Human Activity Recognition

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Engineering

in

Computer Science & Engineering

by

Charu Arora

Examination Roll No.: M4CSE19023
Class Roll No.: 001710502016
Registration No.: 140755 of 2017-18

Under the Guidance of

Dr. Sanjoy Kumar Saha
Professor
Department of Computer Science and Engineering
Faculty Council of Engineering and Technology
JADAVPUR UNIVERSITY
Kolkata - 700032.

May 2019

Declaration of Authorship

I, Charu Arora, declare that this thesis titled, "Human Activity Recognition" and the work presented in it are my own. I confirm that this work was done wholly or mainly while in candidature for a Master degree in Computer Science and Engineering at this University

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Examination Roll Number: M4CSE19023

Class Roll Number: 01710502016

Registration Number: 140755 of 2017-18

Signature :

Date :

Faculty Council of Engineering and Technology

JADAVPUR UNIVERSITY, KOLKATA 700032

Certificate of Recommendation

This is to certify that the thesis entitled **"Human Activity Recognition"** is a bonafide record of work carried out by Charu Arora, Examination Roll No.: M4CSE19023, University Registration No.: 140755 of 2017-2018 in partial fulfillment of the requirements for the award of the degree of Master of Engineering in Computer Science and Engineering from the Department of Computer Science and Engineering, Jadavpur University for the academic session 2017-2019. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Dr. Sanjoy Kumar Saha
Professor (Thesis Supervisor)
Dept of Computer Science and Engineering
Jadavpur University
Kolkata - 700 032

Prof.(Dr.) Mahantapas Kundu(HOD)
Dept. of Computer Science and Engineering,
Jadavpur University,
Kolkata – 700 032

Prof.(Dr.) Chiranjib Bhattacharjee(Dean)
Faculty Council of Engineering and Technology,
Jadavpur University,
Kolkata – 700 032

Faculty Council of Engineering and Technology

JADAVPUR UNIVERSITY, KOLKATA 700032

Certificate of Approval

The foregoing thesis is hereby approved as a creditable study of Master of Engineering in Computer Science Engineering and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion therein but approve this thesis only for the purpose for which it is submitted.

_____ Date:

(Signature of the examiner)

_____ Date:

(Signature of the examiner)

Abstract

Activity recognition pertains to recognize meaningful expressions of motion by a human, involving the hands, arms, face, head, and/or body. It involves finding out activities performed by humans in an image or a video file like a man jumping, a girl playing a musical instrument etc. The motive of the project is to successfully label all the mainstream gestures. One core problem behind these applications is automatically recognizing low-level actions and high-level activities of interest.It has uses in applications like human surveillance whereby looking at the action of a person doing that activity we can take necessary measures, etc. It also has its applications in entertainment environments and healthcare systems. We propose an action recognition scheme based on motion and appearance. Firstly, we subtracted frame by frame so as to get the mask of the moving part. Then we calculated the vector of the centroid and used it's mean and standard deviation as features. Classification was done using SVM. The analysis was done on KTH and Weizmann data set.

Keywords: Activity Recognition, Computer Vision, Optical Flow

Acknowledgements

First of all I would like to express my earnest gratitude and regards towards Dr. Sanjoy Kumar Saha, who gave me the opportunity to work under his guidance and his invaluable advice and encouragement throughout my course of studying M.E in Computer Science and Engineering and my final year thesis.

Also, I thank Prof Mahantapas Kundu, Head, Department Of Computer Science and Engineering, for his assistance in allowing me to work in the departmental laboratory without which my work would have been incomplete.

In addition I would like to thank my friends, especially Dibyadip Chatterjee ,without whose relentless help and encouragement I would really found it difficult in completing my M.E thesis. I would like to thank all the teaching and non-teaching staffs who helped till the end.

I have always found my parents by my side, whenever I have faced any difficulties in my life. I will be in debt life long with them. No words for them to express gratitude, love and respect towards them. In the end I would like to express gratitude to whom whoever helped in my work..

Contents

List of Figures

Chapter 1

Introduction

Human Activity Recognition (HAR) deals with the problem of predicting what a person is doing based on a trace of their movement. HAR can be seen as a general machine learning problem which comprises of feature selection and feature extraction. An action can be viewed as movement of human body in a sequential fashion concurrently. In the context of computer vision, action recognition means to relate the observation(in this case video) with some pre-defined patterns and then accurately assign a label to it based on the action type. Depending on complexity, human activities can be categorized into four levels: gestures, actions, interactions and group activities [1], and much research follows a bottom-up construction of human activity recognition.

The activity recognition system consists of number of major steps like segmentation of region of interest, feature extraction, action learning and classification [2]. In simple terms, the process consists of three steps, namely detection of human and/or its body parts, tracking, and then recognition using the tracking results. For instance, to recognize "shaking hands" activities, two person's arms and hands are first detected and tracked to generate a spatial-temporal description of their movement. This description when compared with existing patterns in the training data determines the action type.

There are several levels to understand the video:

- Object Level Understanding to identify the locations of objects and persons
- Tracking Level Understanding which correspond to object trajectories.
- Pose Level Understanding to recognize human body parts and finally,
- Activity Level Understanding to detect human activities and events.

Human activity refers to collection of human/object movements with a particular semantic meaning while recognition of that activity is the search for segments in that video that display the properties of the movements.

Human activity can be categorized as follows.

- Atomic movements which are gestures.
- Actions where there is just a single actor.
- Interaction which refer to human-object interactions or multiple human interactions.
- Group activities that correspond to physical/conceptual group

Action representation can be categorized as: tracking based approaches [6], spatio-temporal shape template based approaches [4, 5], flow based approaches [3], and interest points based approaches [7]. Spatio-temporal shape template based approaches treat the action recognition problem as a 3D object recognition problem and extracts features from the 3D volume. The extracted features are very huge so the computational cost is unacceptable for real-time applications. In flow based approaches optical flow computation is used to describe motion, it is sensitive to noise and cannot reveal the true motions. Tracking based approaches suffer from the same problems. Interest points based approaches have the advantage of short feature vectors; hence low computational cost.

Movements often include normal indoor activities such as standing, sitting, jumping etc. This has been an active area of research in computer vision applications. The goal of activity recognition is an automated analysis of ongoing events and their context from video data. It's applications include surveillance systems, patient monitoring systems, and a variety of systems that involve interactions between persons and electronic devices such as human-computer interfaces. In a surveillance environment, the automatic detection of abnormal activities can be used to alert the related authority of potential criminal or dangerous behaviours, such as automatic reporting of a person with a bag loitering at an airport or station. Similarly, in an entertainment environment, the activity recognition can improve the human computer interaction (HCI), such as the automatic recognition of different player's actions during a tennis game so as to create an avatar in the computer to play tennis for the player. Furthermore, in a healthcare system, the activity recognition can help the rehabilitation of patients, such as the automatic recognition of patient's action to facilitate the rehabilitation processes. There have been numerous research efforts reported for various applications based on human activity recognition, more specifically, home abnormal activity [8], ballet activity [9], tennis activity [10, 11] , soccer activity [12], human gestures [13], sport activity [14][15], human interaction [16], pedestrian traffic [17] and simple actions [18, 19], and healthcare applications.[8, 20]

1.1 Overall structure of a human activity recognition system

Generally speaking, human activity recognition can be separated into three levels of representations, individually the low-level core technology, the mid-level human activity recognition systems and the high-level applications as shown in Figure 1. In the first

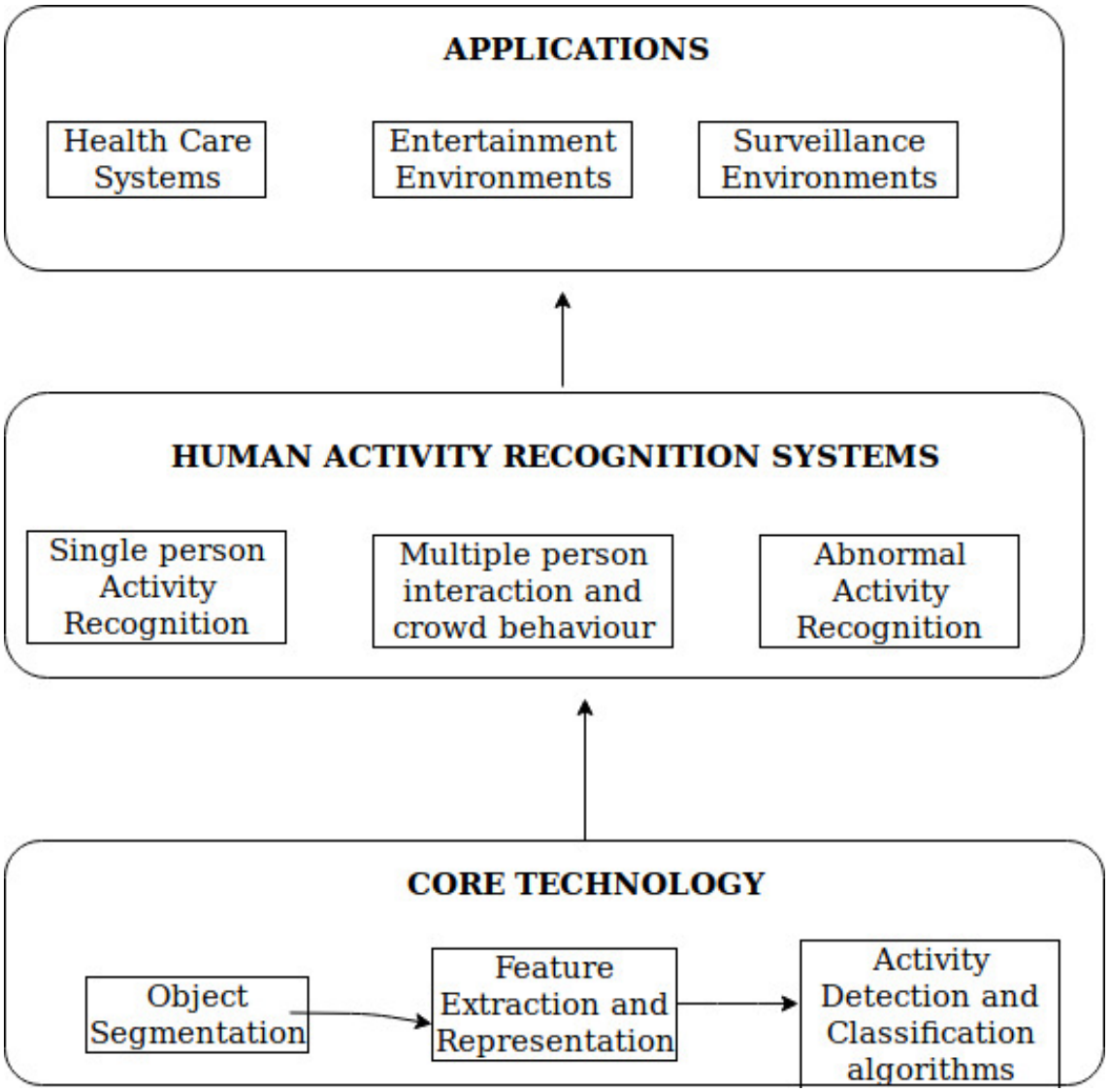


FIGURE 1.1: The overview of a general system for human activity recognition.

level of core technology, three main processing stages are considered, i.e., object segmentation, feature extraction and representation, and activity detection and classification algorithms. The human object is first segmented out from the video sequence. The characteristics of the human object such as shape, silhouette, colors, poses, and body motions are then properly extracted and represented by a set of features. Subsequently, an activity detection or classification algorithm is applied on the extracted features to recognize the various human activities. Moreover, in the second level of human activity recognition systems, three important recognition systems are discussed including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. Finally, the third level of applications discusses the recognized results applied in surveillance environments, entertainment environments or healthcare systems. This work deals with activities considering a single person performing some action and our goal is to accurately identify that task. In this work, first is the object level classification where we detect the object in our case the human body. This is the first level in the recognition to fix our object/objects of interest done by cascading the human body and segmentation to avoid unnecessary noise in the video. With segmentation we

are able to separate the image into background and foreground and the movement of human can be interpreted as the movement of the human and then it can be tracked. The next level is tracking which is done by following the motion pattern between consecutive frames by means of optical flow of the moving object. Post tracking, the features are formed by plotting histograms of magnitude and direction obtained from optical flow . The histograms of magnitude are taken in the range of 0 and 1 while that of direction in the range 0 to 180 . The entire video at frame level is divided into several grids and taking the mean and standard deviation of 3*3 grids gives the most optimal result.

1.2 Datasets

In this section we discuss and describe datasets in use since 2009. Datasets that have been utilized earlier than 2009 can be found in [1] in more detail. We focus on new datasets collected and we further analyze and compare them across several aspects.

1.2.1 The KTH Dataset

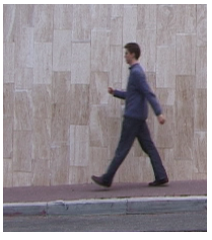
The current database covers six actions – walking, jogging, running, boxing, hand waving and hand clapping performed several times by 25 subjects in four different scenarios outdoors, outdoors with scale variation, outdoors with different clothes and indoors. It contains a total of 2391 sequences. All sequences are taken with a static camera with 25fps frame rate, down sampled to the spatial resolution of 160x120 pixels. In the original paper [21], sequences were divided into a training set (eight persons), a validation set (eight persons) and a test set (nine persons). The dataset does not provide background models and extracted silhouettes.

1.2.2 The Weizmann Dataset

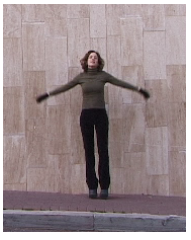
The database covers 10 natural actions – running, walking, skipping, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping sideways, waving-two-hands, waving one- hand and bending performed by nine subjects [25]. It contains a total of 93 sequences. All sequences are 3taken with a static camera with 25fps frame rate, down sampled to the spatial resolution of 180x144 pixels. The dataset also has ten additional sequences of walking captured from a different viewpoint varying between 0 and 81 relative to the image plane. The extracted masks after background subtraction and background sequences are provided.

Sequences from Weizmann Dataset

Running



Walking



Jumping Jack

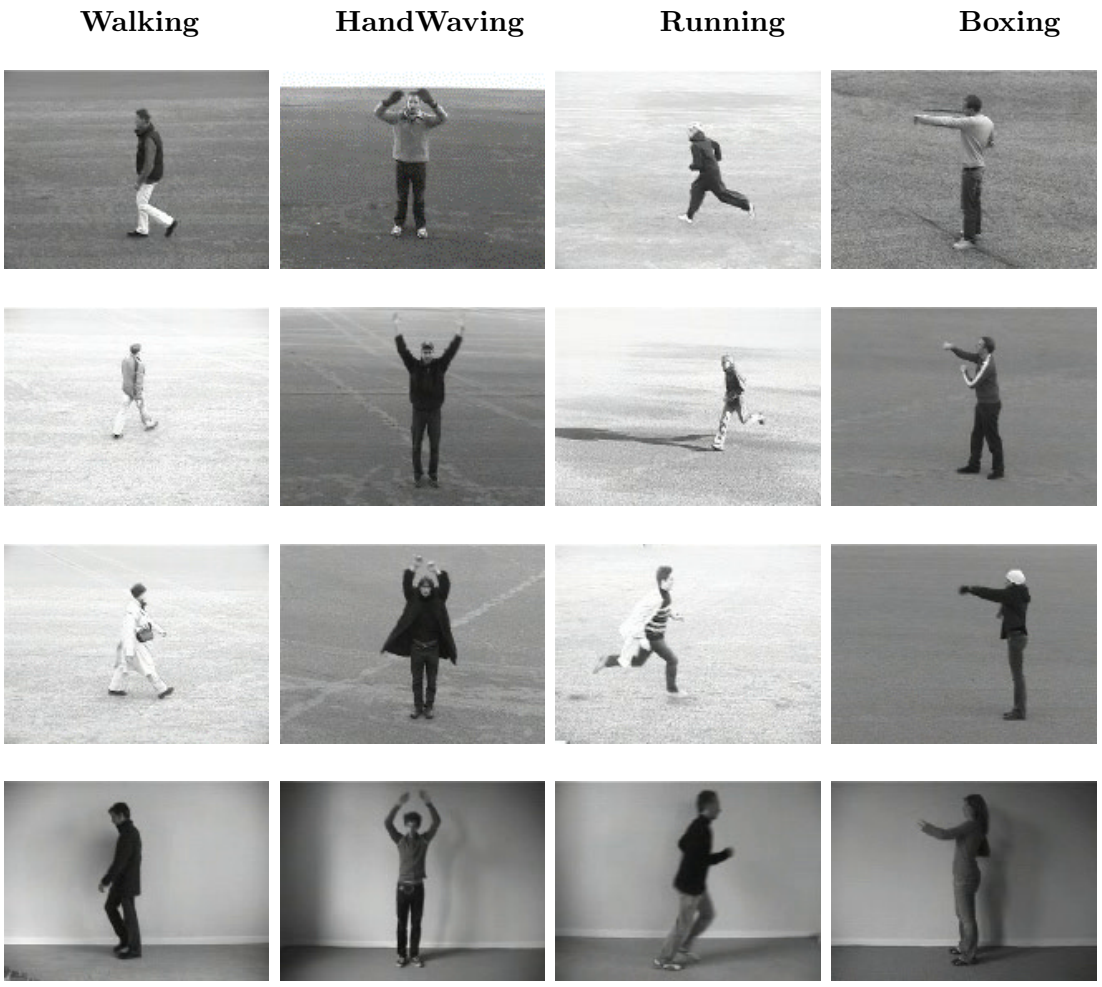


TABLE 1.1: **Sequences from KTH Dataset**

Chapter 2

Past Work

In this chapter we provide an overview of some of the main contributions in these areas with emphasis on motion-based recognition and motion representation.

Liu et al. [23] suggested descriptors that jointly encode shape and motion, while Liu and Shah [24] suggested a method to automatically find the optimal number of visual word clusters through maximization of mutual information (MMI) between words and actions. Chen and Hauptmann [25] proposed MoSIFT which models the local motion and detects interest points that encodes their local appearance. First SIFT algorithm is applied to find visually distinctive components in the spatial domain and thereafter to detect interest points with 'sufficient' amount of optical flow around the points. On the other hand, Laptev and Lindeberg [26] considered Harris and Forstner operators to detect interset points and also detected local structures in time-space where the intensities have significant local variations in both space and time. Then computed the scale-invariant spatio-temporal descriptors.

Trajectory-based approaches considers that the tracking of joint positions is sufficient for humans to recognize actions[52]. Trajectories are usually constructed by tracking joint points or other interest points on human body. Various representations and corresponding algorithms match the trajectories for action recognition. Wang et al. [55] proposed an approach to describe videos by dense trajectories. They sampled dense points from each frame and tracked them based on displacement information from a dense optical flow field. Local descriptors of HOG, HOF and MBH (motion boundary histogram) around interest points were computed.

Bregonzio et al. [36] exploited only the global distribution information of interest points. In particular, holistic features from clouds of interest points accumulated over multiple temporal scales are extracted. A feature fusion method is formulated based on Multiple Kernel Learning. Chen and Hauptmann [25] proposed modified SIFT to detect interest points then encodes their local appearance and motion.

Sadanand and Corso [38] presented a high-level representation of video where individual detectors in this action bank capture example actions, such as "running-left" and "biking-away," and are run at multiple scales over the input video; it represents a video as the collected output of many action detectors that each produces a correlation volume. Being a template-based method, there is actually no training of the individual bank detectors, the detector templates in the bank are selected manually. This method requires using a number of action templates as detectors, which is computationally expensive.

Tran et al. [39] combined local and global representations of the human body parts, encoded the relevant motion information. It represented motion of body parts in a sparse quantized polar space as the activity descriptor. Fathi and Mori [40] constructed a mid-level motion features built from low-level optical flow information. But it is sensitive to noise. These features are based on local regions of the image sequence. Mid-level shape features were constructed from low-level gradient features using the AdaBoost algorithm.

Kovashka and Grauman [41] first extracted local motion and appearance features from training videos, quantizes them to a visual vocabulary, and then forms candidate neighborhoods consisting of the words associated with nearby points and their orientation with respect to the central interest point. Considering multi-scale higher level vocabularies were formed maintaining space-time hierarchy. The most intuitive space-time volume approach would use the entire 3-D volume as feature or template, and match unknown action videos to existing ones to obtain the classification. However, the method suffers from the noise and meaningless background information, and therefore, some effort has been made to model the foreground movement.

Based on Bobick and Davis's [43] work on movement, various efforts are made to extend it. Hu et al. [44] proposed to combine both motion history image (MHI) and appearance information for better characterization of human actions. Two kinds of appearance-based features were proposed. The first appearance-based feature is the foreground image, obtained by background subtraction. The second is the histogram of oriented gradients feature (HOG), which characterizes the directions and magnitudes of edges and corners. SMILE-SVM (simulated annealing multiple instance learning support vector machines) was proposed for classification. It aims to obtain a global optimum via simulated annealing method without relying on model initialization to avoid local minima. Qian et al. [45] also combined global features and local features to classify and recognize human activities. The global feature was based on binary motion energy image (MEI), and its contour coding of the motion energy image was used instead of MEI as a better global feature because it overcomes the limitation of MEI where hollows exist for parts of human blob are undetected. For local features, an object's bounding box was used. The feature points were classified using multi-class support vector machines. Roh et al. [46] also exended Bobick and Davis's [43] MHI from 2-D to 3-D space, and proposed volume motion template for view-independent human action recognition using stereo videos.

Motivated by a gait energy image [47], Kim et al. [48] proposed an accumulated motion image (AMI) to represent spatio-temporal features of occurring actions. The AMI was the average of image differences. A rank matrix was obtained using ordinal measurement of AMI pixels. The distance between rank matrices of query video and candidate video was computed 8 using L1-norms, and the best match, spatially and temporally, was the candidate with the minimum distance.

Various researchers tried to incorporate person models such as silhouettes or skeletons for action recognition. Ikizler and Duygulu [49] proposed a new pose descriptor called histogram of oriented rectangles(HOR) for action recognition. They represented each human pose in an action sequence with oriented rectangular patches extracted over the human silhouette, which then formed spatial oriented histograms to represent the distribution of these rectangular patches. The local dynamics was captured with

the summation of the HOR within a sliding window. Four matching methods were performed for classification, namely nearest neighbor, global histogram matching, SVM and dynamic time warping.

Fang et al. [50] first mapped the high dimensional silhouettes to low dimensional points as spatial motion description using locality preserving projection. This low-dimensional motion vector was assumed to describe the intrinsic motion structure. Then three different temporal information, i.e. temporal neighbor, motion difference and motion trajectory, was applied to the spatial descriptors to obtain the feature vectors, which were fed with k -nearest neighborhood classifier.

Ziaeefard and Ebrahimnezhad [51] proposed the cumulative skeletonized image (CSI) across time as features, and constructed 2-D angular/distance histograms based on it. A hierarchical SVM was used for the matching process. First a coarse classification of CSI histograms using an SVM classifier was obtained with dissimilar actions, and then a second SVM was applied to confused actions using salient features among similar actions.

Messing et al. [53] extracted feature trajectories by tracking Harris3D interest points using a KLT tracker [54], and the trajectories were represented as sequences of log-polar quantized velocities. It used a generative mixture model to learn a velocity-history language and classified video sequences. A weighted mixture of bags of augmented trajectory sequences was modeled for action classes. These mixture components can be thought of as velocity history words, with each velocity history feature being generated by one mixture component, and each activity class has a distribution over these mixture components. Further, they showed how the velocity history feature can be extended, both with a more sophisticated latent velocity model, and by combining the velocity history feature with other useful information, like appearance, position, and high level semantic information.

Bregonzio et al. [58] proposed clouds of space-time interest points to overcome the limitations of the Dollar detector [57]. Using the detected interest points from [57], this was achieved through extracting holistic features from clouds of interest points accumulated over multiple temporal scales followed by automatic feature selection. SVMs and Nearest Neighbor Classifiers (NCCs) were employed for classification.

The survey paper by Aggarwal and Shangho [60] focuses on modeling of motion and recognition of actions and interactions. Moeslund et al. [62] emphasize the human motion capture and analysis, including human model initialization, tracking, pose estimation and action recognition. Krüger et al. cite64 pay more attention to the recognition of actions at different levels of complexity. Turaga et al. [64] focused on the recognition of actions and activities at higher levels. They did not emphasize on the lower-level processing modules like detection and tracking. Enzweiler and Gavrila [65] concentrated on the detection of pedestrians. Candamo et al. [66] focus on the recognition of human behaviors in transit scenes. The most recent survey by Aggarwal and Ryoo [67] emphasizes the recognition of human actions, interactions and group activities. Jiang et al. [68] worked on event recognition, rather than on human activity or crowd behavior.

Chapter 3

Methodology

As discussed in previous chapter, in human activity recognition there are three major modules – segmentation, feature extraction and representation, activity detection. In our work also, these steps are followed. The overall block diagram is shown in Fig. 3.1. Steps are elaborated in the following sections.

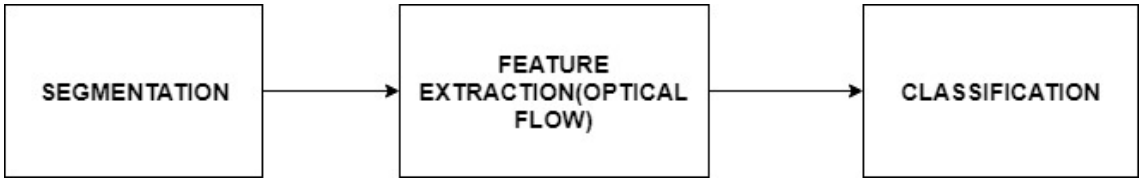


FIGURE 3.1: Overall block diagram.

3.1 Segmentation

For human activity recognition, the first stage is object segmentation, i.e., the human objects are extracted from the background image. It is assumed that the camera is fixed at a specific position and angle. Since the background is fixed (almost static), it is natural to build a background model in advance, so that the foreground object can be segmented from the image.

3.1.1 Background Subtraction

The background model represents the stationary background scene without any foreground object. It is assumed that any change is caused by moving objects. Hence the foreground object can be obtained by subtracting the current image of the background image, followed by a magnitude thresholding to obtain the segmentation mask. The segmentation mask often contains rough and fractional foreground object(s) and usually requires some post-processing, such as closing and opening morphological operations. This is all done in the frame-level. We have used Gaussian Mixture-based Background/-Foreground Segmentation Algorithm. On that mask, we applied median blurring to remove noise from the video. Such noise reduction is a typical pre-processing step to improve the results of later processing (for example, edge detection on an image). On

that we have dilated the mask obtained by blurring we applied dilation. Dilation is one of the two basic operators in the area of mathematical morphology, the other being erosion. It is typically applied to binary images, but there are versions that work on gray-scale images. The basic effect of the operator on a binary image is to gradually enlarge the boundaries of regions of foreground pixels (i.e. white pixels, typically). Thus areas of foreground pixels grow in size while holes within those regions become smaller. The steps are summarized as follows.

- Background/Foreground Segmentation Algorithm applied on frames.
- Successively, median blurring is done on the mask obtained by the first step to remove noise.
- Dilation is done on the mask to complete the edges and thereby obtain a proper enclosed regular mask.

Some clips after background subtraction

Running

Walking

Jumping Jack

3.2 Feature Extraction and Representation

The second stage is feature extraction and representation. This is to compute the important characteristics of segmented region in the frames and to represent the same to formulate the descriptor. It has crucial influence in the performance of recognition. For our work, we have adopted three approaches. As optical flow is utilized in these approaches, a brief description of the same is provided.

Optical Flow

Another category for segmentation on moving camera is optical flow, which denotes a displacement of the same scene in the image sequence at different time instant. The pixel-based local optical flow in image sequence can be robustly evaluated by the Lucas-Kanade-Tomasi (LKT) feature tracker [27, 28], which effectively selects corner feature points of the reference image patch. In [29], Daniilidis et al. apply an FIR-kernel based LKT feature tracker to estimate the optical flow and to infer the motion of objects. The spatial FIR-kernels are binomial approximations to the first derivatives of the Gaussian function. Moreover, Huang et al. [30] also apply the LKT feature tracker to obtain the

optical flow. Those features points with similar optical flows (similar magnitude and orientations) are then grouped together. Finally, the detected moving object patch is validated by target's color histogram as well as contour outlier removing. Even though the optical flow can be estimated by the LKT feature tracker, which robustly captures the local descriptor, it will perform poorly when the reference image patch is occluded by the moving target, the feature points originally located at the background will be moved with the target and result in inaccurate estimation of the optical flow. To overcome this issue, the LKT feature points need to be updated every few frames.

Optical flow works on several assumptions:

1. The pixel intensities of an object do not change between consecutive frames
2. Neighbouring pixels have similar motion.

Consider a pixel I*(x,y,t)* in first frame (Check a new dimension, time, is added here. Earlier we were working with images only, so no need of time). It moves by distance *(dx,dy)* in next frame taken after dt time. So since those pixels are the same and intensity does not change, we can say,

$$I(x,y,z) = I(x+dx, y+dy, z+dz)$$

Then take taylor series approximation of right-hand side, remove common terms and divide by dt to get the following equation:

$$\boldsymbol{f_x u + f_v y + f_t = 0}$$

where,

$$\mathbf{f_x} = \frac{df}{dx}; \quad f_\mathbf{y} = \frac{df}{dy}$$
$$\mathbf{u} = \frac{dx}{dt}; \quad v = \frac{dy}{dt}$$

Above equation is called Optical Flow equation. In it, we can find f_x and f_y, they are image gradients. Similarly f_t is the gradient along time. But (u,v) is unknown. We cannot solve this one equation with two unknown variables. So several methods are provided to solve this problem and one of them is Lucas-Kanade.

- **Approach 1**

  Global representations extract global descriptors directly from original videos or images and encode them as a whole to obtain the feature. Background is subtracted from each frame. On the foreground mask Haar-based cascade classifier is applied to obtain the human body. Motion is estimated using optical flow. Finally, motion vectors of the points falling on the human body is used in developing the feature. It enables us to find the movement of the body and thereby successfully detect the activity. We get a 2-channel array with optical flow vectors, (u,v) where u and v are the magnitudes along two orthogonal direction. We find the magnitude and direction from the motion vector. Magnitude is normalized within [0,1] a 10 -bin histogram is formed with an interval of 0.1. The direction (ranging from 0 to 180 degree) is also divided in to 8 bins and a histogram is formed. Both the histograms are then normalized. Finally, mean and standard deviation of bin numbers computed from the two histograms are computed to generate four dimensional feature vector. We took their mean and standard deviation as features. Thus for of a sequence of N frames the feature vector will be of 4xN dimension. Classification of the video sequence was done using linear SVM. In summary, the steps are as follows.

  1. Optical flow based motion vector generation for extracted human body.
  2. Normalize motion magnitude. Quantize magnitude and direction into smaller number of bins.
  3. Compute mean and standard deviation of the magnitude and direction histograms to form frame level feature vector.
  4. Concatenate frame level feature vector to form sequence descriptor.
  5. Classify using SVM.

  This feature vector was not robust, and accuracy was not satisfactory. So we looked for other approaches.

- **Approach 2**

  In this approach for motion estimation, Lucas Kanade dense optical flow [27] has been applied. Dense one is followed to obtain more interest points. Finally, interest points falling within the detected human body are considered. To make the descriptor a bit detailed the frame is divided into number of blocks (in our work, it is experimentally decided as 9). As in approach 1, for each such blocks magnitude and direction histograms are formed considering interest points lying on the extracted object. As in approach 1, for each bloc a four dimensional feature vector is formed and block level features are concatenated to generate frame level vector. These are again combined to have sequence descriptor which is fed to SVM classifier. The steps are summarized as follows.

  1. Apply dense Optical flow algorithm.
  2. Divide the frame into number of blocks.

3. Consider only the interest points on the extracted human body.
4. Prepare block level histograms and compute block level features.
5. Concatenate block level feature to have frame level vector. Combine those to have sequence descriptor.
6. Classify based on the descriptor.

- **Approach 3**

  Sometimes, a gesture is affected by context of preceding as well as following gestures. With this objective, in this approach, along with the grid features we concatenated contextual pattern information.Unlike the previous approaches this was more of a video level approach where we take in consideration a window of say k=10 for every frame and try to record the contextual pattern of the video. We assume that in the video of a person doing an activity, the frames record similar pattern and thereby the activity is detected. This approach was congruent to the ROI's of the previous approach, i.e., the frame level grids and the video based context information. We took a window of k frames(k=10) and took the histogram (magnitude and direction) intersection by means of chi-square distance to observe the pattern and thereby to store the information of context along with the motion flow. Just this feature vector failed to improve the accuracy as compared to the previous approaches. So, here along with the contextual feature vector we took into consideration the feature vector of approach 2. Finally, we applied ensemble learning approach where we train the first feature vector(contextual pattern) with a weight of 0.3 and the second feature vector(the grid based magnitude and direction mean and standard deviation) with a weight of 0.7. We trained it using two different SVM's. The grid based feature vector was trained using linear Svm with test set size as 0.2 while the contextual based feature vector was trained using rbf kernel and finally the classifier ensembled gave the best approach. The steps are summarized as follows.

  1. Optical flow based segmentation giving two feature vectors-magnitude and direction.
  2. Ensemble based learning with the feature vectors trained on two different weighted classifiers.
     (a) Compute frame level feature using approach 2.
     (b) Context feature vector taking into account the window of 10 frames and saving the context by means of histogram intersection of magnitude and direction vectors.
  3. Then weighted ensemble learning was done on these two feature vectors giving the best classification.

3.3 Activity Detection and Classification Algorithms

After selecting proper features from image or video, activity detection and classification algorithms are the next stage under consideration for human activity recognition. To

achieve good recognition performance, it is essential to choose a proper classification algorithm using the selected feature representation. In our work, we have classified using SVMs. For all the three approaches Svm was used.
Approach 1: SVM with linear kernel and test size =0.3 while training size was 0.7.
Approach 2: SVM with linear kernel with test size =0.3 and training size was 0.7.
Approach 3: Two Svm's - for ensemble learning. The Contextual Feature vector was trained using svm with "rbf" kernel while the second grid level feature vector was trained with linear svm. For the first feature vector test size was 0.2 and training size was 0.8 while for second feature vector test size was 0.3 and training was 0.7.

3.3.1 Support Vector Machine (SVM) explained

The SVM [31–33] is one of the most popular margin-based supervised classifier in the pattern recognition. It is used to separate a data set into two classes. The goal of designing an SVM is to find the optimal dichotomic hyperplane which can maximize the margin (the largest separation) of two classes. The data points on the margin of the hyperplane are called support vectors. Schuldt et al. [31] apply SVMs to recognize human activities by extracting local space-time features in a video.Moreover, Laptev et al. [31] use a nonlinear SVM with a multi-dimensional Gaussian kernel for recognition of various natural human activities, including AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp and StandUp by building spatial-temporal bag-of-features (space-time grids). The main drawback of an SVM is its higher computation burden for the constrained optimization programming used in the learning phase.

Chapter 4

Experimental Results

APPROACH	DATASET	TRAINING ACCURACY(%)	TESTING ACCURACY(%)
APPROACH 1	**WEIZMANN**	**100**	**82.16**
APPROACH 1	KTH	83.57	72.26
APPROACH 2	WEIZMANN	100	67.85
APPROACH 2	KTH	100	85.56
APPROACH 3	**KTH**	**100**	**90.81**

TABLE 4.1: Comparison of accuracies with different datasets using different approaches.

Table 4.1 shows the accuracy for different approaches on KTH and Weizman datasets. KTH dataset has six classes and Weizmann dataset has ten classes. For KTH dataset maximum accuracy obtained is **90.81%** and it is obtained with the grid leveland contextual feature (Approach 3). For the Weizmann dataset, highest accuracy of **82.16%** is obtained with just taking mean and standard deviation of histogram (Approach 1). More detailed features are used in approach 2. But it fails to improve the performance. Possibly it becomes sensitive to minor changes in the spatial distribution of interest points. However in approach 2 when those features are supplemented by preceding and following contexts performance improves. However, as the number of frames in the sequences of Weizmann dataset is quite less approach 3 could not be applied on it. Confusion matrices for the best cases on the two datasets are shown in Table 4.2 and 4.3.

	WALKING	JOGGING	RUNNING	BOXING	WAVING	CLAPPING
WALKING	**1**	0	0	0	0	0
JOGGING	0.07	**0.84**	0.09	0	0	0
RUNNING	0.08	0.08	**0.84**	0	0	0
BOXING	0	0	0	**0.90**	0.05	0.05
WAVING	0	0	0	0.13	**0.87**	0
CLAPPING	0	0	0	0	0	**1**

TABLE 4.2: The confusion matrix for Approach3 on KTH Datset.

	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE1	WAVE2
BEND	**1**	0	0	0	0	0	0	0	0	0
JACK	0.33	**0.51**	0	0	0	0.16	0	0	0	0
JUMP	0	0	**1**	0	0	0	0	0	0	0
PJUMP	0	0	0	**1**	0	0	0	0	0	0
RUN	0	0	0	0	**1**	0	0	0	0	0
SIDE	0	0	0	0	0	**1**	0	0	0	0
SKIP	0	0	0	0	0	0	**0.75**	0	0	0.25
WALK	0	0	0	0	0	0	0	**1**	0	0
WAVE1	0	0	0	0	0	0	0	0	**1**	0
WAVE2	0	0	0	0	0.33	0	0	0	0	**0.67**

TABLE 4.3: The confusion matrix for Approach1 on Weizmann Datset.

METHOD	KTH(Testing accuracy%)	WEIZMANN(Testing accuracy%)
Proposed Method(Approach 1)	72.26	82.16
Proposed Method(Approach 3)	90.81	—-
Liu and Shah[24]	94.2	——
Schuldt et al.[31]	71.72	——
Scovanner et al.[19]	-	82.6
Dollar et al. [18]	81.17	85.2

TABLE 4.4: Comparison with other methods

Table 4.4 shows the comparison of performance of the proposed work and few other works. Liu and Shah[24], outlined a general conceptual framework which unifies the tasks of selecting probing locations and test plans based on the concept of mutual information.Schuldt et al.[31] demonstrated that action recognition can be achieved using local measurements in terms of spatio temporal interest points (local features) . They explored the combination of local space-time features and SVM and applied the resulting approach to the recognition of human actions. Scovanner et al.[19] introduced a 3-dimensional (3D) SIFT descriptor. They used a bag of words approach to represent videos, and present a method to discover relationships between spatio-temporal words in order to better describe the video data. Dollar et al. [18] showed that direct 3D counterparts to commonly used 2D interest point detectors are inadequate for detection of spatio-temporal feature points and proposed an alternative.They developed and tested a number of descriptors to characterize the cuboids of spatio-temporally windowed data surrounding a feature point. Cuboids extracted from a number of sample behaviors from a given domain were clustered to form a dictionary of cuboid prototypes. The only information kept from all subsequent video data is the location and type of the cuboid prototypes present. From Table 4.4 it is observed that performance of the proposed methodology is comparable with those works. However, we have relied on much simpler descriptors.

Chapter 5

Conclusion

In this work, we have presented a simple approach based on global features. We have finally proposed the descriptor that combines frame level features along with its neighbourhood context. Optical flow proved to be a very powerful tool for localizing moving human object and its output are further processed into good features. The magnitude and direction feature vectors were converted to histograms and mean and standard deviation (either for full frame or grid wise) was computed to generate frame level features. With the context information in mind, final feature set was computed taking the window of 10 frames and recording the surrounding context of a particular activity. The feature set was trained using SVM.

Although progress in recent video-based human activity recognition has been encouraging, there are still some apparent performance issues that make it challenging for real-world deployment. Some the aspects are as follows.

- The viewpoint issue remains the main challenge for human activity recognition. In real world activity recognition systems, the video sequences are usually observed from arbitrary camera viewpoints; therefore, the performance of systems needs to be invariant from different camera viewpoints.
- Since most moving human segmentation algorithms are mostly based on background subtraction, which requires a reliable background model. It is better to have a background model that can be adaptively updated and can handle some moving background or dynamic cluttered background, as well as inconsistent lighting conditions.
- Natural human appearance can change due to many factors such as walking surface conditions (e.g., hard/soft, level/stairs, etc.), clothing (e.g., long dress, short skirt, coat, hat, etc.), footgear (e.g., stockings, sandals, slippers, etc.), object carrying (e.g., handbag, backpack,briefcase, etc.) [34]. The change of human action appearance leads researchers to a new research direction, i.e., how to describe the activities that are less sensitive to appearance but still capture the most useful and unique characteristics of each action.
- Unlike speech recognition systems, where the features are more or less unified to be the mel-frequency cepstral coefficients (MFCCs) for HMM classifiers, there are still no clear winners on the features for human activity recognition, nor the

corresponding classifier designs. It can be expected that 3D viewpoint invariant modeling of human poses would be a good starting point for a unified effort.

In on-going work, we are planning to extend the proposed histogram-based image descriptors for a single person doing that activity to multiple people doing that activity and also, trying to find out the feature vector which is uniform so that it can be applied for all datasets. It will enhance the robustness and make the classification much more generalized.

Bibliography

[1] Aggarwal, J., Ryoo, M., 2011. *Human activity analysis: A survey.* ACM Computing Surveys 43, 1–43.

[2] Poppe, R., 2010. *A survey on vision-based human action recognition.* Image and Vision Computing 28, 976–990.

[3] Fathi A, Mori G. *Action recognition by learning mid-level motion features.* Comput Vision Pattern Recogn, CVPR IEEE 2008:1–8.

[4] Blank M, Gorelick L, Shechtman E, Irani M, Basri R. *Actions as space-time shapes.* Int Conf Comput Vision, ICCV IEEE 2005;2:1395–402.

[5] Ke Y, Sukthanka R, Hebert M. *Efficient visual event detection using volumetric features.* Int Conf Comput Vision, ICCV IEEE 2005;1:166–73.

[6] Sheikh Y, Sheikh M, Shah M. *Exploring the space of a human action.* Int Conf Comput Vision, ICCV IEEE 2005:144–9.

[7] Chen MY, Hauptmann AG. MoSIFT: *recognizing human actions in surveillance videos.* Technological report, CMU-CS-09-161, Carnegie Mellon University; 2009. p. 9–161.

[8] Duong, T.V.; Bui, H.H.; Phung, D.Q.; Venkatesh, *S. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model.* In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego,CA, USA, 20–25 June 2005; Volume 1, pp. 838–845

[9] Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, *R. Actions as Space-time Shapes.* InProceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing,China, 17–21 October 2005; Volume 2, pp. 1395–1402.

[10] Ke, Y.; Sukthankar, R.; Hebert, M. *Spatio-temporal Shape and Flow Correlation for Action Recognition.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.

[11] Yamato, J.; Ohya, J.; Ishii, K. *Recognizing Human Action in Time-sequential Images using Hidden Markov Model.* In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Champaign, IL, USA, 15–18 June 1992; pp. 379–385.

[12] Lu, W.; Little, J.J. *Simultaneous tracking and action recognition using the PCA-HOG descriptor.* In Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, Quebec, PQ, Canada, 7–9 June 2006; p. 6.

[13] Brand, M.; Oliver, N.; Pentland, *A. Coupled hidden Markov Models for Complex Action Recognition.* In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Juan, PR, USA, 17–19 June 1997; pp. 994–999.

[14] Luo, Y.; Wu, T.; Hwang, J. *Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks.* Comput. Vis. Image Underst. 2003, 92,196–216.

[15] Lu, X.; Liu, Q.; Oe, *S. Recognizing Non-rigid Human Actions using Joints Tracking in Space-Time.* In Proceedings of the IEEE International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, NV, USA, 5–7 April 2004; Volume 1; pp. 620–624.

[16] Du, Y.; Chen, F.; Xu,t W. *Human interaction representation and recognition through motion decomposition.* IEEE Signal Process. Lett. 2007, 14, 952–955.

[17] Bodor, R.; Jackson, B.; Papanikolopoulos, N. *Vision-based Human Tracking and Activity Recognition.* In Proceedings of the 11th Mediterranean Conference on Control and Automation, Rhodes, Greece, 18–20 June 2003; Volume 1, pp. 18–20.

[18] Dollár, P.; Rabaud, V.; Cottrell G.; Belongie, S. *Behavior Recognition via Sparse Spatio- Temporal Features.* In Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15– 16 October 2005; pp. 65–72.

[19] Scovanner, P.; Ali, S.; Shah, M. *A 3-dimensional SIFT Descriptor and Its Application to Action Recognition.* In Proceedings of the 15th International Conference on Multimedia, ACM, Augsburg, Germany, 23–28 September 2007; pp. 357–360.

[20] Kuo, Y.; Lee, J.; Chung, P. *A visual context-awareness-based sleeping-respiration measurement system.* IEEE Trans. Inf. Technol. Biomed. 2010, 14, 255–265.

[21] Sch, C., Barbara, L., . *Recognizing human actions : A local SVM approach*

[22] Chen MY, HauptmannAG. MOSIFT: *Recognizing human actions in surveillance videos.* Technological report, CMU-CS-09-161, Carnegie Mellon University 2009, p.9-161

[23] Lin Z Jiang, Z Davis LS *Recognizing actions by shapemotion prototype trees.* Int Conf Computer Vision, ICCCVIEEE. P:1-8.

[24] Liu J Shah M.*Learning human actions via information maximization*, Comput Vision and Pattern Recognition, CVPR IEEE 2008; 1-8.

[25] Chen MY, HauptmannAG. MOSIFT: *Recognizing human actions in surveillance videos.* Technological report, CMU-CS-09-161, Carnegie Mellon University 2009, p.9-161.

[26] Ivan Laptev and Tony Lindeberg in Proc ICCV'03, Nice , France: *Space Time Interest Points* pp I:432-439

[27] Lucas, B.D.; Kanade, T. *An Iterative Image Registration Technique with An Application to Stereo Vision.* In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, B.C., Canada, 24–28 August 1981.

[28] Shi, J.; Tomasi, C. *Good Features to Track.* In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 21– 23 June 1994; pp. 593–600.

[29] Daniilidis, K.; Krauss, C.; Hansen, M.; Sommer, *G. Real-time tracking of moving objects with an active camera.* Real-Time Imaging 1998, 4, 3–20.

[30] Huang, C.; Chen, Y.; Fu, L. *Real-time Object Detection and Tracking on a Moving Camera Platform.* In Proceedings of IEEE ICCAS-SICE, Fukuoka, Japan, 18–21 August 2009; pp. 717–722.

[31] Schuldt, C.; Laptev, I.; Caputo, B. Recognizing Human Actions: *A Local SVM Approach.* In Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR), Cambridge, UK, 23–26 August 2004; Volume 3, pp. 32–36

[32] Vapnik, V . *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1999.

[33] Vapnik, V.; Golowich, S.E.; Smola, *A. Support vector method for function approximation, regression estimation, and signal processing.* Adv. Neural Inf. Process. Syst. 1997, 9, 281 287.

[34] Gafurov, D. A survey of biometric gait recognition: *Approaches, security and challenges.* In Proceedings of Norwegian Symposium on Informatics 2007 (NIK 2007), Oslo, Norway, 19– 21, November, 2007.

[35] Heng Wang, Alexander Kläser, Cordelia Schmid, Liu Cheng-Lin. *Action Recognition by Dense Trajectories.* CVPR 2011 - IEEE Conference on Computer Vision

[36] Bregonzio M, Xiang T, Gong S. *Fusing appearance and distribution information of interest points for action recognition.* Pattern Recogn 2012;45(3):1220–34.

[37] Niebles J, Wang H, Fei-Fei L. *Unsupervised learning of human action categories using spatial-temporal words.* Int J Comput Vision 2008;79(3):299–318.

[38] Sadanand S, Corso J. Action bank: *a high-level representation of activity in video.* Comput Vision Pattern Recogn, CVPR IEEE 2012:1234–41

[39] Tran KN, Kakadiaris IA, Shah SK. *Modeling motion of body parts for action recognition.* British Mach Vision Conf, BMVC 2011.

[40] Fathi A, Mori G. *Action recognition by learning mid-level motion features.* Comput Vision Pattern Recogn, CVPR IEEE 2008:1–8.

[41] Kovashka A, Grauman K. *Learning a hierarchy of discriminative space-time neighborhood features for human action recognition.* Comput Vision Pattern Recogn, CVPR IEEE 2010:2046–53.

[42] Enzweiler, M.; Gavrila, D.M. *Monocular pedestrian detection: Survey and experiments.* IEEE Trans. Pattern Anal. Mach. Intell. 2009, 31, 2179–2195.

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

[44] Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T., 2009. *Action detection in complex scenes with spatial and temporal ambiguities,* in: IEEE International Conference on Computer Vision (ICCV), pp. 128–135.

[45] Qian, H., Mao, Y., Xiang, W., Wang, Z., 2010. *Recognition of human activities using SVM multi-class classifier.* Pattern Recognition Letters 31, 100–111.

[46] Roh, M.C., Shin, H.K., Lee, S.W., 2010. *View-independent human action recognition with Volume Motion Template on single stereo camera.* Pattern Recognition Letters 31, 639–647.

[47] Han, J., Bhanu, B., 2006. Individual recognition using gait energy image. IEEE Transaction Pattern Analysis and Machine Intelligence 28.

[48] Kim, W., Lee, J., Kim, M., Oh, D., Kim, C., 2010. *Human Action Recognition Using Ordinal Measure of Accumulated Motion.* EURASIP Journal on Advances in Signal Processing 2010, 1–12.

[49] Ikizler, N., Duygulu, P., 2009. *Histogram of oriented rectangles: a new pose descriptor for human action recognition.* Image and Vision Com-puting 27, 1515–1526.

[50] Fang, C.H., Chen, J.C., Tseng, C.C., Lien, J.J.J., 2010. *Human Action Recognition Using Spatio-temporal Classification* , 98–109.

[51] Ziaeefard, M., Ebrahimnezhad, H., 2010. *Hierarchical human action recognition by normalized-polar histogram,* in: International Conference on Pattern Recognition (ICPR),pp. 3720–3723.

[52] Johansson, G., 1975. *Visual motion perception.* Scientific American 232, 76–88.

[53] Messing, R., Kautz, H., 2009. *Activity recognition using the velocity histories of tracked . keypoints,* in: IEEE International Conference on Computer Vision (CVPR), pp. 104–111.

[54] Lucas, B.D., Kanade, T., 1981. *An iterative image registration technique with an application to stereo vision,* in: The 7th International Joint Conference on Artificial Intelligence - Volume 2, pp. 674–679.

[55] Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L., 2011. *Action recognition by dense trajectories,* in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA. pp. 3169–3176.

[56] Laptev, I., Lindeberg, T., 2003. *Space-time interest points*, in: IEEE International Conference on Computer Vision (ICCV), pp. 432–439.

[57] Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. *Behavior recognition via sparse spatio-temporal features*, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.

[58] Bregonzio, M., Gong, S., Xiang, T., 2009. *Recognising action as clouds of space-time interest points*, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 1948–1955.

[59] Jones, S., Shao, L., Zhang, J., Liu, Y., 2012. *Relevance feedback for real-world human action retrieval.* Pattern Recognition Letters 33, 446–452.

[60] Aggarwal, J. K.; Park, S. Human Motion: *Modeling and Recognition of Actions and Interactions.* In Proceedings of IEEE 2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), Thessaloniki, Greece, 6–9 September 2004; pp. 640–647.

[61] Valera, M.; Velastin, S.A. *Intelligent distributed surveillance systems: A review.* IEEE Proc. Vis. Image Signal Process. 2005, 152, 192–204.

[62] Moeslund, T.B.; Hilton, A; Krüger, V. *A survey of advances in vision-based human motion capture and analysis.* Comput. Vis. Image Underst. 2006, 104, 90–126.

[63] Krüger, V.; Kragic, D.; Ude, A.; Geib, C. *The meaning of action: A review on action recognition and mapping.* Adv. Robot. 2007, 21, 1473–1501.

[64] Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. *Machine recognition of human activities: A survey.* IEEE Trans. Circuits Syst. Video Technol. 2008, 18, 1473–1488.

[65] Enzweiler, M.; Gavrila, D.M. *Monocular pedestrian detection: Survey and experiments.* IEEE Trans. Pattern Anal. Mach. Intell. 2009, 31, 2179–2195.

[66] Candamo, J.; Shreve, M.; Goldgof, D.B.; Sapper, D.B.; Kasturi, R. *Understanding transit scenes: A survey on human behavior-recognition algorithms.* IEEE Trans. Intell. Transp. Syst. 2010, 11,206–224.

[67] Aggarwal, J.K.; Ryoo, M.S. *Human activity analysis: A review.* ACM Comput. Surv. (CSUR) 2011, 43, 16.

[68] Jiang, Y.; Bhattacharya, S.; Chang, S.; Shah, M. *High-level event recognition in unconstrained videos.* In International Journal of Multimedia Information Retrieval, 2013, 2, 73–101.

[69] Enzweiler, M.; Gavrila, D.M. *Monocular pedestrian detection: Survey and experiments.* IEEE Trans. Pattern Anal. Mach. Intell. 2009, 31, 2179–2195.