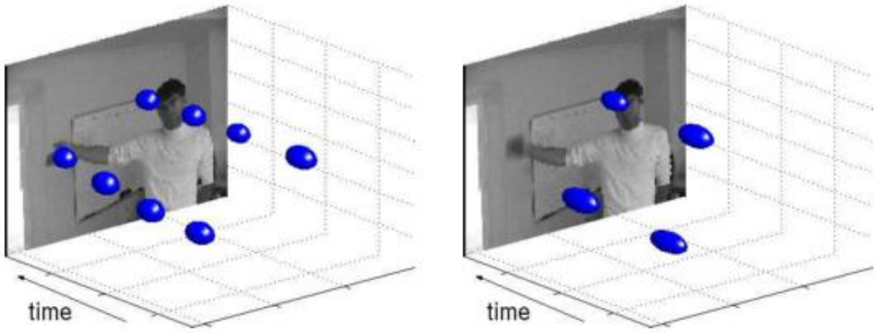


Figure 2.8 Result of interest point detection [41] for waving hand sequences: (a) Interest points for hand gestures with high frequency and (b) Interest points for hand gestures with low frequency.

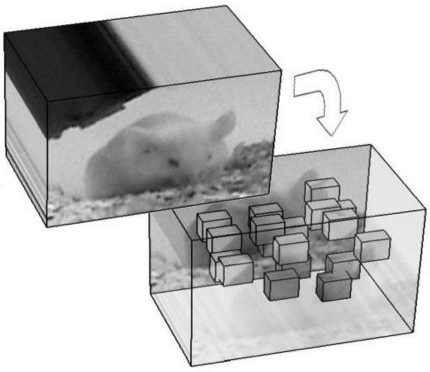


Figure 2.9 Cuboids detection using the spatio-temporal interest point detector [44].

Dollár et al. [44] developed the sparse spatio-temporal feature detector, by using application of separable linear filters to extract cuboids which contain the spatio-temporal windowed pixel values as shown in Figure 2.9. The local cuboids are detected by applying 2D Gaussian smoothing kernel in the spatial dimensions and applying 1D Gabor filters in the temporal dimension. Niebles et al. [45] proposed an unsupervised learning method by using latent topic models on the spatial-temporal words which are extracted by using reference [44]. Lui and Yang [46] proposed the multiple features based on the Affine-SIFT key point trajectories and hybrid/discriminative model for action recognition. Zhang et al. [47] proposed the manifold-constrained sparse representation based recognition to utilize cuboids feature [44]. Zhang and Tal [36] applied Slow Feature Analysis (SFA) on random sampling cuboids which located at motion boundary to extract slow feature function. Ji et al. [37] proposed 3D Convolutional Neural Networks (CNNs) to extract features from the spatial and temporal domain of multiple contiguous frames.

In many researches, the motion information is used to construct feature vector by analyzing the motion pattern in the longer period of time. The motion flow in space-time approaches contain more information to classify action than the detected motion in the consecutive frames. Lui et al. [48] proposed the object motion detection based on the spatio-temporal information saliency map to provide additional object saliency information which can be used in event recognition. Efros et al. [23] recognized human actions at the distance by smoothing and aggregating noisy optical flow to construct spatio-temporal motion descriptors. Ali and Shah [25] derived kinematic features which extract dominant kinematic trends of dynamics of the optical flow pattern from an image sequence. Shechtman and Irani [49] applied 2D image correlation into 3D space-time to estimate motion flow. In reference [34], Poisson equation was applied to extract space-time features from 2D volumetric shape.

Bobick and Davis [30] construct MEI and MHI within a space-time volume as shown in Figure 2.6. Hu and Boulgouris [32] classified human actions by using posture information, which based on centered MEI and motion information. Tian et al. [35] proposed the hierarchical filtered motions, which filtered HOG of MHI [30] by interested points [42] to classify actions in crowded videos. There are several approaches based on MHI applied in many applications [50, 51] by using the advantage of simplicity and low computation. However, the limitation of MHI has to be considered e.g., motion self-occlusion, speed variation, and dynamic background.

2.2.3 Key Frame Approach

The Key Frame approach is a method to find the discrimination characteristic of features. The extracted features are grouped or manually selected to find the representations for each action instead of using all of the extracted features to classify actions. Also the action representation is considered to extract more compact content. For sparse and compact action representation, the works in references [17, 18, 50, 52, 53] extracted key frame(s) to represent the discriminative features for classifying actions. References [17, 53] found key poses of each action by using K-mean clustering with Euclidean distance. To extract a set of the representative, K-mean clustering is used to separate data into groups and find means of the clusters. The standard algorithm of K-mean [54] uses iterative refinement technique to adjust the data in each cluster.

Given a set of observation $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ and defined K clusters $(K \leq n)$. The initial mean vectors and cluster spaces are $(\mu_1, \mu_2, \ldots, \mu_K)$ and $(S_1, S_2, \ldots, S_K)$ respectively. In each iteration, the samples are assigned to the cluster space S_i which μ_i is nearest to the sample. After the classifying, every cluster recomputed μ_i. The process is running until there is no change in μ_i or the number of iteration reach a maximum number of iteration which we set. The objective of K-mean clustering is to minimize the within-cluster sum of squares defined as

$$\underset{S}{argmin}\sum_{i=1}^{K}\sum_{x_j\epsilon S_i}\left\|x_j-\mu_i\right\|^2 \quad (8)$$

The disadvantage of K-mean clustering is its sensitivity to noise and outlier. Also, the number of cluster K has to be defined before clustering and usually based on the empirical testing.

Baysal et al. [18] clustered line-pair features by using K-medoids and selected the candidate of the top rank as key frames. K-medoids clustering is computed in the same way as K-mean clustering. But the representative of each group is a data point instead of a mean.

There is other method for selecting key frames or key poses. In [50], the approach found a set of the most discriminative key frames by measuring the entropy of the generated visual words. Raptis and Sigal [52] modeled actions by using local key frames, which gathered partial key poses of performers. However, the approaches did not concern the characteristic of biological motion and a relationship between motion and key frames.

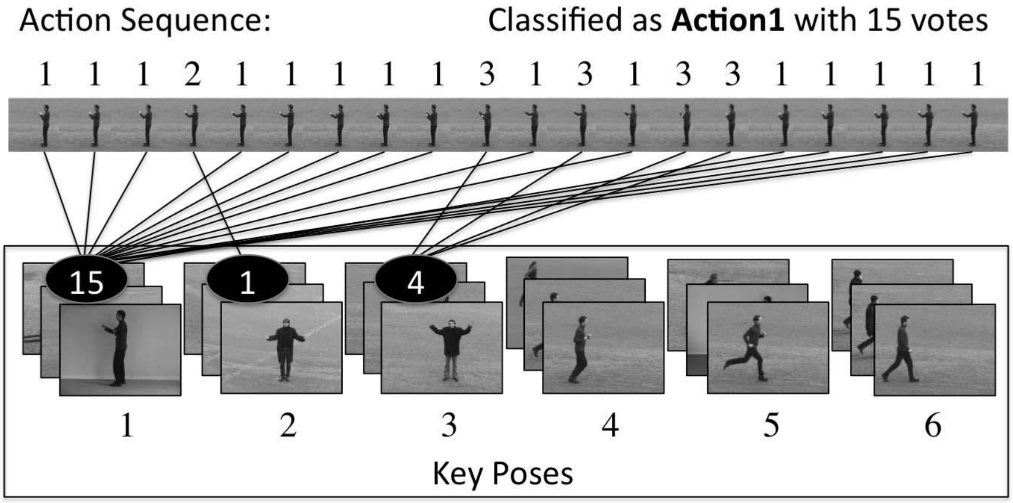


Figure 2.10 Action classification using key poses [18]

CHAPTER 3
HUMAN ACTION CLASSIFICATION

In biological perception [55, 56], the motion cues can be used to segment event boundary, where one movement ends and another begins. The change in motion is one of the low-level visual cues that has an ability to activate the brain responses. So, the distinctive changes in motion are important to detect key frames and primitive action intervals.

In visual analysis, the normal activities consist of periodic and non-periodic motions. The repeatability of postures can be used to classify actions as well. The similarity matrix of extracted foreground of walking, running, jumping-jack, bending, and waving with one hand are shown in Figure 3.1. The brightness indicates the correlation distance between extracted foreground frames. The similarities between frames in the same action indicate that there is a period which the distance becomes small and the self-similarity evolves in time. The periodicity and the self-similarity pattern do not occur in the different actions at the same time. A characteristic of the periodic motion which appear in action such as walking, running, jumping-jack, and waving with one hand, is the self-similarity. Even though, the actions are periodic, the period of action cycles is different. For non-periodic motion such as bending, although it contains a single bend down and rise up. So, there are repeatability of postures in some non-periodic motion.

For the periodic motion, the motion of a point [15] can be defined as $\vec{X}(t+p) = \vec{X}(t) + \vec{T}(t)$, where $\vec{X}(t)$ is a motion of the point at time t and $\vec{T}(t)$ is a translation of the point at time t. The period p is a period of a motion cycle. Our work is to find the period p that specifies an event boundary of the motion perception. Also, the period should be automatically adapted based on actions and performers. By finding the local extrema, we obtain an event boundary and a key posture. The boundary is used to find period p of a primitive action and the number of frames to extract features. The key postures can be used to discriminate actions.

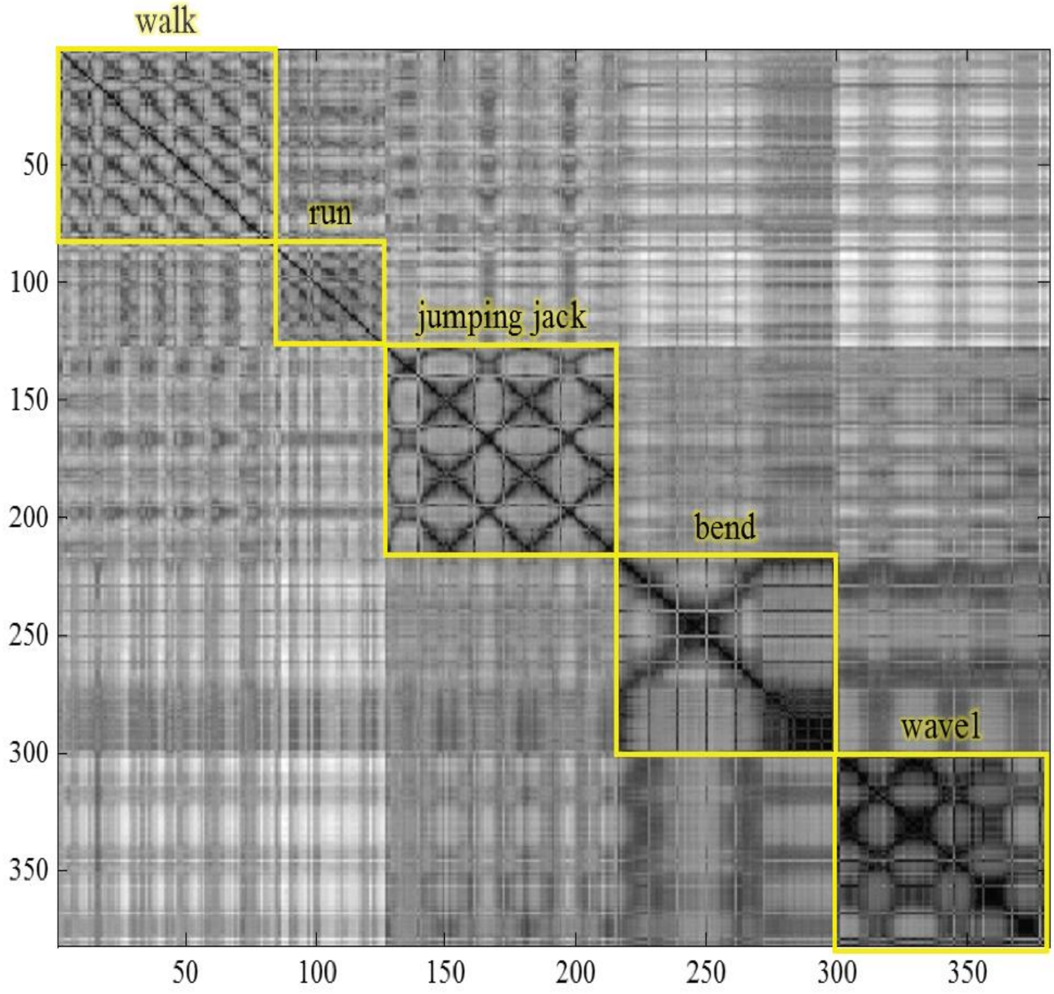


Figure 3.1 Similarity matrix of walking (frame 1-84), running (frame 85-126), jumping-jack (frame 127-215), bending (frame 216-299), and waving with on hand (frame 300-381) of a video sequence from Weizmann dataset

In this work, the automatic event boundary detection by a low-level vision variation is used to segment the meaningful features for human action recognition. By detecting the beginning and ending of an event, the system adjusts an appropriate number of frames for different actions and performers. The adaptive interval also solves a problem of speed variation and extracts informative representation of human primitive actions. The issue of selecting appropriate number of frames for extracting features using speed variation in actions and performers is considerable important. Since the movement of different actions may have different velocities such as walking and running, the number of frames has to be adjusted to capture a whole cycle of action.

The extracted features should base on the difference in time domain to extract the discriminative and informative representation among actions. The window size in temporal domain also causes motion overwriting problem of the Motion History Image (MHI) [30] which occurs when a time span is too large [32]. Figure 3.2(a) shows the MHI of bending for the whole sequence duration. The information of bending down is lost as can be seen from the extracted features. Figure 3.2(b) and Figure 3.2(c) then show the informative and compact representation of the actions.

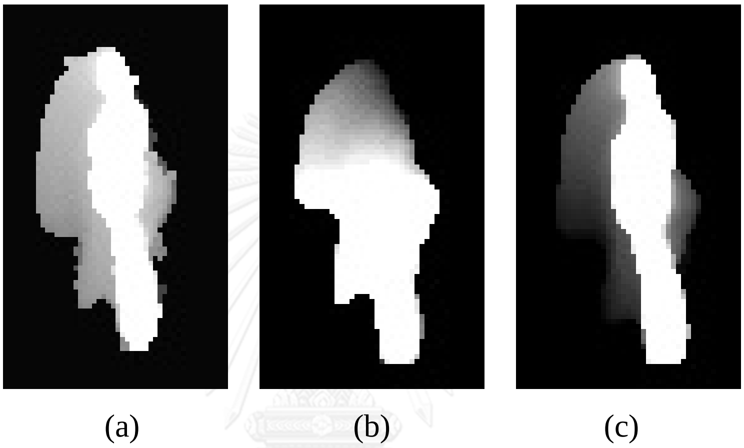


Figure 3.2 MHI of bending contains 2 steps: bending down and rising up.
(a) MHI of maximum duration of bending. (b) MHI of bending down interval,
(c) MHI of rising up interval.

3.1 System Overview

The overview of our system is shown in Figure 3.3. The process of extracting features starts with localizing and tracking human. In this work, we use the extracted foreground and bounding box information which provided by the dataset publisher. The silhouette and the difference between consecutive frames are used in both key frame detection and feature extraction stages.

To detect key frames as shown in Figure 3.4, motion variation is computed at each consecutive frames. The number of motion pixels is then used to compare with the number of motion pixels from previous frames to specify the critical point in the temporal domain. The key frame image shows the similarity of postures, which are identically within the same action class, as shown in Figure 3.5. The number of key frame is not fixed and automatically adjusted based on the characteristic of the action. So, the number of detected key frames in a sequence also varies. For instance, the numbers of key frames of walking (84 frames), running (42 frames), jumping-jack (89 frames), bending (84 frames), and waving with one hand (82 frames), are 5, 4, 5, 3, 7 frames, respectively.

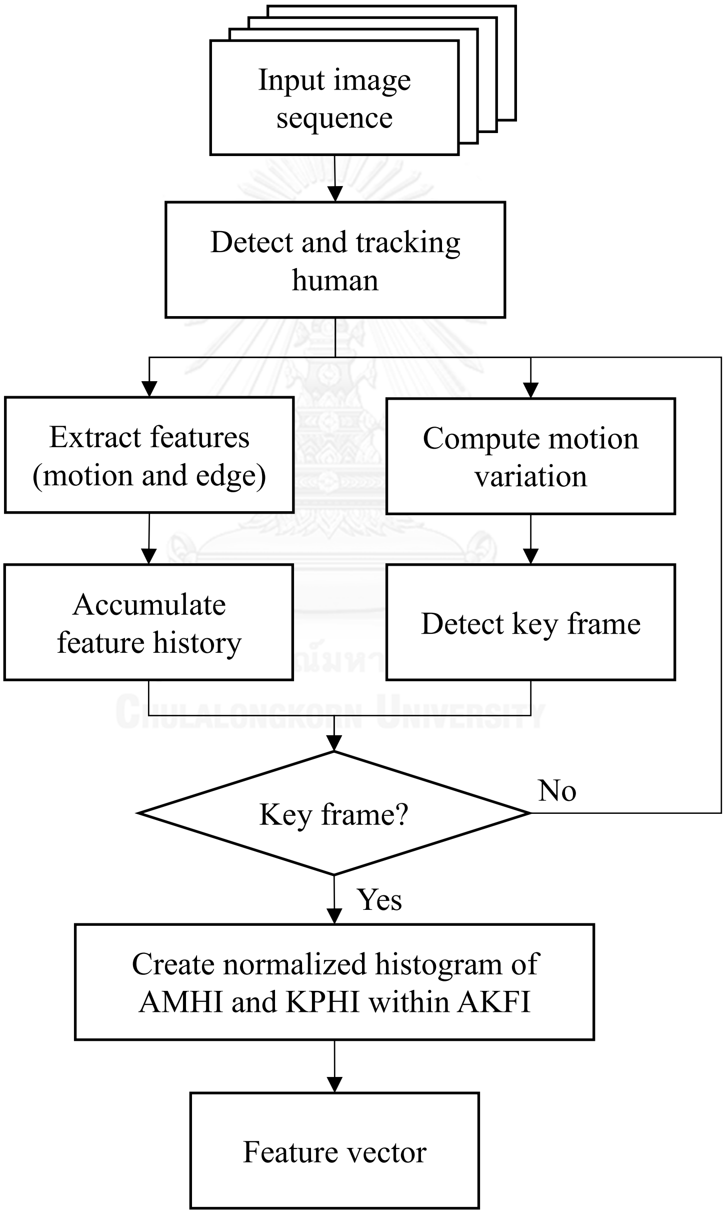

Input image sequence
Detect and tracking human
Extract features (motion and edge)
Compute motion variation
Accumulate feature history
Detect key frame
Key frame?
No
Yes
Create normalized histogram of AMHI and KPHI within AKFI
Feature vector

Figure 3.3 Overview of the proposed system.

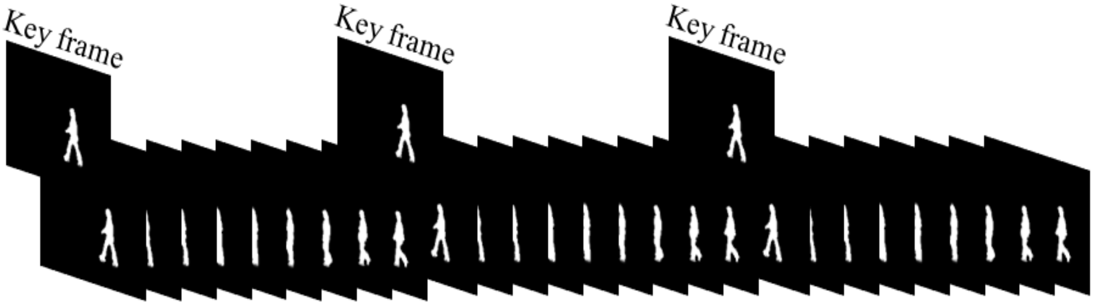
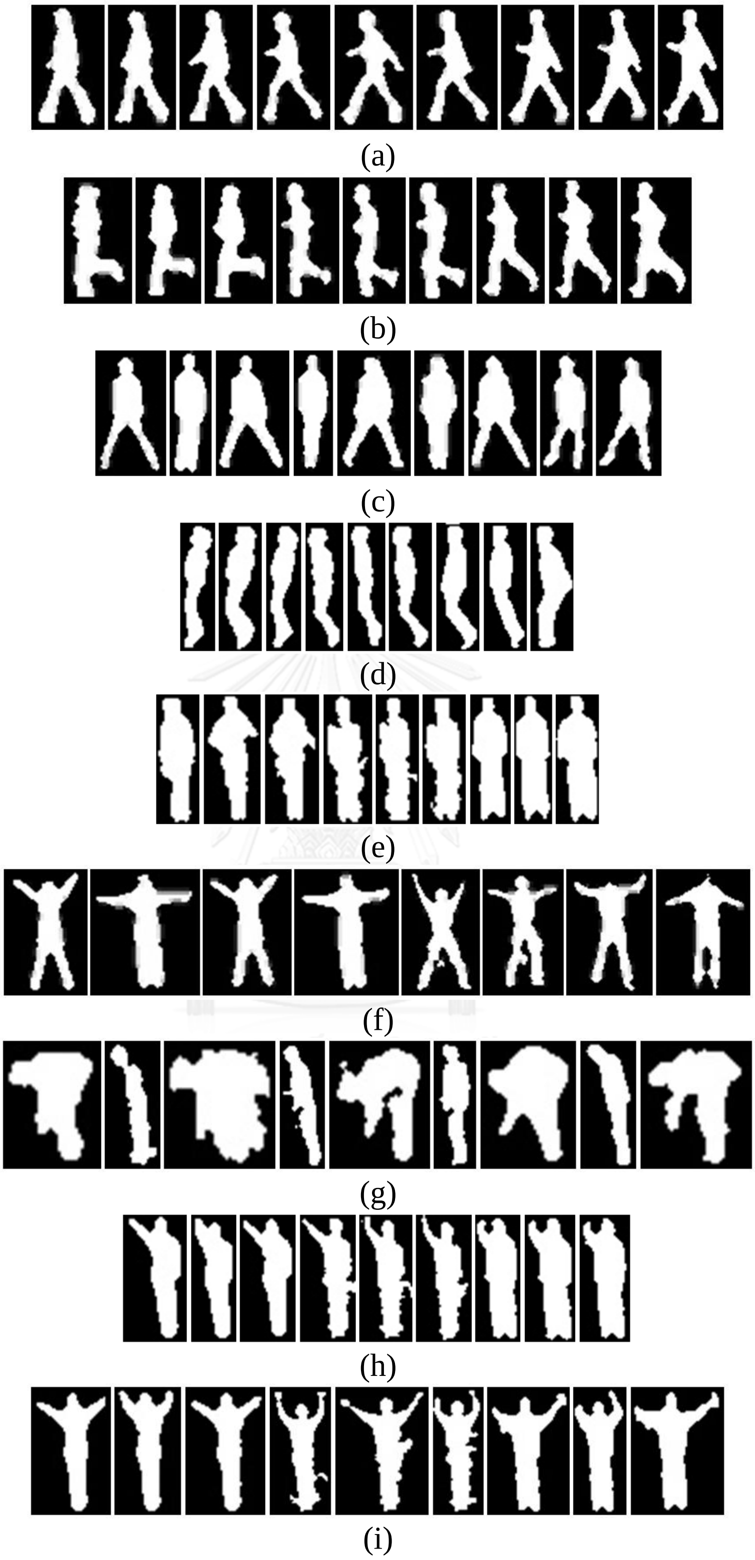


Figure 3.4 Key frame extraction

Simultaneously, the system accumulates features and constructs feature history images. At each key frame, the history images are normalized by the duration of Adaptive Key Frame Interval (AKFI) to construct Adaptive Motion History Image (AMHI) and Key Pose History Image (KPHI), as shown in Figure 3.6. AMHI is a layered silhouette image which is used to extract motion direction during AKFI instead of computing motion direction in every frame. KPHI is an aligned layered silhouette image similar to references [32, 50] within AKFI. The features are accumulated through time until another key frame occurred. Then, the local HOG [20] are created within sub-regions of AMHI and KPHI. A feature vector is constructed by concatenating the local oriented histograms. The primitive actions are classified for each AKFI by finding *k* nearest neighbors from the training data.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 3.5 Extracted key frame from Weizmann dataset. (a) walk, (b) run, (c) gallop sideways, (d) jump- forward-on-two-legs, (e) jump-in-place-on-two-legs, (f) jumping-jack, (g) bend, (h) wave-one-hand, and (i) wave-two-hands.

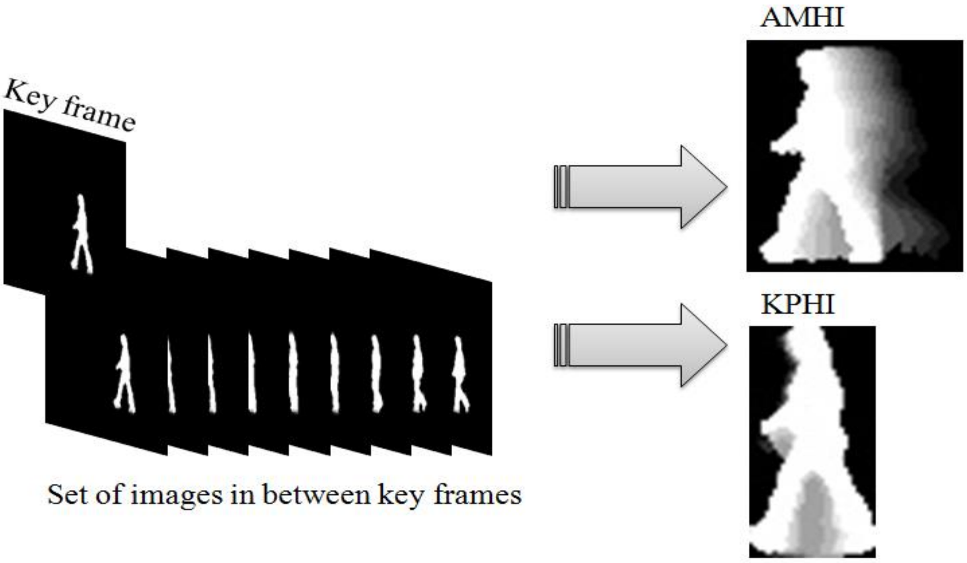


Figure 3.6 AMHI and KPHI image constructed from an AKFI.

3.2 Adaptive Key Frame Interval (AKFI) Extraction

In this work, we consider the variation of motions, which is produced by the articulated body over time, for segmenting a subsequence action. We observed that human postures are similar during the period of increasing or decreasing speed. While changing posture, the speed of the movement is stable or slightly changed, as shown in Figure 3.7.

To extract a key frame from a sequence, we use a number of motion pixels and a number of silhouette pixels to observe the variation. For inter-frame motion, frame differencing is used to estimate a binary foreground image, as shown in Figure 3.8(a). A binary foreground image is extracted by using background subtraction, as shown in Figure 3.8(b). The number of foreground pixels is used to normalize motion due to scaling.

Given the preprocessed binary foreground image $F(x, y, t)$ and binary motion image D(x, y, t) at time t. The motion variation is defined as,

$$C_m(t) = \frac{\sum_{(x,y)\in I} D(x, y, t)}{\sum_{(x,y)\in I} F(x, y, t)} \tag{9}$$

, where I is the spatial extent of pixels in an image. Before using $C_m(t)$ to detect key frame at time t, the values are smoothed by averaging $C_m(t)$ over the duration τ. Since a number of motion pixels from frame differencing varies due to the speed of movement, the value of $C_m(t)$ can varied depending on the motion speed, as shown in Figure 3.7.

From Figure 3.7, the chart indicates the variation of the number of motion pixels to the number of foreground pixels through time for walking sequence [34]. Key frames are then identified at the local minima or the local maxima of the $C_m(t)$ plot. By taking the first derivative of $C_m(t)$, a slope of the signal is computed. The local minima and local maxima locate at points which slopes switch from decreasing to increasing and from increasing to decreasing, respectively. The extracted silhouettes at key frames periodically repeat at the maxima and the minima. By detecting key frames, the segmented interval of primitive movement locates in between the two contiguous key frames.

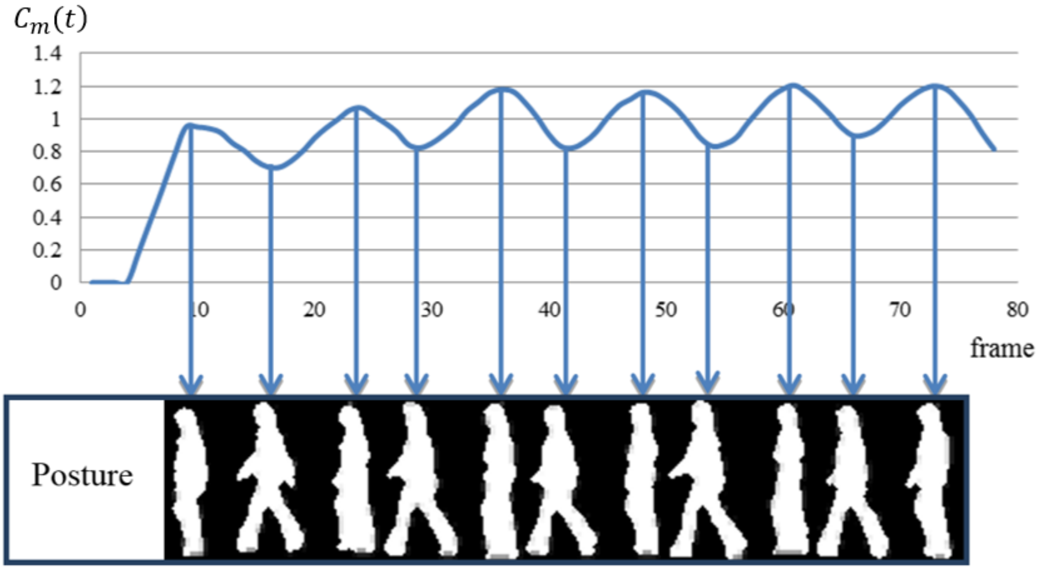


Figure 3.7 Description of motion variation over time (top) and silhouette at the local minima and the local maxima motion of walking (bottom).

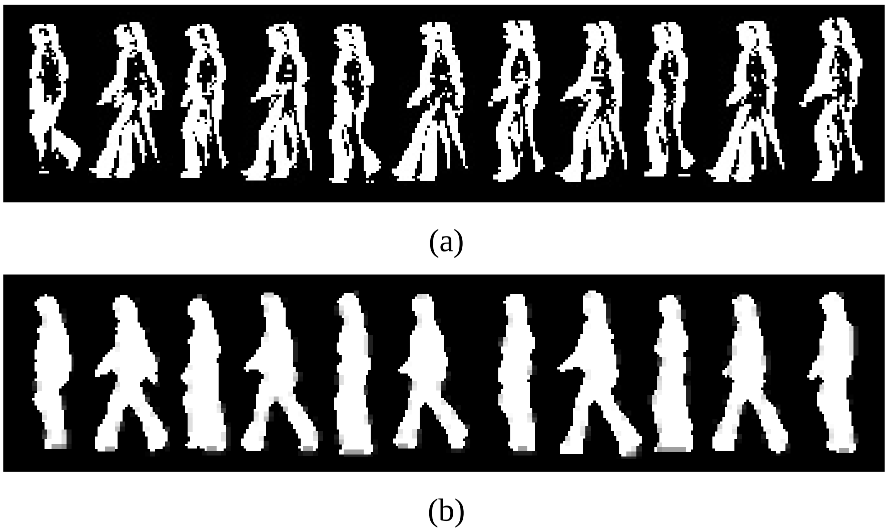


Figure 3.8 Extracted information at the key frame (a) Frame differencing of walking (b) Foreground of walking

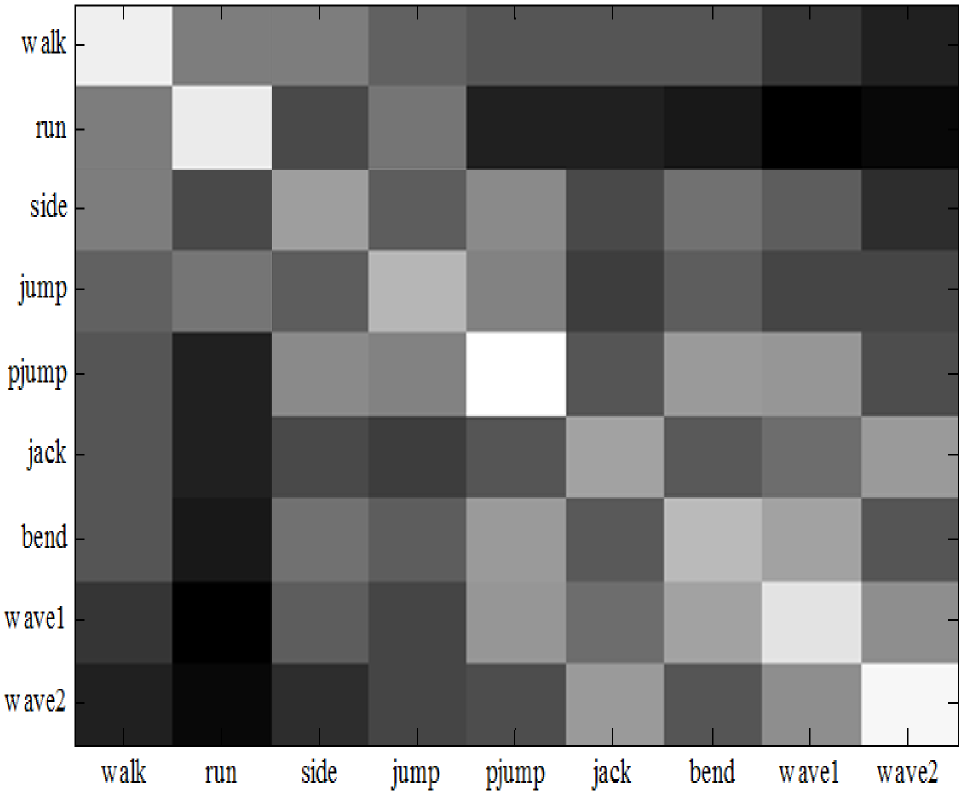


Figure 3.9 A confusion matrix of average correlation of histogram of oriented gradient of key pose.

In this work, we use only the local minima for detecting key frames. Since The motion variation rapidly increases when illumination and noise occur. Then, the detected local maximum point is not based on the human motion but based on the amount of environment variation which is not desirable for our system. Although the system uses only local minima point as key frame, the appearance and motion information at the local maxima are included within the AKFI.

Figure 3.5 illustrated key frames of 9 actions: walk, run, gallop sideways, jump-forward-on-two-legs, jump-in-place-on-two-legs, jumping-jack, bend, wave-one-hand, and wave-two-hands from Weizmann dataset [34]. The number of key frames can be varied due to the number of primitive movement of each action. Although the extracted silhouettes indicate the similarities within the class, there are some possible confusion between classes such as walk and run, gallop sideways and jump-in-place-on-two-legs, wave-one-hand and wave-two-hands, as those actions share similar key postures.

In Figure 3.9, the confusion matrix indicates the actions classification results by using the silhouette of the detected key frame as shown in Figure 3.5. By normalizing the size of the silhouette, the distance between testing key frames and training key frames are compared. The key frames are classified by using K-nearest neighbor. The grayscale intensity illustrated that the brighter shade implies the higher value of classifying than that of the darker shade. The correct classifying is illustrated in the main diagonal line of the matrix. From the first row of the confusion matrix, the brightest is in the first column which means most of walking key frames can be correctly classified as walking. While the results of side (gallop sideways) in the third rows, the intensity in the third column is not outstanding compare to the other columns. In addition, the incompleteness of the extracted silhouette sometimes occurs. So, using only key poses is not enough to achieve accurate action classification. Thus, we propose to use information in AKFI to construct the feature in spatial and temporal domains. A set of images between key frames, as shown in Figure 3.4, are thus used to construct AMHI and KPHI, as shown in Figure 3.6.

Once the key frame is detected, our system collects movement information which consists of motion and posture. During AKFI, both features, i.e., AMHI and KPHI, are extracted from each frame until the stopping point occurs, i.e., at the next key frame. Next, the image oriented gradient of both features are computed and used to construct feature vector.

3.3 Adaptive Motion History Image (AMHI)

AMHI represents motion occurred during key frame interval. The concept of AMHI is similar to MHI [30]. However, the maximum duration of AMHI is not fixed. Time duration is adapted to the speed variation which is specified by key frame interval. The successive layered silhouette, i.e., motion indicated by frame differencing, is constructed and normalized at each segmented time. At the end of the key frame interval or at the key frame, the layered image is normalized by the number of frame within the key frame interval.

Given, t_{pk} is a timestamp of a previous key frame and t_{ck} is a time stamp of a current key frame. The feature extraction process of each feature starts at t_{ck} and ends at t_{pk}. $AMHI_{t_{ck}}(x, y)$ at t_{ck} is constructed by using foreground region $F_{\mathrm{t}}(x, y)$ from t_{pk+1},which is a starting time of an AMHI image, to t_{ck} , which is the ending time of accumulating AMHI. $AMHI_{\mathrm{t}}(x, y)$ at time t, while t is not equal to t_{ck} and $t > \ t_{pk}$, is defined as,

$$AMHI_t(x, y) = \begin{cases} t - t_{pk} & , if \ F_t(x, y) > 0 \\ max(0, AMHI_{t-1}(x, y) - \ t_{pk}) & , otherwise \end{cases} \tag{10}$$

From eq. (10), the white silhouette of foreground area, $AMHI_t$ is equal to t, as shown in Figure 3.6. While, the other areas contain the value from the previous frame $AMHI_{\mathrm{t-1}}(x, y)$, or the value is set to 0 if $AMHI_{\mathrm{t-1}}(x, y) > \ \ t_{pk}$.

At time t_{ck}, $AMHI_{t_{ck}}(x, y)$ is normalized by the number of frame within the consecutive key frames interval $t_{ck} - t_{pk}$. The normalized $AMHI_t$, defined as $nAMHI_{t_{ck}}$, which contains successive layered silhouette and motion trajectory are used to extract local feature. AMHI calculated from walking sequence is shown in Figure 3.6. The image indicates the successive motion gradient within the adaptive cycle. The pixel intensity varies with the timestamp. The brighter pixels correspond to the recent silhouette which is similar to MHI [30].

3.4 Key Pose History Image (KPHI)

KPHI represents human posture over a period of time. The center of a silhouette in each frame is aligned to create a successive layered of silhouette image within a key frame interval. By aligning the center, KPHI emphasizes the body part movement. In Figure 3.6, KPHI indicates the movement of legs and arms while other parts of the body have no movement.

To construct KPHI, foreground region $F_{\mathrm{t}}(x, y)$ is segmented and aligned at the center of $KPHI_t(x', y')$, where $x' = x + dw$, $y' = y + dh$, dw and dh are aligned parameters for adjusting the center of foreground region to the center of $KPHI_t$. $KPHI_t(x', y')$ at time t, while t is not equal to t_{ck} and $t > \ t_{pk}$, is defined as,

$$KPHI_t(x', y') = \begin{cases} t - \ t_{pk} & , if\ F_t(x, y) > 0 \\ max(0, KPHI_{t-1}(x', y') - \ t_{pk}) & , otherwise \end{cases} \tag{11}$$

Similar to $AMHI_{t_{ck}}(x, y)$ and $KPHI_{t_{ck}}(x', y')$ are normalized by the period $t_{ck} - t_{pk}$. The normalized $KPHI_{t_{ck}}$, defined as $nKPHI_{t_{ck}}$, which contains the layered silhouette are used to extract local features for constructing feature vector.

3.5 Feature Vector

The layered images of $nAMHI_{t_k}$ and $nKPHI_{t_k}$ are used to compute image oriented gradient by convoluting with Sobel filters separately. The gradient of the images is orthogonal with the boundary. While, $G_{Kx}(x, y)$ and $G_{Ky}(x, y)$ are spatial derivative in x and y dimensions of $nKPHI_{t_k}$, respectively. Gradient orientation of $nKPHI_{t_k}, \theta_{\mathrm{K}},$ at each pixel is defined as,

$$\theta_K(x, y) = arctan \frac{G_{Ky}(x, y)}{G_{Kx}(x, y)} \tag{12}$$

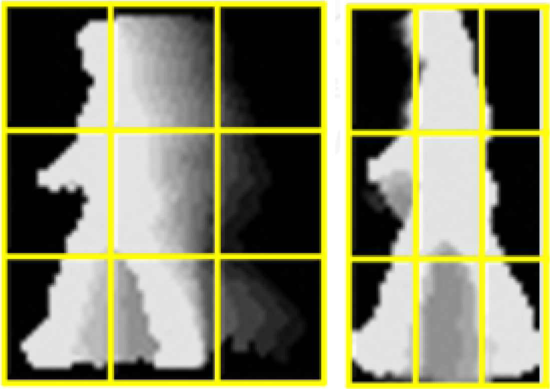
, where $G_{Mx}(x, y)$ and $G_{My}(x, y)$ are spatial derivative in x and y dimensions of $nAMHI_{t_k}$, respectively. Gradient orientation of $nAMHI_{t_k}, \theta_{\mathrm{M}}(x, y)$, at each pixel is defined as,

$$\theta_M(x,y) = arctan\frac{G_{My}(x,y)}{G_{Mx}(x,y)} \tag{13}$$

For AMHI, the gradients indicate the normal optical flow for specifying motion direction occurrences during primitive movement.

Feature vector is represented by local orientation histogram of $\theta_{\mathrm{K}}(x,y)$ and $\theta_{\mathrm{M}}(x,y)$. AMHI and KPHI images are equally divided into non-overlapped $n \times n$ regions over a region of interest, which layered silhouettes occur, as shown in Figure 3.10. From the empirical experiment, we varied the number of regions in the same setting experiment and found that $n = 3$ gives the best results for our representation. In this works, the region of interest is divided into 3×3 regions (when $n = 3$).

At each sub-region, the histogram of orientation is created and normalized. Feature vector $H = \{h_1, h_2, h_3, \dots, h_{nxn}\}$ is a concatenating of local histogram of $nAMHI$ and $nKPHI$. The histogram values in sub-region h_i are weighted by a ratio of the number of sub-region pixels to the total number of pixels $W = \{w_1, w_2, w_3, \dots, w_{nxn}\}$, which satisfies $\sum_{i=1}^{nxn} w_i h_i = 1$.



(a) (b)

Figure 3.10 The 3x3 regions of (a) AMHI and (b) KPHI.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSIONS

In this chapter, the experimental results of human action classification are presented. We tested our system with datasets that have different scenarios. The datasets include Weizmann dataset, KTH dataset, and UT Interaction dataset. Weizmann dataset [34] is used for testing actions with periodic and non-periodic movements of a single person. The results of key frame detection indicate the similarity of posture through time, as shown in Figure 3.5. The Weizmann dataset includes sequences for testing robustness in both deformation and viewpoint variants. KTH dataset [57] contains an amount of human action sequences with different scenarios. UT Interaction dataset [58], contains non-periodic human-human interactions in two scenarios. The datasets contain both normal and abnormal actions in surveillance scenarios.

4.1 Weizmann dataset

The Weizmann video database [19], as shown in Figure 4.1, is a public human action database containing 81 sequences of 9 human actions: walk, run, gallop sideways, jump- forward-on-two-legs, jump-in-place-on-two-legs, jumping-jack, bend, wave-one-hand, and wave-two-hands, performed by 9 people. From the environments of the scenario, we can identify normal actions which are walk, run, wave-one-hand, and wave-two-hands. The gallop sideways, jump- forward-on-two-legs, jump-in-place-on-two-legs, jumping-jack, and bend can be used as anomaly. The sequences have the spatial resolution of 180x144 pixels and have a length of four seconds each. The background is static and uniform. This dataset is included the extracted foreground by background subtraction.

To test robustness, Weizmann robustness dataset [19] contains videos of walking in different variations of both deformations and viewpoints. The deformation datasets were used to test the system sensitivity in the case of partial occlusions, body deformation, and irregular performance, as shown in Figure 4.2. Also, the robustness dataset for different viewpoints varying from 0° to 81° of a person walking is included, as shown in Figure 4.3.

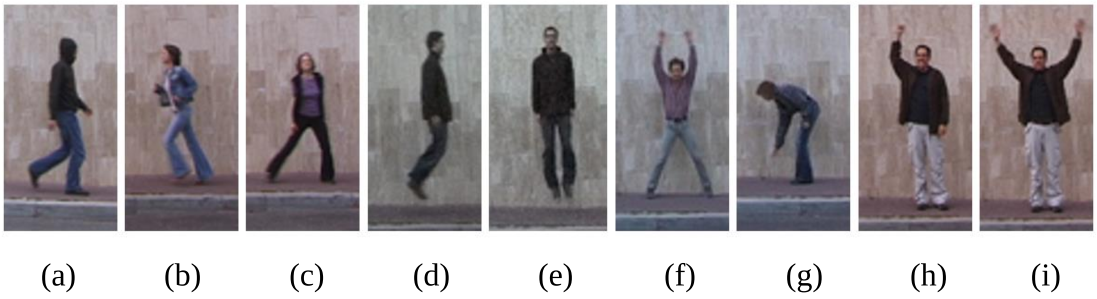


Figure 4.1 Weizmann dataset contains 9 actions: (a) walk, (b) run, (c) gallop sideways, (d) jump- forward-on-two-legs, (e) jump-in-place-on-two-legs, (f) jumping-jack, (g) bend, (h) wave-one-hand, and (i) wave-two-hands.

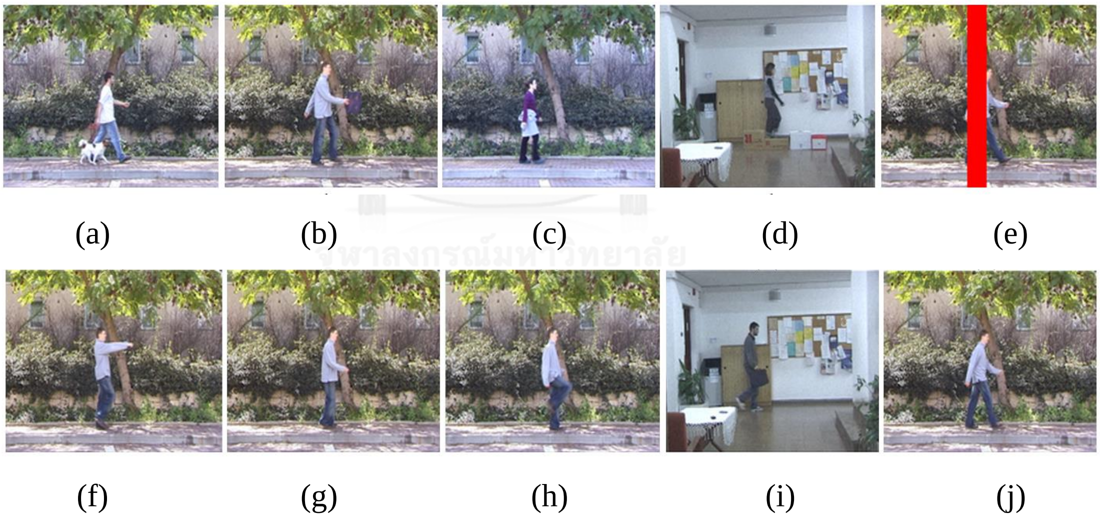
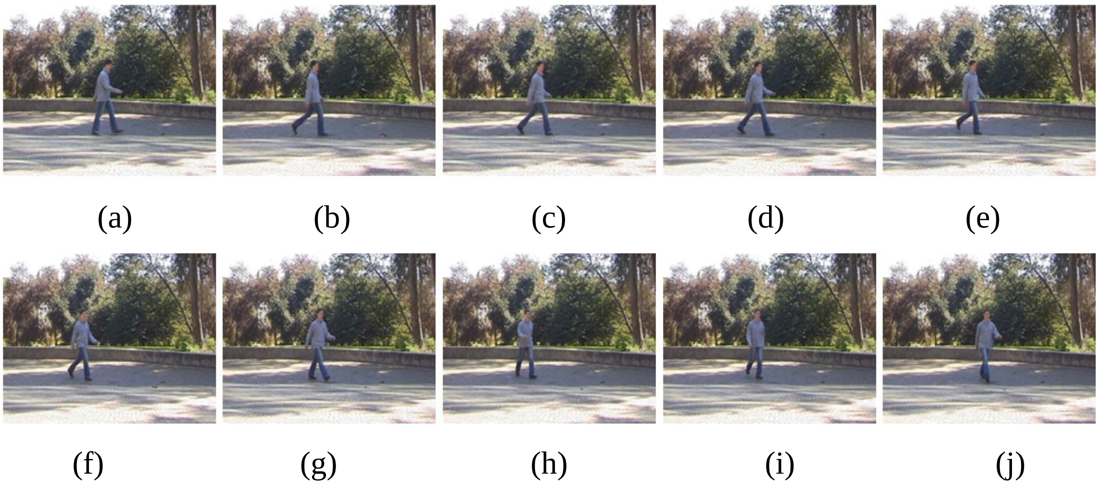


Figure 4.2 Robustness dataset contains 10 walking videos in the different scenarios, (a) walk with a dog, (b) swinging a bag, (c) walk in a skirt, (d) occluded feet, (e) occluded by a pole, (f) moonwalk, (g) limp walk, (h) walk with knee up, (i) walk with briefcase, and (j) normal walk.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 4.3 Robustness dataset contains 6 walking videos in the different viewpoints, (a) 0°, (b) 9°, (c) 18°, (d) 27°, (e) 36°, (f) 45°, (g) 54°, (h) 63°, (i) 72°, and (j) 81°.

For feature extraction, we use foreground masks provided by [19], to construct the AMHI and KPHI and use a number of foreground pixels to detect key frames as explained in Eq. (9). Based on our empirical results, 8-bin histogram and 3x3 regions of human body give the best results in Weizmann dataset. So, we use 8-bin histogram of oriented gradient at each non-overlapped 3x3 regions. The width and height of the regions vary based on the size of the region of interest of an image. The proposed features are tested in 3 modes: using AMHI only as a feature, using KPHI only as a feature, and a concatenated of AMHI and KPHI as a feature. The leave-one-out cross validation is used to measure the precision of the classification. In the training process, the testing sequence is separated from other sequences while the system constructs the database. The testing key frames of the sequence are then compared with the training set and are categorized by K-nearest neighbors. The majority voting is used to classify action in sequence.

From Table 4.1, the classification results based on the three features which are extracted within key frame interval are presented. The overall number of AKFI for Weizmann dataset is 369 intervals. While the overall number of features extracted with the fixed 5 frames windows (snippets of 5 frames) scheme is 1287 intervals. K-nearest neighbor (K = 1) is used to classify actions. The results of feature which is the concatenation of AMHI and KPHI information have the highest in term of accuracy among others. According to the key frame interval, the extracted feature contains more discriminative information compared to the feature which is constructed from the fixed length window. We used a sliding window of 5 frames, which is an average of the key frame interval among periodic actions, to compare with the adaptive key frame extraction. The recognition rate obviously increases in all features when using AKFI window length comparing to fixed sliding window. As mentioned in CHAPTER 3, window length effects the result of action discrimination. The feature, which is extracted within a large time interval, might have the motion overwriting and might be affected by the speed variation. On the other hand, a short interval causes the extracted feature lacking of information to classify action.

Table 4.1 A comparison of recognition rates among features on Weizmann dataset.

Feature	Oriented gradient of AMHI (2)	Oriented gradient of KPHI (1)	(1) and (2)
Accuracy (%) K-nearest neighbor (per AKFI)	95.10	92.84	97.46
Accuracy (%) K-nearest neighbor (snippets of 5 frames)	84.39	79.73	86.16
Accuracy (%) K-nearest neighbor (per AKFI) Concatenated sequence	94.52	92.80	97.44
Accuracy (%) SVM (per AKFI)	95.10	92.84	97.46
Accuracy (%) Sparse representation (per AKFI)	90.78	92.80	96.83

The accuracy of the concatenated sequence in Table 4.1 implies that we concatenated sequences of all actions from the same performer into a new sequence. Then, training and testing are based on the concatenated sequences to test the accuracy in the situation that there are several actions in a video. The result of the concatenated sequence indicates that the AKFI can be used to separate actions within the sequence. So, the extracted features can be used to classify action as well as features which are extracted from one action sequence.

The extracted features are also tested with the different classifier. The classification results from K-nearest neighbor, support vector machine (SVM), and sparse representation are shown in Table 4.1. The K-nearest neighbor classifier assigns action label by finding the training action feature which has the minimum distance from the testing feature. The SVM algorithm constructs action models by separating training data categories with gaps that are as wide as possible. While the sparse representation uses dictionary learning to find a basis which represents actions. For the testing process, a testing descriptor is decomposed and reconstructed by using Orthogonal Matching Pursuit (OMP) method. The residual of the sparse representation is used to classify action by finding the minimum residual from the dictionary basis. From the classification results, K-nearest neighbor, SVM, and sparse representation classifier results are similar. The extracted features distribution are heterogeneous and compact. Also, it can be used to classify actions by using the different classifier while preserving the accuracy.

The comparison of features on each action of Weizmann dataset are shown in Figure 4.4 which indicates that AMHI and KPHI are capable to classify different actions. For instance, walking and running have similar motions while their key frame postures are different, so KPHI is considered a better feature than AMHI. While running and jumping are similar in posture, however, they are different in motion. By combining AMHI and KPHI, the features that contain both posture and motion information can improve the accuracy rate.

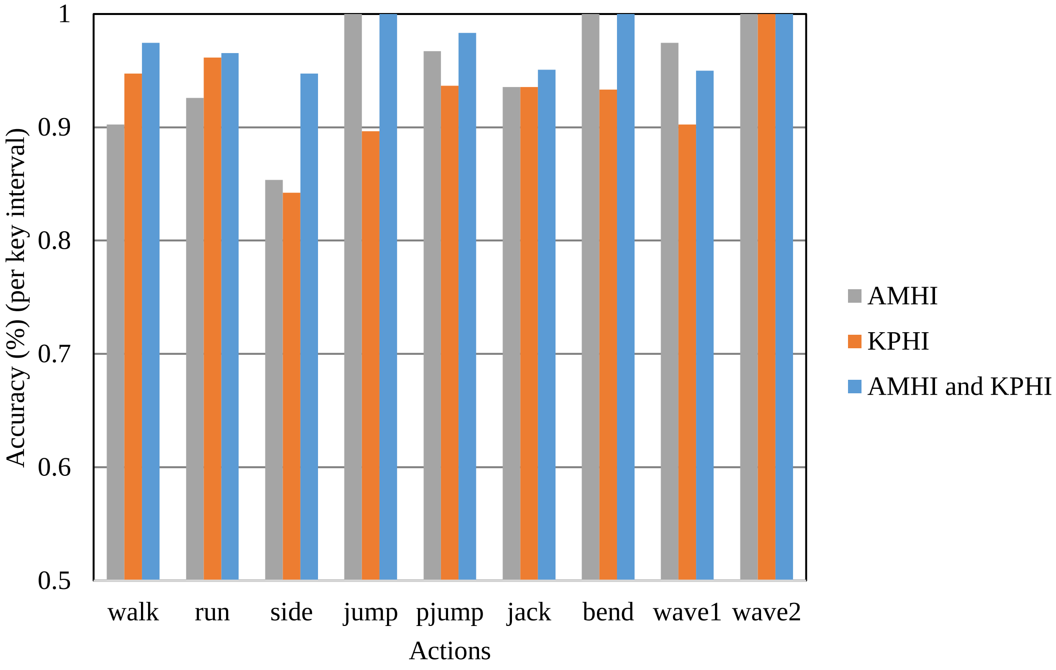


Figure 4.4 Comparison of features on each action of Weizmann dataset.

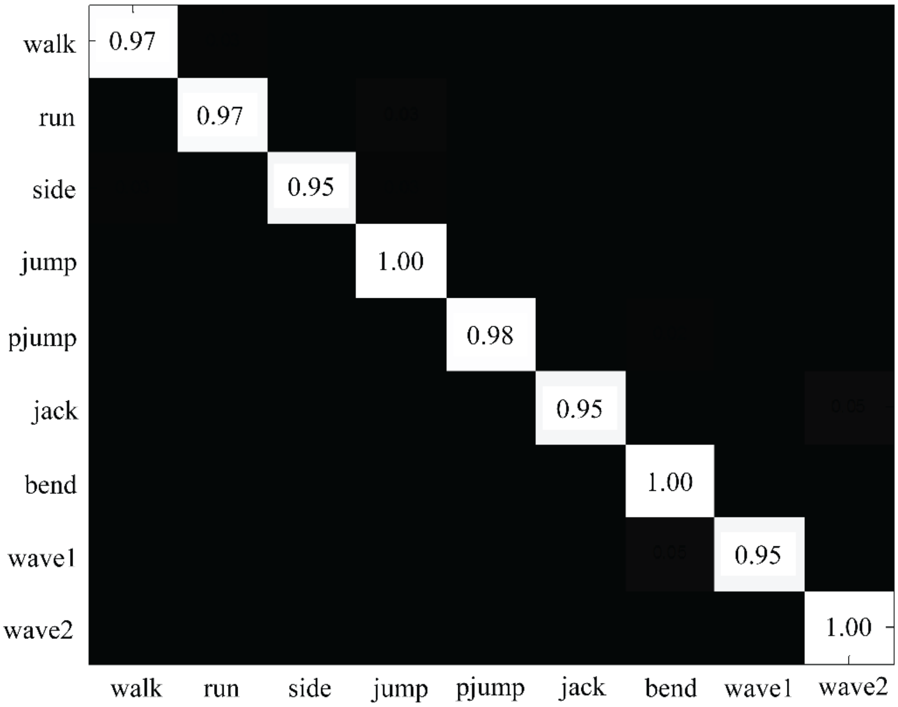


Figure 4.5 Confusion matrix of action classification (per AKFI) by using the concatenation of AMHI and KPHI features

Most of the misclassification locate at the first or at the last key frame of a sequence. Because a cycle of action is not completed, therefore, features are not properly constructed at that time. The other misclassified results are from the confusion of the similar movement actions such as walk and run. Moreover, the similarity of the partial movement such as arm movement of jack and wave2, the upper body of bend, and arm movement of wave1, affects the classification rate, as shown in Figure 4.5.

From Table 4.2, the classification results for the entire sequence of our proposed method and other comparative works, References [17, 18, 21, 22, 34, 59], are presented. We use leave-one-out and the same number of actions for testing recognition similar to other comparative works. The testing sequences are classified by K-nearest neighbor for each AKFI. Then, majority voting is used to identify action based on the results of K-nearest neighbor. Our method can classify actions as well as References [22, 34], [40], in which Reference [40] used DTW for matching for the entire sequence, and Reference [34] used space-time cube with fixed frame sliding window. The works in References [18, 21, 59] used a still image for classification. Some approaches [17, 18] used extracted key poses in the training. However, the appropriate number of key poses per action still needs to be investigated as it is not equal among different datasets. In addition, the approach using frame by frame action classification does not consider temporal information at all in the recognition process. The results demonstrate that using AKFI could considerably increases classification accuracy. By using AKFI, we do not have to specify the number of frames for extracting and matching features. Moreover, our extracted features are more compact with 369 features, while Reference [34] used 549 cubes of space-time features and used 10 frames sliding window with 5 overlapping frames to handle the incomplete movement period. In [36], 2,000 cuboids are collected to learn 200 slow feature functions for each action. Then, the dimension of each feature is 200 x (number of action). Compared to References. [18, 21, 22, 59] in which the features are extracted frame by frame, our feature contains temporal information within subsequences and performs more effectively.

Table 4.2 A comparison of recognition rates among methods on Weizmann dataset.

Method	Accuracy (%) (entire sequence)
Our method	100.00
Blank et al. [34]	100.00
Niebles et al. [59]	72.80
Thurau et al. [21]	94.40
Ikizler et al. [22]	100.00
Baysal et al. [18]	92.60
Jiang et al. [40]	100.00
Chaaraoui et al. [17]	92.77
Zhang et al. [36]	89.33

In detecting anomaly based on the normal events training data, we use the common assumptions of the anomaly [60] which are infrequent occurrence and different characteristic compared to the normal events. From the environments of the Weizmann scenario, we can identify normal actions which are walk, run, wave-one-hand, and wave-two-hands. The gallop sideways, jump- forward-on-two-legs, jump-in-place-on-two-legs, jumping-jack, and bend can be used as anomaly. In this experiment, we separate the dataset into 2 classes which are normal actions and anomaly. The features AMHI and KPHI of normal actions are clustered by using K-mean clustering [54]. Then, the representatives of each group are used to compute distance and assign event label. The thresholds of each cluster is based on the standard deviation (SD). If the distance is less than threshold, the testing sequence is a normal action, otherwise it is assigned as anomaly. The leave-one-out cross validation is used to measure the precision of the classification. In testing process, every actions including normal actions and anomalies are tested. We vary the number of K and threshold to find the optimal parameters for classification. From the experiment, K=16 with threshold = 2SD gives the best result. The anomaly detection accuracy is 89.81 (per AKFI) and 96.67 (entire sequence). Since the training data contains only normal actions, the anomaly which has similar posture and motion pattern is misclassified. From the results, only jumping-jack is misclassified as a normal action. Since the upper body of human in the action are similar to wave-two-hands.

In order to test the sensitivity of the system, the irregular performance and viewpoint variation of walking were tested by using Weizmann human dataset. The robustness dataset which contains only walking sequences is used to test several challenges including occlusion, body deformation, irregular movement, and viewpoint variation. The results of various difficult scenarios in Table 4.3 indicate that our proposed technique can classify action when there are occlusion and irregular movement. The misclassification occurs in knees up, limping man, and occluded legs which have the lower part of body features distorted in both appearance and motion. By using majority voting, the system can correctly classify all sequences as walking. So the proposed method could accurately classify walking in the irregular situations including partial occlusions, non-rigid deformations such as swinging bag, carrying briefcase, walking with a dog. According to the viewpoint variation, the system can accurately classify walking between 0 degree and 36 degrees as shown in Table 4.4. Since the posture of the view over 36 degrees are similar to the other actions which have the frontal face to the camera such as gallop sideways, jump-in-place-on-two-legs, and jumping-jack. The system can classify actions, which based on the single view training, within the view variation less than 36 degrees. Moreover, the clustered background causes the foreground extraction to extract shadow as foreground pixels which deform the human shape for feature extraction. According to the robustness result, our system is robust against the partial occlusion, body deformation, and some angle of viewpoint variations.

Table 4.3 Robustness experimental results with irregularities performance of walking.

Test sequence	Accuracy (%) (per AKFI)
Swinging a bag	100.00
Carrying briefcase	100.00
Walking with a dog	100.00
Knees Up	71.43
Limping man	66.67
Sleepwalking	100.00
Occluded Legs	50.00
Normal walk	100.00
Occluded by a "pole"	100.00
Walking in a skirt	100.00

Table 4.4 Robustness experimental results with varying view point of walking.

Test sequence	Accuracy (%) (per AKFI)
walk in 0 degree	100.00
walk in 9 degree	100.00
walk in 18 degree	83.33
walk in 27 degree	83.33
walk in 36 degree	83.33
walk in 45 degree	33.33
walk in 54 degree	0.00
walk in 63 degree	0.00
walk in 72 degree	0.00
walk in 81 degree	0.00

4.2 KTH dataset

The KTH video database [57], as shown Figure 4.6, is a set of public sequences containing 599 videos of 6 types of human actions: boxing, handclapping, hand waving, jogging, running, and walking. Boxing can be used as an anomaly in this environments. The actions were performed by 25 performers in 4 different scenarios: outdoor S1, outdoor with scale variation S2, outdoor with different clothes S3, and indoor S4. The sequences have the spatial resolution of 160x120 pixels with grayscale and have a length of four seconds in average with 25fps frame rate. The bounding boxes information are provided by [40].

In this dataset, we use the same setting as Weizmann dataset which are 3x3 sub-regions and 8-bin histogram of the oriented gradient. For feature extraction, we do not use foreground mask, but use frame differencing to construct AMHI and KPHI. Also, instead of using a number of foreground pixels to detect key frame, the size of the bounding box is used to detect key frame in Eq. (9). The setting indicates that the system can classify actions where there is only a bounding box provided without a foreground mask. The leave-one-person-out cross validation is used for measuring the accuracy.

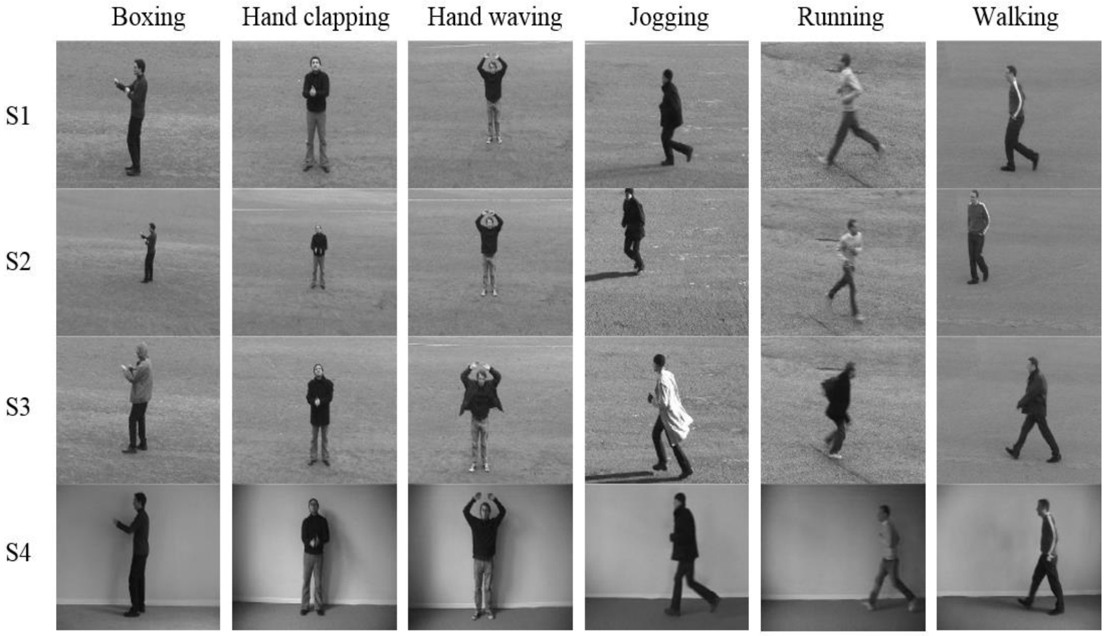


Figure 4.6 Sample of KTH dataset

In Table 4.5, the comparison of the recognition accuracy is separated into actions. The results indicate that the system can classify actions compared to the state of the art methods. The misclassified actions are jogging and running, due to the similarity in posture and movement. Also, our method cannot distinguish between hand clapping and hand waving since our features are constructed by motion pixels of frame differencing. The extracted features lose some information of human body position. So, the misclassification of hand clapping and hand waving is due to the lack of shape information and the similarity of movement. Other approach [37, 44, 45, 57] suffer from the similarity of the actions.

Table 4.5 A comparison of recognition rates among methods on KTH dataset.

Method	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking	Average
Our method	97.00	90.91	100.00	94.00	76.00	95.00	92.15
Dollar et al. [44]	93.00	77.00	85.00	57.00	85.00	90.00	81.20
Niebles et al. [45]	98.00	86.00	93.00	53.00	88.00	82.00	83.30
Jiang et al. [40]	96.00	99.00	96.00	91.00	85.00	93.00	93.43
Schuldt et al. [57]*	97.9	59.70	73.6	60.40	54.90	83.80	71.70
Baysal et al. [18]*	90.00	96.00	94.00	87.00	98.00	84.00	91.50
Ali and Shah[25]*	88.50	86.44	84.46	86.20	91.51	89.11	87.70
Ji et al. [37]*	90.00	94.00	97.00	84.00	79.00	97.00	90.20
Zhang et al. [36]*	96.00	94.00	99.00	78.00	87.00	93.00	91.17

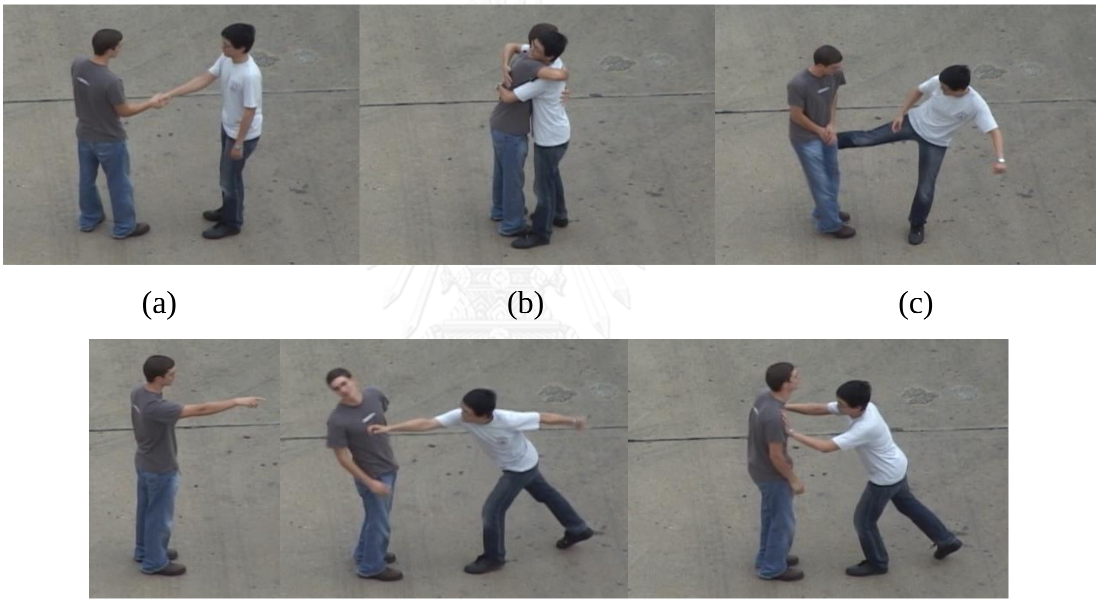
* The approaches used another method to split data for training and testing

The spatio-temporal features from detecting the spatio-temporal interested point approaches [44, 45, 57] can be extracted features with sparsity and focus on the characteristic of local motion. However, the sparse features discard some global information that is important to identify actions. For kinematic features [25], the features capture the motion pattern in spatio-temporal. However, the method cannot segment subsequence by itself and heavily relies on the quality of optical flow. In reference [37], CNNs, which require a large-scale dataset for training the network, did not perform well in this dataset due to insufficient of training samples. The results in [40] are slightly better than our results since they used foreground mask to construct shape descriptor and computed the alignment in sequence matching.

From the Weizmann and KTH experiments, we found that the approaches [17, 18], [40] need preliminary experiments to specify the optimal parameters which give the best recognition rate from the training data. Reference [17, 18] used the information of key poses to construct feature vectors. Because the approach clusters the training data by K-medoids, the classification accuracies are varied base on the number of key poses per action. The work in [40] also has to find the optimal K in K-mean clustering and the optimal k in K-nearest neighbor classification. Since the optimal parameter is trained from a specific dataset, the parameters need to redo the preliminary experiments before applying to another dataset. For the feature dimension, our AMHI and KPHI dimension is 144 at every key frame interval. While the shape-motion descriptor [40] is 512 dimensions. In [36], the number of cuboids for each class is 3,000 for learning feature functions and the feature dimension in the KTH dataset is 1,200 in each frame. In addition, the segmented foreground is not required for detecting key frames and extracting AMHI and KPHI. Since, the system does not use the information of a whole sequence, the system is well adapted to the scaling, noise, and small view variation without the need to fine-tune the parameters.

4.3 UT Interaction dataset

UT Interaction dataset [58], contains 6 classes of human-human interactions: hand shaking, hugging, kicking, pointing, punching, and pushing, as shown in Figure 4.7. There are two scenarios. The first scenario is captured in a parking lot with the static background. The second scenario is captured on a lawn in a windy day. The scenarios contain non-periodic actions in the realistic surveillance environment. The dataset provides 60 activity executions for training and testing the classification in each set.



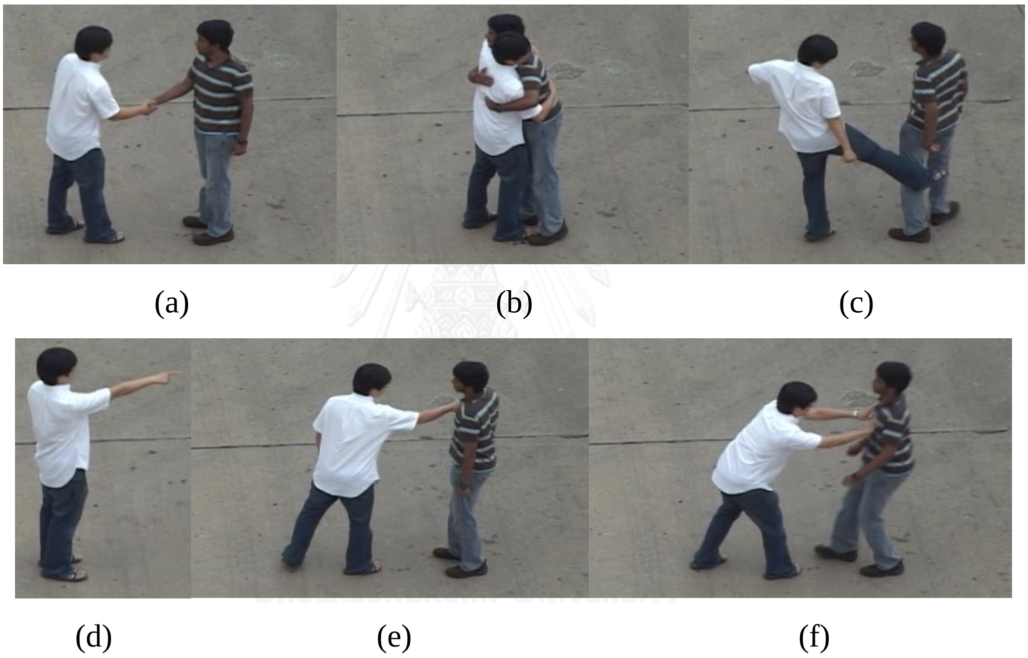
(a) (b) (c)

(d) (e) (f)

Figure 4.7 UT Interaction dataset contains 6 actions: (a) hand shaking, (b) hugging, (c) kicking, (d) pointing, (e) punching, and (f) pushing.

(a) (b)

Figure 4.8 UT Interaction dataset scenarios: (a) video sequences taken on a parking lot, (b) video sequences taken on a lawn in a windy day.



(a) (b) (c)

(d) (e) (f)

Figure 4.9 The detected key frame for each action: (a) hand shaking, (b) hugging, (c) kicking, (d) pointing, (e) punching, and (f) pushing

In this dataset, we use the same setting as Weizmann and KTH which are 3x3 sub-regions and 8-bin histogram of oriented gradient. For feature extraction, we do not use foreground mask, but use frame differencing to construct AMHI. Also, instead of using a number of foreground pixels to detect key frame, the size of segmented sequences is used to detect key frame in Eq. (9). Since we do not have the information of each performer such as separated bounding box, we use the oriented gradient of key frame instead of KPHI. The 10-fold leave-one-out cross validation is used to measure the accuracy. The testing feature is classified by nearest neighbors and the testing sequence is classified by majority voting.

The system detected key frames by using the variation of motion pixels which obtained from frame differencing, as shown in Figure 4.9. In Table 4.6, the comparison of the recognition rates on UT Interaction dataset is separated by actions. The results indicate that the system can classify non-periodic actions compared to the state of the art methods. The results of “Laptev + SVM (best)” and “Cuboid + SVM (best)” are taken from [61] which are implemented based on spatio-temporal features [57] and cuboid [44].

The misclassified actions between punching and pushing are due to the similarity in posture and movement. The spatio-temporal features are used in “Laptev + SVM (best)”, “Cuboid + SVM (best)”, and [62-64] to find the relationship between the testing and the training. Ryoo et al. [62] measured the structure similarity of the spatio-temporal features between sequences. While Yu et al. [63] proposed pyramidal spatio-temporal relationship match for more robust matching. However, the number of atomic actions is too small to distinguish similar actions. Reference [65] extracted key pose doublet which represents the corresponding actions. They selected key poses for each performer and combined the key poses of two performers to construct key pose doublets. The accuracy was based on the number of key pose doublets in each action. In [36], accumulated squared derivative feature extracted from the slow feature analysis results outperform the other approaches in this dataset. However, the number of extracted cuboids for learning is 5,000 per action which is more than a number of cuboids use in Weizmann (2,000 cuboids) and KTH (3,000 cuboids). The number of extracted cuboids is based on the characteristic and the spatial resolution of the dataset.

In our proposed system, the number of key frames is not fixed since key frames are based on the action characteristic. The system can automatically vary the number of key frame per action in each dataset without changing parameters. The system can accurately classify actions under the different environments by using the same setting.

Table 4.6 A comparison of recognition rates among methods on UT Interaction dataset.

Method	Hand shaking	Hugging	Kicking	Pointing	Punching	Pushing	Average
Our method	90.00	80.00	80.00	100.00	60.00	90.00	83.33
Laptev + SVM (best)	50.00	75.00	75.00	85.00	55.00	60.00	66.67
Cuboid + SVM (best)	80.00	85.00	75.00	95.00	70.00	60.00	77.50
Ryoo et al. [62]	75.00	87.50	62.50	50.00	75.00	75.00	70.80
Yu et al. [63]	100.00	65.00	100.00	85.00	75.00	75.00	83.33
Mukherjee et al. [65]	75.00	85.00	85.00	80.00	65.00	85.00	79.17
Yuan et al. [64]	70.00	80.00	85.00	100.00	55.00	80.00	78.20
Zhang et al. [36]	100.00	100.00	100.00	100.00	95.00	100.00	99.17

CHAPTER 5
CONCLUSION AND FUTURE WORKS

For human action classification, an AKFI for feature extraction is proposed. AKFI can be detected by analyzing motion variation through time. So the system can automatically detect the starting time, the ending time, and the primitive action duration for extracting space-time features. The appropriate number of frames, which is required for classifying action, is automatically adapted based on speed and motion variation of the actions. The results of classifying action by extracting features, which are AMHI and KPHI, within AKFI indicate good performance compared to the feature extraction from fixed sliding window method when using the same features. The Experiment results indicate that the system can effectively classify actions for Weizmann dataset, KTH dataset, and UT Interaction dataset. For robustness testing, the system can handle the partial occlusions, body deformation of walking. In addition, AKFI can be used to extract features in the situation that noise and scaling occur.

However, AKFI requires the accurate motion detection for detecting key frames. If the scenario contains a background which is similar to human appearance or difficult to detect motion, the key frame detection cannot perform accurately. View variation is also a limitation of AMHI and KPHI as shown in the robustness dataset. So, the accurate motion detection and human body orientation are considered for the future improvement.

For future works, the accurate motion detection is important for detecting key frames and AKFI. Applying accurate background segmentation, human detection, and human tracking can increase the key frame detection performance. In human action recognition, body orientation and occlusion are still the problems as shown in the Weizmann robustness. So, the method cannot apply in the scenario that covering a wide area and complex background. The human body orientation and occlusion can be resolved by using local features, which contain part of human body, instead of training the whole human body in each action. However, training the model by parts of human body is required more accurate labeled dataset and accurate body part localization.

In addition, the results of the anomaly detection can be used to create a dataset for action classification which includes the unusual events for the specific scenario such as UT interactions dataset. The dataset can be used to train the model for action classification by extracting AMHI and KPHI within AKFI. If the database contains enough data to train unusual actions, the anomaly detection can use this information to classify anomaly in more specific actions.

REFERENCES

[1] H. Weiming, T. Tieniu, W. Liang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 34, pp. 334-352, 2004.

[2] D. M. Gavrila, "The Visual Analysis of Human Movement," *Comput. Vis. Image Underst.*, vol. 73, pp. 82-98, 1999.

[3] J. K. Aggarwal and P. Sangho, "Human motion: modeling and recognition of actions and interactions," in *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, 2004, pp. 640-647.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing,* vol. 28, pp. 976-990, 6// 2010.

[5] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, pp. 1-43, 2011.

[6] N. T. Nguyen, H. H. Bui, S. Venkatsh, and G. West, "Recognizing and monitoring high-level behaviors in complex spatial environments," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2003, pp. II-620-5 vol.2.

[7] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* 2005, pp. 838-845 vol. 1.

[8] M. F. a. J. S. M. J. Nascimento "Segmentation and Classification of Human Activities," presented at the Workshop on Human Activity Recognition and Modelling, 2005.

[9] A. G. Hochuli, A. S. Britto, and A. L. Koerich, "Detection of non-conventional events on video scenes," in *2007 IEEE International Conference on Systems, Man and Cybernetics*, 2007, pp. 302-307.

[10] J. L. Patino Vilchis, H. Benhadda, E. Corvee, F. Bremond, and M. Thonnat, "Extraction of activity patterns on large video recordings," *IET Computer Vision,* vol. 2, pp. pp 108-128, 2008-06-17 2008.

[11] H. I. Suk, A. K. Jain, and S. W. Lee, "A Network of Dynamic Probabilistic Models for Human Interaction Analysis," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 21, pp. 932-945, 2011.

[12] G. Lavee, M. Rudzsky, E. Rivlin, and A. Borzin, "Video Event Modeling and Recognition in Generalized Stochastic Petri Nets," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, pp. 102-118, 2010.

[13] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 555-560, 2008.

[14] K. Schindler and L. v. Gool, "Action snippets: How many frames does human action recognition require?," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.

[15] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 781-796, 2000.

[16] N. Ikizler, R. G. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[17] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters,* vol. 34, pp. 1799-1807, 11/1/ 2013.

[18] S. Baysal, M. C. Kurt, and P. Duygulu, "Recognizing Human Actions Using Key Poses," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 1727-1730.

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as Space-Time Shapes," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 29, pp. 2247-2253, 2007.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886-893 vol. 1.

[21] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.

[22] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image Vision Comput.,* vol. 27, pp. 1515-1526, 2009.

[23] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 726-733 vol.2.

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

[25] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 288-303, 2010.

[26] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," presented at the Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2, Vancouver, BC, Canada, 1981.

[27] G. Zhu, C. Xu, W. Gao, and Q. Huang, "Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine," in *Computer Vision in Human-Computer Interaction: ECCV 2006 Workshop on HCI, Graz, Austria, May 13, 2006. Proceedings,* T. S. Huang, N. Sebe, M. S. Lew, V. Pavlović, M. Kölsch, A. Galata, *et al.,* Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 89-98.

[28] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters,* vol. 31, pp. 1369-1376, 8/1/ 2010.

[29] M. Lucena, N. Pérez de la Blanca, and J. M. Fuertes, "Human action recognition based on aggregated local motion estimates," *Machine Vision and Applications,* vol. 23, pp. 135-150, 2012// 2012.

[30] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 257-267, 2001.

[31] G. R. Bradski and J. W. Davis, "Motion segmentation and pose recognition with motion history gradients," *Mach. Vision Appl.,* vol. 13, pp. 174-184, 2002.

[32] J. Hu and N. V. Boulgouris, "Fast human activity recognition based on structure and motion," *Pattern Recogn. Lett.,* vol. 32, pp. 1814-1821, 2011.

[33] G. R. Bradski and J. Davis, "Motion segmentation and pose recognition with motion history gradients," in *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.*, 2000, pp. 238-244.

[34] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, pp. 1395-1402 Vol. 2.

[35] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical Filtered Motion for Action Recognition in Crowded Videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 42, pp. 313-323, 2012.

[36] Z. Zhang and D. Tao, "Slow Feature Analysis for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 436-450, 2012.

[37] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 221-231, 2013.

[38] S. Park and J. K. Aggarwal, "Event semantics in two-person interactions," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 227-230 Vol.4.

[39] N. C. A. Rosani, and F. G. B. D. Natale, "Human behavior recognition using a context-free grammar," *J. Electron. Imaging,* vol. 23, 2014.

[40] Z. Jiang, Z. Lin, and L. Davis, "Recognizing Human Actions by Learning and Matching Shape-Motion Prototype Trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 533-547, 2012.

[41] I. Laptev and T. Lindeberg, "Space-time interest points," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 432-439 vol.1.

[42] C. Harris and M. Stephens, "A Combined Corner and Edge Detection," in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147-151.

[43] A. Gilbert, J. Illingworth, and R. Bowden, "Action Recognition Using Mined Hierarchical Compound Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 883-897, 2011.

[44] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65-72.

[45] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int. J. Comput. Vision,* vol. 79, pp. 299-318, 2008.

[46] J. Liu and J. Yang, "Action recognition using spatiotemporal features and hybrid generative/discriminative models," *Journal of Electronic Imaging,* vol. 21, pp. 023010-1-023010-10, 2012.

[47] X. Zhang, Y. Yang, L. C. Jiao, and F. Dong, "Manifold-constrained coding and sparse representation for human action recognition," *Pattern Recognition,* vol. 46, pp. 1819-1831, 7// 2013.

[48] C. Liu, P. C. Yuen, and G. Qiu, "Object motion detection using information theoretic spatio-temporal saliency," *Pattern Recogn.,* vol. 42, pp. 2897-2906, 2009.

[49] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 405-412 vol. 1.

[50] M. Ju and B. Bir, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 316-322, 2006.

[51] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications," *Machine Vision and Applications,* vol. 23, pp. 255-281, 2012// 2012.

[52] M. Raptis and L. Sigal, "Poselet Key-Framing: A Model for Human Activity Recognition," presented at *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[53] S. Hu, Y. Chen, H. Wang, and Y. Zuo, "Human action recognition based on spatial-temporal descriptors using key poses," in *International Symposium on Optoelectronic Technology and Application 2014: Image Processing and Pattern Recognition*, Beijing, China 2014, pp. 930132-930132-6.

[54] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inf. Theor.,* vol. 28, pp. 129-137, 1982.

[55] N. K. Speer, K. M. Swallow, and J. M. Zacks, "Activation of human motion processing areas during event perception," *Cogn Affect Behav Neurosci,* vol. 3, pp. 335-45, Dec 2003.

[56] J. M. Zacks, K. M. Swallow, J. M. Vettel, and M. P. McAvoy, "Visual motion and the neural correlates of event perception," *Brain Res,* vol. 1076, pp. 150-62, Mar 3 2006.

[57] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* , 2004, pp. 32-36 Vol.3.

[58] M. S. a. A. Ryoo, J. K., "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," ed. http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[59] J. C. Niebles and F.-F. Li, "A Hierarchical Model of Shape and Appearance for Human Action Classification," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[60] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A Review of Anomaly Detection in Automated Surveillance," *IEEE Transactions on Systems, Man,*

and Cybernetics, Part C (Applications and Reviews), vol. 42, pp. 1257-1272, 2012.

[61] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury, "An Overview of Contest on Semantic Description of Human Activities (SDHA) 2010," in *Recognizing Patterns in Signals, Speech, Images and Videos: ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports*, D. Ünay, Z. Çataltepe, and S. Aksoy, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 270-285.

[62] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1593-1600.

[63] T.-H. a. K. Yu, Tae-Kyun and Cipolla, Roberto, "Real-time Action Recognition by Spatiotemporal Semantic and Structural Forest," in *the British Machine Vision Conference*, 2010, pp. 52.1--52.12.

[64] F. Yuan, V. Prinet, and J. Yuan, "Middle-Level Representation for Human Activities Recognition: The Role of Spatio-Temporal Relationships," in *Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I*, K. N. Kutulakos, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 168-180.

[65] S. Mukherjee, S. K. Biswas, and D. P. Mukherjee, "Recognizing interaction between human performers using 'key pose doublet'," in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, Arizona, USA, 2011.

VITA

Kanokphan Lertniphonphan was born in Bangkok, Thailand, in 1985. She recieved her B.Eng. degree in Electrical Engineering from Chulalongkorn University, Thailand, in 2007. She was supported by National Telecommunication Commission of Thailand (NTC)'s Ph.D. Scholarship from 2009-2011 to pursue her Doctoral degree in Electrical Engineering at the Digital Signal Processing Research Laboratory, Department of Electrical Engineering, Chulalongkorn University. This research is partly supported by the 90th Anniversary of Chulalongkorn University, Rachadapisek Sompote Fund. Her research interests include computer vision and pattern recognition.