International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:11, 2015

Human Action Recognition System Based on Silhouette

S. Maheswari, P. Arockia Jansi Rani

***Abstract*—**Human action is recognized directly from the video sequences. The objective of this work is to recognize various human actions like run, jump, walk etc. Human action recognition requires some prior knowledge about actions namely, the motion estimation, foreground and background estimation. Region of interest (ROI) is extracted to identify the human in the frame. Then, optical flow technique is used to extract the motion vectors. Using the extracted features similarity measure based classification is done to recognize the action. From experimentations upon the Weizmann database, it is found that the proposed method offers a high accuracy.

***Keywords*—**Background subtraction, human silhouette, optical flow, classification.

I. Introduction

THE analysis of human body movements can be applied in a variety of application domains, such as video surveillance, video retrieval, human computer interaction systems and medical diagnoses. In some cases, the result of human action analysis can be used to identify people acting suspiciously and other unusual activities directly from videos.

Monitoring activities of daily living is gaining interest because of the growing population of elderly people and their need for care. A system that contributes to the safety of elderly at home is therefore more than needed. The analyzing of human behavior and looking for the changes in the activities of the daily living is essential for the medical professionals to detect emerging physical and mental health problems, before they become critical particularly for elderly. The human action recognition is necessary in shop surveillance, city surveillance, airport surveillance and in other places where security is the prime factor.

The presented method can be applied for sports video analysis like race walking. Race walking is an Olympic athletic event and it is different from running. Using this method the system can recognize whether the race walker is walking or running.

The remainder of this paper is the discussion of the presented scheme. Section II discusses the related literature. Section III explains the presented method. Section IV discusses the experimental results. Finally, Section V concludes the paper.

S. Maheswari, Research Scholar, and Dr. P.Arockia Jansi Rani, Assistant Professor, are with the Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India (e-mail: jan20mahi91@yahoo.com, jansimsuniv@gmail.com).

II. Literature Review

Bobick et al. [1] presented temporal templates through projecting frames onto a single image, namely motion history image (MHI) and motion energy image (MEI). MHI indicates how motion happens and MEI records where it is. This representation gives satisfactory performance under the circumstance where the background is relatively static.

Oikonomopoulos et al. [2] focused on the problem of human action recognition using spatiotemporal events that are localized at points that are salient both in space and time. The spatiotemporal points are detected by measuring the variations in the information content of pixel neighborhoods not only in space but also in time. The classification scheme uses Relevance Vector Machines and on the chamfer distance measure. The classification results are presented for two different types of classifiers, displaying the efficiency for the representation in discriminating actions of different motion classes.

Oikonomopoulos et al. [3] developed a new set of visual descriptors that provide a local space-time description of the visual activity. The descriptors are extracted at spatiotemporal salient points detected on the estimated optical flow field for a given image sequence.

Danielweinland et al. [4] presented an overview and categorization of the approaches used. Feature extraction, action learning, action segmentation, action classification are the stages involved in action recognition. Feature extraction is used to extract the postures and the motion cues from the video. Action learning is the process of learning statistical models from the extracted features. The statistical models are used to classify new feature observations. Action segmentation is used to cut the streams of motions into a single action instances that are consistent to set of initial training sequences used to learn the models.

Droogenbroeck et al. [5] proposed a technique for motion detection that incorporates several innovative mechanisms. This technique stores, a set of values for each pixel taken in the past at the same location or in the neighborhood. It then compares this set to the current pixel value in order to determine whether that pixel belongs to the background, and adapts the model by choosing randomly which values to substitute from the background model.

Laptev et al. [7] combined the histograms of optical flow (HOF) and histograms of oriented gradients (HOG) as a descriptor, which is demonstrated to be better than either of HOG or HOF as a single descriptor.

Aggarwal et al. [8] gave an overview of various methods used prior to 1995, in articulated and elastic non rigid motion.

After a good overview of various motion types, the approaches within articulated motion with or without a priori shape models are described. Then the elastic motion approaches are described in two categories with and without a shape model.

Yamato et al. [9] proposed the silhouette images. This representation computes a grid over the silhouette and computes for each cell the ratio of foreground to background pixels.

K. Schindler and van Gool [10] used optical flow information and Gabor filter responses in a human-centric framework. For each frame, both types of information are weighted and concatenated. PCA over all pixel values is applied to learn the most discriminative feature information. Majority voting yields a final class label for a full sequence in multi-class experiments. Results are carried out on the KTH and Weizmann dataset.

Keigo Takahara [11] proposed a robust background modeling technique. Firstly, a background model is established according to the temporal sequence of the frames. Secondly, the moving objects are detected based on the difference between the current frame and the background model. Object detection helps to identify pixels in the video frame that cannot be adequately explained by the background model, and outputs them as a binary candidate foreground mask. Finally, the background model is updated periodically to adapt the variety of the monitoring scene.

M. Blank et al. [12] represented an action by considering the shape carved by its silhouette in time. Local shape descriptors based on the Poisson equations are computed, and then aggregated into a global descriptor by computing moments.

M. Rodriguez et al. [13] introduced a template-based method in which a maximum average correlation height (MACH) filter is generalised to analyse videos as 3D Spatio temporal volumes in the frequency domain. MACH is capable of capturing intra-class variability by synthesizing a single action MACH filter for a given action class.

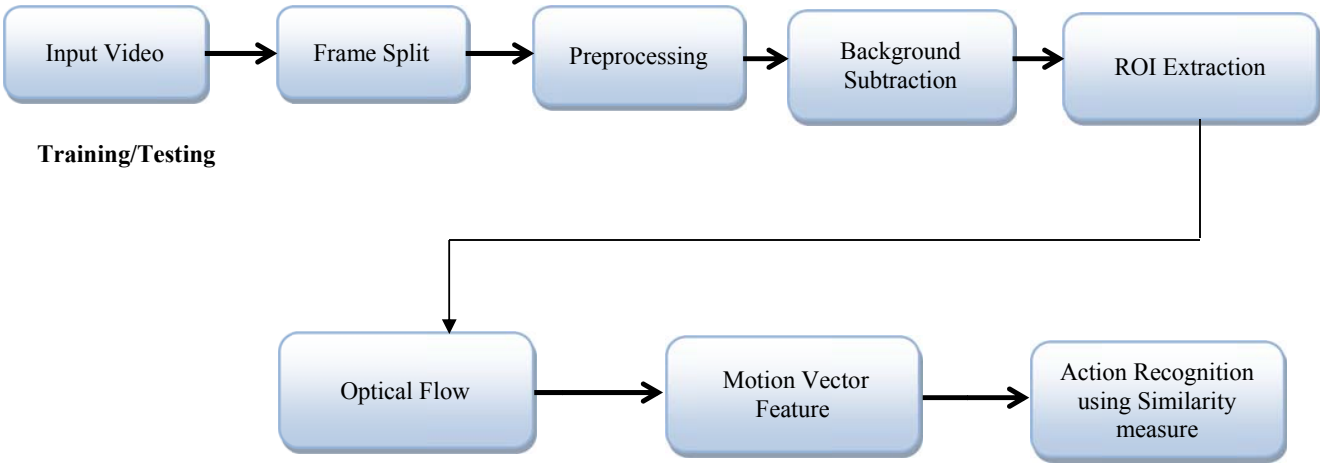

Fig. 1 Architecture of Human Action Recognition System

Pantic et al. [14] defined human computing as "ubiquitous computing" or "ambient intelligence". The key for human computing is to sense certain behavioral cues of the users and to adapt automatically to their typical behavioral patterns and the context in which they act. The technical challenges in understanding the human behaviour are initialization, robustness, speed, training and validation issues.

Deepak et al. [15] recognized the human actions by tracking the selected object over the consecutive frames of gray scale image sequences. Initially the background motion of the input video stream is subtracted and its binary images are constructed. The object which is needed to be monitored is selected by enclosing the required pixels with bounding rectangle and by using spatiotemporal interest points. The obtained results after subtraction are compared with the selected threshold value to predict the type of human action using Linear Prediction technique.

R. Polana and R. Nelson [16] proposed a human tracking framework along with an action representation using spatio-temporal grids of optical flow magnitudes. The action descriptor is computed for periodic motion patterns. By matching against reference motion templates of known periodic actions the final action can be determined.

III. Proposed Work

The presented human action recognition system processes every frame as follows. First, the input video is split into frames and then it is pre-processed to improve the brightness of the image. Background Subtraction is done to extract the ROI (i.e., human silhouette) of every frame. Then morphological processing is done to remove the artifacts or noise from the human silhouette. Optical flow is employed to extract the motion features. Then using similarity measure action is recognized. This process is described in Fig. 1.

A. Input Selection

The input videos are taken from Weizmann Dataset. The input videos are recorded in a homogeneous background with

a static camera. The input videos include walk, run, jump, jack, wave, and jump with run actions. The Weizmann Dataset is downloaded from: www.wisdom.weizmann.ac.il/vision/ SpaceTimeActions. The properties of the input videos are:

- Type: VLC media file (. avi)
- Resolution: 180 x 144
- Frame rate: 25 frames per second.

B. Frame Split

A movie frame is a single picture or still shot, that is shown as a part of a larger video or movie. Here, the input video is split into frames. Frame rate refers to the number of frames that are projected or displayed per second.

C. Preprocessing

The preprocessing is performed to improve the brightness of the image. The brightness of the image is improved to provide better results. The Gaussian filter is used in this phase to remove the noise in the input.

D. Background Subtraction

Background subtraction is a method used to identify and extract ROI namely the human from the input video and this information is used for further processing. There are two phases in background subtraction namely, background modeling and foreground detection. The current frame and the background model are differenced to create the binary foreground object (i.e. the human).

$$f(u,v) = f(x,y) - p(x,y) \quad (1)$$

where p(x, y) is the background pixel, f(x, y) is the pixel in the input frame.

The resultant binary foreground object contains black background with white silhouette information. The silhouettes are analyzed to recognize the human actions.

E. Morphological Processing

ROI (i.e., human) detected after background subtraction is noisy. In order to remove that noise morphological processing is applied. Morphological erosion is used for removing the unwanted portions other than the ROI. Then Morphological closing is used for filling the holes in the ROI. Finally, the human silhouette is extracted clearly.

Fig. 2 Silhouette after Morphological Processing

As silhouettes describe the outer contours of a person, it provides strong cues for action recognition.

F. Optical Flow

Motion estimation is required for solving many computer vision problems. Optical flow is used for motion estimation. Here the optical flow is applied after clearly extracting the ROI (i.e., human silhouette). Horn Schunk method [6] is employed for computing optical flow. The optical flow is computed between the current frame and the n^{th} frame back.

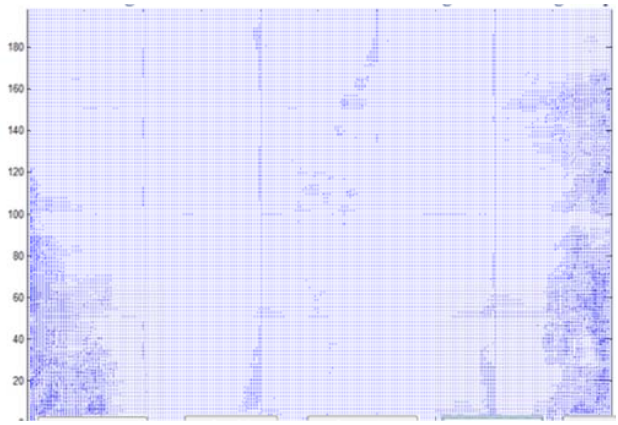

Fig. 3 Optical Flow

The optical flow uses an iterative process to calculate the optical flow between the frames. The motion vector feature is obtained in this phase.

G. Training Phase

In the training mode, videos representing different human action are introduced to the action recognition system. The process carried out in this action recognition system is described through the following steps.

1) Read the input video in matrix M of size (m x n x k), where m and n represent the number of rows and columns for every sequence respectively and k is the number of frames in the video.
2) The Gaussian filter is applied to enhance the input video.
3) Background subtraction is done to detect the ROI (i.e., the human) in every frame of the input video.
4) Morphological processing is done to extract the ROI clearly.
5) Optical flow is computed for the ROIs of the entire video.
6) As a result of optical flow, motion vectors are obtained for the input video.
7) The mean, standard deviation and variance is computed for the obtained Motion vectors.
8) The features mean, standard deviation and variance is concatenated to produce the feature vector.
9) Store the feature vector with their labels representing each action.
10) Repeat the steps 1 to 9 for every input.

H. Testing Phase

In the testing mode, the input video is tested according to the following steps:

1) Read the input video in matrix M of size (m x n x k), where m and n represent the number of rows and columns for every sequence respectively and k is the number of frames in the video.
2) Repeat the steps 2 to 8 in the training phase to obtain the feature vector of the input action.
3) Classification: The distance between the resultant feature vector and the stored feature vector is measured using Euclidean distance. The Euclidean distance measure is calculated using the formula

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

The minimum distance corresponds to the estimated action of the input video.

IV. Experimental Results

In order to evaluate the proposed method, the experiment is conducted on Weizmann database. Table I shows the confusion matrix. Confusion matrix assesses the accuracy of the action recognition system. The confusion matrix shows that actions such as skip and jump with run are confused.

TABLE I
Confusion Matrix

Actions	Run	Walk	Jump	skip	Double sided wave	Jump with Run
Run	10	0	0	0	0	0
Walk	0	10	0	0	0	0
Jump	0	0	10	0	0	0
Skip	0	0	0	9	0	1
Double sided wave	0	0	0	0	10	0
Jump with Run	0	0	0	1	0	9

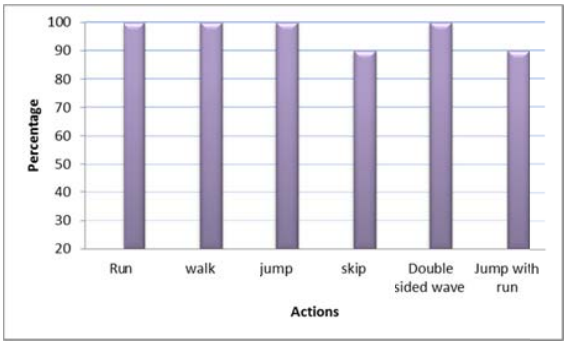

Fig. 4 Chart for Recognition Rate

Recognition rate is calculated for each action on the basis of number of true matches and number of false matches. The confusion matrix shows the number of true matches and the false matches. Recognition Rate (RR) is calculated using the formula

$$RR(\%) = \frac{Number\ of\ correct\ matches}{Total\ number\ of\ actions}\ X\ 100 \quad (3)$$

Recognition Rate is computed for all the actions based on the Recognition Rate formula and the recognition rate for each action is shown in Fig. 4.

The accuracy of the proposed work is 96.6%. There may be a chance of misclassifying the actions run and walk. But in this action recognition system walk action and run action produces 100% result.

V. Conclusion

In this paper, action is recognized using distance based similarity measure. The processing time is fast enough which is useful for real time applications, such as visual surveillance and medical diagnoses. In this paper, optical flow is applied after extracting ROI to extract the motion features and a distance based similarity measure is used for recognizing the human actions. The experiments on Weizmann datasets show the performance of the proposed work. The experimental result demonstrates the potential of the proposed work that it is able to differentiate the usually confusing actions (i.e., walking and running) with 100% accuracy. Experiments can be conducted on other datasets like KTH dataset, HOHa, UCF dataset which evaluates the robustness in different scenarios. In future work, this approach may be extended for action recognition in crowded environments and may be applied in human interactive behaviour recognition.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3):257–267, 2002.
[2] Antonios Oikonomopoulos, Ioannis Patras and Maja Pantic. "Spatiotemporal Salient Points For visual Recognition of Human actions." IEEE Transactions on Image Processing Vol. 36, no. No. 3. (2006).
[3] Antonios Oikonomopoulos, Maja Pantic, Ioannis Patras "Sparse B-spline polynomial descriptors for human activity recognition" in 2009.
[4] Danielweinland, RemiRonfard and Edmond Boyer, "A Survey of Vision- Based Methods for Action Representation, Segmentation and Recognition", October 18, 2010.
[5] Droogenbroeck, O. Barnich and M. Van. "ViBe: A universal background subtraction algorithm for video sequences." IEEE Transactions on Image Processing, June 2011. 20(6):1709-1724.
[6] Enric Meinhardt-Llopis, Javier Sanchez, Daniel Kondermann , "Horn–Schunck Optical Flow with a Multi-Scale Strategy", Image Processing On Line, 3 (2013), pp. 151–172.
[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008.
[8] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, Articulated and elastic non-rigid motion: a review, in Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, 1994, pp. 2–14.
[9] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In CVPR, 1992.
[10] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In CVPR, 2008.
[11] Keigo Takahara, Takashi Toriu and Thi Thi Zin. "Making Background Subtraction Robustto Various Illumination Changes." IJCSNS International Journal of Computer Science and Network Security, March 2011: VOL.11 No.3.
[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. "Actions as space-time shapes." In ICCV, 2005.
[13] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008.

[14] Maja Pantic, Alexpentland and Thomas Huang, "Human Computing and Machine Understanding of Human Behaviour", ICMI'06, November 2-4, 2006.
[15] N.A. Deepak and U.N.Sinha, "Silhouette Based Human Motion Detection and Recognising their Actions from the captured Video Streams", IntJ. Advanced Networking and Applications Volume: 02, Issue: 05, Pages: 817-823(2011).
[16] R. Polana and R. Nelson. Low level recognition of human motion. In IEEE Workshop onNonrigid and Articulate Motion, 1994.

Maheswari received her BE in Computer Science and Engineering from Dr.G.U. Pope College of Engineering, Thoothukudi, TamilNadu, India in 2011 and ME in Computer Science and Engineering from Manonmanium Sundaranar University, Tirunelvelli, Tamilnadu, India in 2013. Her research interests include Digital Image Processing, Neural Networks.

Dr. P. Arockia Jansi Rani graduated B.E in Electronics and Communication Engineering from Government College of Engineering, Tirunelveli, Tamil Nadu, India in 1996 and M.E in Computer Science and Engineering from National Engineering College, Kovilpatti, Tamil Nadu, India in 2002. She has been with the Department of Computer Science and Engineering, Manonmaniam Sundaranar University as Assistant Professor since 2003. She has more than ten years of teaching and research experience. She completed her Ph. D in Computer Science and Engineering from Manonmaniam Sundaranar University, Tamil Nadu, India in 2012. Her research interests include Digital Image Processing, Neural Networks and Data Mining.