Human Action Recognition using DFT

Sonal Kumari
Birla Institute of Technology and Science,
Pilani,
Rajasthan, India
sonal.kumari1910@gmail.com

Suman K. Mitra
Dhirubhai Ambani Institute of Information and
Communication Technology,
Gandhinagar, Gujarat, India
Suman_mitra@daiict.ac.in

Abstract—**Action is any meaningful movement of the human and it is used to convey information or to interact naturally without any mechanical devices. *Human action recognition* is motivated by some of the applications such as video retrieval, Human robot interaction, to interact with deaf and dumb people etc. In any Action Recognition System, some pre-processing steps are done for removing the noise caused because of illumination effects, blurring, false contour etc. Background subtraction is done to remove the static or slowly varying background. In this paper, multiple background subtraction algorithms are tested and then one of them is selected for the further process of action recognition. Background subtraction is also known as foreground/background segmentation or foreground extraction. The next step is the feature extraction which deals with the extraction of the important feature (like corner points, optical flow, shape, motion vectors etc.) from the image frame. The proposed novel action recognition algorithm uses discrete Fourier transform (DFT) of the small image block.**

Keywords-human action recognition; background subtraction; GMM; feature extraction; DFT; K-NN.

I. Introduction

Action recognition, also known as gesture recognition, pertains to recognize meaningful expressions of motion by a human, involving the hands, arms, face, head, and/or body. Action Recognition in video is one of the most promising applications of computer vision. [1, 2, 3] gave the extensive survey on action recognition. Cedras and Shah [1] presented a survey on motion-based approaches to recognition as opposed to structure-based approaches. They argue that motion is a more important cue for action recognition than the structure of the human body. Mitra and Acharya [2] provided a survey on action recognition with particular emphasis on hand gestures and facial expressions. They discussed applications involving hidden Markov models, particle filtering and condensation, finite-state machines, optical flow, skin color, and connectionist models in detail. They also highlighted existing challenges and future research possibilities. Turaga et al. [3] discussed the problem at two major levels of complexity: 1) "actions" and 2) "activities." "Actions" are characterized by simple motion patterns typically executed by a single human. "Activities" are more complex and involve coordinated actions among a small number of humans. They discussed several approaches and classify them according to their ability to handle varying degrees of complexity.

M. Ahmad and Lee [4, 5] presented HMM based view independent human action recognition system using multiview image sequences. They used Lukas-Kanade's optical flow and shape features. They also used PCA to reduce shape feature dimension. The recognition rate [4, 5] was found 87.5% for the combined feature which is higher than the rate obtained by individual features. J. Alon et al. [6] proposed a method which consists of three novel contributions: a spatiotemporal matching algorithm that can accommodate multiple candidate hand detections in every frame, a classifier-based pruning framework that enables accurate and early rejection of poor matches to gesture models, and a subgesture reasoning algorithm that learns which gesture models can falsely match parts of other longer gestures. Kaaniche and Bremond [7] selected corner points in order to compute and track HoG (Histogram of Oriented Gradients) descriptors and learn local motion descriptors using K-means with PCA. Finally, they classified the gestures using the K-nearest neighbour's algorithm. Hamed et al. [8] have selected features from the gait of a human for recognition. They have also used Key Fourier Descriptors and PCA techniques for high correct classification and reducing features space. They demonstrated action classification by a simple Nearest Neighbour classifier.

The paper is organized as follows. In Section 2 background subtraction techniques are discussed and in Section 3 feature extraction and classification techniques are described. Section 4 describes experimental results and in the last Section, conclusion is presented.

II. Background subtraction techniques

A. Background Subtraction Techniques

Background subtraction is needed to reduce the number of pixels and speed up the processing time. In background subtraction, redundant information from each frame is removed and only object of interest is kept. Following four available background subtraction methods are discussed: (1) Frame Difference, (2) Background Subtraction, (3) Adaptive Gaussian Mixture Model [9], and (4) Improved Adaptive Gaussian Mixture Model [10]. Frame Difference is a primitive type of background subtraction algorithm and may be defined as following:

$$|f_i - f_{i-1}| > T \tag{1}$$

Where, f_i, f_{i-1}, T, i represents current frame, previous frame, threshold, and frame number respectively. This method fails to identify some slowly varying background object. Background Subtraction is an advanced background subtraction algorithm and it may be given as following:

$$|f_i - bg| > T \tag{2}$$

978-0-7695-4599-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NCVPRIPG.2011.58

Where, f_i, bg, T, i represents current frame, stable background image, threshold and frame number respectively. The main disadvantage of this method is the availability of stable background image. If only lighting changes over time, a single, adaptive Gaussian per pixel would be sufficient. In practice, multiple surfaces often appear in the view frustum of a particular pixel and the lighting conditions change. Thus, multiple, adaptive Gaussians per pixel are necessary. Adaptive GMM [9] is a probabilistic approach where each pixel is modelled as a separate mixture model. The probability of observing the current pixel value is:

$$P(X_t) = \sum_{i=1}^{K} \left(\omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t})\right) \qquad (3)$$

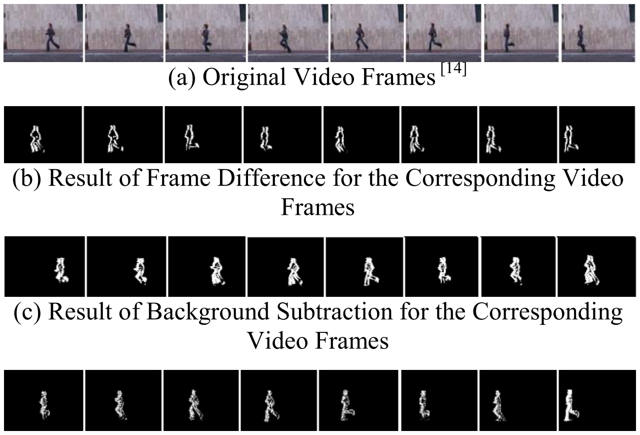
where K is the number of distributions, $\omega_{i,t}$ is an estimate of the weight of the i^{th} Gaussian in the mixture at time t, $\mu_{i,t}$ is the mean value of the i^{th} Gaussian in the mixture at time t, $\sum_{i,\,t}$ is the covariance matrix of the i^{th} Gaussian in the mixture at time t, and where η is a Gaussian probability density function and given by:

$$\boldsymbol{\eta(X_t, \mu, \Sigma)} = \frac{\mathbf{1}}{(\mathbf{2\pi})^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \boldsymbol{e}^{-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1}(X_t - \mu_t)} \qquad (4)$$

Improved Adaptive GMM [10] is similar to Adaptive GMM which starts with GMM with one component and simultaneously selects the appropriate number of components for each pixel. Processing time is reduced in comparison to Stauffer and Grimson method. Segmentation result is also slightly improved.

B. Background Subtraction Result

All background subtraction methods, which are mentioned in this paper, are tested on the same video sequences to find out the best method among all four. Selection of the background subtraction is a very important task in action recognition technique. Since the result of action recognition is based on this background subtraction result.

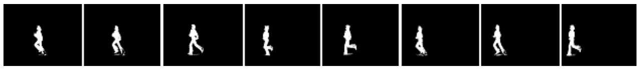


(a) Original Video Frames [14]

(b) Result of Frame Difference for the Corresponding Video Frames

(c) Result of Background Subtraction for the Corresponding Video Frames

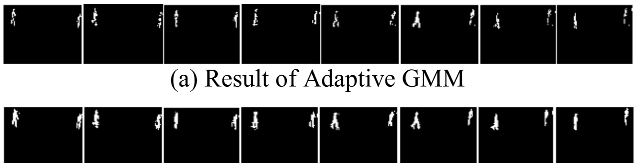
(d) Result of Adaptive GMM for the Corresponding Video Frames



(e) Result of Improved Adaptive GMM for the Corresponding Video Frames

Fig. 1: Result of all four background subtraction algorithms

Improved Adaptive Gaussian Mixture Model gives the best and desired result. The results obtained in Fig. 1(e) show an improvement in the fraction of the actual foreground detected as compared to the other three background subtraction methods. Here the parameter optimization problem is also solved. It is robust method which also updates the number of components. A synthetic low contrast video obtained from *Advanced Computer Vision GmbH – ACV* [15] has been taken for performance evaluation and parameter optimization of the last two robust methods. The low contrast video is taken to measure the performance in challenging conditions which contains 150 numbers of frames.



(a) Result of Adaptive GMM

(b) Result of Improved Adaptive GMM for the Corresponding Video Frames

Fig. 2: Result of Adaptive GMM and Improved Adaptive GMM for the same Low contrast Video Frames

For performance measure of the proposed algorithm, sensitivity (fraction of actual foreground detected) and false alarm rate (fraction of pixels incorrectly classified as foreground) tests have been carried out. Formulae for the same are as given below:

$$S = \frac{\mathrm{TP}}{(\mathrm{TP+FN})} \qquad (5)$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{(\mathrm{FP+TN})} \qquad (6)$$

For Improved Adaptive GMM, the sensitivity value is much higher and false alarm rate is much lower for each video, which signifies that the performance of the algorithm is quite satisfactory.

III. Feature Extraction and classification

In video frame sequences, the object of interest is detected by background subtraction algorithm. Action is converted into a feature vector which will be used for the classification. The concept of feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application. The proposed approach is based on the average discrete Fourier transform (DFT) feature which is extracted from the image block. For feature classification K nearest neighbor (K-NN) algorithm is used.

A. Action Region Extraction

The action region is defined as a rectangular area where action is occurred in the image frame. This step is basically to remove redundant data and to extract the object of interest. This step helps in reducing the processing time. But the feature of the same cardinality cannot be found by using these frames. So area of the rectangle should be normalized for the further process. For that bi-cubic interpolation method is employed, shown in Fig. 3(d) for DA-IICT data. Finally at the end of this step, normalized sequences of frames are selected as shown in Fig. 4 which will be used for the further process.

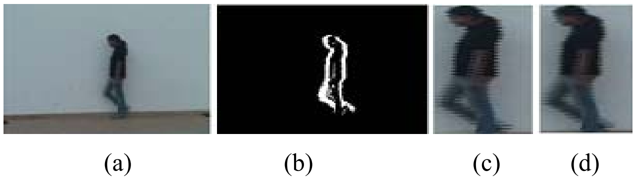

(a) (b) (c) (d)

Figure 3: Image Processing Steps of an image in WalkTurnBack action video of DA-IICT data. (a) Sample frame (350 × 450 pixels) (b) background subtracted frame (350 × 450 pixels) (c) action region (211 × 106 pixels) (d) normalized image frame (100 × 70 pixels)

Figure 4: Normalized frame sequences of walk

B. Feature Extraction

One of the important steps in action recognition is feature selection. Proposed approach to the action recognition problem is based on the proposed novel average DFT feature. The Fourier domain image has a much greater range than the image in the spatial domain. Hence, to be sufficiently accurate, its values are usually calculated and stored in float values. Also, the Fourier transform preserves information from the original signal, and ensures that important features are not lost as a result of the FFT. In this approach, DFT of image frame is calculated. It is used to gain information about the geometric structure (shape) of the spatial domain foreground object provided the foreground object intensity value should be different from the background object intensity value. Here the normalized image frame is divided into small size blocks. DFT is calculated block wise and then taken the average of all the DFT values present in a single block. For a block of size M × N, the two-dimensional DFT is given by:

$$F(k,l) = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} f(i,j)\, e^{-j2\pi(\frac{ki}{M}+\frac{lj}{N})} \qquad (7)$$

Where, $f(i,j)$ is the block in the spatial domain and the exponential term is the basis function corresponding to each point $F(k,l)$ in the frequency domain. If there is n number of blocks, it will produce n dimensional feature space.

C. Classification

K-nearest neighbour algorithm is a method for classifying objects based on closest training examples in the feature space. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, K is a user-defined constant, and an unlabelled vector is classified by assigning the label which is most frequent among the K training samples nearest to that query point. An object is classified by a majority vote of its neighbours. The neighbours are taken from a set of objects for which the correct classification is known. K-NN is well suited for multi-modal classes as its classification decision is based on a small neighbourhood of similar objects. So, even if the target class is multi-modal, it can still lead to good accuracy. Classifier can be updated online at very little cost as new instances with known classes are presented. K-NN is chosen for this application due to the following advantages: 1) Very simple model, 2) Few parameters to tune, 3) Test time is independent of the number of classes, 4) Robust with regard to the search space; for instance, classes don't have to be linearly separable.

IV. EXPERIMENTAL RESULTS

For experimentation purpose, following databases have been used: 1) MuHaVi database and 2) DA-IICT database. MuHAVi data [13] contains 17 action classes. Six actions among 17 actions are taken from the database, which includes WalkTurnBack, RunStop, JumpOverGap, Kick, CrawlOnKnees and WaveArms. These six actions are represented as W, R, Ju, K, C, and W1 respectively in the tables. The size of video frames of each action is 576 × 720 and frame rate is 14 frames per second. In addition with MuHaVi database, an another database is also built at open air theatre, DA-IICT, Gandhinagar by using digital camera which contains 4 actors, and eight following classes of actions: WalkTurnBack, RunStop, JumpOverGap, Kick, SitStandBack, Jack, TwoHandWave, and OneHandWave. These actions are represented as W, R, Ju, K, S, Ja, W1 and W2 respectively. The image sequences have 350 × 450 pixel resolution and bit rate of 1725 kbps for each action. In the DA-IICT data, a single video contains multiple actions one by one performed by a single actor. Multiple actions are not separated into separate videos to test the robustness of the proposed action recognition algorithm. There is a transition from one action to another action in a single video data.

Tables 1 and 2 show the confusion matrices using K-NN for DFT feature on DA-IICT data. Here the block size taken is 16 by 16. The dimension of the DFT feature is 36 which is based on the number of blocks. Each column in the table represents the best match for each testing action video sequences. Some sequences are misclassified, such as running and walking. This is due to the higher degree of similarity in the viewing direction. In table 2, the only difference is made by combining four consecutive frames to consider third dimension which is time. Four continuous frame sequences are appended to preserve changes in two continuous frame sequences. The performance is increased from 94.25% to 96%. This is due to the inclusion of third dimension which is time. In tables 3 and 4, the experiment is similar to tables 1 and 2 respectively. These tables contain result for MuHaVi data instead of DA-IICT. Overall recognition rate using optical flow feature [4], DFT feature and combined feature is 65.13, 94.25 and 94.13 respectively on DA-IICT data. Reduced result of combined features is due to more influence of optical flow [4] in comparison to the DFT. Since dimension of optical flow [4] is just double to the dimension of DFT. Fig. 5 shows graph for the recognition rate for different values of K using the K-NN classification algorithm on DA-IICT data. The experiment is performed for single frame and 4-appended frames. The graph indicates that as the value of K increases, the performance of the recognition model degrades. In all the case, the best recognition rate is achieved when the value of K is one.

Table 1: Confusion matrix using DFT for single frame on DA-IICT data (94.25%)

	W	R	Ju	K	S	Ja	W1	W2
W	93.3333	0	0	0.9524	2.8571	2.8571	0	0
R	0	100	0	0	0	0	0	0
Ju	0	0	96.9388	0	3.0612	0	0	0
K	1.8868	0	0	88.6792	7.5472	1.8868	0	0
S	0	0	3.3708	1.1236	91.0112	2.2472	2.2472	0
Ja	0	0	0	2.1739	0	97.8261	0	0
W1	0	0	1.8182	1.8182	3.6364	1.8182	89.0909	1.8182
W2	0	0	0	0	1	1	0	98

Table 2: Confusion matrix using DFT for four appended frames on DA-IICT data (96%)

	W	R	Ju	K	S	Ja	W1	W2
W	92.5926	0	0	3.7037	0	3.7037	0	0
R	0	100	0	0	0	0	0	0
Ju	0	0	100	0	0	0	0	0
K	0	0	0	91.6667	4.1667	0	4.1667	0
S	0	0	7.4074	3.7037	88.8889	0	0	0
Ja	0	0	0	0	0	100	0	0
W1	0	0	0	4	0	0	96	0
W2	0	0	0	0	0	0	0	100

Table 3: Confusion matrix using DFT for single frame on MuHaVi data (83.1667%)

	W	Ju	K	C	W1
W	73.0769	7.0513	17.3077	0	2.5641
Ju	0	100	0	0	0
K	2.5424	26.2712	71.1864	0	0
C	0.7463	2.9851	3.7313	89.5522	2.9851
W1	1.6260	4.0650	3.2520	0	91.0569

Table 4: Confusion matrix using DFT for four appended frames on MuHaVi data (82.6667%)

	W	Ju	K	C	W1
W	75	7.5	15	0	2.5
Ju	0	93.75	6.25	0	0
K	0	28.5714	71.4286	0	0
C	0	3.0303	6.0606	90.9091	0
W1	0	9.0909	3.0303	0	87.8788

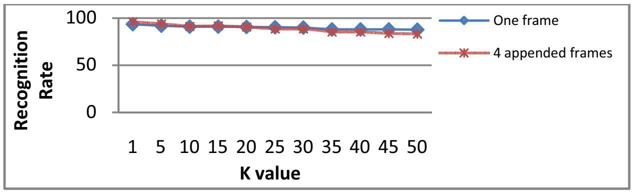


Figure 5: Effect of K values in K-NN on DA-IICT data

V. Conclusion

Improved adaptive Gaussian mixture model gives the best result in terms of speed and result quality both. [8] has used DFT of contour points but in this paper DFT of image itself is used. The proposed approach ensures efficiency and accuracy of the result. Here combination of frames is done for the trajectory motion. The K-NN classification technique is used for classification purpose due to its high performance with large databases as stated in [8].

There are some assumptions like single moving object in the video frames; object position and direction etc. Camera may capture the object in the various directions like front-view, back-view, side-view, top-view etc. Some of these assumptions (object position and direction) may be relaxed by using multiple cameras, kept in all viewing directions to capture front-view, back-view, side-view and top-view of the object. Human action recognition using multiple cameras may give more robust result in occlusion condition.

References

[1] C. Cedras and M. Shah, “Motion-based recognition: A survey”, Image Vis. Comput., Vol. 13, No. 2, pp. 129–155, 1995.

[2] S. Mitra and T. Acharya, “Gesture Recognition: A Survey”, IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews, pp. 311-314, Vol. 37, No. 3, May 2007.

[3] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, “Machine Recognition of Human Activities: A Survey”, IEEE Transactions on In Circuits and Systems for Video Technology, Vol. 18, No. 11, pp. 1473-1488, Nov. 2008.

[4] M. Ahmad and S. W. Lee, “Human Action Recognition Using Multi-view Image Sequences Features”, Proc. 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, pp. 523-528, April 2006.

[5] M. Ahmad and S. W. Lee, “HMM-based Human Action Recognition Using Multiview Image Sequences”, Proc. 18th IAPR/IEEE International Conference on Pattern Recognition, Hong Kong, China, Vol. 1, pp. 263-266, Aug. 2006.

[6] J. Alon, V. Athitsos, Y. Quan and S. Sclaroff, “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 31, No. 9, pp.1685-1699, Sept. 2009.

[7] M. B. Kaaniche and F. Bremond, “Tracking HoG Descriptors for Gesture Recognition”, Advanced Video and Signal Based Surveillance, In AVSS’09: 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 140-145, Genoa, Italy, Sept. 2009.

[8] H. Nassar, G. EL-Taweel and E. Mahmoud, “A Novel Feature Extraction Scheme for Gait Recognition”, International Journal of Images and Graphics, Vol. 10, No. 4, pp. 575-587, 2010.

[9] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking”, Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on, In Proc. CVPR, Vol. 2, pp. 246-252, Jun. 1999.

[10] Z. Zivkovic, “Improved Adaptive Gaussian Mixture Model for Background Subtraction”, In Proc. ICPR, Vol. 2, pp. 28-31, Aug. 2004.

[11] T. M. Cover and P. E. Hart, “Nearest Neighbor Classification”, IEEE Transactions on Information Theory, Vol. 13, pp. 21-27, Jan. 1967.

[12] K. Teknomo, “K-Nearest Neighbors Tutorial”, http:\\people.revoledu.com\kardi\ tutorial\KNN\

[13] MuHAVi-MAS: Multicamera Human Action Video and Manually Annotated Sihouette Data, REASON Project, http://dipersec.king.ac.uk/MuHAVi-MAS/.

[14] “Action database”, http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html, 2005.

[15] Advanced Computer Vision GmbH, “Motion detection video sequences”, http://muscle.prip.tuwien.ac.at/data here.php.