An Optical Flow-based Action Recognition Algorithm

Upal Mahbub*, Hafiz Imtiaz*, and Md. Atiqur Rahman Ahad**
*Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh
E-mail: omeecd@eee.buet.ac.bd; hafiz.imtiaz@live.com
** Kyushu Institute of Technology, Kitakyushu, Japan
Email: {atiqahad}@univdhaka.edu

***Abstract*— This paper proposes a new technique for motion-based representation on the basis of optical flow analysis and random sample consensus (RANSAC) method. The method is based on the fact that an action can be characterized by the frequent movement of the optical flow points or interest points at different areas of the human figure. A combination of optical flow and RANSAC to identify a human body within a frame and the average percentage of change of interest points at various positions around the body are used as features for a particular action. Using the extracted features, a distance-based similarity measure and a support vector machine (SVM)-based classification technique have been exploited for recognition. From experimentations upon a standard motion database, it is found that the proposed method offers a very high degree of accuracy.**

I. Introduction

Recognizing the identity of individuals as well as the actions, activities and behaviors performed by one or more persons in video sequences are very important for various applications, such as surveillance, robotics, biomechanics, medicine, sports analysis, film, games, mixed reality, etc.[1] [2]. In this paper, a novel action clustering-based human action recognition algorithm is presented. Here, optical flow is employed in order to detect the presence and direction of motion, whereas, RANSAC is used for further localization and identification of the most prominent motions within the frame. The density of optical flow interest points indicates the probable position of the person along the horizontal direction. More localization is done next based on statistical evaluation of the positions of the interest points both horizontally and vertically. A small bounding box around the subject is thus obtained. The area of the box is then divided into a number of small blocks and the percentage of change in number of interest points within each block is calculated frame by frame. All the matrices formed this way from the similar actions are averaged and used as a feature for that respective action. Finally, classifiers are utilized for the classification task.

II. Proposed Method

Figure 1 shows the overall flow diagram of the proposed system, which consists of feature extraction, learning and recognition phases. The objective of the proposed method is to extract the variations, present in different human actions, by developing a successful measure to follow the movement of different body parts at different directions during an action. For this, the optical flow analysis is employed to track any movement in multiple frames. As the optical flow method is prone to slightest background movement or camera movements, the RAndom SAmple Consensus (RANSAC) [3] algorithm is next applied for further purification of the motion detection. This way, some interest points are obtained which seem to be tagged with the moving body part and move towards the same direction. As can be seen in Fig. 2, most of the interest points (in red points), after performing RANSAC, are gathered around the moving hand. Moreover, almost all the interest points are now gathered around the whole body, which provides the scope to detect the position of the human subject in the scene. A moving window of a fixed length is thus taken along X-axis (preferably a window wider than the width of the human body in the scene), and the number of interest points inside the window is calculated. This process returns the position of the window along X-axis with maximum number of interest points depicting position of the human subject.

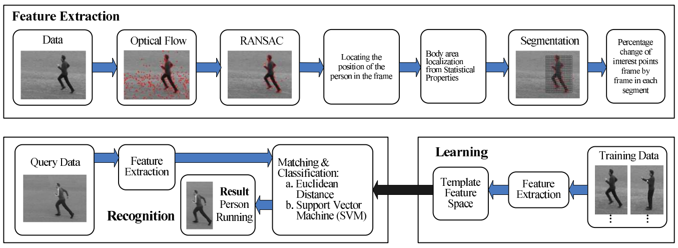
Feature Extraction
Data
Optical Flow
RANSAC
Locating the position of the person in the frame
Body area localization from Statistical Properties
Segmentation
Percentage change of interest points frame by frame in each segment
Query Data
Feature Extraction
Matching & Classification:
a. Euclidean Distance
b. Support Vector Machine (SVM)
Result
Person Running
Recognition
Learning
Template Feature Space
Feature Extraction
Training Data

Fig. 1. The main components of the feature extraction and the human action recognition system.

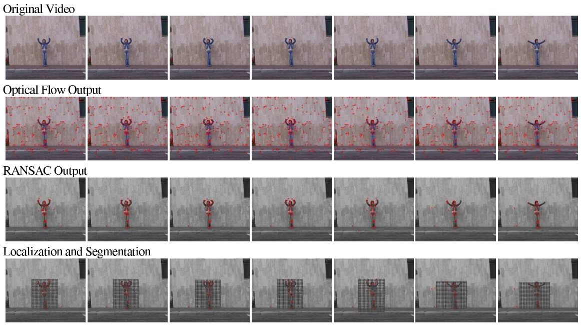

Fig. 2. Feature extraction for a person waving both hands (Weizmann Dataset)

Next, the mean (ϑ_y) and standard deviation (σ_y) of the Y-axis values of all the points in the chosen window are

TABLE I

COMPARISON OF RECOGNITION RATES OF VARIOUS ACTIONS OF THE WEIZMANN DATASET

Action	Run	Side	Skip	Jump	Pump	Bend	Jack	Walk	Wave1	Wave2
ED	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SVM	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%	91.67%

calculated. The Y-axis is divided into n segments starting from $\vartheta_y - \sigma_y - \delta_y$ to $\vartheta_y + \sigma_y + \delta_y$. The value δ_y is an adjustment constant chosen empirically based on the probable height of a human subject in the frame. Within the window along X-axis and the limit imposed along Y-axis, the mean (ϑ_x) and standard deviation (σ_x) of the X-axis values of all the points are calculated. Then the X-axis is also divided into n segments starting from $\vartheta_x - \sigma_x - \delta_x$ to $\vartheta_x + \sigma_x + \delta_x$. The value δ_x is another adjustment constant chosen empirically. Thus, the human subject is now encapsulated within a bounding box which is divided into $n \times n$ smaller blocks or segments.

The procedure is repeated in each frame. If the number of interest points within block k at the i-th frame is IP_k^i, then the change in number of interest points in each block φ_k^i is calculated frame by frame by, $\varphi_k^i = \varphi_k^i + \mid IP_k^i - IP_k^{i-1} \mid$, where, $k = 1, 2, ..., n^2$ and $i = 2, 3...., r$ and r is the total number of frames. Also, the total change in interest points ψ^i for a given frame i is calculated and cumulated throughout the operation by,

$$\psi^i = \sum_{k=1}^{n^2} \mid IP_k^i - IP_k^{i-1} \mid . \qquad (1)$$

Finally, the percentage of change in number of interest points in all the blocks is calculated employing,

$$PercentChange_k = \sum_{i=1}^{r} (\frac{\varphi_k^i}{\psi^i}) \times 100\%. \qquad (2)$$

The above operation is done for several persons performing the same action and the matrices of percentage change of interest points obtained from all the persons are averaged to obtain a feature vector for the respective action. The whole process is repeated for several actions and a feature vector table is constructed. Next, for action classification, a distance-based similarity measure and an support vector machine (SVM)-based similarity measure are utilized.

III. Experimental Results and Analysis

In this experiment, the Weizmann Database [4], consisting of 90 low-resolution (180 × 144) video sequences showing nine different people, each performing 10 natural actions has been considered. The classification task were performed following the leave-one-out test rule. The localization was mostly accurate for different actions and the results in terms of recognition accuracies obtained by the proposed method for the Weizmann-database are listed in Table I. It can be seen that, the recognition accuracies were 100% with simple Euclidean distance measure while with SVM the accuracy rate averaged 91.67%. Here, the Euclidean Distance classifier uses the average values of all the feature vectors obtained from different persons performing a certain action as the feature for that action, while SVM-based classifier considers each person to be a different entity. As the averaged feature matrix contains more generalized information about the movement pattern of body during an action, the Euclidean distance measure surpluses the SVM method. Table II shows the comparison of performance of the proposed method with some other promising methods.

TABLE II

COMPARISON OF PROPOSED METHOD WITH OTHER ALGORITHMS IN TERMS OF AVERAGE ACCURACY

Method	Weizmann dataset
Proposed: Euclidean Distance	100%
Proposed: SVM	91.67%
Ali *et al* [5]	94.75%
Seo *et al* [6]	97.5%
Laptev *et al* [7]	91.8%

IV. Conclusions

This paper presents a novel motion-based human action representation approach by evaluation of the statistical properties of a combination of optical flow and RANSAC algorithms. The major advantage of the proposed method is that it is simple but efficient. It tries to identify any action by tracking the movement of the body parts, quite like how the human brain differentiates between actions. Because of person localization, the method is robust enough to identify the same action performed anywhere else within the frame. However, the method is yet to be tested for more complex actions in cluttered outdoor environment and its performance against the self occlusion problem needs to be investigated. Finally, based on the experiments, it can be strongly claimed that the proposed method can be useful for various applications related to gesture and action understanding in the future.

References

[1] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa, “Motion history image: its variants and applications,” *Machine Vision and Applications*, pp. 1–27, 2010.

[2] ——, “Human activity recognition: various paradigms,” *Int’l Conf. Control, Automation and Systems*, pp. 1896–1901, 2008.

[3] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, pp. 381–395, 1981.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE PAMI*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[5] S. Ali and M. Shah, “Human action recognition in videos using kinematic features and multiple instance learning,” *IEEE PAMI*, pp. 288–303, Feb 2010.

[6] H. J. Seo and P. Milanfar, “Action recognition from one example,” *IEEE PAMI*, vol. 33, no. 5, May 2011.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” *IEEE CVPR*, 2008.