A Review on Human Actions Recognition Using Vision Based Techniques

Muhammad Hassan, Tasweer Ahmad, Nudrat Liaqat, Ali Farooq, Syed Asghar Ali, and Syed Rizwan hassan
Department of Electrical Engineering, Government College University, Lahore, Pakistan
Email: muhammadhassan@gmail.com, tasveer.ahmad@gcu.edu.pk, {nudratliaqat, engg.ali.farooq, asghar.tarmazi, syedrizwanhassan} @gmail.com,

***Abstract*—A lot of efforts have been rendered to recognize human actions such as face recognize, pose estimate and motion analysis. The major problems in action recognition are: 1. Scale variation. 2. Articulated Pose variation. 3. Illumination variation. The various motion patterns and the poses formed by a single human being are termed as actions. Human Actions are identified by the pose of the body such as standing, sitting etc. The main objective of this work is to recognize different human actions using various vision based techniques. We examine different approaches and compare their merits and demerits. In this paper, a detailed review of the latest vision based techniques consists of: 1. Methods. 2. Systems. 3. Quantitative evaluation.**

***Index Terms*—face recognize, pose estimate, vision based recognition, quantitative evaluation**

I. Introduction

Recognizing human actions have been very challenging for computer vision scientists and researcher since the last two decades. Human Action Recognition systems have a wide spread applications in surveillance, pedestrian tracking and Human Machine Interaction (HMI) [1]**.** These systems often employ techniques such as robust foreground segmentation, people tracking and occlusion handling. But event then these technologies are not matured enough to be fully deployed somewhere. The vision-based action recognition systems entail low resolution task and high resolution tasks [2]. The low resolution tasks incorporate human body as a whole e.g. center of mass, area, volume and velocity of entire body [3], [4]. On the other hand, high resolution tasks manifests measurement and relation of individual body parts e.g. human pose, silhouette, position and velocity of hands, feet etc. [5], [6], [7].

There exist two types of shape-based features 1) silhouette 2) contour. The first technique performs on all pixel areas in the polygon [8]. While second technique works well for outer boundary of objects in an image. Contour-based methods work on Fourier Transform, Artificial Neural Network and Hough Transform.

The activity recognition has also been carried out by researchers using micro sensor-based systems. The main reason for deployment of such system is that they are low-power, cost-effective and privacy-aware. The

Activity based recognition system should have the following key features: 1. Robustness: Accurate extraction of features from sequences 2. Discriminative: it should be discriminative instead of generative model 3. Rejection: Activity recognition should not be exhaustive for foreseeable future. The activity-based recognition systems work in a hierarchical fashion. At primary level, object detection and tracking is done. At mid-level, atomic actions are recognized using various classification techniques. At high level, these atomic actions are linked together to figure out an activity [9].

The goal is to recognize the activities from a series of image frames. In a simple scenario, a set of images contains a human action, the target is to recognize the action correctly. In general scenarios the continuous actions of humans are recognized with start and end time of the action in the set of frames of image. In this review paper a number of approaches have been discussed to recognize human actions and classify them. In this discussion, we refer the simple motion patterns performed by single person and duration lasted for short intervals of time are Actions. Some examples of different actions are; bending, walking, running etc. A hierarchical approach is used for the human actions recognition system, where the lower levels are on segmentation, detection and tracking, middle level on action recognition and higher level are the classifiers.

Human actions are categorized into four different categories such as: gestures, poses, interactions and activities. Gestures are the basic and atomic motions of a body part such as waving a hand etc. Poses are not atomic motions of a human body parts instead it is comprised of a multiple motions of a human body parts of a single person such as walking etc. Interaction is consists of more than one person actions such as fighting etc. Activities are performed by a group of people trying to perform actions together like meeting etc. The main goal of this paper is to give a review on various methodologies to recognize human actions. We talk about various techniques made to recognize various levels of actions performed by humans. A survey performed by Aggarwal and Cai in 1999 [10] has achieved the goal to recognize lower level actions such as: tracking and body posture analysis. However this analysis was not sufficient enough and we have to describe the higher level actions. So we

Manuscript received January 6, 2013; revised May 12, 2014.

doi: 10.12720/joig.2.1.28-32

conduct this survey to perform higher level actions to be recognized.

The vision based recognition becomes the primary goal to recognize the actions automatically. After recognizing and classifying the actions of human, we can identify the abnormal and suspicious actions from normal actions and this will be helpful in predicting the threats and terrorist attacks.

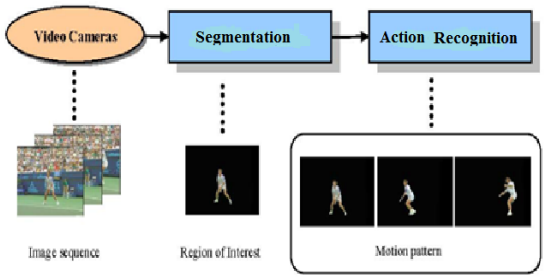

Figure 1. Flowchart for action recognition

As shown in Fig. 1, the lower level aims to find the regions of interest from the image sequences while the higher level focuses on the recognition of temporal motion patterns. Our primary focus in this paper is to briefly explain the existing techniques used to recognize human actions in vision based technology.

II. Related Work

Cedras and Shah in 1995 [11], Gavrila in 1999 [12] and Aggarwal and Cai in 1999 [10] produced the review on recognizing human actions. Kruger et al. in 2007 also discussed the recognition of human actions by arranging them on the basis of the complex features. Turaga *et al.* in 2008 [13] also furnished a survey related to vision based action recognition techniques. In all above surveys they categorize different techniques on the basis of complexity in actions.

Since most of the previous reviews contain only introduction and summaries of various methods to recognize human actions but they do not compare various approaches. In our review, various methods require to recognize human actions. We used a standard taxonomy to compare every method. Furthermore our review not only covers recognizing methods but also the complex human actions in group form. And in the end we discuss the results received after applying various methods on the datasets available in public for the purpose to recognize various human actions.

III. General Overview

The recognition of any action can be done generally by following the series of steps shown below:

- Image sequences
- Low level feature Extration
- Mid-level action descriptions from low level features.
- Semantic interpretation using High level features

Background subtraction helps in finding the region of interest i.e., moving parts, and isolates them. The human silhouette will be a key factor to recognize actions and we can isolate it using background subtraction (see Fig. 2). The recognition after background subtraction is achieved by the help of global method; boundary method and skeletal descriptors. In global method, the entire shape region is used to compute the shape descriptor. In Boundary method, only contour is used to define the characteristic of the shape. Skeletal method deals with the complex shape and identifies them as set of one dimensional skeletal curve.

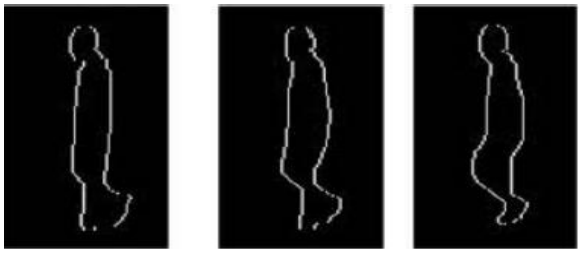

Figure 2. Silhouettes extracted from the walking sequence

Segmentation is applied on each image to find the region of interest and the segmentation is based on the camera situation either it is moving or stationary. The two types of segmentation are shown in Fig. 3.

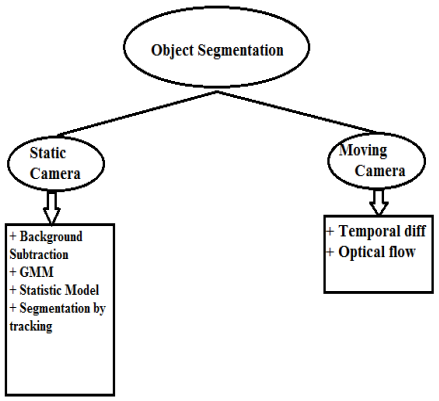

Figure 3. Object segmentation methods

Since the camera is stationary in the static case so the viewpoint and the background should be at rest and to extract the area of interest. There are many methods we can apply to retrieve the region of interest, such as background subtraction, GMM, statistical modeling, by tracking models and many more. The simplest one is by the help of background subtraction. Now if we have the camera in motion, then the background is continuously varying and also the target is also changing its position as well. So for this purpose we use temporal difference and optical flow technique to acquire the region of interest. Template matching method isolates the motion features and produces certain motion patterns and then we obtain the human motion templates which represent already defined activities. The normal activities are categorized by matching the templates. State space approach formulates a statistical model through training and this will help in recognizing human activities. In this method, every static posture is considered as a single state and we correlate it with various models. For motion sequence we

calculate the joint probability and see when it has maximum value.

Now once the segmentation is done and we obtained the region of interest in form of features which are categorized in the four groups shown below in Fig. 4.

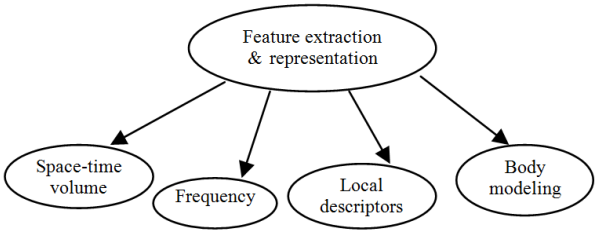

Figure 4. Different categories for feature extraction and representation

From frame sequences, consecutive silhouette of objects are accumulated along time-axis, this representation is called as Space Time Volume. The extraction is a 3D model and represented by XYT volume and this scheme is not good for periodic activities. Now we work on the frequency domain by using Discrete Fourier Transform (DFT) and this will be helpful in getting information of the geometric structure. Since DFT and STV are global functions and applied on the full image while if we needed something local, we use the descriptors. The local descriptors contain the characteristics of image patch. Scale Invariant Feature Transform (SIFT) and histogram of oriented gradient (HOG) are two types of local feature extraction techniques widely used. Human modeling methods are used for pose estimation.

Once the area of interest is isolated then the next stage is detection and classification. The algorithms used for this purpose are: dynamic time warping (DTW), generative models and discriminative models. The DTW is used to calculate the similar patterns in the temporal sequences but it is not good for large amount of data. Generative models are dynamic classifiers are based on probabilities of the system. Some important types of the generic models are Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN). If we are working on the discriminative models which are known as static classifiers can also be used when we are not dealing with probabilities. Different types of discriminative models are Artificial Neural Network (ANN), Support Vector Machine (SVM) and Relevant Vector Machine (RVM). There were other methods are introduced as well like Kalman filter, binary tree, multidimensional indexing, and K nearest neighbor (K-NN) to classify and detect different activities.

IV. METHODS

The main step after segmentation is feature extraction and representation and it has a great significance in recognition of human actions. If the features contain the information of space and time relationship, then they are space time volumes (STV) and if we add discrete Fourier transform (DFT) image frames which contain the image intensity variation spatially. STV and DFT take the whole image for extraction of features but they are sensitive to noise etc. Since the local features are more effective against noise and occlusion.

A. Local Descriptors:

Local descriptors are configured to be more effective against occlusion and noise. Some of the local descriptors commonly used are I) Scale invariant feature transform. II) Histogram-of-Oriented Gradient (HOG) III) Nonparametric weighted feature extraction (NWFE) IV) Kanade-Lucas Tomasi (KLT) Tracker V) Shape-based features. VI.) Appearance-based features.

There are four major steps to obtain the SIFT feature of the image frame, which are as follows: firstly we detect the extreme values in the form of invariant interest points of scale space using difference-of-Gaussian (DoG) function. Then we have to calculate the keypoint localization which is most suitable among interest points which are calculated in the first step. The next step is to apply the orientation assignment according to the gradient directions and lastly keypoint descriptor extraction as shown in Fig. 5. The main defects in SIFT descriptor is high dimensionality in matching features and also it is not discriminative technique instead it is based on probabilities.

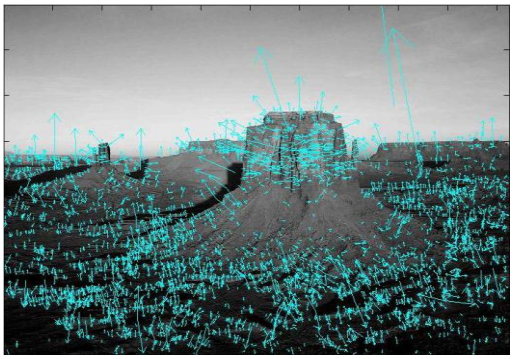

Figure 5. 2D SIFT features results

Histogram of oriented gradient (HOG) descriptors is used to evaluate the normalized local histograms of image gradient. This evaluation is done in the dense grid using the fine scale gradient and is represented in the Fig. 6. We apply HOG here first and then Principal Component Analysis on linear subspace. This sort of PCA-HOG descriptor is less susceptible to illuminations, poses etc. variation. The main issue in HOG is that it depends on the size of the human i.e., local descriptors only evaluated on fixed scale.

Figure 6. (a) Image gradients of local grids (b) HOG descriptors with nine bins

Shape analysis is more likely the same method like silhouette extraction process and is helpful as it is invariant to rotation, translation and scale. In this human silhouette extraction, the feature describes the human gaits which are helpful in recognition and the similarity measures are done by the help of dynamic time warping (DTW) algorithm. We can apply both unsupervised and supervised methods for recognizing human actions. The main issue we face here is when there is a body part moving in the silhouette region. Also it requires the exact segmentation of silhouette which is not possible.

B. Pose Estimation

In the Pose Estimation technique, we use Higher-order Local Auto-Correlation (HLAC) image descriptor. The local appearance context descriptors are computed in a HOG-based image where human objects are present. On this processed image, PCA is applied for dimensionality reduction before computing HLAC. We finally retrieve the viewpoints of shape-skeleton models of the image frames. We apply them on the training plane and we finally obtain the required features from the image given. Extensive training of dataset is required to obtain better viewpoints performance.

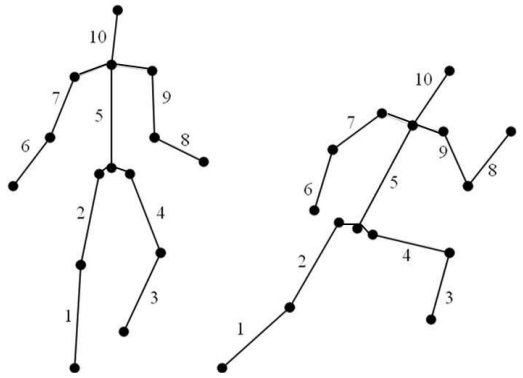

Figure 7. Human bodies shown joints points

Human pose estimation contains a model which detects and tracks the torso and after that converted into a normalized feature space and we identify different key poses using nearest mean classifiers (NMC). For 3D human pose estimation, we have to follow three stages 1) Foreground blobs tracking. 2) Two Dimensional joints and body parts tracking. 3) Data-driven Markov chain Monte Carlo (DD-MCMC) for estimation of the pose. These projected 3D postures are converted into feature points and then classified as human actions using HMMs.

Once the features are extracted and detected the next stage will be classification of the recognized activity. For classification, the following are the most common and good in performance algorithms, such as, dynamic time warping (DTW), hidden Markov model (HMM), support vector machines (SVM), relevance vector machines (RVM), artificial neural networks (ANN), dynamic Bayesian network (DBN) and many more. DTW is a dynamic programming algorithm used to calculate the distance among two sequences. The issue with DTW is that it requires extensive templates. HMM and DBN calculate the joint probability distribution while SVM, RVM and ANN are discriminative models which work on the conditional probability distribution.

C. Motion History

The main idea in space time volume recognition is calculation of similarities between two volumes and system must describe the similarities in two volumes. To calculate the correct similarities we focus on the silhouette instead of taking the whole image and to find shape changes easily. The over segmented volumes computes set of 3-D XYT volume segments that relates to a human in motion and instead of 3-D space-time volume of every action we use a template composed of two 2-D images. In this template, one image is of motion energy image (MEI) and the other one is of motion history image (MHI). These images are made from a sequence of foreground images which contains 2-D projections of the original 3-D XYT space-time volume. After template matching technique is executed on the pair of both images, we are able to recognize motions like sitting, waving etc. as shown in Fig. 8.

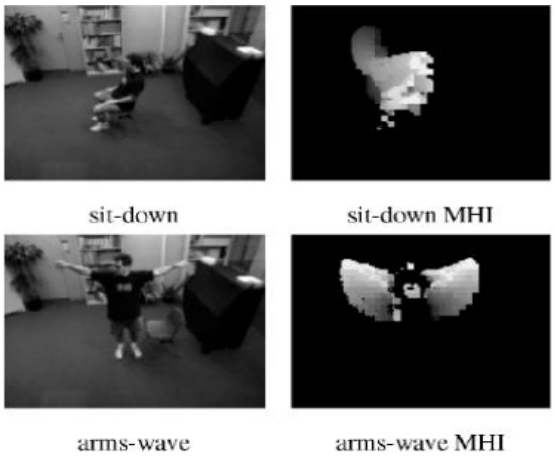

Figure 8. Space time action representation

We can also design a 3-D space time video template correlation method that works on the correlation between given video and template volumes and is termed as hierarchical space-time volume correlation. Every volume patch contains exact local motion and after correlation result will be local match score to the system and at the end overall correlation is computed. A support vector machine (SVM) is applied for human action recognition. We also analyze the 3D space time volumes using synthesizing filters in which maximum average correlation height (MACH) filters are used for recognition. For every action, a synthesized filter is generated and action is classified by MACH filter. The problems faced in this technique are recognition of actions, when more than one person is in the scenario and actions which are not spatially segmented.

V. Applications

Some of the areas where the activity recognition is quite beneficial are highlighted here. Biometrics contains the techniques which are based on uniquely recognizing humans like fingerprint, face or iris detection. Content-

based video summarization and content-based image retrieval (CBIR) has been progressing with great interest now a days and sports is one of the most widely applied domain. Security and surveillance systems using vision-based techniques can detect automatically the anomalies in a camera field of view. The interaction between a machine and human can be done by visual cues such as gestures and activity recognition. Ubiquitous environments like smart rooms are the fruits from vision based methods. The animation and gaming industry depend on producing artificially humans and human motion. If improvement is done in algorithms then more realistic motion-synthesis will be achieved and will be helpful in training of military soldiers, fire-fighters and others in simulated situations.

In surveillance environment, the main focus is on tracking abnormalities in crowds so that recognition of the criminal activities can be accurately made. Since the main focus of video based surveillance systems is to detect suspicious events but they can also be used to ensure safety of swimmers in pools. A system is equipped with a network of cameras mounted above the surface of water and thus can increase the chance of saving a life from drowning.

In entertainment activities like sports, dances and games many video based recognition systems have been proposed. Sport activities are recognized by applying HMMs to recognize the time-sequential images of tennis scenes, forehand volley, backhand volley, smash and serving and many more.

In healthcare systems, the analysis then understanding a patient activity is the primary concern and can be achieved by the help of vision based systems to recognize actions. It will be helpful to diagnose, treat and care for patients.

VI. Conclusion

Human actions recognition is assumed to be the fundamental step of understanding the human behavior and this review gives a brief survey of existing techniques based on actions recognition using vision systems. The critical modes of any recognition system like object segmentation, feature extraction plus representation, and action classification are discussed briefly. Also the applications of such recognition system are explained including surveillance, entertainment and healthcare. Although great progress has been made but still technical issues are still unresolved and need to be resolved for practical deployment.

Making a machine able to see and understand like humans do requires long fascinated scientists, engineers etc. Research disciplines like Computer Vision, Artificial Intelligence, Neuro-sciences, and Linguistics etc have provided us very close results which is a great victory so far. But still we have to face technical challenges and many issues are to be resolved to get a perfect model. The advancements achieved are should be consolidated in terms of real time conditions and performance. Then we have a firm ground for further research.

References

[1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, March 1999.

[2] L. Fiore, "Multi-camera human activity monitoring," *Journal of Intelligent and Robotic Systems*, vol. 52, no. 1, May 2008.

[3] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. 11th Mediterranean Conf. on Control and Automation*, June 2003.

[4] B. Maurin, O. Masoud, and N. Papanikolopoulos, "Monitoring crowded traffic scenes," in *Proc. the IEEE 5th Int. Conf. on Intelligent Transportation Systems*, Singapore, September 3–6, 2002, pp. 19-24.

[5] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continuous detection," in *Proc. the Intl. Conference on Computer Vision*, 1999.

[6] R. Fablet and M. J. Black, "Automatic detection and tracking of human motion with a view-based representation," in *Proc. European Conf. on Computer Vision*, May 2002.

[7] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, August 2000.

[8] A. Aker and R. Gaizauskas, "Generating image description using dependency relation patterns," in *Proc. of 48th Annual Meeting of Association for Computational Linguistics*, Uppsala Sweden, July 2010, pp. 1250-1258.

[9] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, October 2008.

[10] J. K. Aggarwal and Q. Cai., "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, March 1999.

[11] C. Cedras and M. Shah, "Motion-based recognition: A survey," *Image and Vision Computing*, vol. 13, March 1995

[12] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, January 1999.

[13] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, Nov. 2008.

Muhammad Hassan, received his degree BEng from Government College University, GCU, Lahore. He is also doing his MEng from GCU Lahore. His area of interest is image processing, machine learning.

Tasweer Ahmad received his BEng Degree from University of Engineering and Technology, Taxila, Pakistan and MEng from University of Engineering and Technology, Lahore, Pakistan. His area of interest is Computer Vision and Machine Learning.

Nudrat Liaqat, received her degree BEng from UET, Lahore. Currently, she is also doing his MEng from GCU Lahore. His area of interest is image processing, machine learning.

Ali Farooq, received his degree BEng from UCP, Lahore. Currently, he is also doing his MEng from GCU Lahore. His area of interest is image processing, machine learning.

Syed Asghar Ali, received his degree BEng from GCU, Lahore. Currently, he is also doing his MEng from GCU Lahore. His area of interest is image processing, machine learning.

Syed Rizwan Hassan, received his degree BEng from GCU, Faisalabad. Currently, he is also doing his MEng from GCU Lahore. His area of interest is image processing, machine learning.

