University of Alberta

SHADOW REMOVAL FOR ACTION RECOGNITION IN A SMART CONDO ENVIRONMENT

by

Samaneh Eskandari

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2014
Edmonton, Alberta

Abstract

Depending on the method we choose for human action recognition, some algorithms require the localization of human in action videos as a preprocessing step. This preprocessing is more challenging in real environments with noises or varying illumination, and may include background subtraction, shadow removal and noise removal.

This thesis concerns with shadow removal problem with the goal of extracting human silhouette properly. Detecting shadows can be considered as a labeling problem and can be transformed into an energy minimization problem in a Markov Random Field framework. The advantage of the algorithm is that it considers the neighborhood information as well as chromaticity values of the pixels.

Also a new dataset is provided containing several daily actions of people living in a smart condo. The dataset is used in order to evaluate the proposed method. Moreover, it can be useful in action recognition and motion analysis studies with health care applications.

Acknowledgements

First and foremost, I would like to praise and thank God for providing me this opportunity and granting me the wisdom, capability and strength to proceed. The following document summarises two years worth of study, research and experiments. The contributions of many different people have made this project possible. I would therefore like to extend my appreciation and offer my sincere thanks to each and every one of them.

First of all, I am grateful to my supervisor, Dr. Herbert Yang for being abundantly helpful throughout this project. Without his patience, warm encouragements, thoughtful guidance, critical advises and invaluable assistance this achievement would have not been possible.

I would also like to express my gratitude to the committee members, Dr. Eleni Stroulia and Dr. Pierre Boulanger for their contributions and helpful comments on this work. Certainly, their suggestions are greatly valuable to improve my thesis.

Last but not least, I would like to thank my beloved family for their unconditional support, both financially and emotionally throughout my academic trajectory. I am also grateful to all my friends and my group members in the Computer Graphics Lab at University of Alberta who were always by my side.

Table of Contents

1 Introduction 1

2 Background and Related Work 4

2.1 Human Action Recognition 4

2.2 Background Subtraction 10

2.3 Shadow Removal 12

2.4 Dataset 16

3 The Proposed Method 19

3.1 Motivation 19

3.2 Shadow detection as a labeling problem 20

3.2.1 Data Term 21

3.2.2 Smoothness Term 22

4 Experiments 24

4.1 Human Action Recognition in a Smart Condo 24

4.2 Smart Condo 24

4.3 Human Action Dataset 26

4.4 Implementation 29

4.4.1 Mixture of Gaussian Background Subtraction 29

4.4.2 Shadow Removal 29

4.4.3 Chromaticity-based Method 29

4.4.4 The Proposed Method 32

4.4.5 Other methods 32

4.4.6 Video Alignment 32

4.4.7 Action as Space-time Shape 34

4.5 Results 35

4.5.1 Experiment 1: The Recognition Performance on the Smart Condo Action Data 35

4.5.2 Experiment 2: The Effect of MRF-based Shadow Removal Method in Action Recognition 36

4.5.3 Experiment 3: Comparing Other Shadow Removal Methods with the MRF-based Method in Action Recognition . . 38

5 Conclusion and Future Works 40

5.1 Conclusion 40

5.1.1 Human Action Dataset 40

5.1.2 Shadow Removal 41

5.2 Future Works 41

Bibliography 43

List of Tables

4.1 Action confusion in classification using our method 35

4.2 Action classification accuracies using our shadow removal method . 37

4.3 Action classification accuracies using the chromaticity-based shadow removal method. 37

4.4 Action classification accuracies without a shadow removal. 37

4.5 Action classification accuracies using each one of MRF-based, physical-based, and chromaticity-based methods. 38

List of Figures

2.1 (a) A frame of a soccer video (b) The strongest interest point detected in the sequence on a player heading the ball (c) Resulting interest points for a video frame with the action hand shake (from [1]). 5

2.2 Comparison of MEI and MHI for an action. (a) One example frame of an aerobic action. (b) The MEI image represents the shape of the region in which the motion occurs. (c) The MHI image represents the temporal history of an motion (from [2]). 7

2.3 (a) Optical flow of bend action at frame 20. (b) Optical flow of hand wave action at frame 10 (from [3]). 8

2.4 (a) The space-time shape of “jumping-jack” and “walk” actions, and (b) their corresponding solutions of Poisson equation (from [4]). 8

2.5 The top row, from left to right: The original frame of jumping-jack action, the extracted foreground mask, solution of Poisson eq. The bottom row, from left to right: space-time saliency, measure of “plateness”, measure of “stickness” (from [4]). 9

2.6 Comparison of trajectory based SSM matrices for actions (a) bend and (b) kick, taken from two different views (from [5]). 10

2.7 an object-shadow pair and its contour information (from [6]). . . . 14

2.8 KTH action dataset: examples of sequences of different types of actions and scenarios (from [7]). 16

2.9 Weizmann action dataset: examples of action sequences (from [4]). 17

2.10 INRIA Xmas Motion Acquisition, example action sequence (from [8]). 18

4.1 The work flow of the action recognition in the smart condo 25

4.2 The smart condo from different views, captured by mounted cameras. 25

4.3 A single view is used to capture all the actions. 26

4.4 The human action dataset. The top row from left to right: bend, lie. The middle row from left to right: limp, run. The bottom row from left to right: sit, stand-up. 27

4.5 The human action dataset. The top row from left to right: turn, walk. The bottom row: walker(1), walker(2). 28

4.6 Background Subtraction (BGS) results: the first two rows show sample images from the raw videos. The last two rows show the results of BGS. 30

4.7 Chromaticity-based shadow removal results: the first two rows show the results of BGS. The last two rows show the results of the algorithm proposed by Cucchiara et al. [9]. 31

4.8 MRF shadow removal results: the first two rows show the results of BGS. The last two rows show the results of our shadow removal algorithm. 33

4.9 Shadow removal results after applying different methods: the first row shows the results of BGS. The second row shows the results of physical-based shadow removal, the third row shows the results of geometrical-based Shadow removal, and the fourth row shows the results of texture-based shadow removal. 39

Chapter 1

Introduction

Human action recognition considers the task of labeling videos which contain different human action classes. In a simple case a video is segmented into single actions, and the recognition algorithm assigns one of the learned labels to each video segment. In a more general form, labeling a sequence of actions is needed, which requires detecting actions as well. The detection and recognition of human actions have been motivated by many applications such as video indexing, video retrieval, human computer interaction, visual surveillance and security applications. As an example, automated video surveillance in public places is often employed nowadays. Detecting suspicious activities like “a person leaving a bag” in an airport, or detecting unusual activities such as “an old person falling” in a senior care center, are some of the ultimate goals of a surveillance system.

Humans have innate abilities in recognizing and distinguishing different actions, or in detecting desired activities. This can be due to our perception of the scene and human poses that are achieved by either our immediate visual evidence, our visual experience, our interaction with the real world or a combination of them. In fact, humans have an advantage of understanding the physical and geometrical properties of scenes and actors (the shape of the body, the depth information and the color and texture details of the scene) in order to determine the relationship among them. However, for a computer, considering all of these factors, properties and the relationship between them make recognition quite difficult. Thus, almost all of the existing methods consider the task of action recognition with a particular application, and use assumptions and constraints to simplify the problem. For instance, one may assume the camera and/or the scene are static, each action is assigned to exactly one predetermined label and the whole body of the action performer falls within the camera view.

Despite the fact that various approaches with different environmental settings have been used for action recognition and motion analysis, very few of them have addressed the problem in the context of health care applications in real environments. For instance, Osmani et al. [10] analyze the activities performed by doctors and nurses in a health-care environment in order to improve the efficiency and quality of medical services. In another study, Robinovitch et al. [11] analyze the real-life falls using long-term health care data in order to prevent fallings. In this thesis, we consider human actions in a smart condo with health care applications.

A smart condo is a health care environment accompanied with assisted living devices and technologies for patients or elderly people in order to increase the quality of treatment and of their lives. It is expected to reduce the risk of injuries during their stay. A smart condo is facilitated with intelligent technologies such as wireless sensors and cameras for remote monitoring the place. This remote monitoring can provide health care professionals a better understanding of the daily activities of patients. This understanding enables the professionals to analyze the patients movements, to identify the need of treatment or rehabilitation and to discriminate normal and risky activities in order to reduce the risk of injuries.

The ability to recognize human actions automatically in a smart condo offers many advantages, such as providing useful information for analyzing the daily activities of patients, detecting unusual or risky activities and designing facilities. These automatic analysis of the captured data can ease and reduce the efforts of the medical staff. There are some challenges with this task such as illumination changes, inhomogeneous background, variations in performances and clothing, shadows and highlights, occlusion by furniture, and so on.

Depending on the method, many action recognition or more generally motion analysis algorithms need preprocessing of the captured data. The preprocessing may include foreground/background segmentation [4], shadow removal, noise removal, video alignments [12], etc. Most of the methods assume that the proper silhouette is available and remove the burden of these preprocessing steps from the recognition task, which may not be possible in a real application.

In this thesis, the contributions include a new dataset containing basic daily actions captured in a smart condo environment and a new shadow removal algorithm. The goal of the dataset is to provide common challenges in a real situation with inhomogeneous background, shadows, illumination changes and noises. The dataset includes considerable variation in action performances and in the actors themselves. The shadow removal methodology can be applied to the results of the background subtraction step. It is based on chromaticity values. The unique feature of the algorithm is that it considers the neighborhood information of pixels, and provides an improvement compared with that of existing methods. Additionally, it is fast enough and gives reasonably good foreground without removing informative parts compared to other algorithms. We assume that the camera and the background are static, and that all the actions are captured from the same view for simplicity. Given an image containing the moving human and its shadow, the output of the shadow removal stage includes the human body as foreground and the shadow pixels as background. Then extracted silhouettes can be used as inputs to a recognition task. Furthermore, the proposed method can be used by any motion analysis algorithm as a preprocessing step.

The remaining chapters of this thesis are organized as follows. Chapter 2 discusses the previous works in action recognition, background subtraction, and shadow removal. Then some existing action datasets are introduced. Chapter 3 presents the motivation of the study and explains the proposed method. Chapter 4 describes the works flow of human action recognition in a smart condo and gives an introduction about the smart condo facility of the university of Alberta. Then the proposed human action dataset is introduced, and the implementation details and experimental results are presented. Finally, chapter 5 provides a conclusion and recommendations for future works.

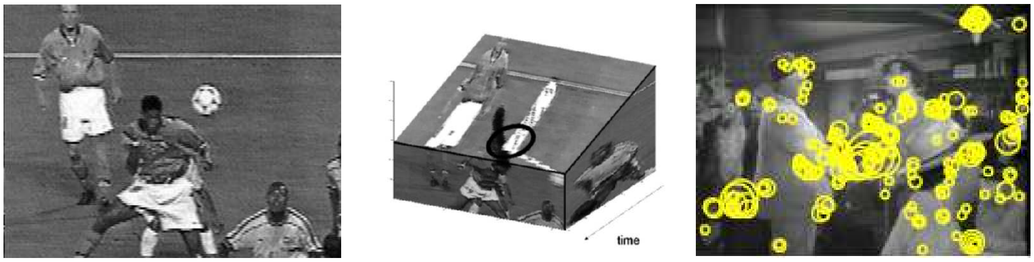
Chapter 2

Background and Related Work

2.1 Human Action Recognition

Vision based activity recognition is the process of analyzing and labeling human movements in a video or a sequence of images. The recognition of an activity can be performed at various levels. A gesture or action primitive is defined as the movement of a human's body part and can be described as an atomic component of an action. For instance, putting a leg forward or raising a hand may be referred to gestures, whereas walking is an action. So an action is composed of several atomic gestures and may contain the whole body's movement. Actions are defined as simple unique activities which may be possibly cyclic, such as running or jumping. However, an activity can be described as a complex sequence of consecutive actions, such as playing soccer that includes running, kicking, jumping, etc. Moreover, an activity may include interactions between two or more people or perhaps between a human and an object.

Early works on motion analysis and activity recognition were conducted back in the 1990's, as Aggarwal et al. stated based on their survey [13], whereas these analysis were mainly concentrated on primitive actions and motions with simple structures [14] [15]. However, in recent years, many studies have been done in areas with more complex motion analysis in a wide range of applications under different environments such as human tracking, pose estimation, body structure analysis and action recognition. This thesis is focused on recognition of relatively simple actions, and interactions with objects or group activities are not considered. Usually,



(a) (b) (c)

Figure 2.1: (a) A frame of a soccer video (b) The strongest interest point detected in the sequence on a player heading the ball (c) Resulting interest points for a video frame with the action hand shake (from [1]).

it is assumed that actions are segmented in time and this assumption removes the extra process of segmentation away from the recognition task, although it is not always realistic. Action detection algorithms address this issue [16].

In general, vision-based human action recognition can be considered as feature extraction and then classification of these representations. Various studies have approached the problem differently. Some methods [7] [17] [18] represent an action by a collection of local descriptors. They encode the sequence of images using local patches and represent information using a collection of interest points. The bag of words is then used to summarize the whole action as a set of local descriptors. An advantage of such methods is that usually there is no need for preprocessing, such as human localization and background subtraction in videos.

Laptev et al. [1] extracted local 3D information from action videos by extending the Harris [19] Forstner [20] interest point detector. The idea is based on the observation that pixel values with significant local variations in space and time contain useful information in encoding the video and in understanding the motion. After extracting 3D interest points, local spatio-temporal neighborhoods of the points are described using spatio-temporal Gaussian derivatives for action detection in videos. Later the authors use the same interest points and compute the 3D Histogram of Gradients(HOG) and Optical Flow(OF) features for each cuboid around the points for action recognition [21]. Figure 2.1 illustrates interest points detected in two different action videos. One drawback of such methods is that the detected interest points are sparse and unstable. As an improvement, Williems et al. [22] propose a method to detect dense and scale invariant spatio-temporal interest points. In addition to using Harris detector, they use the determinant of the Hessian matrix of interest points as the saliency measure to localize the features. Their detector is more efficient and robust to scale changes. They also use the SURF [23] descriptor to describe the interest points in various space and time scales. Other algorithms use the local HOG and HOF descriptors [21] or the SIFT descriptor [24] for action recognition. One disadvantage of representing actions with local features is that the relative information between local features is not kept. Usually a bag-of-words or a histogram of features is used to represent the whole action. Works done by Niebles et al. [17] and Park et al. [25] are examples of such an approach.

On the other hand, some methods use the global representation of actions for recognition. The basic idea is to encode the region of interest (person) as a whole, which contains more information about the structure of the body. A useful representation is the silhouette of the person. Bobick and Davis [2] propose one of the earliest works on movement representation and action recognition using silhouettes. The idea is to consider the shape of the region of interest to find and characterize movements. They define two simple templates to demonstrate the motion properties of pixels. A binary Motion Energy Image (MEI) that shows the occurrence of motion is constructed by computing the difference of two consecutive frames. Then the Motion History Image (MHI) is defined as a function of the history of motion for each pixel, such that a recent movement is shown with a higher intensity. Figure 2.2 shows the MEI and MHI images for an aerobic move. Finally, using moment-based features the statistical information of the both templates (MEI and MHI) are extracted and used for comparison of aerobic moves in action recognition experiments.

Apart from concentrating on the physical features of a person, some methods consider motion information [26] [27]. Shah et al. [3] extract some discriminative features from optical flow (OF) for action recognition. The OF fields for two different actions are shown in figure 2.3. The “Divergence” feature is defined to focus on the expanding OF caused by the movements of limbs. Also the “Vorticity” feature emphasizes on circular motion. There is also another feature, which represents

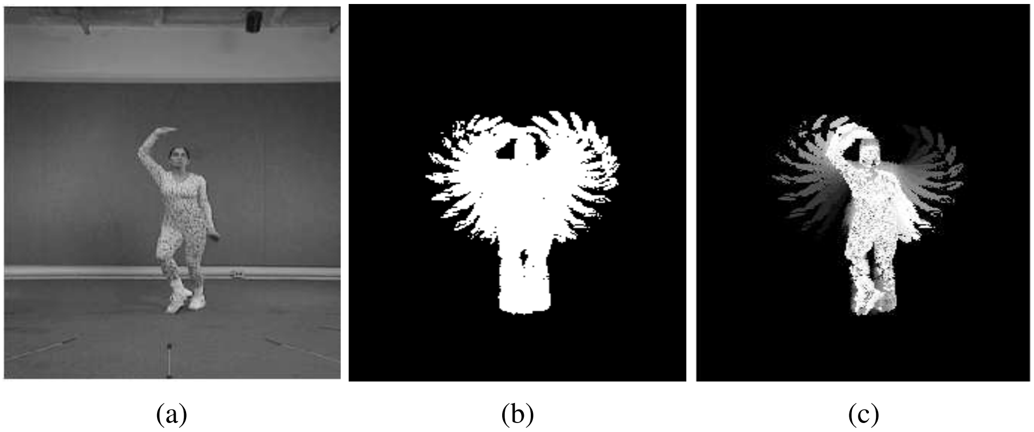


Figure 2.2: Comparison of MEI and MHI for an action. (a) One example frame of an aerobic action. (b) The MEI image represents the shape of the region in which the motion occurs. (c) The MHI image represents the temporal history of an motion (from [2]).

the symmetry in time of a human action around the diagonal axis of the image. Moreover, they define several kinematic features to encode small scale motions of limbs and to show deviation of the action performance from rigid body motion. Finally, dominant kinematic modes are extracted by performing Principal Component Analysis (PCA) of the above features.

Furthermore, some algorithms consider an action as a 3D Spatio-Temporal Volume (STV) and try to represent and compare the STVs by some informative features [28] [29]. Blank et al. [4] use appearance based features by extending the idea of Gorelick et al. [30] to 3D shapes. Gorelick et al. introduce a 2D shape representation by solving a Poisson equation of the shape that is interpreted as a mean distance of each point to the boundary undergoing a random walk. Blank et al. extract the 3D space-time shape from each action video, and solve the of Poisson equation for the 3D shape. Figure 2.4 shows the 3D STV and the solution of Poisson equation for actions “jumping jack” and “walk”. The space-time saliency feature is defined to emphasize on fast moving body parts relative to the torso. Also the “stickness” and the “plateness” features which are inspired by an old work by Rivlin et al. [31] are extracted to represent the local orientation and aspect ratios of the 3D shape. Finally, an action sequence is represented by a set of global features as the weighted moments of the 3D shape and the above mentioned local features

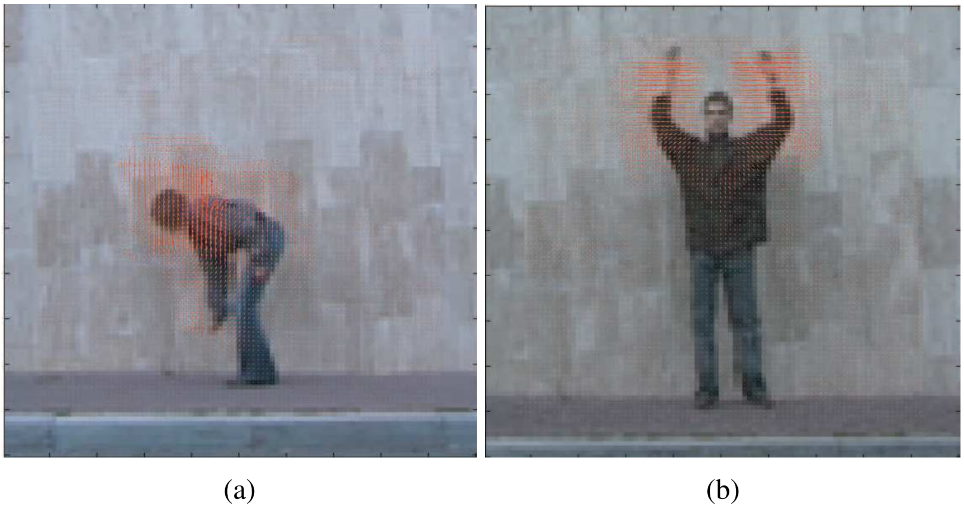
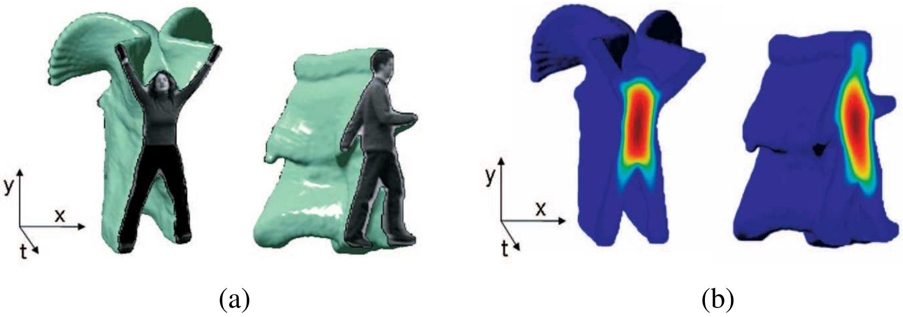

(a) (b)

Figure 2.3: (a) Optical flow of bend action at frame 20. (b) Optical flow of hand wave action at frame 10 (from [3]).

are used as the weights. The nearest neighbor classification method is used to compare the actions. Figure 2.5 shows the features used in the paper [4] for "jumping jack" action.

Some other algorithms interpret an action as a set of space-time trajectories of points [32] [33]. Junejo et al. [5] construct a self-similarity matrix (SSM) by computing pairwise distances between features extracted from a moving person in different time frames. So the SSM captures structural similarities and dissimilarities of the body pose within an action sequence. It is independent to the view and robust to performance variations. Figure 2.7 demonstrates the SSM matrix constructed from the 3D position of the body joints for two actions, each taken from a different



(a) (b)

Figure 2.4: (a) The space-time shape of "jumping-jack" and "walk" actions, and (b) their corresponding solutions of Poisson equation (from [4]).

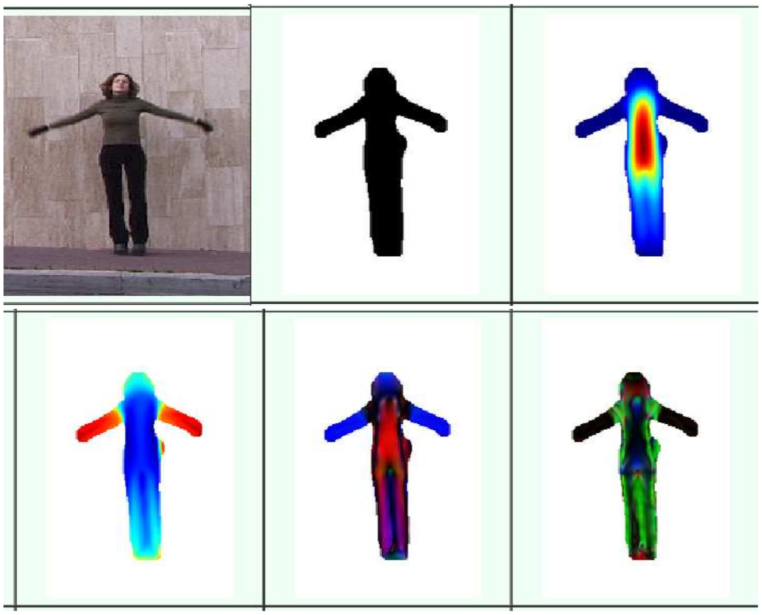


Figure 2.5: The top row, from left to right: The original frame of jumping-jack action, the extracted foreground mask, solution of Poisson eq. The bottom row, from left to right: space-time saliency, measure of "plateness", measure of "stickness" (from [4]).

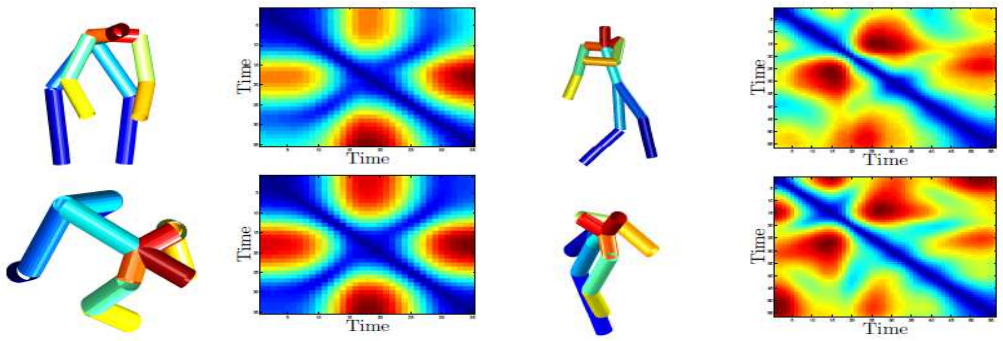
view. It can be seen that the self-similarity matrix represents similar patterns for an action taken in different view points. Different features are used to construct the SSM matrices including the 3D position of body joints, Histogram of Gradients (HOG) vectors extracted from each frame and the concatenation of the optical flow vectors within a bounding box around the person. These different SSMs are then evaluated by the authors for the task of action recognition.

As a whole, global representation methods are dependent on preprocessing, such as background subtraction and tracking; and although they encode useful general information of the action, they are more sensitive to noise and partial occlusions. Therefore, if a proper preprocessing method is available, usually global representation methods perform well.

Finally, another group of approaches use hierarchical methods to recognize high level and more complicated activities or human interactions with objects. They assume that a complex activity is composed of simpler actions, and those simpler actions can be recognized easily using previous methods. For instance, fighting is a high level activity composed of punching and kicking. These approaches need less training data to learn activities with complex structures compared to single layered approaches that were introduced before. As examples of such approaches, Oliver et al. [34] and Nguyen et al. [35] use multi-layered state-based models such as Hidden Markov Model (HMM) to learn activities with sequential structures. Moreover, Ivanov and Bobick [36] and Minnen et al. [37] use context free grammars, which consider activities as a set of atomic actions that can be described by some production rules. The drawbacks with such algorithms are the difficulty of finding the productions rules and none-uniqueness in decomposing activities into simpler actions.

2.2 Background Subtraction

Background subtraction in a video is the process of detecting moving objects from the static scene. The early study of background subtraction took place more than thirty years ago [38]. Recently it has had a wide applications in different areas of computer vision such as tracking, object detection, recognition, etc. Depending on the application and the environment, different approaches have been developed for removing the background. In this process, one might usually face with challenges such as varying background in outdoor scenes (trees, the sea surface, etc.), gradually or sudden changes in illumination, color similarities between the background and foreground pixels, light reflection and shadows. In recent stud-



(a) (b)

Figure 2.6: Comparison of trajectory based SSM matrices for actions (a) bend and (b) kick, taken from two different views (from [5]).

ies mostly statistical methods have been used to model the background. Sobral et al. [39] have provided a C++ framework named BGSLibrary for background subtraction and have implemented and adapted several existing algorithms [40]. The Mixture of Gaussian (MOG) models [41] is a widely used approach for background modeling and numerous improvements of the original method have been proposed recently [42] [43] [44]. In this thesis the MOG algorithm by Stauffer et al. [41] is employed because it gives reasonably good results in our application and is fast enough. The algorithm's code is available in the LBMixtureOfGaussians class of the BGSlibrary package and is adapted from the "Scene" application by Bender et al. [45].

The original MOG was introduced by Stauffer et al. [41]. In the RGB color space, the probability of observing the current value of pixel X at time t can be expressed as a mixture of N Gaussian distributions:

$$P(X_t) = \sum_{i=1}^{N} \omega_{i,t}.\eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \tag{2.1}$$

where N is the number of distributions, $\omega_{i,t}$ a weight of the i^{th} Gaussian at frame t with mean $\mu_{i,t}$ and standard deviation $\Sigma_{i,t}$, and η is the Gaussian probability density function:

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(X_t-\mu)\Sigma^{-1}(X_t-\mu)}. \tag{2.2}$$

The weights and the parameters $\mu_{i,t}$ and $\sigma_{i,t}$ employed by Stauffer et al. [41] are initialized using the K-means algorithm [41]. K determines the multi dimensionality of the background and is set to 4 in BGSLibrary [39]. After foreground detection for the first frame, the parameters are then updated. For detection, all N Gaussians are ordered using the ratio $r_j = \omega_j/\sigma_j$. As one expects the background to happen more often than the foreground for a pixel, the weight should be bigger and the variance should be lower for the background pixels. Hence, using a simple threshold, the B distributions with higher ratio values are considered as the background and the others are considered to be the foreground. Therefore, to label a pixel, the Mahalanobis distance defined as follows is used to find the distance of the pixel to each distribution.

$$sqrt((X_{t+1} - \mu_{i,t})^T . \Sigma_{i,t}^{-1} . (X_{t+1} - \mu_{i,t})) < \tau \sigma_{i,t}. \tag{2.3}$$

Note that τ is a constant threshold. If the nearest distribution to the pixel belongs to the background distributions, the pixel is labeled as background; otherwise it is labeled as foreground. After each labeling the distribution parameters are updated. Although MOG can remove the static background pretty well, it cannot remove the moving shadows. Hence, there is a need to apply an extra step to remove shadows.

2.3 Shadow Removal

Shadow removal is a critical preprocessing step for object detection and tracking. In action recognition, depending on the method, we may need to remove the shadow pixels in order to obtain the silhouette. Although there are many useful methods for subtracting the background, such as the Gaussian mixture models introduced before [40], a major problem of such methods is that shadows tend to be classified as a part of the foreground, because shadows have the same movement patterns and intensity changes as the foreground pixels. Therefore, an extra process is usually needed to remove the shadow. We can classify the features used by shadow removal algorithms into intensity-based, chromaticity-based, physical-based, geometry-based and textural features as Sanin et al. [46] suggested. Methods using intensity features use the fact that regions under shadow are darker because they are blocked from the light source. The chromaticity based methods keep the previous fact and also use the fact that the chromaticity of pixels under shadow does not change. Intensity-based or chromaticity-based features are usually used as the first step by many algorithms such as Hsieh et al. [6].

A very effective and simple chromaticity-based method is provided by Cucchiara et al. [9]. The HSV color space is used in this paper, since it separates the luminosity and chromaticity of the pixels. The *value* (V) is the measure of intensity, which is lower for pixels under the shadow. On the other hand, the *hue* (H) is directly related to chromaticity and does not change in shadow pixels, whereas the authors noted that the *saturation* (S) of pixels is lower under the shadow. Using these facts for a pixel p in a frame F and its corresponding pixel in the background reference image B, if the following three conditions are satisfied, the pixel is labeled as shadow:

$$\beta_1 \leq (F_p^V / B_p^V) \leq \beta_2, \tag{2.4}$$

$$(F_p^S - B_p^S) \leq \tau_S, \tag{2.5}$$

$$|F_p^H - B_p^H| \leq \tau_H. \tag{2.6}$$

Note that $\beta_1 = 0.3$, $\beta_2 = 1$, $\tau_H = 48$, and $\tau_S = 40$ as they are tuned experimentally by the Sanin et al. in their survey [47].

On the other hand, physical approaches try to model or learn the specific appearance of the shadow regions. An example of such methods is given by Huang et al. [48]. The color change of a pixel p, from shadow to background values is represented by $v(p)$. Then the feature vector $x(p) = [\alpha(p), \theta(p), \phi(p)]$ is constructed containing three factors $\alpha(p)$ as the illumination attenuation, and $\theta(p)$ and $\phi(p)$ as the directions of $v(p)$ in spherical coordinates. At first, a basic shadow detection is applied and pixels in the foreground with lower luminosity and different saturation from the background are selected. Then the above features of the primary shadow pixels are used to learn Gaussian Mixture Models for the cast shadows. Finally, an observed pixel is classified by deriving the posterior probabilities of the cast shadow and foreground for the pixel and thresholding these posteriors.

Geometry-based approaches try to locate the shadow with geometric analysis of the shape, such as in the method by Hsieh et al. [6]. At first, the authors extract a set of person-shadow pairs by applying a simple background subtraction technique, followed by a histogram projection method. Then for each individual region following steps are performed. Each pixel is represented by its coordinates as (x, y). Then the center of gravity $(\bar{x}, \bar{y})$, the central moments $(\mu_{p,q})_R$ and the orientation θ of region R are defined as:

$$(\bar{x}, \bar{y}) = (\frac{1}{|R|} \sum_{(x,y) \in R} x, \frac{1}{|R|} \sum_{(x,y) \in R} y), \tag{2.7}$$

$$(\mu_{p,q})_R = \sum_{(x,y) \in R} (x - \bar{x})^p (y - \bar{y})^q, \tag{2.8}$$

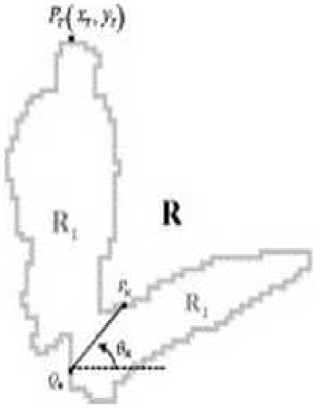


Figure 2.7: an object-shadow pair and its contour information (from [6]).

$$\theta = \frac{1}{2} arctan(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}). \tag{2.9}$$

Note that $|R|$ is the area of the region. The authors use the assumption that the shadow will touch the person's body at the feet and the shadow and the body can be separated using a straight line $Q_R P_R$ as it is shown in figure 2.8. The point P_R is found below the center of gravity, where the maximum vertical change happens. The vertical change is considered between two adjacent points in the region from an object pixel to a none-object pixel or vice versa. Then knowing the orientation θ and P_R the line $Q_R P_R$ is found to roughly separate the shadow from the foreground. In the next step, the candidate shadow pixels are modeled by a Gaussian mixture model using the coordinates of the pixels in addition to the intensity information. If (x, y) are the Cartesian coordinates of a pixel in region R, then the elliptic coordinates of the pixel is defined as follows:

$$\begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} cos\theta_R & -sin\theta_R \\ sin\theta_R & cos\theta_R \end{pmatrix} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}, \tag{2.10}$$

where θ_R is the major orientation of the region R, and μ_x and μ_y are, respectively, the means of x and y coordinates in R. The Gaussian object model is then defined as:

$$G(s, t, g) = exp\left[-\left(\frac{\omega_s s^2}{\sigma_s^2} + \frac{\omega_t t^2}{\sigma_t^2} + \frac{\omega_g (g - \mu_g)^2}{\sigma_g^2} \right) \right]. \tag{2.11}$$

Note that g is the intensity of a pixel, ω_s , ω_t and ω_g are, respectively, the weights of the s coordinate, the t coordinate, and the intensity component, and σ_s , σ_t and σ_g are, respectively, the variances of s, t, and the intensity component. The above model summarizes the intensity and coordinates of the pixels in the region. Finally, for each pixel in the region, it is classified as shadow if it agrees with the model or non-shadow otherwise.

Finally, texture-based approaches assume that regions under shadow keep most of their texture information. Leone and Distante in their paper [49] extract gradient based texture features from the shadow regions. Similar to many other algorithms, a big enough candidate shadow region is first found by applying the method described by Cuchiara et al. [9]. Then using the idea that a shadow preserves texture, they remove the incorrectly classified pixels.

The primary shadow parts are segmented into regions using edge information and the texture correlation between each region in the frame and the background reference is computed. Denote $\bigtriangledown_y$ as the vertical gradient, $\bigtriangledown_x$ as the horizontal gradient, F as the frame and B as the background reference, the difference between the gradient directions of B and F for a pixel p is defined as:

$$\Delta\theta_p = arccos\frac{\bigtriangledown_x^F \bigtriangledown_x^B + \bigtriangledown_y^F \bigtriangledown_y^B}{\sqrt{(\bigtriangledown_x^F 2 + \bigtriangledown_y^F 2)(\bigtriangledown_x^B 2 + \bigtriangledown_y^B 2)}}. \tag{2.12}$$

Then the correlation between the two regions is computed as:

$$c = \frac{\sum_{p=1}^{n} H(\tau_a - \Delta\theta_p)}{n}. \tag{2.13}$$

It is noteworthy that n is the number of pixels in the region, H is a unit step function. When the angular difference is less than or equal to the threshold τ_a, H is equal to 1; otherwise it is equal to zero. Meanwhile, c represents the fraction of pixels in the frame region in which their gradients are similar to the corresponding background. Using another threshold τ on c can lead to a final decision for the region. if c is greater than τ then the region is classified as shadow, otherwise it is labeled as non-shadow.

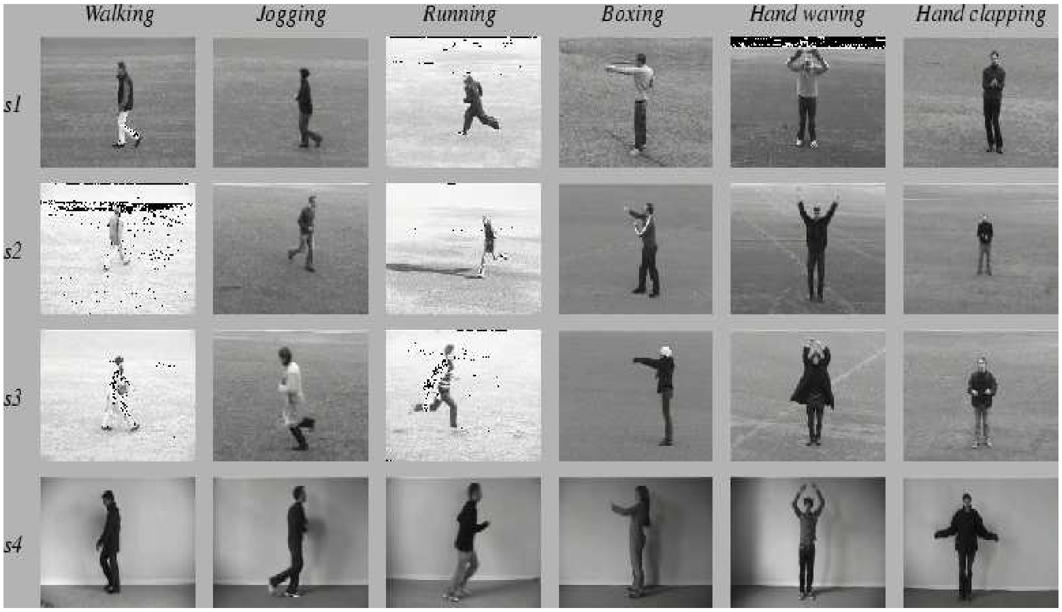


Figure 2.8: KTH action dataset: examples of sequences of different types of actions and scenarios (from [7]).

2.4 Dataset

There are several publicly available datasets for action recognition. They have been used by many people for evaluations and comparisons of different approaches. Here we review some of the most widely used datasets. The KTH human motion dataset [7] contains six types of actions including walking, running, jogging, boxing, hand waving and hand clapping (actions are performed by 25 different actors in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors). Having about 2300 video sequences, there is a considerable variation in the performance and duration of the actions and slight differences in the camera viewpoint. However, the background is simple and homogeneous.

The Weizmann human action dataset [4] contains 90 low-resolution videos including 10 daily actions such as walking, running, skipping, jumping-jack, jumping forward on two legs, jumping in place on two legs, galloping sideways, waving two hands, waving one hand or bending. 9 different actors perform the actions. All the scenarios are outdoors, with a simple and homogeneous background. The foreground silhouettes are also provided in the dataset.

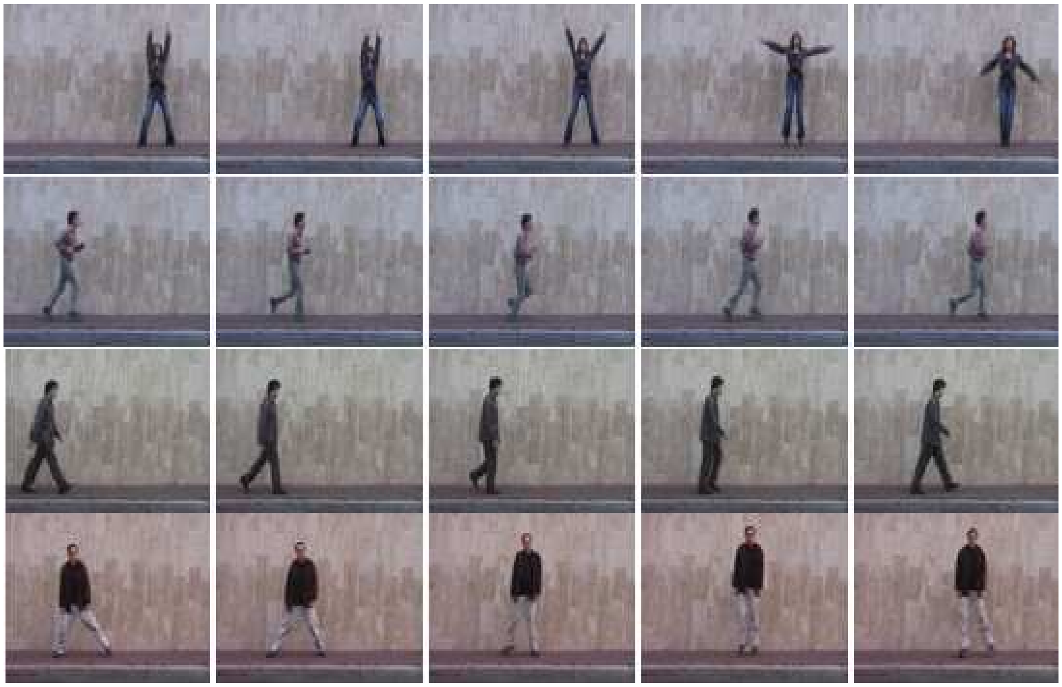


Figure 2.9: Weizmann action dataset: examples of action sequences (from [4]).

The INRIA Xmas Motion Acquisition Sequence (IXMAS) [8] is a multi view dataset from 11 different actors containing 14 actions (checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking up, throwing overhead, throwing from bottom). The background is static and the illumination is controlled. The dataset also contains the foreground silhouette and the reconstructed mesh geometry of the frames.

There are also other action datasets such as the UCF [50] sports action dataset which contains 150 video sequences of sport motions, and the Hollywood human action dataset [21] which contains 8 realistic human actions extracted from movies.

Although existing datasets try to provide natural and realistic human actions, there are many limitations. First, all existing datasets are limited in the number of action classes and may not be suitable to all applications. Therefore, new datasets for particular applications such as the smart condo is much needed. Moreover, the environments in which the data are captured are usually controlled environments that simplify the preprocessing of the data and extracting the human silhouettes. In a more realistic situation, the illumination of a scene may change during video

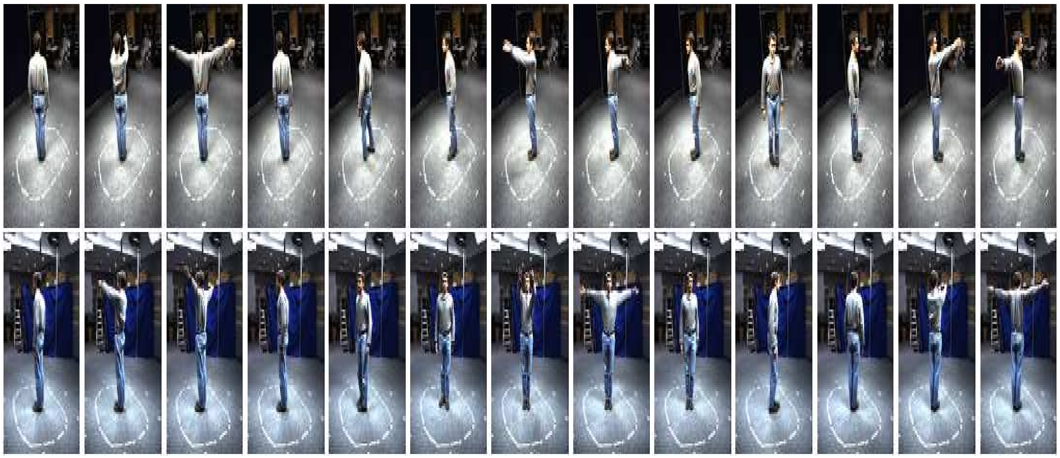


Figure 2.10: INRIA Xmas Motion Acquisition, example action sequence (from [8]).

capturing and the background regions are more complex. Hence, background subtraction is more challenging in such environments. For indoor scenes, we may have shadows, and hence, shadow removal algorithms must be applied in addition to background subtraction. In this thesis, a new indoor dataset for daily actions in a smart condo environment is provided and instead of assuming that preprocessing is already done, some useful techniques for preprocessing of the captured data are proposed in order to extract the human silhouettes. It is anticipated that our preprocessing steps can be applied to other datasets and the results may be used as inputs to any human action recognition algorithm.

Chapter 3

The Proposed Method

3.1 Motivation

As it was mentioned in the previous chapter, if there exist properly preprocessed inputs, global methods show good performances. But in most of the cases the proper input data is achieved manually. Moreover, some datasets have been captured in controlled environments which make preprocessing trivial and less challenging. But in real environments, preprocessing of the data is a challenging task. Particularly, in a scene with normal light sources, there are moving shadows when the light source is blocked by a person. In such situations, shadow removal is needed in addition to background subtraction to localize the person performing an action. Although shadow removal has been studied in tracking and human detection tasks, none of the action recognition techniques has addressed this issue as an inevitable preprocessing step. Moreover, no action dataset includes moving shadows. Our goal is to perform action recognition in a real health care environment such as a smart condo, and we try to address some of the existing challenges. So we perform the whole process of the recognition including collecting action datasets in the smart condo, preprocessing the collected data , and finally applying an existing action recognition method to evaluate the results. Our contributions include a new dataset of human actions in a smart condo environment, and proposing a new shadow removal algorithm. We show that our proposed shadow removal algorithm outperforms some existing methods in the smart condo and similar environments.

3.2 Shadow detection as a labeling problem

The input of the algorithm is the result of applying an existing background subtraction technique on a sequence of action frames. The goal is then to extract the foreground pixels that represents the person's silhouette. In an input image pixels with the value of one are either the foreground or the shadow, and pixels with the value of zero are stationary background. In order to remove the shadow pixels, we propose a new shadow detection technique. Like many other shadow detection algorithms we provide our method as an improvement to a chromaticity-based algorithm. As Sanin et al. [47] show in their survey the chromaticity-based algorithm provided by Chucchiara et al. [9] gets a reasonably good detection rate for indoor scenes which are similar to our application, but the discrimination rate of the algorithm is poor compared to others. It means their method is strong in detecting the real shadow pixels, but it incorrectly labels some foreground pixels as shadow as well. Our goal is to improve the discrimination performance of the algorithm. Like other shadow detection algorithms there are two basic assumptions of our method:

1. The luminosity of a pixel is lower under shadow, and
2. The chromaticity of a pixel does not change under shadow.

Many problems in computer vision have been represented as a labeling problem and solved using energy minimization. For instance Cohen [51] casts the background subtraction into a labeling problem, and uses graph cuts algorithm for the energy minimization. Szeliski et al. in their paper [52] compare the accuracy and computation time of several energy minimization algorithms, such as Iterated Conditional Modes(ICM) [53], graph cuts [54] and Loopy Belief Propagation (LBP) [55]. As they mentioned, such energy minimization problems can be justified in terms of maximum a posteriori probability (MAP) estimation of a Markov Random Field (MRF). The energy function E can be defined as the sum of two terms, $E = E_d + \lambda E_s$. The data term E_d which penalizes values that are inconsistent with observations or assumptions, and the smoothness term E_s which enforces spatial consistency.

The pixel labeling approach has been used in many problems such as background subtraction [51], stereo matching [56], and image segmentation [57]. But no one has applied this approach to shadow detection before. Here motivated by the works of [51] [58] we consider shadow detection as a pixel labeling problem, where the label determines if a pixel is in shadow or not. The label for each pixel is determined by applying the chromaticity-based method, and neighborhood information in the spatial domain. As Szeliski et al. [52] mentioned Boykov et al. [54] introduce two graph cut algorithms called *swap move* and *expansion move* which converge quickly to a strong local minimum. They evaluate both algorithms on different benchmark problems and show that *expansion move* performs better than the other. So we use the *expansion move* graph cuts algorithm to minimize the energy function. The result is a binary image representing shadow pixels with the value one, and other pixels with zero. Finally we subtract the result from the input frame to get the foreground silhouette.

In each frame a pixel p can be expressed in terms of its coordinates as $p(i, j)$. The data term function $D_p(f_p)$ is defined as the cost of assigning label f_p to p. For the smoothness term we consider four neighbors of a pixel in the horizontal and vertical directions, in a way that $p(i, j)$ and $q(s, t)$ are neighbors if $|i-s|+|j-t| = 1$. Then the smoothness term function $V_{pq}(f_p, f_q)$ is defined as the cost of assigning f_p and f_q to the neighbors p and q. We define the energy function E as:

$$E(f_p) = \sum_{p \in P} D_p(f_p) + \sum_{\{p,q\} \in N} V_{p,q}(f_p, f_q). \tag{3.1}$$

Note that P is the set of all pixels p in the frame, N is the set of neighboring pixels, and $\{p, q\}$ stands for an unsorted set of pixels.

3.2.1 Data Term

We define the data term based on the chromaticity-base method provided by Cucchiara et al. [9]. For a pixel p, $X_{p,v} = (F_p^v / B_p^v)$ represents the ratio of the *Value* component of p in the frame to its corresponding pixel in the background. Shadow pixels get the ratio with values between β_1 and β_2. Similarly, $X_{p,s} = (F_p^s - B_p^s)$ represents the difference between the *saturation* of pixel p in the frame and that of its corresponding background pixel. For shadow pixels the value of $X_{p,s}$ is less than the threshold τ_s. In a similar way, $X_{p,h} = |F_p^h - B_p^h|$ denotes the difference between the *hue* of pixel p in the frame and that of the corresponding background. The value of $X_{p,h}$ is less than the threshold τ_h for shadow pixels. We set $\beta_1 = 0.3$, $\beta_2 = 1$, $\tau_h = 48$, and $\tau_s = 40$.

Using these three conditions defined in the paper, the cost function $D_p(f_p)$ is defined as the multiplication of two terms as:

$$D_p(f_p) = G_1(p)^{f_p} \cdot G_2(p)^{(1-f_p)}, \tag{3.2}$$

where $G_1(p)$ is the cost of labeling pixel p as shadow, $G_2(p)$ is the cost of labeling p as none shadow, and are defined as:

$$G_1(p) = \begin{cases} K_{11} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} \leq \tau_s \text{ and } X_{p,h} \leq \tau_h \text{ ,} \\ K_{12} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} > \tau_s \text{ and } X_{p,h} \leq \tau_h \text{ ,} \\ K_{13} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} > \tau_s \text{ and } X_{p,h} > \tau_h \text{ ,} \\ K_{14} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} \leq \tau_s \text{ and } X_{p,h} > \tau_h \text{ ,} \\ K_{15} & \text{else.} \end{cases} \tag{3.3}$$

$$G_2(p) = \begin{cases} K_{21} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} \leq \tau_s \text{ and } X_{p,h} \leq \tau_h \text{ ,} \\ K_{22} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} > \tau_s \text{ and } X_{p,h} \leq \tau_h \text{ ,} \\ K_{23} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} > \tau_s \text{ and } X_{p,h} > \tau_h \text{ ,} \\ K_{24} & \text{if } \beta_1 \leq X_{p,v} \leq \beta_2 \text{ and } X_{p,s} \leq \tau_s \text{ and } X_{p,h} > \tau_h \text{ ,} \\ K_{25} & \text{else.} \end{cases} \tag{3.4}$$

In our experiments we set $K_{11} = 1$, $K_{12} = K_{13} = K_{14} = 3$, $K_{15} = 9$, $K_{21} = 9$, $K_{22} = K_{23} = K_{24} = 7$ and $K_{25} = 1$. These values are obtained by trial-and-error. So the function $D_p(f_p)$ penalizes pixels that do not have the described three conditions. It can be seen that the first condition on the *value* of a pixel is more critical in deciding about the label of that pixel, because shadow keeps the *value* unchanged.

3.2.2 Smoothness Term

In order to remove incorrectly detected shadow pixels the smoothness term is defined. For two neighboring pixels p and q the smoothness term gives a high value if

f_p and f_q do not have the same labels. Therefore, as Szeliski et al. [52] suggested, we define the smoothness term as a function of the difference of the two neighbor's labels:

$$V_{p,q}(f_p, f_q) = w_{pq}.|f_p - f_q|. \tag{3.5}$$

We set w_{pq} the constant value 25, which is found experimentally. The energy minimization of the above energy function is done using the publicly available framework in the C++ programming language by Szeliski et al. [52]. We use the *expansion move* graph cuts algorithm [54] [59] implemented in the framework, which uses the max-flow algorithm provided by Boykov et al. [60]. Later we will show that detecting shadow pixels using chromaticity and neighborhood information improves the results. The next chapter shows the qualitative results of the comparison of our shadow removal algorithm to the chromaticity-based algorithm [9] using different datasets. Also we compare the effect of our algorithm and of other existing shadow removal methods on action recognition.

Chapter 4

Experiments

4.1 Human Action Recognition in a Smart Condo

In this chapter, first the work flow of human action recognition in a smart condo environment is provided. Then a new dataset for action recognition is introduced, and our experimental setup is described. Finally, the proposed shadow removal method is evaluated and compared to other existing methods.

In order to evaluate the proposed shadow removal algorithm quantitatively and to demonstrate its performance in the smart condo environment, we have picked an existing action recognition algorithm proposed by Blank et al. [4] to process the results after shadow removal. The algorithm was explained in the background chapter. It is shown to have a good performance on the Weizmann dataset that contains some common actions with our proposed dataset. The preprocessing steps include background subtraction, shadow removal, noise removal and alignment of the videos. For background subtraction, the Gaussian Mixture Model (GMM) is first applied to a raw action video sequence, from which the moving objects are extracted. Then, in each frame the moving shadow is removed by either the proposed method or one of the existing algorithms for comparison. The flow chart of the whole process is shown in figure 4.1.

4.2 Smart Condo

The smart condo is located in the Edmonton Clinic Health Academy (HCHA). The goal is to provide research opportunities for research around health care facilities

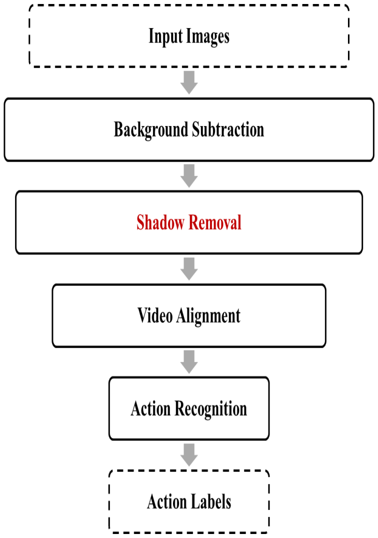


Figure 4.1: The work flow of the action recognition in the smart condo

and in particular "intelligent" home for seniors and patients. The smart condo is equipped with technologies such as wireless sensors and cameras in order to monitor the events. There are four cameras mounted on the ceiling of the living room, two mounted on the ceiling of the kitchen, one in the bedroom and one in the bathroom. Figure 4.2 shows the condo from different views. It can be seen that not all the views are useful for motion analysis. The camera mounted at the end of the living room is used to capture the actions in this thesis, as shown in figure 4.3.

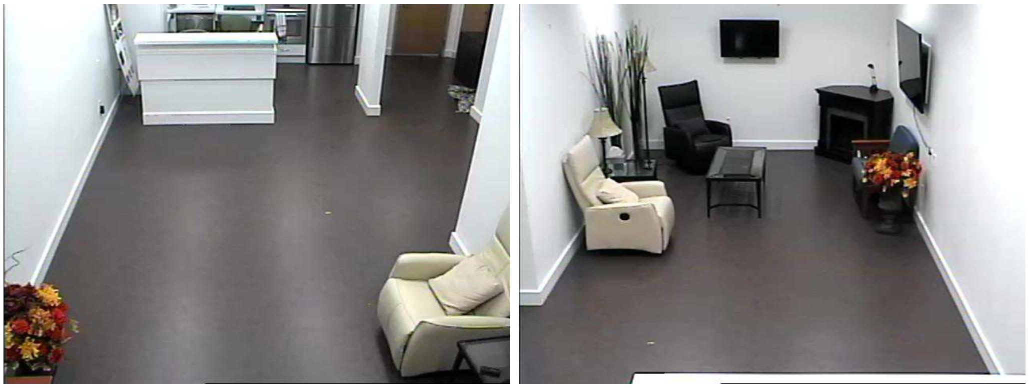


Figure 4.2: The smart condo from different views, captured by mounted cameras.

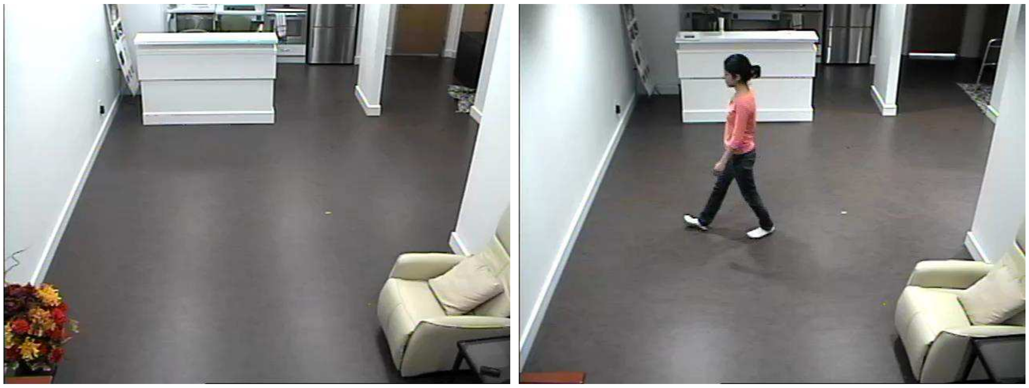


Figure 4.3: A single view is used to capture all the actions.

4.3 Human Action Dataset

We have collected a database of 98 video sequences each with a resolution of 480 × 640 and a frame rate of 30 fps. The video sequences include 10 different actions performed by 7 people. The actions are “bend”, “lie”, “limp”, “run”, “sit”, “stand-up”, “turn”, “walk” and “walker” which means walking with a walker. Figure 4.4 and 4.5 show selected sample frames from the dataset. As it can be seen the actors are of different genders with varying clothing and appearances (overweight, thin, short and tall). Also we have tried to include variations in the performance and the speed of each action. Actions are selected based on simple daily actions people normally perform in their daily lives. Although there are several datasets with focus on health care applications [61] [62], they are restricted to particular uses such as fall detection. None of them considers daily actions of people living in a smart condo environment. Our new dataset will be useful to other researched in future studies on action recognition with focus on health care applications.

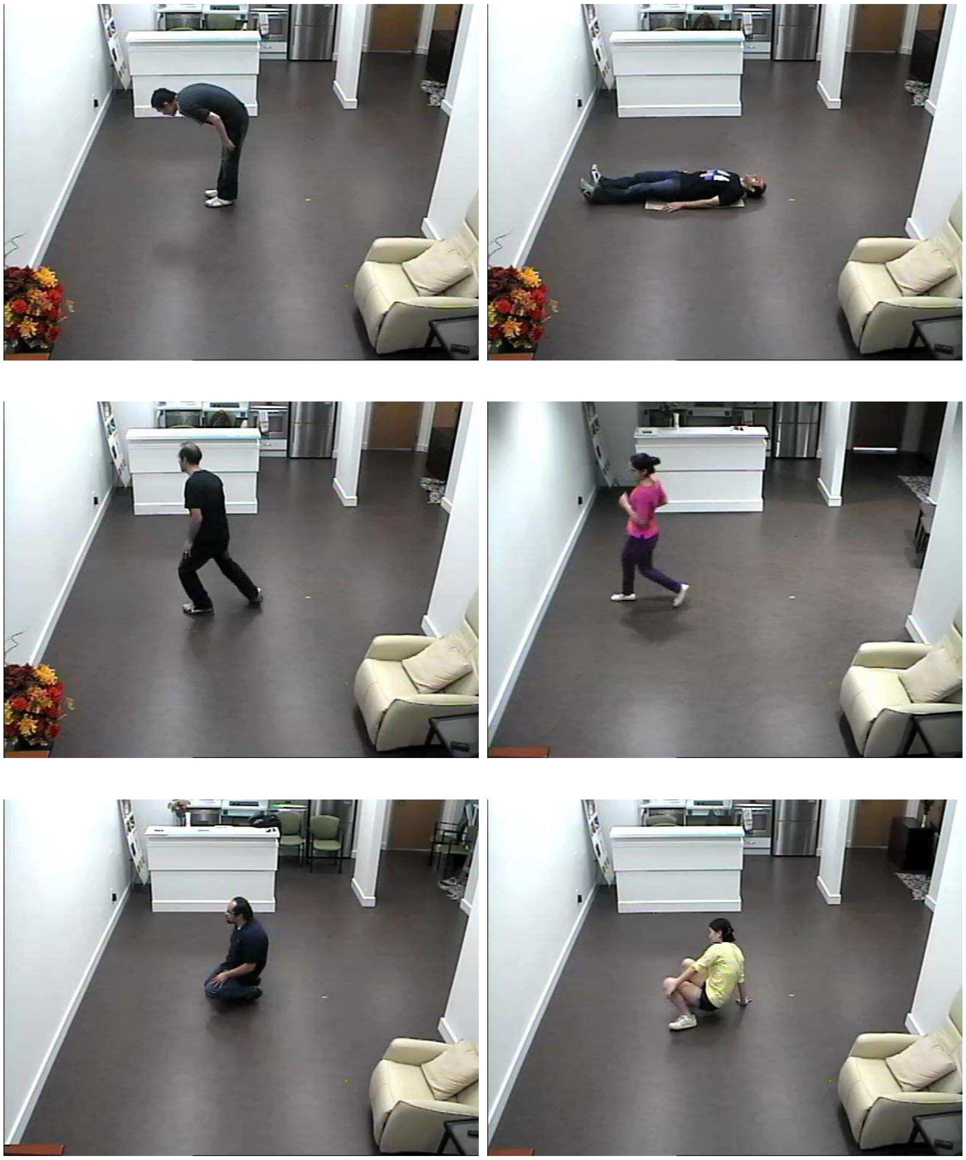


Figure 4.4: The human action dataset. The top row from left to right: bend, lie. The middle row from left to right: limp, run. The bottom row from left to right: sit, stand-up.

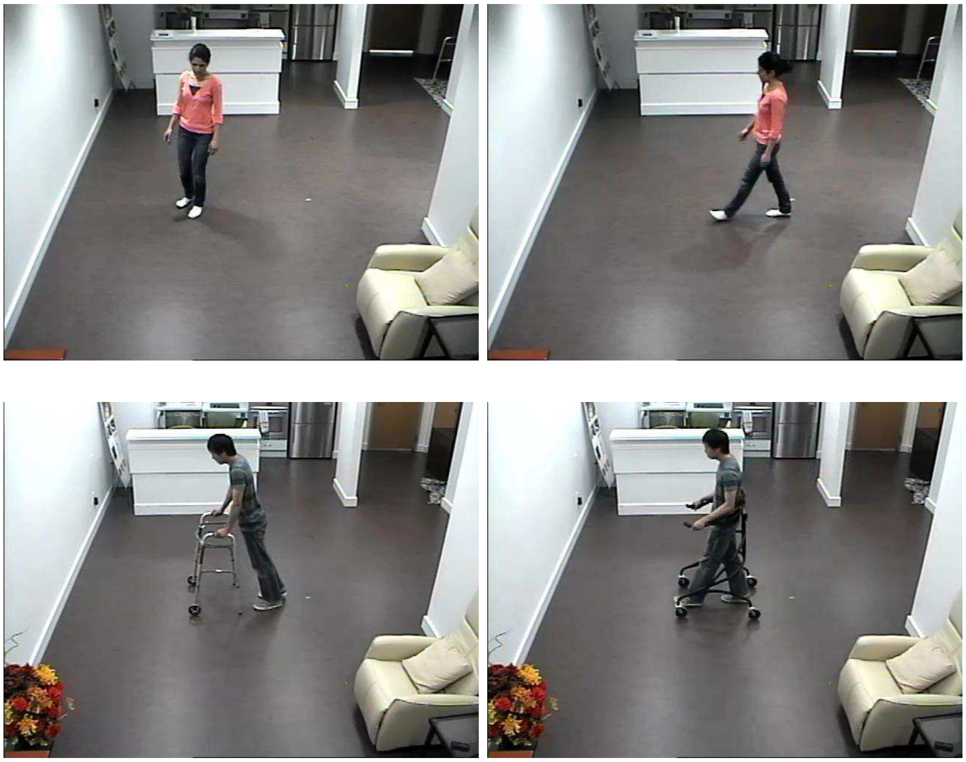


Figure 4.5: The human action dataset. The top row from left to right: turn, walk. The bottom row: walker(1), walker(2).

4.4 Implementation

4.4.1 Mixture of Gaussian Background Subtraction

Sobral and Andrews [39] have provided the BGSLibrary C++ framework for Background Subtraction (BGS). They have implemented a large number of algorithms and compared their results on "Highway Traffic" scenes. As described before, we use the Mixture of Gaussians provided by Staffer et al. [41] which is available in the LBMixtureOfGaussians class of the BGSlibrary package and is adapted from the "Scene" application by Bender et al. [45]. The parameters are set as provided in the package. The "Sensitivity" is set to 81, which describes the sensitivity to changes in the background. The "Learning Rate" is set to 59, which determines the rate at which the model adapts to changes in the video frames. Finally, the "Noise Variance" is set to 206, which specifies the minimum value of the variance in the Gaussian model. The input of the algorithm is the raw human action videos, and the output is the logical images representing the moving foreground. Figure 4.6 shows the results of the BGS algorithm.

4.4.2 Shadow Removal

Sanin et al. [46] have provided the source code for a number of shadow removal algorithms, in addition to the ground truth data for some shadow removal datasets, in order to compare the results. The source code is available in [63]. The input to the following algorithms includes the original frame taken from an action video, the corresponding BGS result, and the background image itself. The output is the moving foreground with the shadow removed.

4.4.3 Chromaticity-based Method

The algorithm proposed by Cucchiara et al. [9] is used here for comparison and as the basis of the proposed method. The parameters are tuned in the work of Sanin et al. [46] as follows: $\beta_1 = 0.3$, $\beta_2 = 1$, $\tau_h = 48$, and $\tau_s = 40$. Figure 4.7 shows the results of the method.

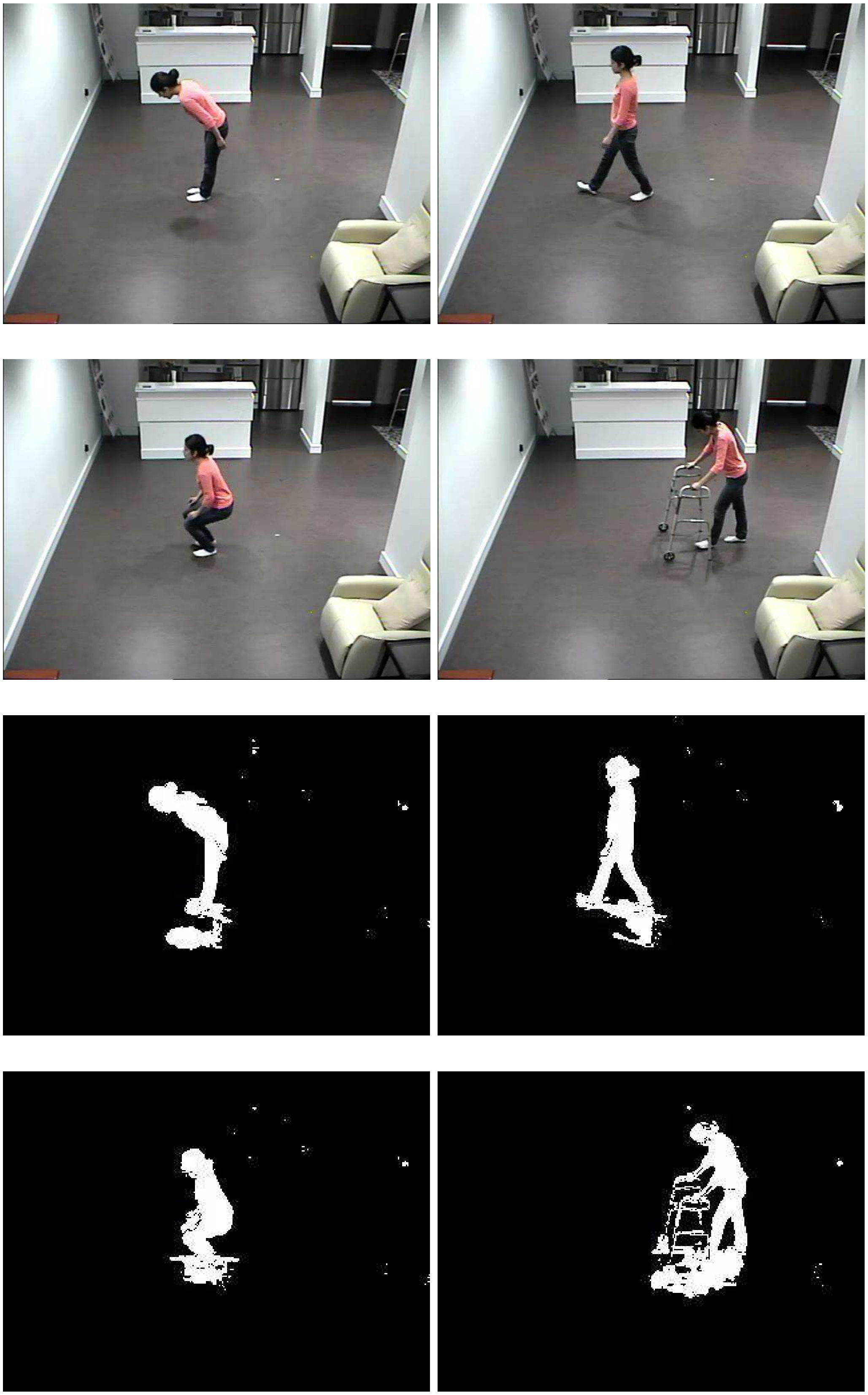


Figure 4.6: Background Subtraction (BGS) results: the first two rows show sample images from the raw videos. The last two rows show the results of BGS.

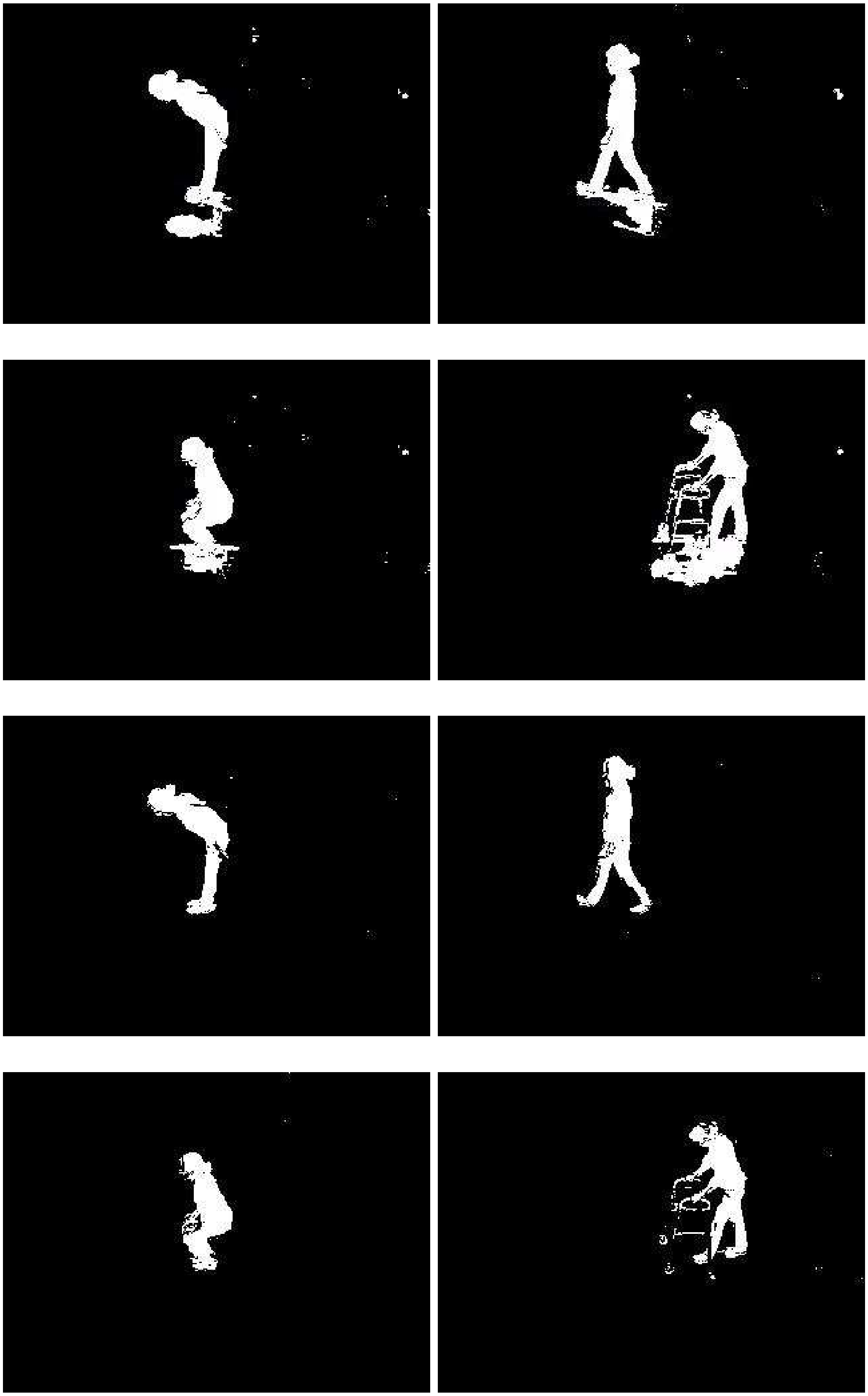


Figure 4.7: Chromaticity-based shadow removal results: the first two rows show the results of BGS. The last two rows show the results of the algorithm proposed by Cucchiara et al. [9].

4.4.4 The Proposed Method

The MRF energy minimization software available in [64] is used for this part. The software that accompanies the paper [52] is provided by Szeliski et al. We use the graph cuts algorithm as the energy minimization method. The energy function is explained in section “The Proposed Method” . The data term is based on the chromaticity-based shadow removal algorithm, while the smoothness function is based on a simple neighborhood information, in order to reduce the false positive labels. The output of the algorithm is shown in figure 4.8.

4.4.5 Other methods

In order to quantitatively compare the results of our method to others in the smart condo application, we have performed action recognition using the output of other shadow removal methods. We have used the physical-based method by Huang et al. [48] , the geometrical-based method by Hsieh et al. [6], and the texture-based method by Leone and Distante [49]. The source code of the above algorithms are provided in [63]. The default values of the parameters are used as they are provided in the source code.

4.4.6 Video Alignment

The frames of each video sequence are aligned in order to compensate for the global motion of the body, and to emphasize on the motion of relative body parts, as suggested by Blank et al. [4]. First, the bounding box of the person is found for each frame. Having the centers of the mass of all the bounding boxes, the best fitting line to them is found, then all the frames are aligned to a reference point.

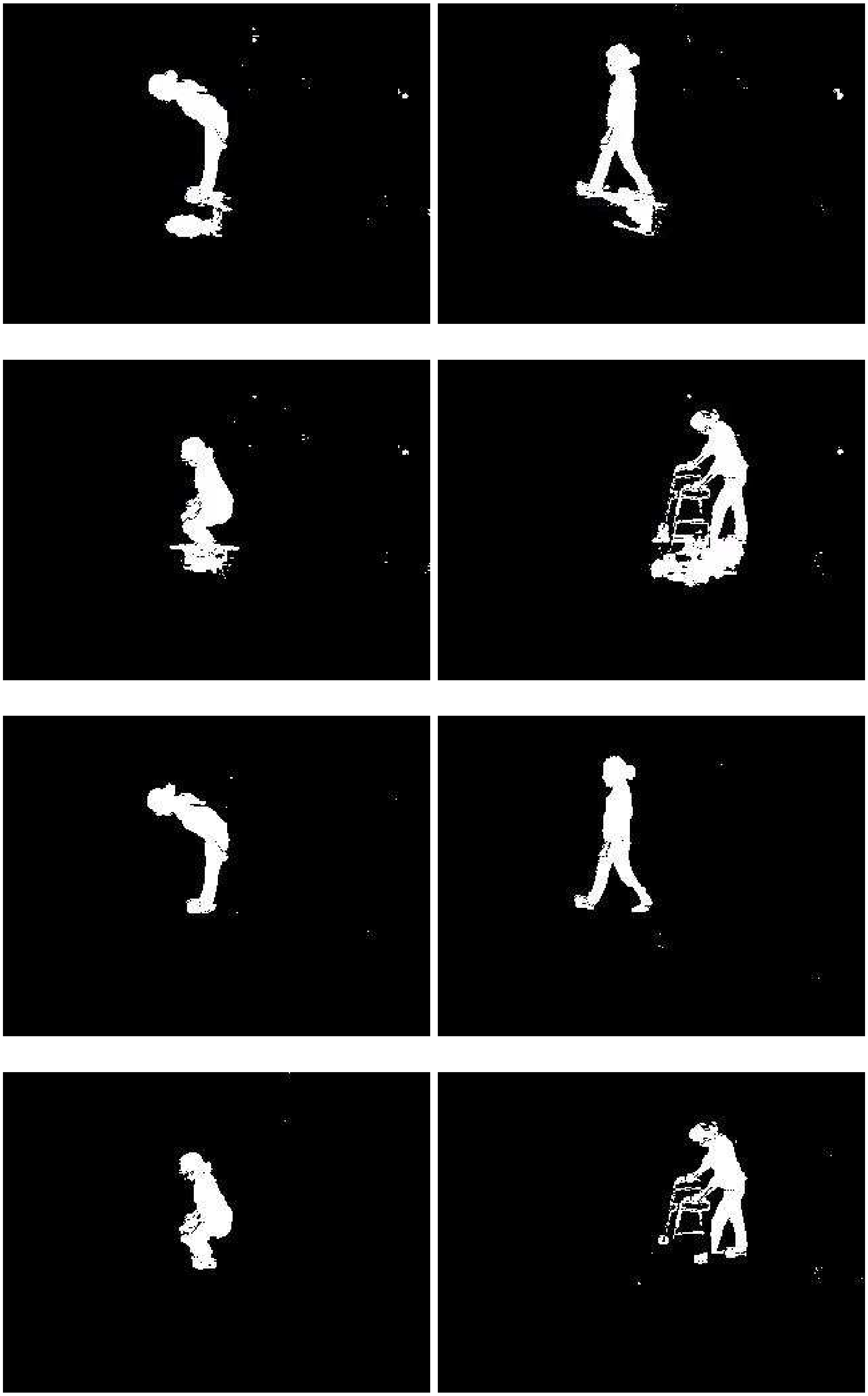


Figure 4.8: MRF shadow removal results: the first two rows show the results of BGS. The last two rows show the results of our shadow removal algorithm.

4.4.7 Action as Space-time Shape

In this thesis, the algorithm proposed by Blank et al. [4] is implemented for action recognition. An action is represented by a 3D space-time shape formed by concatenating all the frames with the background and shadow removed. Then local and global features are extracted from the 3D shape to encode the relative motion of the body to the torso. The algorithm has a good performance on the Weizmann dataset. Because the algorithm concerns the shape of an action, the preprocessing step can affect the results. So in the next section we compare the accuracy of action recognition when different shadow removal algorithms are applied. The input to the method is an aligned sequence of frames for one action, and the output is one of the ten action labels: “bend”, “lie”, “limp”, “run”, “sit”, “stand-up”, “turn”, “walk” and “walker”. All the other parameters are set the same as described in the paper. Also the nearest neighbor classification with Euclidian distance is used for classification.

4.5 Results

4.5.1 Experiment 1: The Recognition Performance on the Smart Condo Action Data

In this experiment, we test the recognition performance on the dataset collected in the smart condo and preprocessed by the proposed shadow removal method. Table 4.1 shows the action confusion results for classification of the videos. The *average accuracy* is 81% which means the average of accuracies for all action labels. On the other hand, we define the *overall accuracy* as the ratio of the number of correctly classified action sequences for all the action labels by the total number of action sequences. The *overall accuracy* is 87% for the recognition algorithm. The best accuracy is obtained for the “crawl” action, because it is quite different from the others. The poorest result belongs to “run” which is often misclassified as “walk”, and “lie” is misclassified as “stand” in many cases, as they share similar body poses in some frames.

	bend	crawl	lie	limp	run	sit	stand	turn	walk	walker
bend	89	1	0	2	0	3	4	0	0	1
crawl	0	100	0	0	0	0	0	0	0	0
lie	0	4	70	0	0	3	23	0	0	0
limp	0	0	0	80	0	1	3	9	1	6
run	0	0	0	0	55	0	4	0	32	9
sit	2	0	2	0	0	89	7	0	0	0
stand	0	0	2	1	0	6	89	0	2	0
turn	1	0	0	9	0	1	3	77	1	8
walk	0	0	0	5	7	0	2	1	71	14
walker	0	0	0	2	0	0	1	1	4	92

Table 4.1: Action confusion in classification using our method

4.5.2 Experiment 2: The Effect of MRF-based Shadow Removal Method in Action Recognition

In this experiment we show the effect of removing shadows from videos in the recognition results. Also the proposed shadow removal algorithm is compared to the state of the art chromaticity-based method quantitatively by comparing their recognition results. Table 4.2 shows the recognition accuracy for each action label when our shadow removal method is used in preprocessing of the data. Table 4.3 shows the accuracies when the chromaticity-based algorithm is used. Finally, table 4.4 shows the results when no shadow removal algorithm is applied to the action sequences.

It can be seen that the action recognition algorithm gives lower performance when the shadow regions are not removed, which suggests the importance of the shadow removal step. On the other hand, the MRF method improves the accuracies for almost all the action labels, and the average accuracy is enhanced by 3%.

Also it can be seen that the proposed method gets low accuracy for the “lie” action, as if no shadow removal is used. A good explanation is that when a person is lying on the ground, the body-shadow pair is seen as a single object with large neighborhood of pixels. So the smoothness term reduce the effect of data term, and therefore, the MRF-based method performs poorly in removing the shadow pixels compared to chromaticity-based method.

	bend	crawl	lie	limp	run	sit	stand	turn	walk	walker	average accuracy	overall accuracy
Accuracy (%)	89	100	70	80	55	89	89	77	71	92	81	87

Table 4.2: Action classification accuracies using our shadow removal method

	bend	crawl	lie	limp	run	sit	stand	turn	walk	walker	average accuracy	overall accuracy
Accuracy (%)	86	100	84	66	53	85	89	66	60	88	78	82

Table 4.3: Action classification accuracies using the chromaticity-based shadow removal method.

	bend	crawl	lie	limp	run	sit	stand	turn	walk	walker	average accuracy	overall accuracy
Accuracy (%)	86	99	69	48	38	77	84	70	60	88	72	78

Table 4.4: Action classification accuracies without a shadow removal.

4.5.3 Experiment 3: Comparing Other Shadow Removal Methods with the MRF-based Method in Action Recognition

We tried other shadow removal methods on our dataset, and compared their results. The parameters of each algorithm are set to their default values as described in the source code [63]. The physical-based method introduced by Huang et al. [48] shows the best result qualitatively. Figure 4.9 shows the sample qualitative results of other shadow removal methods.

In another experiment we compared the recognition results of the MRF algorithm output with the physical-based method output, as physical-based algorithm performs better than texture-based and geometrical-based methods on the smart condo dataset. Table 4.5 shows the recognition accuracy results by using each of them. It can be seen that the proposed MRF shadow removal algorithm is the best among them. Another interesting conclusion is that the chromaticity-based method gives better recognition accuracy than other methods except the MRF shadow removal method.

	bend	crawl	lie	limp	run	sit	stand	turn	walk	walker	average accuracy	overall accuracy
MRF Shadow Removal	89	100	70	80	55	89	89	77	71	92	81	87
Chromaticity-based Shadow Removal	86	100	84	66	53	85	89	66	60	88	78	82
Physical-based Shadow Removal	83	100	85	37	32	84	76	48	36	98	67	76

Table 4.5: Action classification accuracies using each one of MRF-based, physical-based, and chromaticity-based methods.

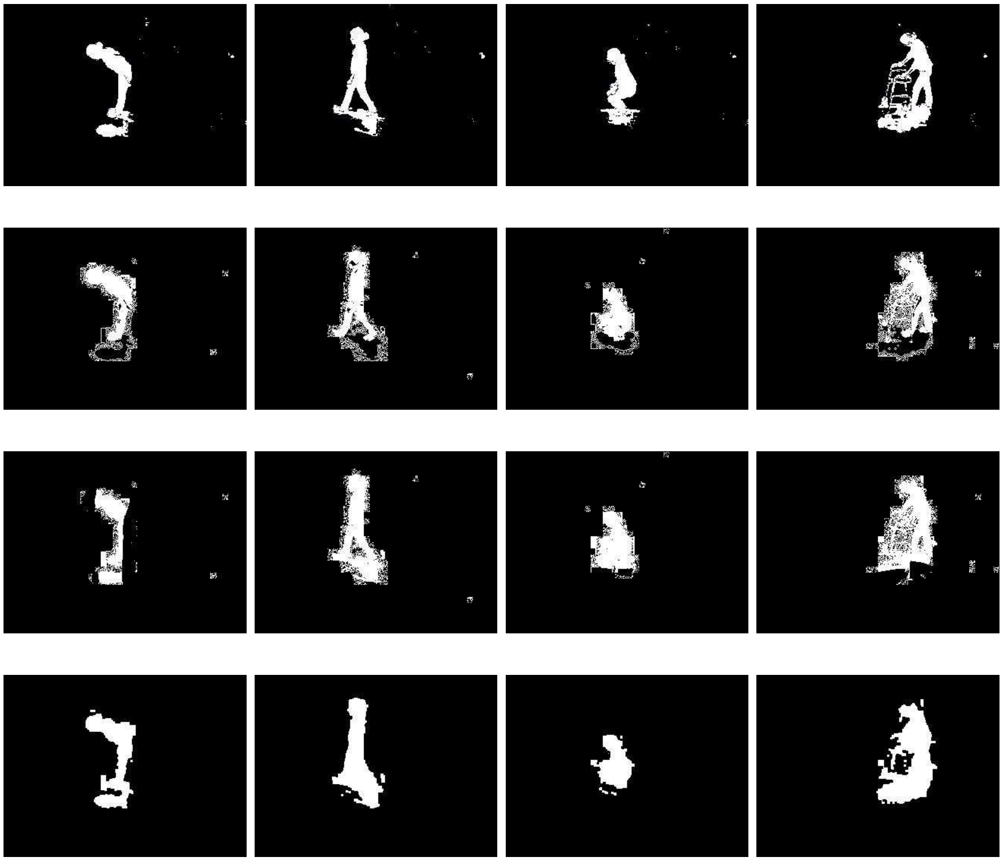


Figure 4.9: Shadow removal results after applying different methods: the first row shows the results of BGS. The second row shows the results of physical-based shadow removal, the third row shows the results of geometrical-based Shadow removal, and the fourth row shows the results of texture-based shadow removal.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In this thesis we consider action recognition with a focus on health-care applications. We first introduce a new human action dataset captured in a smart condo, then we propose a new shadow removal method in order to improve the localization of human body in videos and provide the proper inputs for motion analysis methods.

5.1.1 Human Action Dataset

Although several human action datasets are available for research studies, none of them considers daily actions in a smart condo environment. In recent years, many studies have aimed to improve health care services by automating the remote monitoring of patients and of elderly people in health care environments. The goals in mind include injury prevention, recognition of unusual activities or motion analysis with the aim of improving facility design.

The smart condo facility at the University of Alberta is designed for interdisciplinary research projects related to computer science, computer engineering and rehabilitations. There are several projects by different groups taking place in the smart condo. We collect a simple dataset from daily actions which may occur in a health care environment. This is also useful for future research on action recognition and motion analysis. The dataset contains ten different actions performed by seven different people. We have tried to add enough variations in the action performances such as in gender, clothing and pace. A significant innovation of the dataset is that it includes moving shadows produced by the action performers. None of the action datasets has the shadow problem, although shadow naturally exists in indoor or outdoor videos. We then investigate the shadow removal problem for the action recognition task.

5.1.2 Shadow Removal

The novelty of our shadow removal method is that it considers the neighborhood information to enhance the results of an existing shadow removal algorithm [9]. Based on this approach, the probability of a pixel to be labeled as a shadow is higher when its neighbors are shadows as well. So the problem transforms into an energy minimization problem in a Markov Random Field framework with the goal of minimizing the sum of the costs in an image. We use the chromaticity-based algorithm by Cucchiara et al. [9] for the data term of the energy function, because the algorithm performed better than the other methods on our dataset. Depending on the application the MRF approach may be used with other approaches to contain more information about the pixels such as texture, geometrical or physical information.

We compared the proposed method with other shadow removal algorithms by their effect on the action recognition task. We picked an existing action recognition method by Blank et al. [4], then we prepared its inputs by applying the background subtraction [41] following the shadow removal on the action videos. The videos are divided into a number of segments using sliding windows in time. Each video segment is then classified and labeled as one of the ten available actions. The effect of shadow removal and specifically the advantage of our method is then described.

5.2 Future Works

The dataset can be extended by adding actions or variations in the performances. For instance, in our experiment a single camera is used in the condo, but in order to capture the actions from different views with different background objects, extra cameras can be employed. Moreover, in order to include more real features to the dataset, occlusions or interactions with other objects maybe added, such as “pushing a chair” or “picking up an item from the floor”, etc. In ideal cases, we may use real patients or elderly people in our experiments, as their motion and action performances are not easy to imitate. Also another challenge to consider in future may be the issue of varying illumination in a single action capture, as it may happen in real situations. The challenge is then to improve the background subtraction and shadow removal processes to address this issue.

In the shadow removal part, we have defined a simple smoothness function. Defining a more complex function to include more neighborhood information may improve the results. The data term can be also improved to include more features such as texture information, in addition to color and neighborhood information. In the experimental part, the performance of algorithm may be evaluated by applying different action recognition methods or even by methods for different applications such as object detection, tracking, gait analysis, fall detection, etc. and in different environments, such as outdoor scenes.

Bibliography

[1] I. Laptev, “On space-time interest points,” *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[2] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[3] S. Ali and M. Shah, “Human action recognition in videos using kinematic features and multiple instance learning,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288–303, 2010.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1395–1402, IEEE, 2005.

[5] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez, “View-independent action recognition from temporal self-similarities,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 172–185, 2011.

[6] J.-W. Hsieh, W.-F. Hu, C.-J. Chang, and Y.-S. Chen, “Shadow elimination for effective moving object detection by gaussian shadow modeling,” *Image and Vision Computing*, vol. 21, no. 6, pp. 505–516, 2003.

[7] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 32–36, IEEE, 2004.

[8] D. Weinland, R. Ronfard, and E. Boyer, “Automatic discovery of action taxonomies from multiple views,” in *Computer Vision and Pattern Recognition,*

2006 IEEE Computer Society Conference on, vol. 2, pp. 1639–1645, IEEE, 2006.

[9] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1337–1342, 2003.

[10] V. Osmani, S. Balasubramaniam, and D. Botvich, "Human activity recognition in pervasive health-care: Supporting efficient remote collaboration," *Journal of network and computer applications*, vol. 31, no. 4, pp. 628–655, 2008.

[11] S. N. Robinovitch, F. Feldman, Y. Yang, R. Schonnop, P. M. Lueng, T. Sarraf, J. Sims-Gould, and M. Loughin, "Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study," *The Lancet*, 2012.

[12] F. Liu and R. W. Picard, "Finding periodicity in space and time," in *Computer Vision, 1998. Sixth International Conference on*, pp. 376–383, IEEE, 1998.

[13] J. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," in *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pp. 90–102, IEEE, 1997.

[15] G. Johansson, "Visual motion perception.," *Scientific American*, 1975.

[16] H. J. Seo and P. Milanfar, "Detection of human actions from a single example," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1965–1970, IEEE, 2009.

[17] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[18] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65–72, IEEE, 2005.

[19] C. Harris and M. Stephens, “A combined corner and edge detector.,” in *Alvey vision conference*, vol. 15, p. 50, Manchester, UK, 1988.

[20] W. Förstner and E. Gülch, “A fast operator for detection and precise location of distinct points, corners and centres of circular features,” in *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, pp. 281–305, 1987.

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[22] G. Willems, T. Tuytelaars, and L. Van Gool, “An efficient dense and scale-invariant spatio-temporal interest point detector,” in *Computer Vision–ECCV 2008*, pp. 650–663, Springer, 2008.

[23] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006*, pp. 404–417, Springer, 2006.

[24] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” in *Proceedings of the 15th international conference on Multimedia*, pp. 357–360, ACM, 2007.

[25] S. Park and M. M. Trivedi, “Understanding human interactions with track and body synergies (tbs) captured from multiple views,” *Computer Vision and Image Understanding*, vol. 111, no. 1, pp. 2–20, 2008.

[26] J. Little and J. Boyd, “Recognizing people by their gait: the shape of motion,” 1998.

[27] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," in *Computer Vision, 1998. Sixth International Conference on*, pp. 120–127, IEEE, 1998.

[28] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[29] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–123, IEEE, 2001.

[30] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt, "Shape representation and classification using the poisson equation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 1991–2005, 2006.

[31] E. Rivlin, S. J. Dickinson, and A. Rosenfeld, "Recognition by functional parts [function-based object recognition]," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pp. 267–274, IEEE, 1994.

[32] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 144–149, IEEE, 2005.

[33] A. Yilma and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 150–157, IEEE, 2005.

[34] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pp. 3–8, IEEE, 2002.

[35] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov

model," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 955–960, IEEE, 2005.

[36] Y. A. Ivanov and A. F. Bobick, "Recognition of visual activities and interactions by stochastic parsing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 852–872, 2000.

[37] D. Minnen, I. Essa, and T. Starner, "Expectation grammars: Leveraging high-level expectations for activity recognition," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, pp. II–626, IEEE, 2003.

[38] R. Jain and H.-H. Nagel, "On the analysis of accumulative difference pictures from image sequences of real world scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 2, pp. 206–214, 1979.

[39] A. Sobral, "BGSLibrary: An opencv c++ background subtraction library," in *IX Workshop de Viso Computacional (WVC'2013)*, (Rio de Janeiro, Brazil), Jun 2013.

[40] T. Bouwmans, F. El Baf, B. Vachon, *et al.*, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219–237, 2008.

[41] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, IEEE, 1999.

[42] B. Han and X. Lin, "Update the gmms via adaptive kalman filtering," in *Visual Communications and Image Processing 2005*, pp. 59604F–59604F, International Society for Optics and Photonics, 2005.

[43] H. Yang, Y. Tan, J. Tian, and M. Liu, "Accurate dynamic scene model for moving object detection," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 6, pp. VI–157, IEEE, 2007.

[44] W. Zhang, X. Fang, X. Yang, and Q. J. Wu, “Spatiotemporal gaussian mixture model to detect moving objects in dynamic scenes,” *Journal of Electronic Imaging*, vol. 16, no. 2, pp. 023013–023013, 2007.

[45] L. Bender and I. Guerra, “Scene: An open source multiplatform computer vision framework.” `scene.sourceforge.com`, 2011.

[46] A. Sanin, C. Sanderson, and B. C. Lovell, “Improved shadow removal for robust person tracking in surveillance scenarios,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 141–144, IEEE, 2010.

[47] A. Sanin, C. Sanderson, and B. C. Lovell, “Shadow detection: A survey and comparative evaluation of recent methods,” *Pattern Recognition*, vol. 45, no. 4, pp. 1684 – 1695, 2012.

[48] J.-B. Huang and C.-S. Chen, “Moving cast shadow detection using physics-based features,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2310–2317, IEEE, 2009.

[49] A. Leone and C. Distante, “Shadow detection for moving objects based on texture analysis,” *Pattern Recognition*, vol. 40, no. 4, pp. 1222–1233, 2007.

[50] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action mach a spatio-temporal maximum average correlation height filter for action recognition.,” in *CVPR*, IEEE Computer Society, 2008.

[51] S. Cohen, “Background estimation as a labeling problem,” in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1034–1041, IEEE, 2005.

[52] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, “A comparative study of energy minimization methods for markov random fields with smoothness-based priors,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 6, pp. 1068–1080, 2008.

[53] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.

[54] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[55] J. S. Yedidia, W. T. Freeman, Y. Weiss, *et al.*, "Generalized belief propagation," in *NIPS*, vol. 13, pp. 689–695, 2000.

[56] S. T. Barnard, "Stochastic stereo matching over scale," *International Journal of Computer Vision*, vol. 3, no. 1, pp. 17–32, 1989.

[57] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[58] M. Granados, H.-P. Seidel, and H. Lensch, "Background estimation from non-time sequence images," in *Proceedings of graphics interface 2008*, pp. 33–40, Canadian Information Processing Society, 2008.

[59] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 147–159, 2004.

[60] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[61] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic summarization of video for fall detection using voxel person and fuzzy logic," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 80–89, 2009.

[62] Z. Zhang, W. Liu, V. Metsis, and V. Athitsos, “A viewpoint-independent statistical method for fall detection,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3626–3630, IEEE, 2012.

[63] A. Sanin, C. Sanderson, and B. Lovell, “C++ source code and grouth truth for shadow detection / removal.” `http://arma.sourceforge.net/shadows/`, 2012.

[64] `http://vision.middlebury.edu/MRF/code/`, 2001.