IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

RGB AND GRAY VIDEO ACTION DETECTION AND PREDICTION USING MATLAB

Vipin Kumar Batra*[1] & Mrs. Priyanka Gaur[2]
*[1]M. tech Scholar, E.C.E., Advanced Institute of Technology, and Management, AITM, Palwal
[2]Assistant Professor, E.C.E., Advanced Institute of Technology, and Management, AITM, Palwal

DOI: 10.5281/zenodo.1247285

ABSTRACT

We present a compact representation for human action recognition in videos using line and optical flow histograms.
We introduce a new shape descriptor based on the distribution of lines which are fitted to boundaries of human figures. By using an entropy-based approach, we apply feature selection to identify our feature representation, thus, minimizing classification time without degrading accuracy. We also use a compact representation of optical flow for motion information. Using line and flow histograms together with global velocity information, we show that high-accuracy action recognition is possible, even in challenging recording conditions.
This paper presents a novel feature descriptor for multi view human action recognition. This descriptor employs the region-based features extracted from the human silhouette. To achieve this, the human silhouette is divided into regions in a radial fashion with the interval of a certain degree, and then region-based geometrical and Hu-moments features are obtained from each radial bin to articulate the feature descriptor. A multiclass support vector machine classifier is used for action classification.

KEYWORDS: *computer visions; human action recognition; view-invariant feature descriptor; classification; support vector machines*.

I. INTRODUCTION

Human action recognition has gained a lot of interest during the past decade. From visual surveillance to human computer interaction systems, understanding what the people are doing is a necessary thread. However, making this thread fast and reliable still remains as an open research problem for the computer vision community.

In order to achieve fast and reliable human action recognition, we should first search for the answer of the question “What is the best and minimal representation for actions?”. While there isn’t a current “best” solution to this problem, there are many efforts. Recent approaches extract “global” or “local” features, either on the spatial or on temporal domain, or both. Gavrila present an extensive survey over this subject in [6]. The approaches in genereal, tend to fall into three categories. First one includes explicit authoring of the temporal relations, whereas the second one uses explicit dynamical models. Such models can be constructed as hidden markov models ([3]), CRFs [1], or finite state models [7].

In this paper, we show how we can make use of a new shape descriptor together with a dense representation of optical flow and global temporal information for robust human action recognition. Our representation involves a very compact form, reducing the amount of classification time to a great extent. In this study, we use rbf kernel SVMs in the classification step, and present successful results over the state-of-art KTH dataset [2].

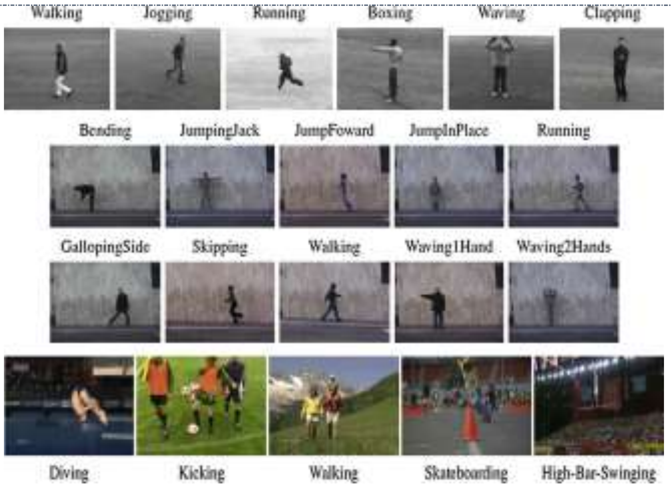

Figure 1: Sample frames from KTH (top), Weizmann (middle), and UCF Sports (bottom) human action datasets.

The major challenges and issues in HAR are as follows:

1) occlusion;
2) variation in human appearance, shape, and clothes;
3) cluttered backgrounds;
4) stationary or moving cameras;
5) different illumination conditions; and
6) viewpoint variations.

Among these challenges, viewpoint variation is one of the major problems in HAR since most of the approaches for human activity classification are view-dependent and can recognize the activity from one fixed view captured by a single camera. These approaches are supposed to have the same camera view during training and testing. This condition cannot be maintained in real world application scenarios. Moreover, if this condition is not met, their accuracy decreases drastically because the same actions look quite different when captured from different viewpoints [3]. A single camera-based approach also fails to recognize the action when an actor is occluded by an object or when some parts of the action are hidden due to unavoidable self-occlusion. To avoid these issues and get the complete picture of an action, more than one camera is used to capture the action—this is known as action recognition from multiple views or view-invariant action recognition [4].

II. RELATEDWORKS

This section presents state-of-the-art methods for multiview action recognition based on a 2D approach. These methods extract features from 2D image frames of all available views and combine these features for action recognition. Then, classifier is trained using all these viewpoints.

After training the classifier, some methods use all viewpoints for classification [10], while others use a single viewpoint for classification of a query action [12]. In both cases, the query view is part of the training data. However, if the query view is different than the learned views, this is known as cross-view action recognition. This is even more challenging than the multiview action recognition [9].

Different types of features—such as motion features, shape features, or combination of motionand shape-based features—have been used for multiview action recognition. In [8], silhouette-based features were acquired from five synchronized and calibrated cameras. The action recognition from multiple views was performed by computing the R transform of the silhouette surfaces and manifold learning. In [2], contour points of the human silhouette were used for pose representation, and multiview action recognition was achieved by the arrangements of multiview key poses. Another silhouette-based method was proposed in [13] for action recognition from multiple views; this method used contour points of the silhouette and radial scheme for pose representation. Then, model fusion of multiple camera streams was used to build the bag of key poses, which worked as a dictionary for known poses and helped to convert training sequences into key poses for a sequence-matching algorithm. In [13], a view-invariant recognition method was proposed, which extracted the uniform rotation-invariant local binary patterns (LBP) and contour-based pose features from the silhouette.

The classification was performed using a multiclass support vector machine. In [4], scale-invariant features were extracted from the silhouette and clustered to build the key poses. Finally, classification was done using a weighted voting scheme.

An optical flow and silhouette-based features were used for view-invariant action recognition in [6], and principal component analysis (PCA) was used for reducing the dimensionality of the data. In [3], coarse silhouette features, radial grid-based features and motion features were used for multiview action recognition. Another method for viewpoint changes and occlusion-handling was proposed in [2]. This method used histogram of oriented gradients (HOG) features with local partitioning, and obtained the final results by fusing the results of the local classifiers. A novel motion descriptor based on motion direction and histogram of motion intensity was proposed in [7] for multiview action recognition followed by a support vector machine used as a classifier. Another method based on 2D motion templates, motion history images, and histogram of oriented gradients was proposed in [8]. A hybrid CNN–HMM model which combines convolution neural networks (CNN) with hidden Markov model (HMM) was used for action classification [7]. In this method, the CNN was used to learn the effective and robust features directly from the raw data, and HMM was used to learn the statistical dependencies over the contiguous subactions and conclude the action sequences.

III. OUR APPROACH

3.1. Line-based shape features

Shape is an important cue for recognizing the ongoing activity. In this study, we propose to use a compact shape representation based on lines. We extract this representation as follows: First, given a video sequence, we compute the probability of boundaries (Pb features [13]) based on Canny edges in each frame. We use these Pb features rather than simple edge detection, because Pb features delineate the boundaries of objects more strongly and eliminate the effect of noise caused by shorter edge segments in cluttered backgrounds to a certain degree. Example images and their corresponding boundaries are shown in Fig 2(a) and Fig 2(b).

After finding the boundaries, we localize the human figure by using the densest area of high response Pb features. We then fit straight lines to these boundaries using Hough transform. We do this in two-fold; first, we extract shorter lines (Fig 2(c)) to capture fine details of the human pose. Second, we extract relatively longer lines (Fig 1(d)) to capture the coarser shape information.

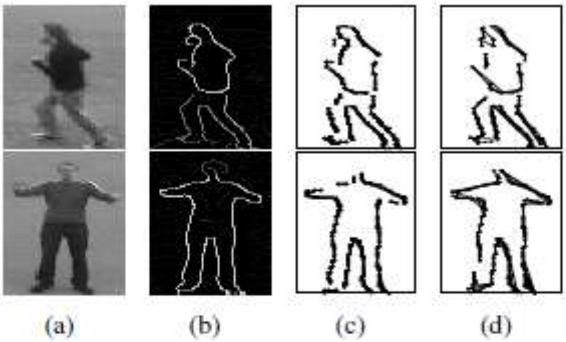

Figure 2. Extraction of line-based features

We then histogram the union of short and long line sets based on their orientations and spatial locations. The lines are histogrammed over 15◦ orientations, resulting in 12 circular bins. In order to incorporate spatial information of the human body, we evaluate these orientations within a $N \times N$ grid placed over the whole body. Our experiments show that $N = 3$ gives the best results (in accordance with [8]). This process is shown in Fig 2. Resulting shape feature vector is the concatenation of all bins, having a length $|Q| = 108$ where Q is the set of all features.

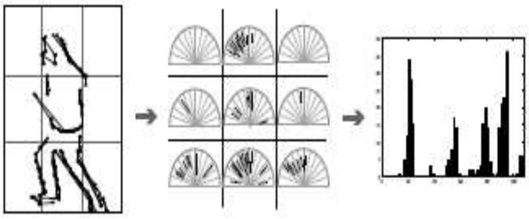

Figure 3. Forming line histograms

3.2. Feature Selection

In our experiments, we observed that, even a feature size of |*Q*| = 108 is a sparse representation for shape. That is, based on the nature of the actions, some of the dimensions of this feature vector are hardly used.

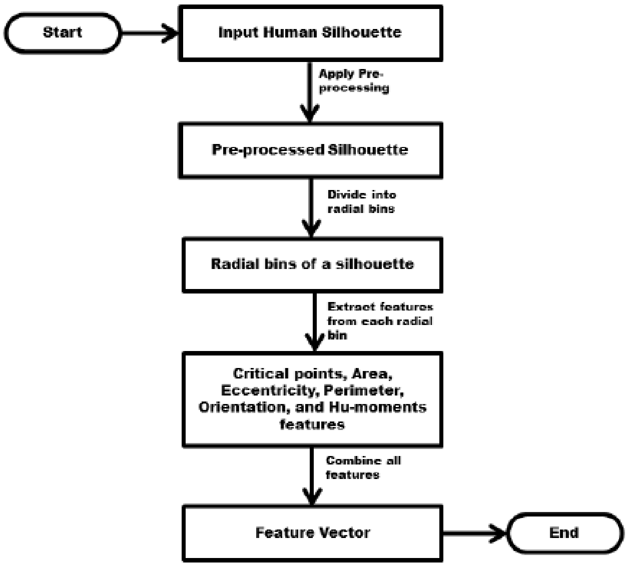
Start
Input Human Silhouette
Apply Pre-processing
Pre-processed Silhouette
Divide into radial bins
Radial bins of a silhouette
Extract features from each radial bin
Critical points, Area, Eccentricity, Perimeter, Orientation, and Hu-moments features
Combine all features
Feature Vector
End

Figure 4. Overview of feature extraction process

To have a more dense and compact representation and to reduce the processing time in classification step, we make use of an entropy-based feature selection approach. By selecting features with high entropy, we are able to detect regions of interest in which most of the change, i.e motion occurs.

We calculate the entropy of the features as follows: Let *fj*(*t*) represent the feature vector of frame at time *t* in video *j* and let |*Vj* | denote the length of the video. The entropy *H*(*fnj*) of each feature *n* over the temporal domain is

$$H(f_j^n) = -\sum_{t=1}^{|V_j|} \hat{f}_j^n(t) log(\hat{f}_j^n(t)) \qquad (1)$$

where ˆ *f* is the normalized feature over time such that

$$\hat{f}_j^n = \frac{f_j^n(t)}{\sum_{t=1}^{|V_j|} f_j^n(t)} \qquad (2)$$

This entropy *H*(*fnj*) is a quantative measure of energy in a single feature dimension *n*. A low *H*(*fnj*) means that the *n*th feature is stable during the action and higher *H*(*fnj*) means the *n*th feature is changing rapidly in the presence of action. We expect that the high entropy features will be different for different action classes. Based on this observation, we compute the entropies of each feature in all training videos separately for each action. More formally, our reduced feature set *Q*_ is

$$Q' = \{f^n | H(f_j^n) > \tau\,,\ \forall j \in \{1,..,M\}, n \in \{1,..,|Q|\}\} \tag{3}$$

where τ is the entropy threshold, M is the total number of videos in training set and Q is the original set of features. After this feature reduction step, our shape feature vector's length reduces to ~ 30. Note that for each action, we now have a separate set of features.

3.3. Motion features

Using pure optical flow (OF) templates increase the size of the feature vector to a great extent. Instead, we present a compact OF representation for efficient action recognition. With this intention, we first extract dense block-based OF of each frame, by matching it to the previous frame. We then form orientation histograms of these OF values. This is similar to motion descriptors of Efros *et al.* [5], however we use spatial and directional binning. For each *ith* spatial bin where $i \in \{1, .., N \times N\}$ and direction $\theta \in \{0, 90, 180, 270\}$, we define optical flow histogram $hi(\theta)$ such that

$$h_i(\theta) = \sum_{j \in B_i} \psi(\tilde{\mathbf{u}}_\theta \cdot \mathbf{F}_j) \tag{4}$$

where Fj represents the flow value in each pixel j, Bi is the set of pixels in the spatial bin i, $\mathbf{u}\tilde{}\theta$ is the unit vector in θ direction and ψ function is defined as

$$\psi(x) = \left\{ \begin{array}{ll} 0 & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{array} \right\} \tag{5}$$

This process is depicted in Fig 5.

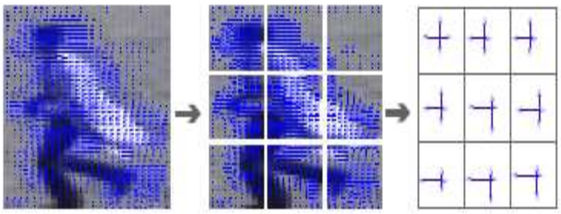

Figure 5. Forming of histograms

IV. CONCLUSION AND FUTURE WORK

4.1 Conclusion

A new motion description method named Shape Features-ME has been proposed to reduce the influence of environment variations for human activity recognition. Shape Features-ME is based on Shape Features features and inherits its advantages such as invariant to environment variations noise, illumination, and camera view angles. Thus, Shape Features -ME can be used as a robust method for motion description of activity recognition in the field of computer vision and pattern recognition. Shape Features -ME features are successfully utilized for describing and recognizing human actions in videos.

The experiment shows that Shape Features -ME outperforms optical flow, 3D Shape Features, and 2D Shape Features features for human activity recognition. Besides, some additional experiments are done to find the relations of recognition result with different GMM components. Shape Features features have the highest recognition rate with lowest GMM components which demonstrate that it is a better action description. On the other hand, compare Shape Features -ME and Shape Features translation, Shape Features -ME outperforms almost 3% accuracy, which means rotation information is also important for activity recognition. Shape Features -ME is another evolution step which improves Shape Features to interpret 2D transformation using a three dimensional vector.

4.2 Future Work

In the future, more research will be done with Shape Features -ME features for querying videos to detect a predefined set of actions such as walking, running, etc. This can be easily achieved by comparing patterns of action. Shape Features -ME is a good feature representation for motion, as long as finding similar motion patterns with queried one, the best match can be achieved and video retrieval can be realized. Shape Features -ME can be improved by using a better matching algorithm as well as incorporating 3D translation and rotation by considering multiple cameras.

As Shape Features -ME is an extension of Shape Features features to represent motion, it does not conflict with Shape Features feature. As well known that Shape Features can be used for object recognition with promising result, so there is possibility to combine Shape Features and Shape Features -ME to do object based activity recognition: recognition activity through the objects people interact with. For example, pick up a gun and pick up an apple are different actions in detail and results in different handling response actions to other people. Recognize activities through objects will extend the application of activity recognition in many fields and increase the capacity of activity recognition.

V. REFERENCES

[1] Xue Li, , Vasu D. Chakravarthy, , Bin Wang, and Zhiqiang Wu, "Spreading Code Design of Adaptive Non-Contiguous SOFDM for Dynamic Spectrum Access" in IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 5, NO. 1, FEBRUARY 2011

[2] J. D. Poston and W. D. Horne, "Discontiguous OFDM considerations for dynamic spectrum access in idel TV channels," in Proc. IEEE DySPAN, 2005.

[3] R. Rajbanshi, Q. Chen, A.Wyglinski, G. Minden, and J. Evans, "Quantitative comparison of agile modulation technique for cognitive radio tranceivers," in Proc. IEEE CCNC, Jan. 2007, pp. 1144–1148.

[4] V. Chakravarthy, X. Li, Z. Wu, M. Temple, and F. Garber, "Novel overlay/underlay cognitive radio waveforms using SD-SMSE framework to enhance spectrum efficiency—Part I," IEEE Trans. Commun., vol. 57, no. 12, pp. 3794–3804, Dec. 2009.

[5] V. Chakravarthy, Z. Wu, A. Shaw, M. Temple, R. Kannan, and F. Garber, "A general overlay/underlay analytic expression for cognitive radio waveforms," in Proc. Int. Waveform Diversity Design Conf., 2007.

[6] V. Chakravarthy, Z. Wu, M. Temple, F. Garber, and X. Li, "Cognitive radio centric overlay-underlay waveform," in Proc. 3rd IEEE Symp. New Frontiers Dynamic Spectrum Access Netw., 2008, pp. 1–10.

[7] X. Li, R. Zhou, V. Chakravarthy, and Z. Wu, "Intercarrier interference immune single carrier OFDM via magnitude shift keying modulation," in Proc. IEEE Global Telecomm. Conf. GLOBECOM , Dec. 2009, pp. 1–6.

[8] Parsaee, G.; Yarali, A., "OFDMA for the 4th generation cellular networks" in Proc. IEEE Electrical and Computer Engineering, Vol.4, pp. 2325 - 2330, May 2004.

[9] 3GPP R1-050971,"R1-050971 Single Carrier Uplink Options for EUTRA: IFDMA/DFT-SOFDM Discussion and Initial Performance Results ",http://www.3GPP.org,Aug 2005

[10] IEEE P802.16e/D12,'Draft IEEE Standard for Local and metropolitan area networks-- Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems', October 2005

[11] 3GPP RP-040461, Study Item: Evolved UTRA and UTRAN, December 200

[12] R. Mirghani, and M. Ghavami, "Comparison between Wavelet-based and Fourier-based Multicarrier UWB Systems", IET Communications, Vol. 2, Issue 2, pp. 353-358, 2008.

[13] R. Dilmirghani, M. Ghavami, "Wavelet Vs Fourier Based UWB Systems", 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, Sep. 2007.

[14] M. Weeks, Digital Signal Processing Using Matlab and Wavelets, Infinity Science Press LLC, 2007.

[15] S. R. Baig, F. U. Rehman, and M. J. Mughal, "Performance Comparison of DFT, Discrete Wavelet Packet and Wavelet Transforms in an OFDM Transceiver for Multipath Fading Channel,", 9th IEEE International Multitopic Conference, pp. 1-6, Dec. 2005.

[16] N. Ahmed, Joint Detection Strategies for Orthogonal Frequency Division Multiplexing, Dissertation for Master of Science, Rice University, Houston, Texas. pp. 1-51, Apr. 2000.

CITE AN ARTICLE

Batra, V. K., & Gaur, P., Mrs. (2018). RGB AND GRAY VIDEO ACTION DETECTION AND PREDICTION USING MATLAB. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY,7*(5), 412-417.