1-1-2010

SIFT-ME: A New Feature for Human Activity Recognition

Guosheng Wu
University of Denver

SIFT-ME: A New Feature for Human Activity Recognition

_____

A Thesis

Presented to

The Faculty of Engineering and Computer Science

University of Denver

_____

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

_____

by

Guosheng Wu

June 2010

Advisor: Richard M. Voyles

Co-advisor: Mohammad H. Mahoor

Author: Guosheng Wu
Title: SIFT-ME: A New Feature for Human Activity Recognition
Advisor: Richard M. Voyles
Degree Date: 06/2010

ABSTRACT

Action representation for robust human activity recognition is still a challenging problem. This thesis proposed a new feature for human activity recognition named SIFT-Motion Estimation (SIFT-ME). SIFT-ME is derived from SIFT correspondences in a sequence of video frames and adds tracking information to describe human body motion. This feature is an extension of SIFT and is used to represent both translation and rotation in plane rotation for the key features. Compare with other features, SIFT-ME is new as it uses rotation of key features to describe action and it robust to the environment changes. Because SIFT-ME is derived from SIFT correspondences, it is invariant to noise, illumination, and small view angle change. It is also invariant to horizontal motion direction due to the embedded tracking information. For action recognition, we use Gaussian Mixture Model to learn motion patterns of several human actions (e.g., walking, running, turning, etc) described by SIFT-ME features. Then, we utilize the maximum log-likelihood criterion to classify actions. As a result, an average recognition rate of 96.6% was achieved using a dataset of 261 videos comprised of six actions performed by seven subjects. Multiple comparisons with existing implementations including optical flow, 2D SIFT and 3D SIFT were performed. The SIFT-ME approach outperforms the other approaches which demonstrate that SIFT-ME is a robust method for human activity recognition.

ACKNOWLEDGMENTS

Life is not easy. There are many kinds of difficulties and challenges interrupting my focus of research and life. I went through all the happiness, sadness, distress of my life in University of Denver, found that without my parents' support, I cannot start the first step and without my wife: Mei Xu's support, I cannot get such achievements. My family is my powerful backing. I would thank them first in this acknowledgement.

At the same time, I would like to thank my advisor Dr. Richard M. Voyles and co-advisor Dr. Mohammad H. Mahoor, for their guidance and support on my research. Dr. Voyles owns a great amount of knowledge and patience to guide me into computer vision and robotics field as well as provide valuable suggestions to improve my thesis work. Dr. Mahoor is an expert in computer vision and he helped me getting through all the difficulties on research and daily life. I feel very lucky to be a student of them.

TABLE OF CONTENTS

I. Introduction

1.1 Background

Activity Recognition technology aims to recognize the actions and goals of an agent from a sequence of observations of the agent's behavior and the environmental conditions. This research field has attracted a lot of researchers' attention [23] since the 1980s due to the broad range of possible applications, like automatic surveillance, human computer interaction, assisted living, etc.

In order to better illustrate human activity recognition, consider the following applications. Human activity recognition can be used in video surveillance systems. For example, during a bank robbery armed gunman force bank tellers to give them money, and make the customers get down on the ground (examples shown in fig.1-1). Many recognizable patterns exist in such a scenario, like gun pointing, customers lying on the ground, etc. As soon as the video surveillance system detects these patterns, it can automatically send an alarm to the police office and let police agents stop the robbery, catch the criminals, and save humans lives.

Another application is Human Computer Interaction (HCI) in a virtual environment system, such as distinguishing two hands making a "zoom" action or a one handed "throw" action on Microsoft's smart table system (Fig. 1-2). Actions of the body can be interpreted as input information for computers. The computer analyzes the input data, extracts pre-defined as commands and executes them based on preprogrammed software. Input devices can be unobtrusive cameras or body-attached sensors, but many

type of sensors are easy to break and hard to synchronize. As technology improves, customers will demand products that offer more freedom. Cameras are non-contact sensors and can meet such requirements without the need to cover the body with sensors.

(1) Bank Robber with large weapon (from FBI).

(2) Bank Robbers with handguns (from FBI: unknown suspects).

(3) Bank Robbers with faces covered (from FBI: unknown suspects).

Fig. 1-1 Bank Rob images

Fig. 1-2 Microsoft smart table system (from Microsoft demo)

The third application is assisting the sick and disabled. For example, Pollack et al. [24] describes work that automatically monitor human activities for home-based rehabilitation of people suffering from traumatic brain injuries. Fig. 1-3 shows Asimo serving customers. Human Activity can be use to recognize the hand waving actions to let Asimo deliver a cup of tea and take orders.

Fig. 1-3 Asimo serves customers (from Chinese Xinhua News)

There are two types of activity recognition, one is sensor based and another is vision based. Sensor-based activity recognition integrates sensor networks with novel data mining and machine learning techniques to model a wide range of human activities [25]. Sensor-based activity recognition researchers believe that they can empower ubiquitous computers and sensors to monitor the behavior of agents. Vision based human activity recognition employs cameras as sensors to track and understand the behavior of agents. Vision-based human activity recognition is one of the most challenging and active research areas in the field of computer vision. There are a broad range of applications for human activity recognition such as automatic surveillance, human computer interaction, video browsing and retrieval.

A great deal of work has been done in vision based human activity recognition during the past 30 years. Researchers have attempted a number of methods, such as optical flow [5, 6 and 16], motion trajectory [7], space time shapes [9], etc., under different modalities such as single camera and stereo camera. Normally, the process of vision based human activity recognition can be divided into four steps as shown in fig. 1-4, namely action description, action representation, action recognition and high-level action evaluation. Action description is some element features which captured from video streams to describe actions. For example, many researchers use optical flow or Histograms of Optical Flow (HOOF) as motion features and silhouettes as shape features to represent actions. Action representation is some models which used to register patterns of action. There are plenty of algorithms that can model actions and classify videos like the Gaussian Mixture Model and the Hidden Markov Model which are often used in this research field. Action recognition is using the similarity of action description and action models to assign labels for each action. High-level action evaluation is a process that utilizes recognition results for knowledge inferring.

Fig. 1-4 Pyramid levels of Human Activity Recognition.

1.2 Problem Statement

Comparing the four stages of human activity recognition, action description is still one of the greatest challenges due to variations in environmental factors and differences in actor's activities due to the variation in both environment and actor behaves. The environment changes include illumination variations, camera view angle difference, and image resolution. Such changes highly influence the performance of human activity recognition. First, illumination variations cause serious problems to non-robust background subtraction, and many research approaches fail due to the lighting problem. Second, videos captured under different camera view angles appear differently. If applying the same approach to different view angle videos, the results could have large differences. Third, image scales can influence the recognition accuracy and high

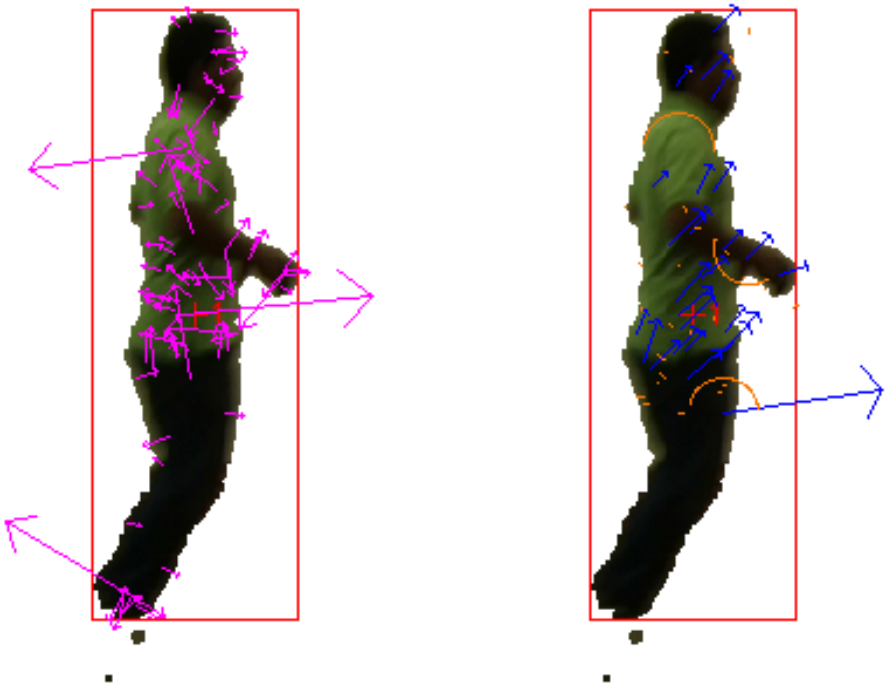
resolution video needs more computation time and obtains more noise than low resolution video.

People look different in different videos and perform similar actions differently, which adds more difficulty for action recognition because similar actions can be easily classified into two different categories. It is also hard to differentiate between some actions such as walking and jogging. Therefore, a human action description method that can represent a wide range of actions performed by different actors under different conditions becomes essential.

1.3 Proposed Approach

A new spatiotemporal feature named SIFT-Motion Estimation (SIFT-ME) is presented in this thesis. The SIFT-ME, which is derived from SIFT correspondences, inherits the SIFT advantages and is invariant to noise, illumination variation, and small view angle changes. The process of estimating SIFT-ME features from videos containing human activities is described as follows. First, robust background subtraction method is applied to videos to isolate the moving subject. Second, SIFT features of the moving subject (foreground) is detected using the approach present by David Lowe [3]. By tracking foreground subject in videos, the motion direction and the body size can be found. Third, translation vector between the SIFT key point correspondences can be calculated by finding the SIFT key point correspondences between two consecutive frames. Afterwards, combine the translation vectors with tracking results that signify the motion direction to represent the subject's motion by a vector containing translation distance, translation direction, and in-plane rotation angle. Fig. 1-5 presents a comparison between SIFT features and SIFT-ME features in a sample video frame. Fig. 1-5(a) shows

the SIFT features with their direction of translation in pink fig. 1-5(b) presents SIFT-ME features with both translation vectors in blue and the rotation arcs around key points in orange. This highlights the important to difference between SIFT features and SIFT-ME features which is the extension from 2D motion of the key points to 3D motion of the key points (translation plus body rotation).



(a) (b)

Fig. 1-5 Comparison between (a) SIFT features and (b) SIFT-ME features.

After extracting the SIFT-ME features which describe the body motion between every two consecutive frames, Gaussian Mixture Model (GMM) utilized to characterize the statistical behavior of SIFT-ME features in different human actions. GMM is a powerful heuristic tool and has become popular among empirical researchers [1]. In the thesis, a GMM trained to represent the extracted SIFT-ME features for each action and then utilize the GMMs for action recognition based on maximum log-likelihood criterion.

1.4 Contribution

The thesis proposed one new feature for human activity recognition named SIFT-ME. This feature is an extension of SIFT and used to represent both translation and rotation in plane rotation for the key features. Compare with other features, SIFT-ME is new as it uses rotation of key features to describe action and it robust to the environment changes.

Most of the approaches focus too much on translation of key features, while not paying enough attention to the rotation information of key features. However, human actions are not just key features translated from one location to another. When analyzing the videos of actions, we can see that many of the actions such as bending, running, walking are not just simple translations. The key features not only have translation variations but also rotation information. The comparison of SIFT-ME and SIFT Translation in chapter 5 clearly illustrates that with rotation information of key features, SIFT-ME possess higher recognition results.

Because SIFT-ME is based on Scale Invariant Feature Transform (SIFT), which is well known for its robustness to environment disturbances such as noise, illumination and view angles, SIFT-ME is also invariant to such kind of environment disturbances. In addition to SIFT, SIFT-ME reveals the key features motion information, which makes it as a space-time feature for action description.

1.5 Organization

The thesis is organized as follows. Literature review is illustrated in chapter 2 and the way to calculate SIFT-ME features is shown in chapter 3. Action Representation and action representation are presented in Chapter 4. Chapter 5 illustrates the experiment results and comparison results. The final conclusion is described in Chapter 6.

II. Literature Review

Human action is a spatiotemporal event since there are spatial features in each frame and information relied on spatial features cross frames. A good action description feature should cover both spatial and temporal information. By reviewing the literature, existing approaches for action representation can be classified into two categories based on whether a static or a dynamic method has been used for the action description.

2.1 Static Method

Static method uses spatial features like pose and shape to describe activities in video [2, 13, 14 and 15]. These features can be easily obtained from each frame without any cross frames relation, which means no time information is contained. There are many ways to represent static information as long as the feature chosen can represent the static information of each action.

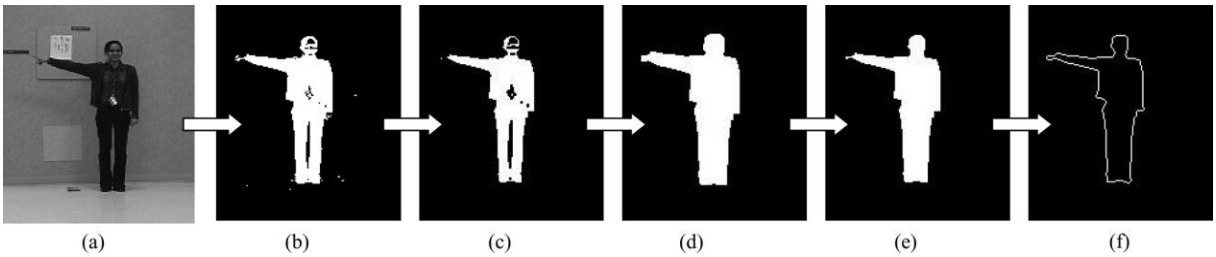


Fig. 2-1 Silhouette extraction processes (reprint from [2])

A silhouette is often used to represent the shape and pose information. In order to obtain the silhouette of each frame, having a few steps for image processing is important. Fig.2-1 shows the normal processing steps to obtain the silhouette information, from left to right: original image, background subtraction, erosion, dilation, erosion again and edge

extraction. Background subtraction can be used to segment a moving object from its scene. The morphology method like erosion and dilation is used to remove noise and fill holes of the body. The Silhouette information can be encoded with many techniques. Cuntoor et al. [21] uses the distance of the silhouette contours from reference vertical and horizontal lines as complex coordinates. Wang et al. [22] computes a complex representation of the silhouette edge using the center of the silhouette as the origin of the complex coordinate system. Singh et al. [2] represents the silhouette boundary as a chain code by travelling eight neighbors of each boundary pixel from highest-leftmost point, which presents in fig. 2-2. As the chain code is cyclic in nature, it can be started from any point. Kellokumpu et al. [13] utilizes affine invariant Fourier descriptors from the contour, and then uses these descriptors as a feature vector to classify the posture with a radial basis SVM. The output of the posture classification module is thus a sequence of discrete postures. After this, they use hidden Markov models to model different activities and calculate the probabilities of the activities based on the posture sequence from posture classification module.

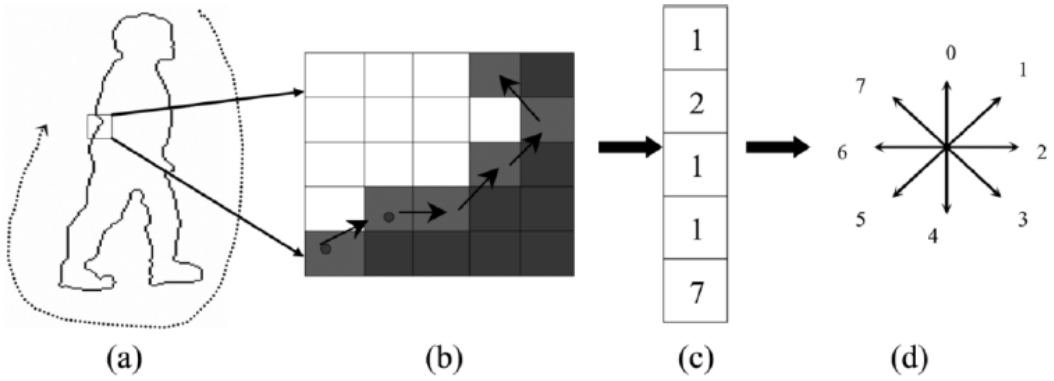


Fig. 2-2 Chain code of Silhouette boundary (reprint from [2])

Scale Invariant Feature Transform (SIFT) descriptor proposed by Lowe [3] has become popular in static features for human activity recognition as it has so many advantages, like being invariant to scale, rotation, robust to illumination, noise and small view angle change. SIFT features exhibit the highest matching accuracies for an affine transform of 50 degrees, outperform other local descriptors on both textured and structured scenes. Many researchers try to directly use SIFT or derive features based on SIFT descriptors for action representation [2, 14, 15 and 17].

Although these approaches have some success in recognizing human actions, they are not capable of capturing the temporal information (dynamics) between frames.

2.2 Dynamic Method

Dynamic method utilizes information cross images such as motion, spatiotemporal shapes, and trajectory to describe actions. Such information should at least cover the time information in order to be classified into dynamic approach, but usually, information should be derived from spatial features which travel through sequence of frames. Most of the dynamic features which covered both space and time information can be considered as spatial temporal features.

Some researchers have used optical flow as a method for body motion estimation (example is shown in fig. 2-3). For example, Efros et al. [5] uses optical flow to match the motion of a player in soccer videos. Chaudhry et al. [6] uses Histogram of Oriented Optical Flow features that are independent of the scale of the moving person as well as the direction of motion to represent human activities. Feng and Abdel-Mottaleb [16] also utilize optical flow and HMM for human activity recognition. However, optical flow is influenced by illumination variation and view angle change. In addition, optical flow is

only suitable for estimating rigid body motion with small displacement which makes optical flow a weak technique for robust motion estimation of non-rigid objects (e.g., human body).

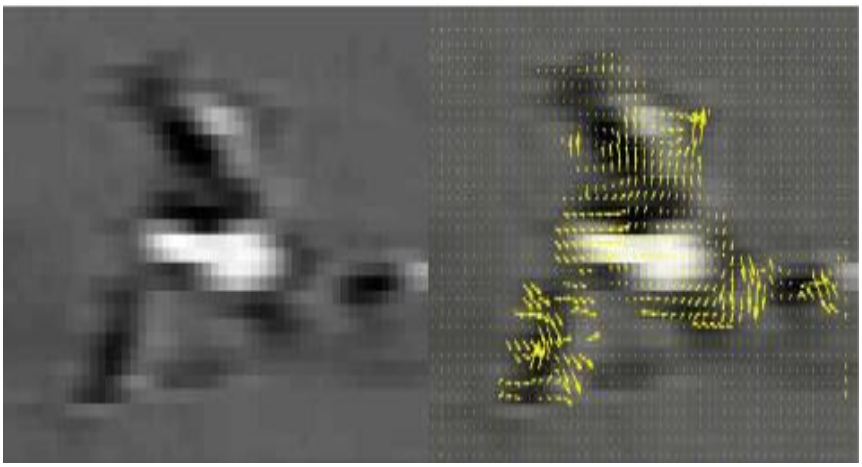


Fig. 2-3 Optical flow example reprint (from [5])

Gorelick et al. [9] uses three-dimensional shapes induced by the silhouettes in a space-time volume to represent action which is presented in fig.2-4. They explain human actions as a moving torso and collection of parts and utilize properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation. Min et al. [7] uses the motion trajectory which is generated from body parts (hand, feet, and joints) based on optical flow magnitude which is shown in fig. 2-5(c). The dominate pixels' trajectories are considered as feature vectors for action representation. Messing et al. [8] applies velocity history of tracked key points to represent motion representation which is shown in fig. 2-6. But, recognition accuracy of trajectory, velocity history and 3D space-time volume rely heavily on the viewing angle.

Fig.2-4 Space-time Shapes (reprint from [9])

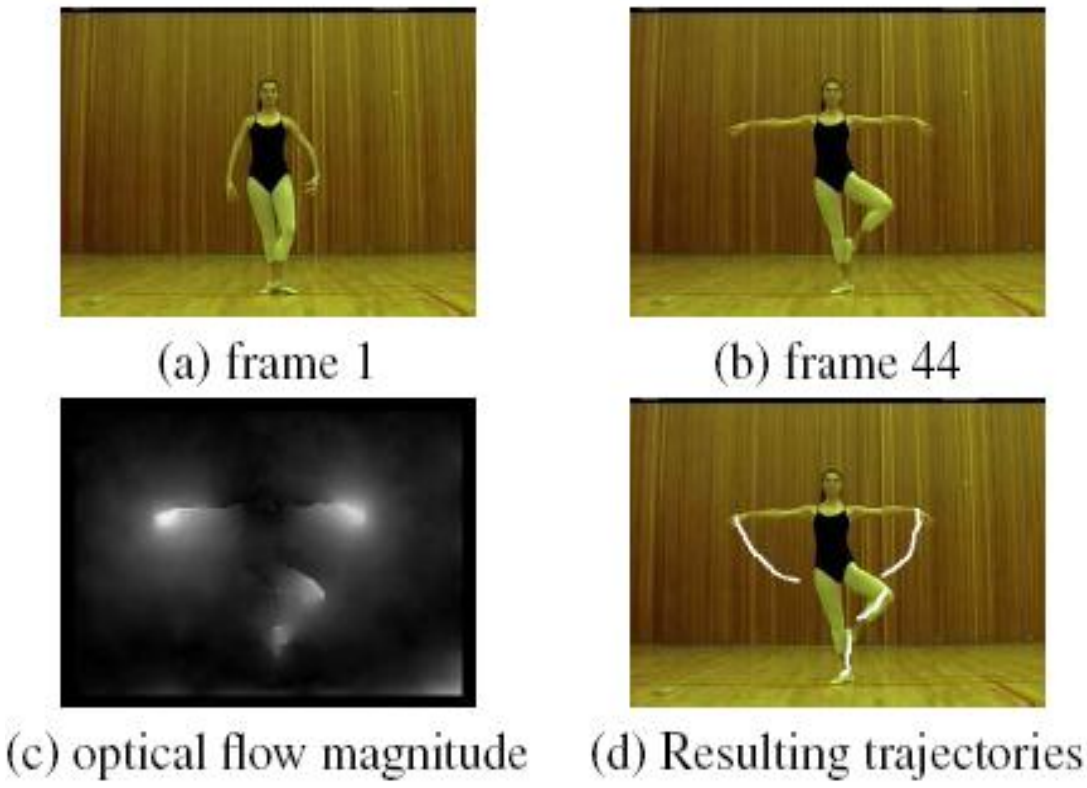


Fig. 2-5 Motion Trajectory (reprint from [7])

Fig. 2-6 Velocity History (reprint from [8])

2.3 SIFT Evolution

Several researchers [2, 14, 15 and 17] have used SIFT in the static category to perform activity recognition. SIFT features are invariant to image scaling and rotation, and partially invariant to illumination change and camera view angles. They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. The algorithm is efficient enough to detect large numbers of features. In addition, the features are highly distinctive, which allows a single feature to be correctly matched with a high probability against a large database of features, providing a basis for object and scene recognition [3]. SIFT Features are presented in fig. 2-7 with yellow arrows.

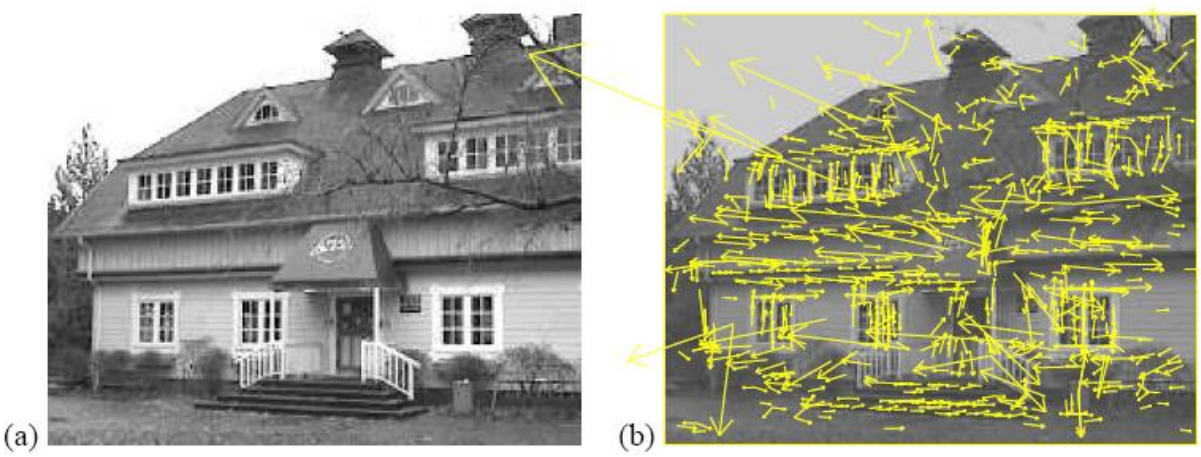


Fig. 2-7 SIFT Features (left is original and right is SIFT features, reprint from [3])

There are four steps to detect SIFT features on an image. First of all, difference-of-Gaussian function applied over all of the image to identify potential points of interest (fig. 2-7) that are invariant to scale and orientation, which named as scale-space extrema. Second, all the candidate points fit into a 3D quadratic function to determine the interpolated location of the maximum and eliminate the edge response. Third, based on the key point location and image gradient direction, each key point assigns one or more orientations. For an image pixel at location $L(x,y)$ at scale σ, the gradient magnitude, $mag(x,y)$, and orientation, $\theta(x,y)$, are pre-computed using pixel differences:

$$\begin{cases} mag(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \\ \theta(x,y) = tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \end{cases} \quad (2\text{-}1)$$

Finally, a key point descriptor is created by computing the gradient magnitude and orientation at each image sample point in a region around the key point location, as shown to the left of fig. 2-8. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms, as shown to the right of fig. 2-8, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region.

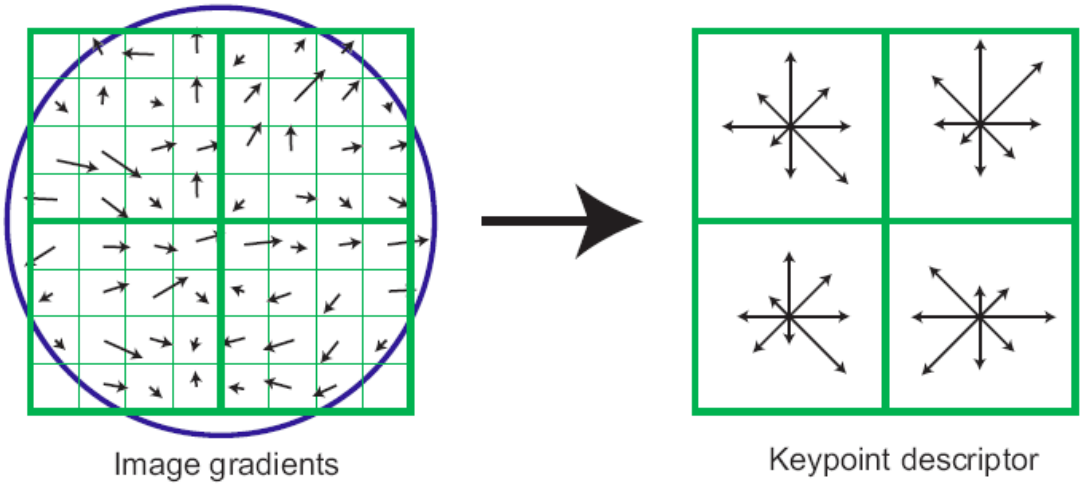


Fig. 2-8 Key descriptors (reprint from [3])

However, action is a spatiotemporal event, and SIFT cannot uniquely reveal broad classes of human action without the aid of temporal analysis. More recently, a few researchers have tracked SIFT features over time to describe human actions [17 and 20] both spatially and temporally. 3D SIFT [17] adds temporal information as a third dimension to the feature to quantify the time variation of the feature itself. This approach considers describing each pixel on image as $L(x, y, t)$. The way to calculate 3D SIFT features is shown in Equation 2-2.

$$\begin{cases} mag(x,y,t) = \sqrt{{L_x}^2 + {L_y}^2 + {L_t}^2} \\ \theta(x,y,t) = tan^{-1}\left(\frac{L_y}{L_x}\right) \\ \phi(x,y,t) = tan^{-1}\left(\frac{L_t}{\sqrt{{L_x}^2+{L_y}^2}}\right) \end{cases} \tag{2-2}$$

Where

$$\begin{cases} L_x = L(x+1,y,t) - L(x-1,y,t) \\ L_y = L(x,y+1,t) - L(x,y-1,t) \\ L_t = L(x,y,t+1) - L(x,y,t+1) \end{cases} \tag{2-3}$$

3D SIFT processing is quite similar to 2D SIFT except considering the relations cross frames, which is shown in fig. 2-9. However, these cross frames' information just

represent the changes over time on the same location, which is clearly different with motion information that represents key point motion in temporal sequence. In other words, 3D SIFT is a spatiotemporal histogram-based representation of image patches, but does not capture human action across the image sequence.

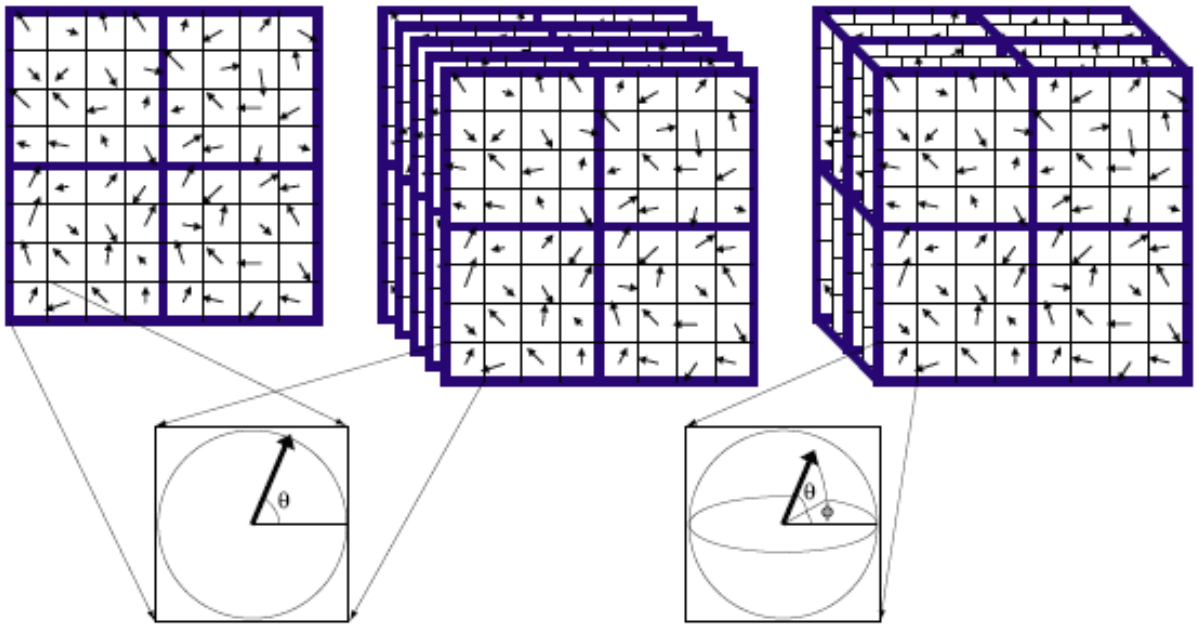


Fig. 2-9 3D SIFT Process (reprint from [17])

MoSIFT [20] was another attempt to improve SIFT for representing motion information by applying optical flow to SIFT key point (detail process is shown in fig. 2-10). In this approach, SIFT features is considered as spatially distinctive interest points, optical flow is applied to these distinctive points to find motion constrain around. However, optical flow has so many weaknesses like variations to scale, rotation and view angles, by combining with SIFT and optical flow, MoSIFT loses these advantages which should be inherent directly from SIFT.

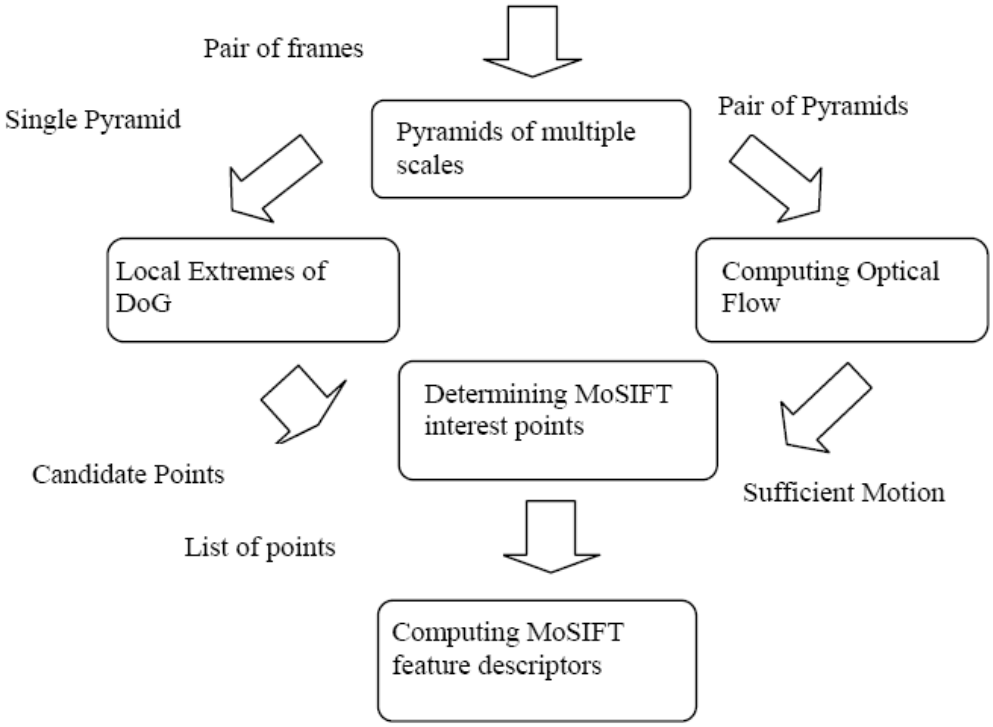


Fig. 2-10 MoSIFT process (reprint from [20])

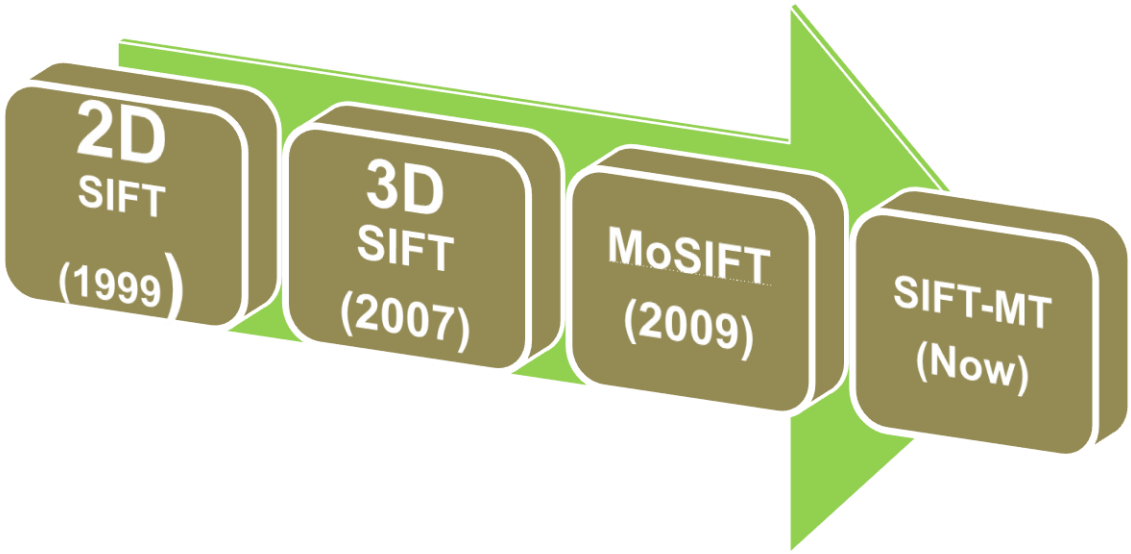


Fig. 2-11 SIFT Evolution

SIFT is moving towards to represent motion as shown in fig. 2-11. In order to take full advantages from SIFT as well as interpret motion information, SIFT Motion Estimation (SIFT-ME) is proposed in the thesis by using an embedded quantifier for SIFT motion, but augmenting it further with temporal tracking across image frames, creating a true spatiotemporal representation. SIFT-ME is inherent invariant to illumination, scale and view angle change. At the same time, it not only can interpret the translation information of key point like optical flow, but can also describe the rotation information, which is a great improvement of SIFT feature.

III. SIFT-ME

Recently, many techniques have been applied to interpret 2D motion, like optical flow, MoSIFT, and SIFT Correspondence. However, the past research on motion interpretation just reveals translation of key points but neglects important information, like rotation. In order to fully reveal the 2D object motion in videos, some research is done on the Motion Estimation of objects and extending the same technique for SIFT features which lead to a completely new motion description feature: SIFT Motion Estimation (SIFT-ME).

3.1 Motion Interpretation

Object's transformation can be performed and represented by a mixture of translation and rotation in the 2D plane. For example, in Fig.3-1 assume there are three objects, A, B, and C, described using different shapes and each object using one dominate SIFT features to represent its key point. The arrows in this figure represent SIFT features at hypothetical key points that are assigned to these objects. Fig 3-1(a) and Fig. 3-1(b) registered the status of the three objects at different times. In order to interpret each object's motion in 2D plane, Fig 3-1(c) compares each object's key SIFT feature. Clearly, Object A in Fig. 3-1(a) is transformed by a pure translation and represented in Fig. 3-1(b). Similarly, Object B is transformed by a pure rotation, and object C is transformed using both translation and rotation. Most of the motion description features such as optical flow and trajectory methods signifies the translation of objects' points without considering the objects' rotation. Which is similar to interpreting the motion of object A

for A, B, C, neglected whether the object has rotation or not. However, human motion is similar to the motion of object C, which is a mixture of translation and rotation. If only utilizing pure translation to estimate the motion of object C, significant information regarding object motion (the rotation information) will be lost. In order to better interpret the motion in a 2D plane, rotation information should be considered. SIFT-ME feature represents both translation and rotation of the points between consecutive frames.

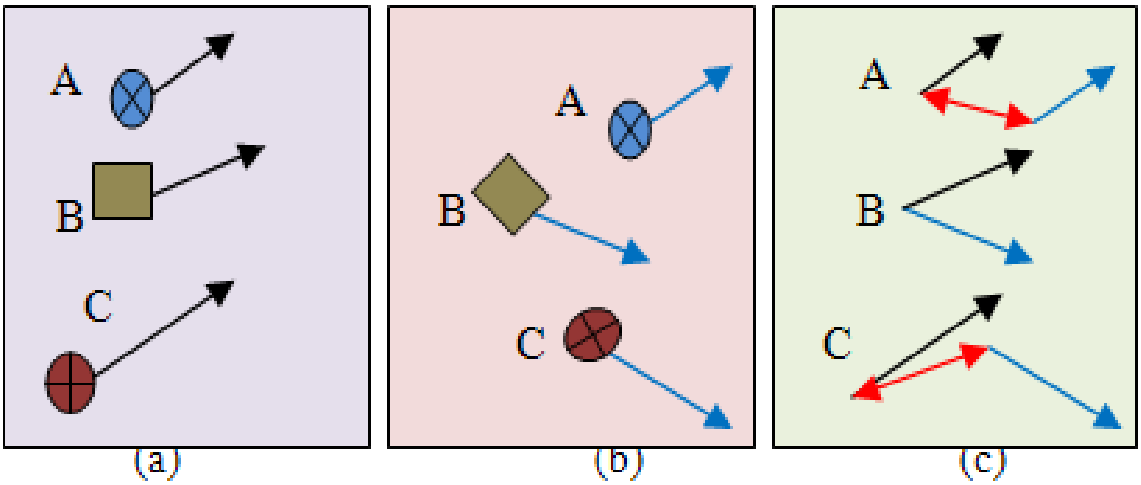


Fig. 3-1 SIFT and SIFT Motion in key point: Different shapes represent objects and the arrow represent SIFT features. (a) The image represents three objects at time t-1 (b) the position of the object after translation, rotation or both at time t. (c) the image shows the motion between objects in images (a) and (b).

Equation 3-1 shows the Euclidean transformation of point, $C: [x_{t-1}, y_{t-1}, 1]^T$ at time $t-1$ to a new location, $C^{'}: [x_t, y_t, 1]^T$ at time t:

$$\begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} = \begin{bmatrix} cos\beta & -sin\beta & \rho cos\alpha \\ sin\beta & cos\beta & \rho sin\alpha \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ 1 \end{bmatrix} \quad (3\text{-}1)$$

where ρ, α are translation distance and translation direction of object. β is the rotation angle with respect to the object center. Vector $< \rho, \alpha, \beta >$ can be used to describe the object's transformation and the proposed approach has similar appearance with it.

The motion of object A and B can be explained by $\beta = 0$ and $\rho = 0$ seperatedly. But the movement of object C needs all three variables in that vector to be explained. In order to describe motion for key points existing on human body, vector $< \rho, \alpha, \beta >$ can be used as a feature for the key points.

As SIFT feature orientation: $\theta \in [-\pi, \pi]$, and the rotation angle $\beta = (\theta_t - \theta_{t-1}) \in [-2\pi, 2\pi]$, there is a large duplicate space which means one to one mapping is no long available. For example, in Fig. 3-2, the rotation angle can be explained with two values: $\phi \; and \; -2\pi + \phi$, both of them are valid values to interpret the rotation angles from vector $\overrightarrow{V1}$ to vector $\overrightarrow{V2}$, the difference is whether following right hand rule or left hand rule. In order to form one to one mapping and shrink β to $[-\pi, \pi]$, only the smallest absolute angle between two vectors is counted. Besides, if rotation direction follows right hand rule, the value of angles is positive, otherwise, the value is negative.

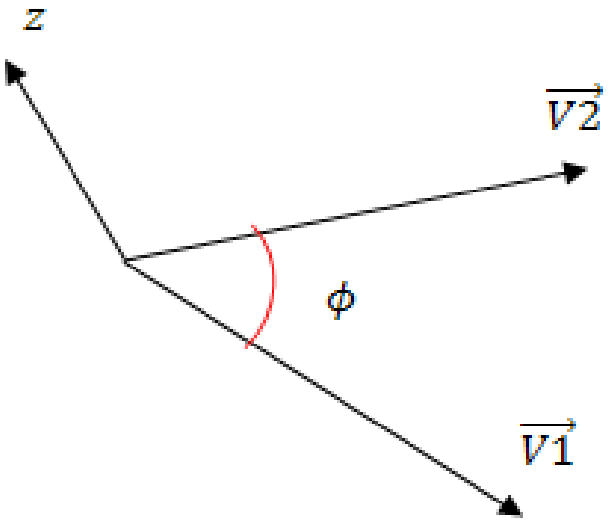


Fig.3-2 Rotation angle

Considering in the vector field, the smallest angle between vector $\overrightarrow{V1}$ and vector $\overrightarrow{V2}$ can be calculated use Equation 3-2.

$$|\beta| = \arccos\left(\frac{\overrightarrow{V1}\cdot\overrightarrow{V2}}{|\overrightarrow{V1}||\overrightarrow{V2}|}\right) = \arccos\left(\frac{|\overrightarrow{V1}||\overrightarrow{V2}|\cos(\theta_2-\theta_1)}{|\overrightarrow{V1}||\overrightarrow{V2}|}\right)$$

$$= \arccos(\cos(\theta_2 - \theta_1)) \tag{3-2}$$

The sign of β can be calculate based on whether smallest rotation angle follow right hand rule or left hand rule.

$$\frac{\overrightarrow{V1}\times\overrightarrow{V2}}{|\overrightarrow{V1}||\overrightarrow{V2}|} = \frac{|\overrightarrow{V1}||\overrightarrow{V2}|\sin(\theta_2-\theta_1)}{|\overrightarrow{V1}||\overrightarrow{V2}|} = \sin(\theta_2 - \theta_1) \tag{3-3}$$

And $\beta \in [-\pi, \pi]$ can be easily got by applying Equation 3-4.

$$\beta = sign(\sin(\theta_2 - \theta_1)) \cdot \arccos(\cos(\theta_2 - \theta_1)) \tag{3-4}$$

$sign(\cdot)$ is a function taking the positive or negative sign of input variables. In this way, β can be shrink into the range of $-\pi \; and \; \pi$, and each value only mapping to one angle in that region. So vector $< \rho, \alpha, \beta >$ represents the motion in 2D plane with one to one mapping relation.

3.2 SIFT-ME Detection

SIFT-ME has similar expression to 2D Euclidean transformation: $< \rho, \alpha, \beta >$, and is obtained using SIFT correspondences and tracking information. Fig. 3-3 shows a diagram for obtaining SIFT-ME features and Fig. 3-4 illustrates images obtained in each step.

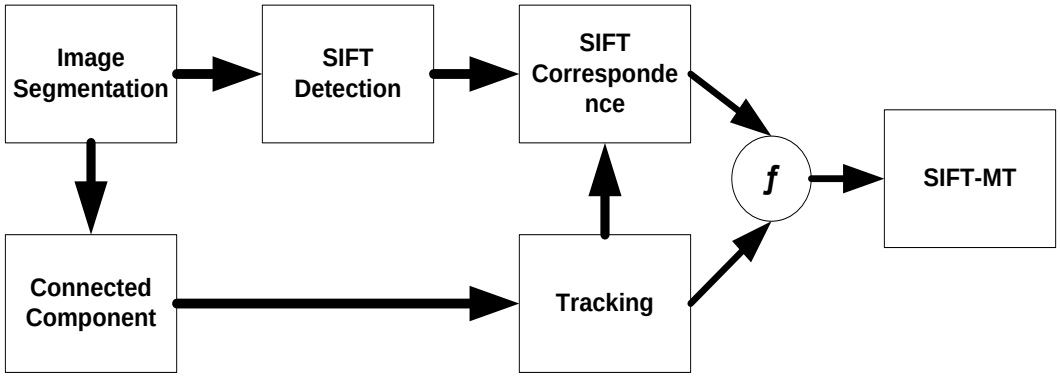


Fig. 3-3 the process of extracting SIFT-ME features.

First, background subtraction is applied on video frames to segment the moving objects (human body) into foreground and background. Second, SIFT feature detection is applied to the foreground image (human body). At the same time, human's motion information can be obtained by applying Kalman filters to the silhouettes. After that, corresponding points between every two consecutive frames are extracted using SIFT feature matching algorithm. Finally, by utilizing the tracking information and corresponding points, the SIFT-ME features are readily driven.

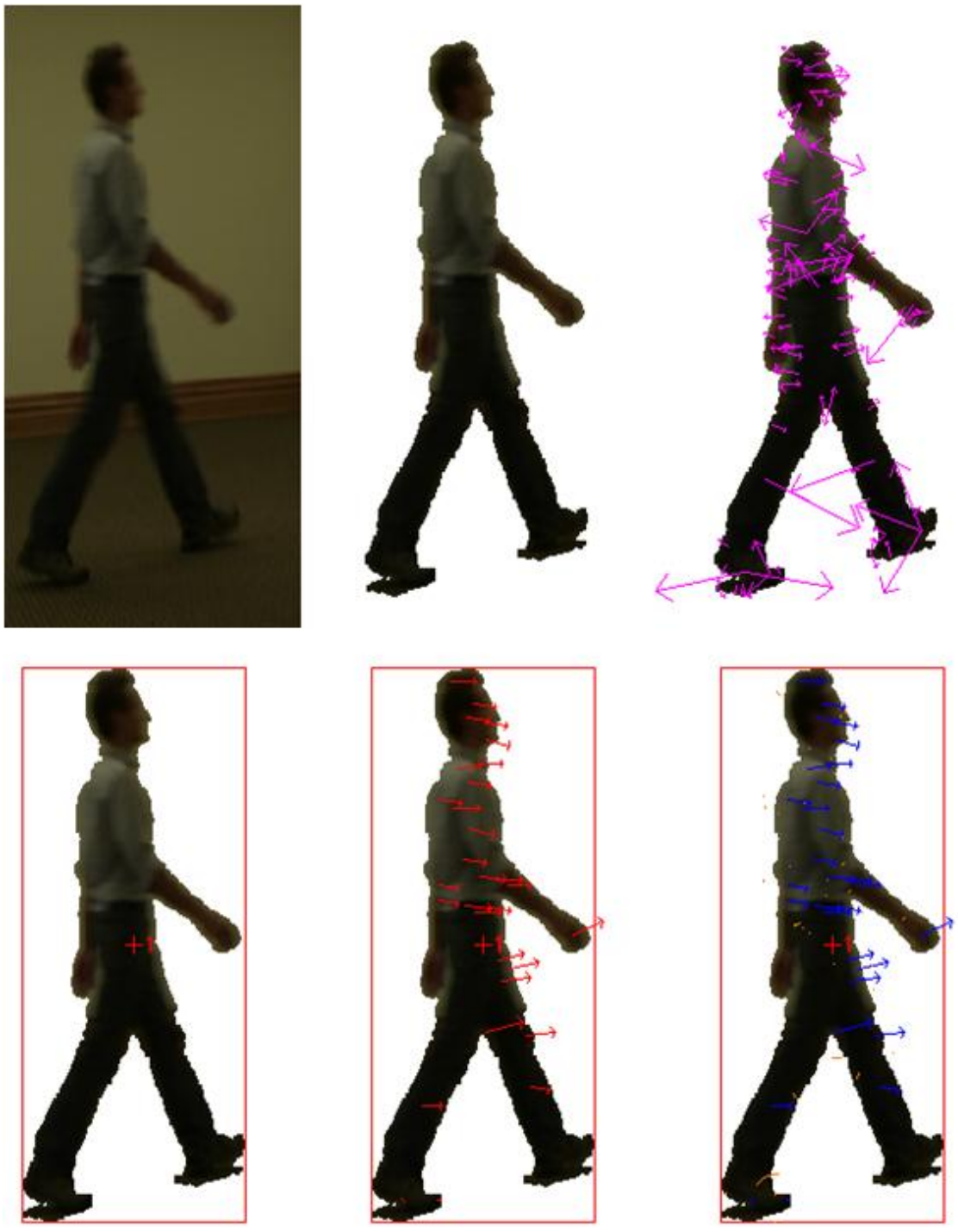


Fig. 3-4 Illustrates the process of extracting SIFT-ME features, from left to right, original image, segmented image, SIFT detection, tracking, SIFT correspondence, SIFT-ME features.

3.2.1 Segmentation

Adaptive Gaussian mixture model [25 and 11] is utilized in this thesis to segment moving objects from static background. Changes in scene lighting can cause problems for many back ground subtraction methods. In order to eliminate the illumination influence, values of a particular pixel is modeled as a mixture of Gaussians and based on the persistence and the variance of each of the Gaussians of the mixture to determine which Gaussians may correspond to background colors. Pixel values that do not fit the background distributions are considered foreground until there is a Gaussian that includes them with sufficient, consistent evidence to support it.

In order to remove noise and fill holes in the foreground image, morphology operators such as open and close are applied to the foreground mask image. The mask image is utilized as a filter to segment out the moving object in the scene. In fig. 3-5, the left image is the original scene from inside building surveillance camera and the right image is the result of background subtraction method. The silhouettes show exactly the actor's positions and shapes.

Fig. 3-5 GMM and background segmentation example (Left: original frame, right: foreground image).

3.2.2 SIFT Detection

SIFT proposed by David Lowe [3] is a technique for robust feature extraction where an image is represented by a large set of features, each of which is invariant to image translation, scale, rotation, partially invariant to illumination changes and robust to local geometric distortion. Each SIFT feature is described by the coordinates of its location, magnitude and orientation of image gradients. In this thesis, x and y denote the position of SIFT key points in frame coordinate system, $mag\ and\ \theta$ represent the magnitude and angle of SIFT key points with $\theta \in [-\pi, \pi]$. Fig. 3-6 shows the result for SIFT detection. SIFT detection is only applied to the segmented moving objects, which can reduce the computation time and filter out noise from the background.

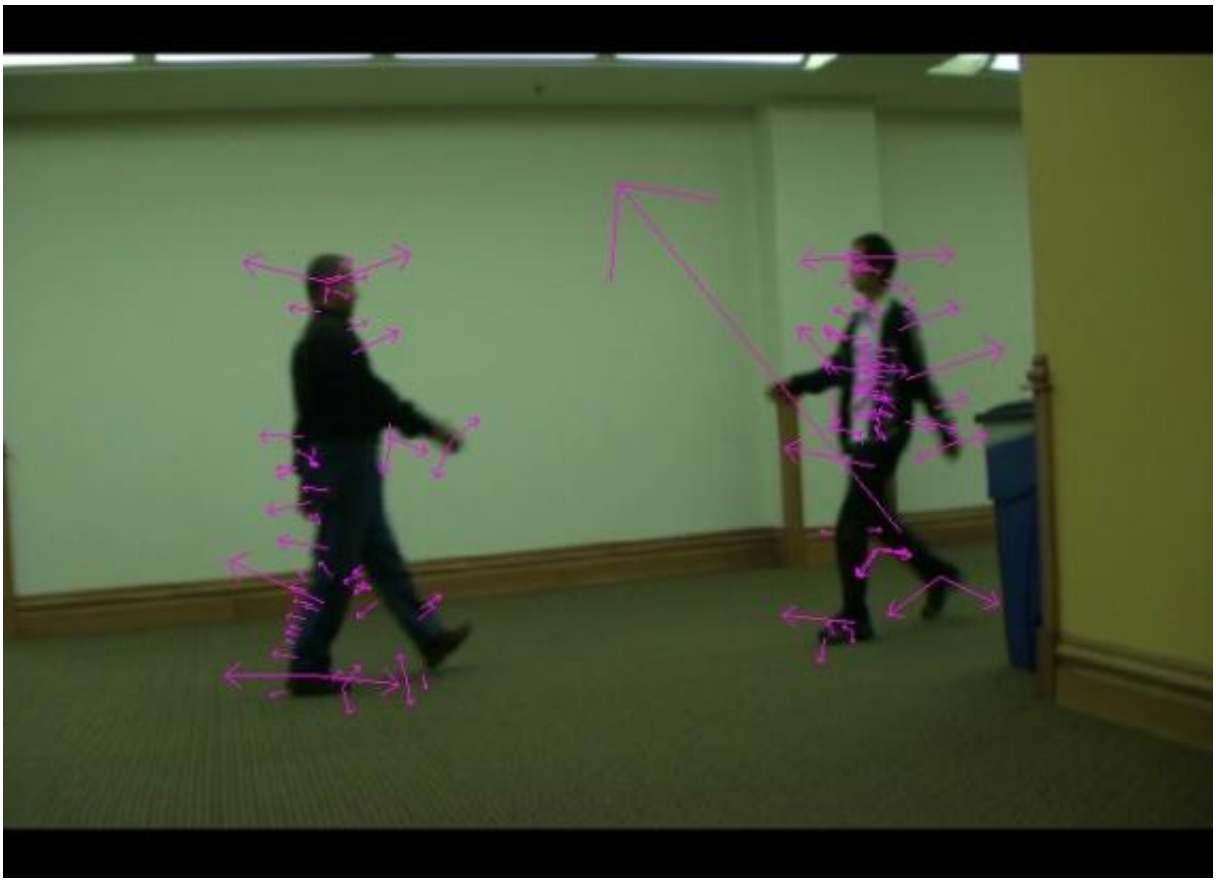


Fig. 3-6 SIFT detection example on the moving objects

3.2.3 Tracking

A simple tracker is applied to the foreground connected components (i.e., human body) detected in the segmentation step. These connected components considered as blobs, which probably are the moving objects or parts of the moving human body. The detected blobs are represented using a bounding box. Kalman filter is applied to these bounding boxes by tracking their position and predict their location in the next frame. Fig. 3-7 shows the tracking result of two different times.

Motion direction, η, is estimated using the difference in horizontal position between the center of the bounding boxes in two consecutive frames under study. Assume that the center position in horizontal axis are $\bar{x}_{t-1}$ and $\bar{x}_t$, the motion direction can be calculated as

$$\eta = sign(\bar{x}_t - \bar{x}_{t-1}) \tag{3-5}$$

Where η represents the action direction (left-to-right or right-to-left) and will be used in calculating SIFT-ME features. One of the reasons that extract motion direction is to make the SIFT-ME feature invariant to the direction of horizontal movement. For example, actions such as walking and jumping can be performed from left-to-right or from right-to-left and there should be no difference if the actions are same.

Fig. 3-7 tracking example

3.2.4 SIFT Correspondence

After extracting the SIFT features in each frame, finding the corresponding points between every two consecutive frames is necessary. A modified k-d tree method called Best-Bin-First [10] that can identify the nearest neighbors with highest probability is applied to every two consecutive frames. Bins in feature space are searched in the order of their closest distance from the query location. The best matched candidate for each key point is found by identifying its nearest neighbor. As a result of this step, corresponding points are detected and utilized for motion estimation.

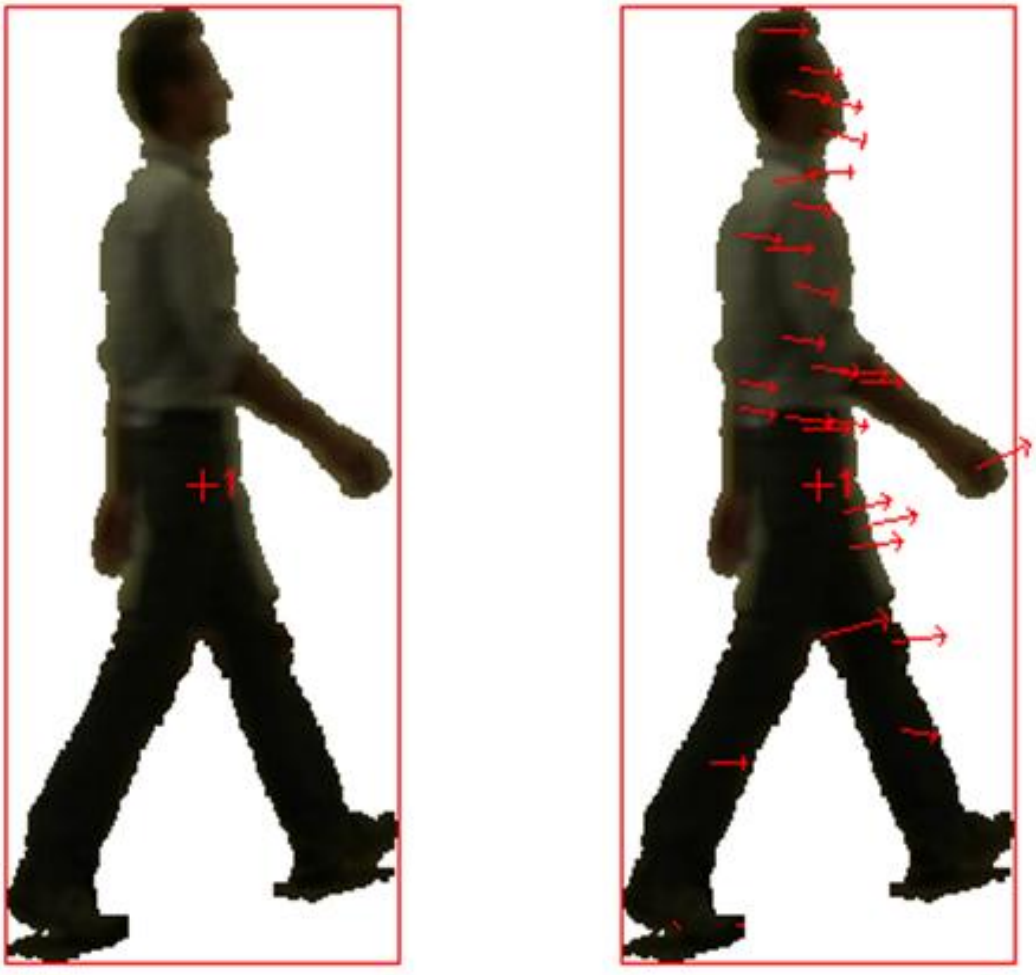


Fig. 3-8 SIFT Correspondence

3.2.5 SIFT-ME Representation

Once the corresponding points are found, the 2D transformation parameters can readily be calculated using the coordinates of the SIFT correspondences and the estimated motion direction. Equation 3-6 presents the 2D transformation parameters:

$$\begin{cases} \rho = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \\ \alpha = atan2\big((y_t - y_{t-1}), \eta(x_t - x_{t-1})\big) \\ \beta = \eta * sign(\sin(\theta_t - \theta_{t-1})) * \operatorname{acos}(\cos(\theta_t - \theta_{t-1})) \end{cases} \tag{3-6}$$

where ρ is the translation distance, $\alpha \in [-\pi, \pi]$ is the direction of translation , and $\beta \in [-\pi, \pi]$ is the rotation angle of the key point. The vector $< \rho, \alpha, \beta >$ is the SIFT-ME features extracted for each pair of corresponding points.

Because of adding motion direction in the translation angle and rotation calculation, SIFT-ME features are invariant to motion direction along x-axis (i.e., horizontal direction). The last image of fig. 3-9 shows SIFT-ME features, where the blue arrows and the orange arcs illustrate the point's translation and the rotation respect to SIFT key points, respectively.

Comparing with optical flow, SIFT-ME has many advantages. It is invariant to illumination, noise and view angle changes which inherited from SIFT and invariant to motion direction and body size which obtained from object tracking. Besides, SIFT-ME not only can interpret the translation of key point, but also can interpret the rotation around key point. Fig. 3-10 shows optical flow and SIFT-ME, SIFT-ME can register rotation information which displayed by orange arcs on the right image.

Fig. 3-9 SIFT-ME Features

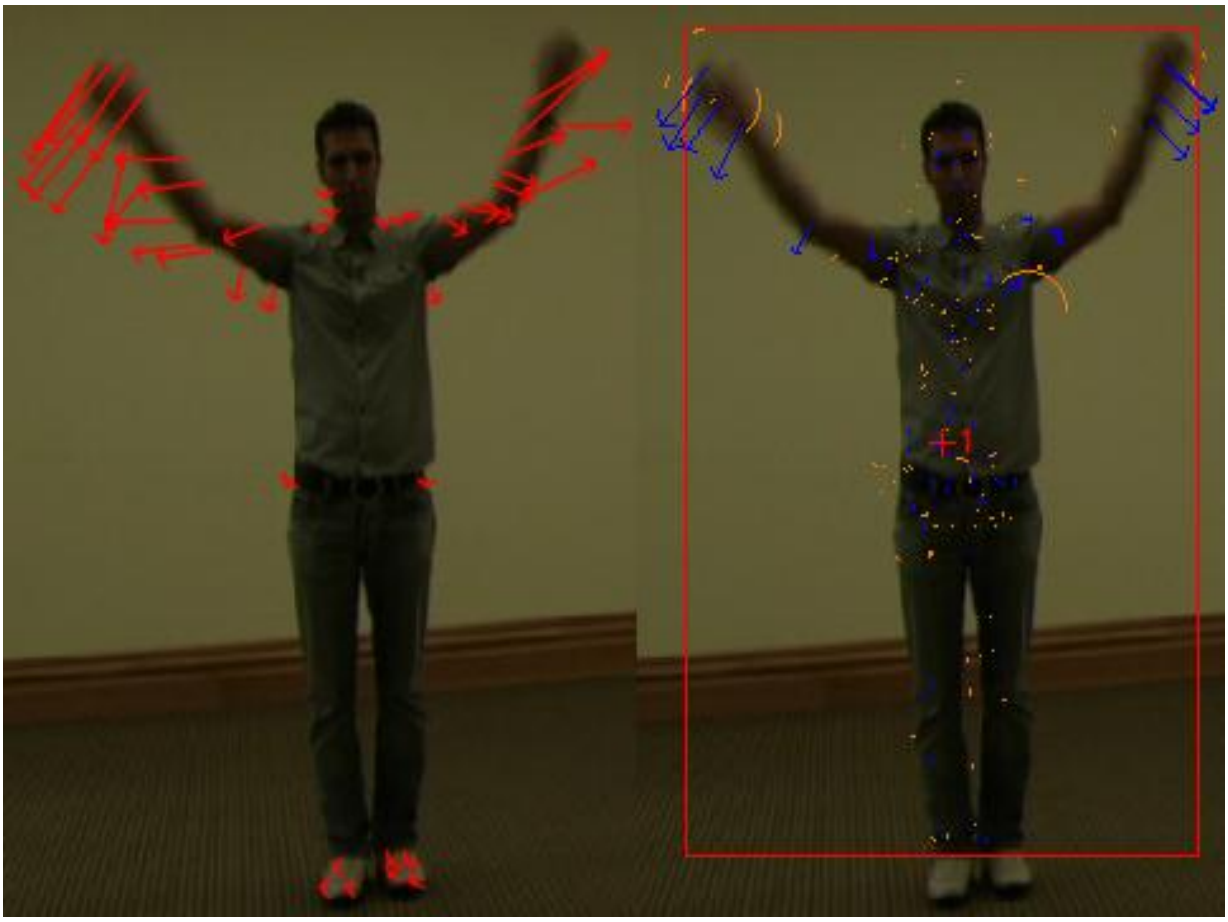


Fig. 3-10: Optical flow extracted using LK method [18] (left) and the SIFT-ME features (right).

3.2.6 Normalization

In order to make SIFT-ME invariant to actors' body size, image resolution, and the distance between camera and the subject, normalization of the translation vector with respect to the height of each actor is necessary. This normalization will make SIFT-ME scale invariant. Body size can be measured by calculating the height of the bounding box in the tracking step. Because the translation and rotation angles are scale invariant, only translation distance need the normalization. All the translation distances of each video frame are divided by the height of each actor in that video. After this step, the SIFT-ME is utilized for action recognition.

IV. Action Representation and Recognition

4.1Action Representation

Each SIFT-ME vector can be considered as action description, and all the SIFT-ME vectors gathered from one action videos were put together as data for classification and recogtion.

Gaussian Mixture Model (GMM) is a powerful method to learn statistical patterns of data [12]. GMM is applied to capture the action patterns and represent actions. Action motion features estimates using SIFT-ME. First of all, all the SIFT-ME features from one video are collected into one file and each SIFT-ME feature save as one line vector. Expected Maximization algorithm [1] applied to train separate GMMs each representing one action using SIFT-ME features extracted from videos of multiple subjects.

4.2 Action Recognition

The GMMs are utilized to classify the action performed in a given video into different classes (e.g., walking, running, etc). Let us assume that $\Psi = \{\psi_1, \cdots \psi_z\}$ is a set of GMMs registering the motion pattern of z actions, $\Theta = \{\vartheta_1, \cdots \vartheta_z\}$, which ϑ_z is the label of each action.

Given one $d(=3)$ dimension SIFT-ME vector $x \in \mathbb{R}^d$, the distribution probability for *N* components GMM model $\psi_k = \{\omega_i, \varphi_i\}_{i=1}^N$ is defined as:

$$f_{\psi_k}(x) = \sum_{i=1}^{N} \omega_i \, p(x|\varphi_i) \tag{4-1}$$

where

1. μ_i, Σ_i are the mean and variance of Gaussian φ_i, $\mu_i \in \mathbb{R}^d$ and Σ_i is $d \times d$ positive

matrix, $i = 1, \cdots N$.

2. ω_i is the weight of ith Gaussian component and $\omega_i > 0$ and $\sum_{i=1}^{N} \omega_i = 1$, $i = 1, \cdots N$.

3. x is a vector with d dimensions , the probability of x over single Gaussian $\varphi_i = \{\mu_i, \Sigma_i\}$ is

$$p(x|\varphi_i) = \frac{1}{\sqrt{2\pi^d |\Sigma_i|}} \exp(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)) \quad (4\text{-}2)$$

After training the GMMs, Maximum Log-likelihood method is applied to classify actions in a given video into different classes. Given an observation sequence, $X = (x_1, x_2, \cdots x_M)^T$, $X \in \mathbb{R}^{M \times d}$, assume that the observations are independent from each other. The likelihood function of Gaussian ψ_k, $(k = 1, \cdots z)$ is defined as:

$$\boldsymbol{\mathcal{L}(\psi_k) = f_{\psi_k}(x_1, x_2, \cdots x_M) = \prod_{j=1}^{M} f_{\psi_k}(x_j)}$$

$$\boldsymbol{= \prod_{j=1}^{M} p(x_j|\psi_k) = \prod_{j=1}^{M} (\sum_{i=1}^{N} \omega_i \, p(x_j|\varphi_i))}$$

(4-3)

The log-likelihood function is calculated by taking the logarithm value of the likelihood function:

$$\boldsymbol{\mathcal{L}^*(\psi_k) = \log \mathcal{L}(\psi_k) = \sum_{j=1}^{M} \log \left(\sum_{i=1}^{N} \omega_i \, p(x_j|\varphi_i) \right)}$$

(4-4)

In order to classify the actions, the maximum log-likelihood criterion utilized to assign

the observation X to GMMs:

$$\widehat{\boldsymbol{\psi}} = \mathbf{arg\ max}_{\boldsymbol{\psi_k} \in \boldsymbol{\Psi}} \boldsymbol{\mathcal{L}}^{*}(\boldsymbol{\psi_k}) \tag{4-5}$$

$\hat{\psi}$ denotes the GMM which has the largest log-likelihood for the observation data sequence. The action label corresponding to this GMM is assigned to the observation sequence:

$$\boldsymbol{Action} = \boldsymbol{\vartheta_k}, \boldsymbol{if}\ \left(\widehat{\boldsymbol{\psi}} = \boldsymbol{\psi_k}\right), \boldsymbol{k} \in \{\mathbf{1}, \cdots \boldsymbol{z}\} \tag{4-6}$$

After this step, each action video can be assigned with a label of action. If the label is same with action performed in the video, then action recognition can be considered success in classification, otherwise, if the labeled name is not the one performed in video, the classification result is wrong.

V. Experiment Results

A dataset of 261 videos of six different actions performed by seven subjects was used for training and test. The videos were captured using a JVC camcorder in 30 fps with resolution of 640x480. The actions include walking, running, turning around, jumping, waving, and picking up. The number of videos per action per subject is between 3 and 7. A few sample frames are shown in Fig.5-1.

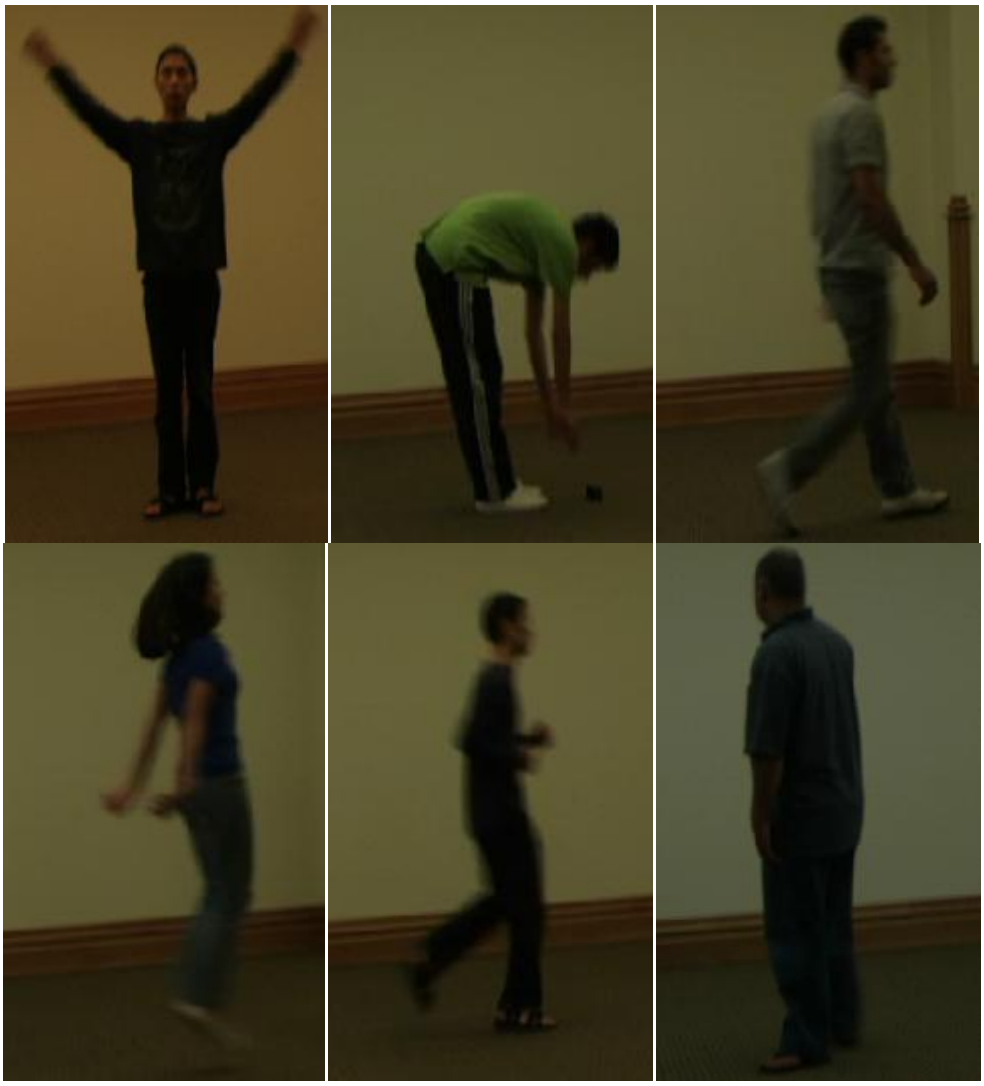


Fig. 5-1: Examples of different actions; from left-top to right bottom, the actions are: waving,

picking up, walking, jumping, running, and turning around.

Because most of the public human action datasets (e.g., Weizmann [9] and KTH [19]) have low resolutions (180x144 and 160x120, respectively), and SIFT-ME requires a handful of SIFT correspondences, therefore dataset which has a higher resolution (640x480) is the best choice. However, to have a fair comparison, the accuracy of existing methods in the literature for human activity recognition (i.e., 2D SIFT, 3D SFIT, optical flow) are tested with the dataset. The result comparison is presented in the following sections.

5.1 SIFT-ME and GMM Results

Leave-One-Subject-Out (LOSO) method is applied in this thesis to verify the performance of proposed approach and other approaches. First, all the videos of one subject from the training set were removed and used the videos of the left subject to train separate GMMs for separate actions (six GMMs for six different actions). Second, the excluded videos from the left subject were used to test the accuracy of the proposed approach in classifying the actions in the videos based on the maximum log-likelihood classification. This process is repeated for all subjects until every subject is used for testing.

Table 5-1 presents the results of classifying actions in videos into different classes based on the LOSO method. Table II shows the percent of classification for each action. As Table II illustrates, the accuracy of proposed approach is 100% for walking and above 93.4% for all the actions. The average recognition rate is 96.6%.

The effect of number of Gaussian components on the accuracy of proposed approach in recognizing human actions is also studied. Fig. 5-1 shows the accuracy of recognizing

different actions while different numbers of Gaussian components are used. The maximum average performance is achieved when the number of GMM components is five (i.e., the results in Tables 5-1 and 5-2).

Actions	Running	Walking	Jumping	Turning	Waving	Picking up
Running	**40**					
Walking	**2**	**47**				
Jumping			**42**			
Turning				**42**	**2**	
Waving				**1**	**42**	**2**
Picking up			**1**		**1**	**39**
Total	**42**	**47**	**43**	**43**	**45**	**41**

TABLE 5-1 CONFUSION MATRIX; RECOGNIZED NUMBER OF VIDEOS; THE NUMBER OF GMM COMPONENTS IS 5.

Actions	Running	Walking	Jumping	Turning	Waving	Picking up
Running	**95.2%**					
Walking	**4.8%**	**100%**				
Jumping			**97.7%**			
Turning				**97.7%**	**4.4%**	
Waving				**2.3%**	**93.4%**	**4.9%**
Picking up			**2.3%**		**2.2%**	**95.1%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

TABLE 5-2 CONFUSION MATRIX, PERCENTAGE OF RECOGNITION; THE NUMBER OF GMM COMPONENTS IS 5.

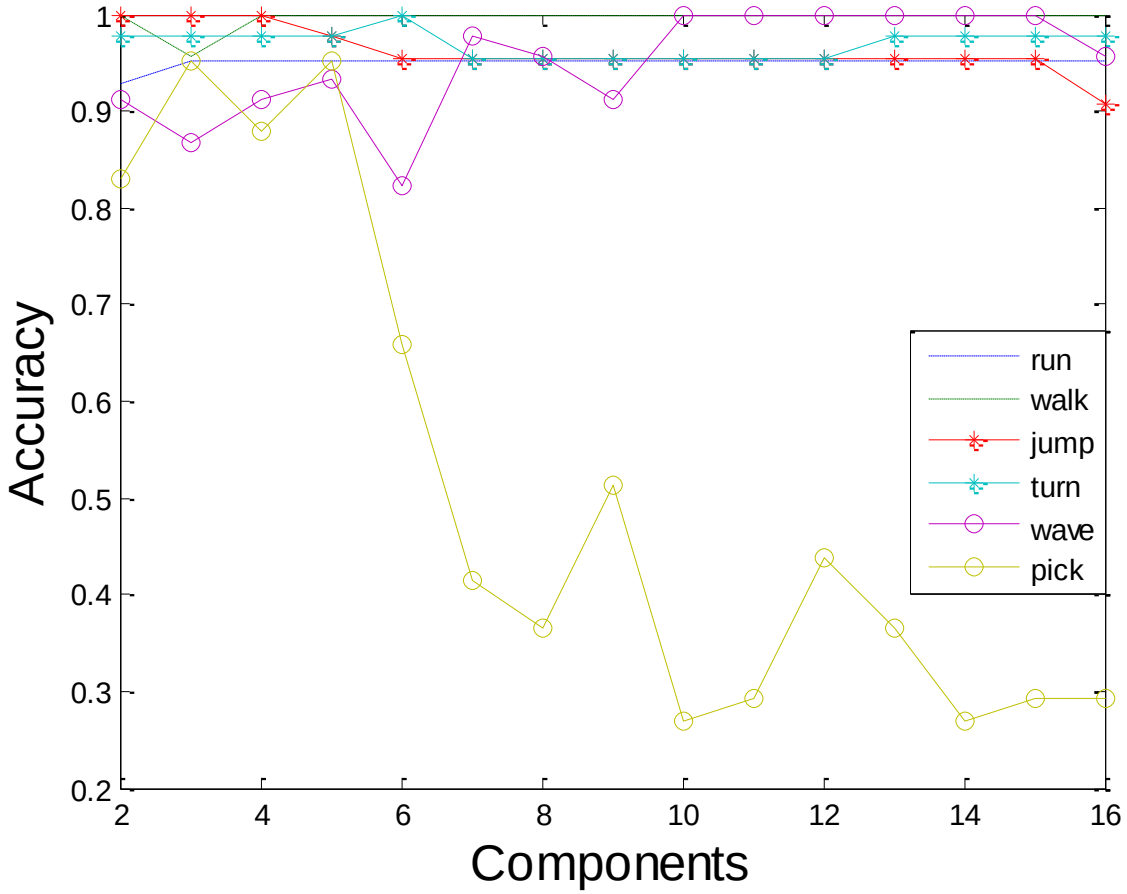


Fig.5-1 Effect of different number of Gaussian components used in GMMs.

From Fig. 5-1, Picking up action has large variations for different GMM components and the other actions have less variation as GMM components increase. At components five, all the action accuracies are the closest, making the recognition result reach the peak of 96.6% total accuracy.

The main reason for the confusion between the picking and waving actions, is the similarity of the rotation pattern between these two actions which makes the recognition task difficult (the picking action is classified as waving). This issue becomes more severe when the number of Gaussian components is larger than five and thus the overall

recognition rate of the system is decreased. One potential solution to this problem is to utilize the location information of the SIFT key points in SIFT-ME representation.

5.2 Compare between Descriptors

For comparison, 2D SIFT features and optical flow are utilized along with GMM and maximum log-likelihood for activity recognition. Based on optical flow [18], the motion features are extracted and used to train GMMs. Based on 2D-SIFT, the magnitude and angle of key points in each frame utilized to train GMMs and then maximum log-likelihood for action classification. Because, different number of components of GMMs influences the classification result, several GMMs with different number of components (2 to 16) are trained and tested. The accuracy of action recognition using different number of GMM components and different feature representations (i.e., 2D SIFT, optical flow, and SIFT-ME) are shown in fig.5-2.

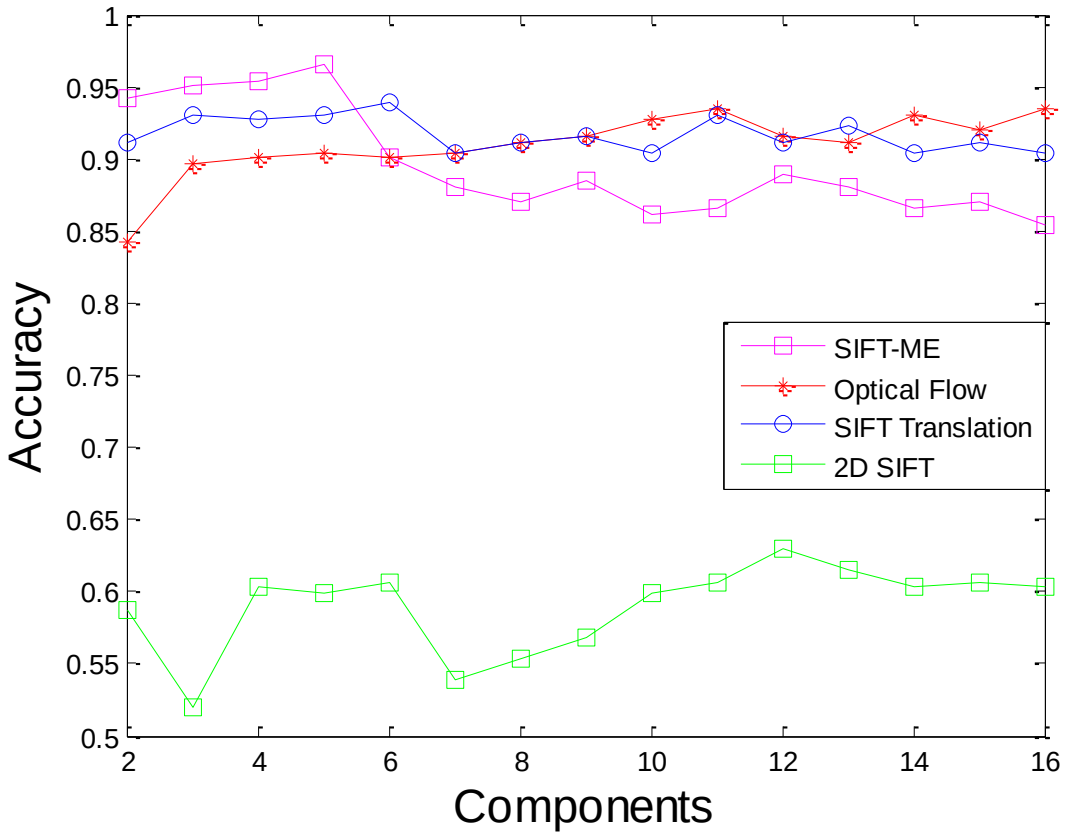


Fig. 5-2 Average performances for SIFT-ME, optical flow, SIFT Translation and 2D SIFT with different number of Gaussian components used.

In order to find the contribution of rotation information, translation features, ρ and α are utilized for action description and recognition, which reports a 93.9% recognition rate obtained (the SIFT Translation in Fig.8 and Table IV). Clearly, from Fig. 5-2, SIFT-ME outperforms all other approaches (2D SIFT, optical flow, and SIFT Translation) for human activity recognition. In addition, the highest accuracy based on SIFT-ME is achieved using fewer GMM components (less than six components). The big drop after components 5 is because the big drop of pick action shows on fig. 5-1.

5.3 Compare with 3D SIFT

Since 3D SIFT is a new technique for action recognition [17], it is applied to the dataset for a comparison. The process includes 3D SIFT features detection, codebooks generation with the Bag of Words technique [4], and then utilizes SVM for action recognition (exactly following the process presents in [17]).

Actions	Running	Walking	Jumping	Turning	Waving	Picking up
Running	**36**	**1**	**5**			
Walking	**3**	**42**	**3**			
Jumping	**3**	**4**	**32**	**1**	**1**	
Turning				**38**	**1**	**1**
Waving			**3**	**3**	**42**	**7**
Picking up				**1**	**1**	**33**
Total	**42**	**47**	**43**	**43**	**45**	**41**

TABLE 5-3 CONFUSION MATRIX OF 3DSIFT WITH MY DATASET IN NUMBERS

Actions	Running	Walking	Jumping	Turning	Waving	Picking up
Running	**85.8%**	**2.1%**	**11.6%**			
Walking	**7.1%**	**89.4%**	**7.0%**			
Jumping	**7.1%**	**8.5%**	**74.4%**	**2.3%**	**2.2%**	
Turning				**88.4%**	**2.2%**	**2.5%**
Waving			**7.0%**	**7.0%**	**93.3%**	**17%**
Picking up				**2.3%**	**2.2%**	**80.5%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

TABLE 5-4 CONFUSION MATRIX OF 3DSIFT WITH MY DATASET IN PERCENTAGE

Table 5-3 and table 5-4 show the confusion matrix using the results of 3D SIFT on the proposed dataset. The highest accuracy that achieves using 3D SIFT is 85.4%, which is close to the accuracy claimed by Scovanner et al. in [17].

Table 5-5 compares the recognition rate of optical flow, 3D SIFT and SIFT-ME on each action, from which one can see that although optical flow obtain is 100% accurate on walking, turning and picking up, however the recognition rate on jumping and waving is very low, which makes the total recognition rate lower at 93.5%. Compare with 3D SIFT and SIFT-ME, SIFT-ME has better performance on all actions than 3D SIFT.

Actions	Optical Flow	3DSIFT	SIFT-ME
Running	90.5%	85.8%	**95.2%**
Walking	100%	89.4%	**100%**
Jumping	81.4%	74.4%	**97.7%**
Turning	100%	88.4%	**97.7%**
Waving	88.9%	93.3%	**93.4%**
Picking up	100%	80.5%	**95.1%**
Total	93.5%	85.4%	**96.6%**

Table 5-5 Compare with optical flow, 3DSIFT and SIFT-ME on each action

Table 5-6 compares the highest accuracy that achieved using SIFT-ME descriptor with the highest accuracy that achieved using 2D SIFT, 3D SIFT, optical flow, and SIFT Translation. Clearly, SIFT-ME has the best recognition rate on human activity recognition among these descriptors.

Descriptor	Accuracy
2D SIFT	**63.0%**
3D SIFT	**85.4%**
Optical Flow	**93.5%**
SIFT Translation	**93.9%**
SIFT-ME	**96.6%**

Table 5-6 Compare the accuracy with all descriptors at the highest accuracy

The comparisons show clearly that SIFT-ME is a robust feature for human activity recognition. SIFT-ME invariant to the environment changes like illumination, small view angle and noise because these advantages inherent form SIFT features as SIFT-ME based on SIFT features and SIFT correspondence. At the same time, SIFT-ME invariant to human behavior difference like motion direction and human body scale as the equation for SIFT-ME already covers these information and make these information as one part of calculation of SIFT-ME. So SIFT-ME is robust to environment variations. Besides, SIFT-ME reveals full parameters of the motion of human body in 2D plane as it adds the rotation information for key point, which is a big improvement compared to existing techniques like optical flow, SIFT Correspondence and MoSIFT.

VI. Conclusion and Future Work

6.1 Conclusion

A new motion description method named SIFT-ME has been proposed to reduce the influence of environment variations for human activity recognition. SIFT-ME is based on SIFT features and inherits its advantages such as invariant to environment variations noise, illumination, and camera view angles. Thus, SIFT-ME can be used as a robust method for motion description of activity recognition in the field of computer vision and pattern recognition. SIFT-ME features are successfully utilized for describing and recognizing human actions in videos.

The experiment shows that SIFT-ME outperforms optical flow, 3D SIFT, and 2D SIFT features for human activity recognition. Besides, some additional experiments are done to find the relations of recognition result with different GMM components. SIFT-ME features have the highest recognition rate with lowest GMM components which demonstrate that it is a better action description. On the other hand, compare SIFT-ME and SIFT translation, SIFT-ME outperforms almost 3% accuracy, which means rotation information is also important for activity recognition. SIFT-ME is another evolution step which improves SIFT to interpret 2D transformation using a three dimensional vector.

6.2 Future Work

In the future, more research will be done with SIFT-ME features for querying videos to detect a predefined set of actions such as walking, running, etc. This can be easily achieved by comparing patterns of action. SIFT-ME is a good feature

representation for motion, as long as finding similar motion patterns with queried one, the best match can be achieved and video retrieval can be realized. SIFT-ME can be improved by using a better matching algorithm as well as incorporating 3D translation and rotation by considering multiple cameras.

As SIFT-ME is an extension of SIFT features to represent motion, it does not conflict with SIFT feature. As well known that SIFT can be used for object recognition with promising result, so there is possibility to combine SIFT and SIFT-ME to do object based activity recognition: recognition activity through the objects people interact with. For example, pick up a gun and pick up an apple are different actions in detail and results in different handling response actions to other people. Recognize activities through objects will extend the application of activity recognition in many fields and increase the capacity of activity recognition.

Reference

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, NO. 1, pp.1-38, 1997.

[2] Singh, M. Basu, A. Mandal, M. K., "Human Activity Recognition Based on Silhouette Directionality," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, pp. 1280-1292, 2008.

[3] Lowe D., "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Computer Vision*, vol. 60, pp.91-110, 2004.

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

[5] Efros A. A., Berg A.C., Mori G., Malik J., "Recognizing Action at a Distance," *Proc. of the 9th IEEE Int. Conf. on Computer Vision*, Nice, France, vol.2,pp. 726-733. Oct. 2003.

[6] Chaudhry R., Ravichandran A., Hager G., Vidal R., "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," *IEEE conf. on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 1932 – 1939, Aug. 2009.

[7] Min J., Kasturi R., "Activity recognition based on multiple motion trajectories," *Proc. of the 17th Int. Conf. on Pattern Recognition*, vol. 4, pp. 199-202, Sep. 2004.

[8] Messing R., Pal C., Kautz H., "Activity Recognition using the Velocity Histories of Tracked Keypoints," *IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, Sep. 2009.

[9] Gorelick L., Blank M., Shechtman E., Irani M., Basri R., "Actions as Space-Time Shapes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2247 – 2253, Nov. 2007.

[10] Beis J. and Lowe D.G., "Shape Indexing Using Approximate Nearest-Neighbor Search in High-Dimensional Spaces," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1000-1006, Jun. 1997.

[11] Stauffer C. and Grimson W.E.L., "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747-757, Aug. 2000.

[12] Calinon, S., Guenter, F. and Billard, A., "On Learning, Representing and Generalizing a Task in a Humanoid Robot," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, pp. 286-298, Apr. 2007.

[13] Kellokumpu V., Pietikainen M., Heikkila J., "Human activity recognition using sequences of postures," *IAPR Conference on Machine Vision Applications*, Tsukuba Science City, Japan, pp. 570-573, May. 2005.

[14] Zhou X., Zhuang X., Yan S., Chang S.-Y., Hasegawa-Johnson, M., Huang T. S., "SIFT-Bag Kernel for Video Event Analysis," *Proc. of the 16th ACM Intern. Conf. on Multimedia*, Vancouver, British Columbia, Canada, pp. 229-238, 2008.

[15] Li L.-J., Fei-Fei L., "What, where and who? Classifying events by scene and object recognition," *IEEE Intern. Conf. on Computer Vision*, Rio de Janeiro, pp. 1-8, Oct. 2007.

[16] Niu F., Abdel-Mottaleb M., "HMM-Based Segmentation and Recognition of Human Activities from Video Sequences," *IEEE Intern. Conf. on Multimedia and Expo (ICME)*, Amsterdam, Netherlands, pp.804-807, July, 2005.

[17] Scovanner P., Ali S., Shah M., "A 3-Dimensional SIFT Descriptor and its Application to Action Recognition," *Proc. of the 15th Int. Conf. on Multimedia*, pp.357-360, 2007.

[18] Lucas B. D., Kanade T., "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proc. of the 1981 DARPA Image Understanding Workshop*, pp. 121-130, April, 1981.

[19] Schuldt C., Laptev I., and Caputo B. "Recognizing Human Actions: A Local SVM Approach", *Proc. Of International Pattern Recognition, Conference (ICPR'04)*, Cambridge, UK, 2004.

[20] Chen M. Hauptmann A., "MoSIFT: Recognizing Human Actions in Surveillance Videos", CMU-CS-09-161, Carnegie Mellon University, Pittsburgh, PA, 2009.

[21] Cuntoor N., Kale A., Chellappa R., "Combining multiple evidences for gait recognition," *Proc. ICASSP*, vol. 3, pp. 33–36, 2003.

[22] Wang L., Ning H., Tan T., Hu W., "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

[23] http://en.wikipedia.org/wiki/Activity_recognition

[24] Pollack, M. E., Brown L., Colbry D., McCarthy C. E., Orosz C., Peintner B., Ramakrishnan S., Tsamardinos I., "Autominder: an intelligent cognitive orthotic

system for people with memory impairment," *Robotics and Autonomous Systems*, PP.1-9, 2003.

[25] Stauffer C., Grimson W. E. L., "Adaptive background mixture models for real-time tracking," *IEEE Conf. on Computer Vision and Pattern Recognition*, Fort Collins, CO, USA, vol.2, PP.246-252, 1999.