Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 6, June -2015

Automated Human Action Recognition Using Machine Learning

Anuradha Maid[1],Asst.Prof. Siddharth B. Bhorge[2]

[1]E&TC Dept., Vishwakarma Institute of Technology, Pune
[2]E&TC Dept., Vishwakarma Institute of Technology, Pune

Abstract — *Human action recognition is one of the most important technology in computer vision . Human action recognition is the process of labeling image sequences with action labels. There are many methods of human action recognition are exist. The human action recognition faces many problems such as intensity variations, dynamic background, aging ,partial occlusion . We first collects the dataset (using KTH dataset) which is having number of classes and video sequences in real time. Interest points and cuboid extraction is done by Dolla'r et al method. Discrete wavelet transform and PCA are used to increase accuracy of the system. It has wide application such as security (pedestrian detection),surveillance (behavior analysis), , control (human-computer interfaces), content based video retrieval, patient monitoring system etc*

Keywords- *STI'Point cuboids, DWT ,PCA, Classification.*

I. INTRODUCTION

In the traditional video surveillance systems, the video captured would be displayed on monitor in control room and it requires continuous attention to monitor the video for any abnormal activities. But due to increasing terrorist attacks and other criminal issues the demand for video surveillance systems for getting more accuracy. For recording daily activities people using digital cameras nowadays and this brings the improvement of video sources on the internet, and also causes the problems of how to classify newly generated video sequences according to their action classes and how to categorize existing video sources. Distributing these videos for processing is a time-consuming if it is done manually. To overcome this problem we used automated video surveillance system.

A computer automatically tells us what is happening in the scene and it identify different human actions and explore the problem of human action categorization in video sequences. Our interest is to design an algorithm that permits computer to learn models for human actions. Then, from novel video, the algorithm should be able to decide which human action is present in the sequence. Furthermore, we look for means to provide a rough indication of where the action is being performed.

Its challenging problem is actions can be performed by subjects of different size, appearance and pose. The problem is compounded by the inevitable occlusion, illumination change, shadow, and camera movement. Action recognition is combination of feature extraction and classification of image representation. Video recognition still requires some improvement, specifically for movies due to variation present in scene like clothing, posture, dynamic background etc.

In our system action is recognized by using features like interest points and cuboids. The project will be used mainly for automated intelligent surveillance systems or applications include behavior recognition and content based searching of videos like healthcare autonomous robotic systems.

II. RELATED WORK

Lots of work have been done in human action recognition and localization to save manual effort and to increase processing efficiency. Existing human action recognition methods can be broadly classified into: flow based , spatio-temporal shape template based , interest points based , and tracking based. Flow based approaches construct action templates based on the optical flow computation [4, 5]. Spatio-temporal shape template based require highly detailed silhouettes to be extracted, which may not be possible given a real-world noisy video input. The computational cost of space-time volume based approaches is not acceptable. Tracking based approaches [16, 1, 15, 19] mostly fail on a noisy dataset such as the KTH dataset, which is featured with strong shadows, low resolution and camera movement provides clean silhouette extraction not possible.

Schuldt et al. [11] tells to represent action using 3-D space-time interest points detected from video. The detected points are clustered to form a dictionary of video-words and the sequence of action is represented by Bag of Words. Zhang et al. [20] proposed the concept of motion context to capture both spatial and temporal distribution of video-words.

To utilize the global information of the subject subtracted from video data, for example the correlating optical flow measurements from low-resolution videos proposed by Efros et al. [4],which segment and stabilize each human figure and annotate each action in the resulted spatial temporal volume. To track the body parts and use the motion

trajectories to discriminate different actions defined by the author A. Yilmaz and Y. Song [6]. Certain feature points are located in a frame-by-frame manner, and the tracks of these points show many discriminative properties, such as position, velocities, and appearance. However, the methods mentioned by Efros et al. and A. Yilmaz et al. are sensitive to partial occlusion and use much unnecessary information that is computationally expensive.

The space-time interest point detectors proposed by Laptev [8] and Dolla´r [1]. Laptev [8] propose a space-time interest point detector to detects local structures in space-time where the image values have significant local variations in both space and time. Dolla´r et al. [1] proposed the system which gives set of separable linear filters detecting interest points with strong motion and respond to complex motion of local regions and the space-time corners. It detects large number of interest points than Laptev's approach, which makes it more reliable with limited frames.Motion history images that capture motion and shape to represent actions proposed by Bobick and Davis. It introduced the global descriptors *motion energy image* and *motion history image*, which matched to stored models of known actions. Blank et al. represent actions as space-time shapes and extract space-time features for action recognition, Similarly, this approach relies on the restriction of static backgrounds which allows them to segment the foreground using background subtraction.

III. PROPOSED SYSTEM

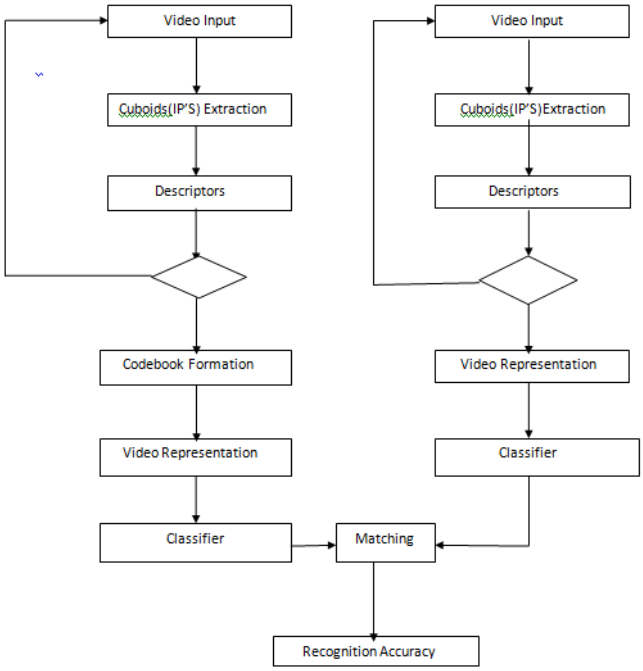

Fig.1 Action Recognition Framwork.

The human action recognition process divided into two phases: training and testing. In training phase as shown in fig. the interest points and cuboids are extracted around the interest points. The descriptors gives structural distribution of interest points. after that descriptors from all sequence are gathers together for clustering which uses Euclidean distance as a clustering matrix. The centers of cluster represented as video words and forms the codebook. The codebook of each feature descriptor is utilized to create model representing the characteristics of each class of training sequence.

In testing sequence, process is same to extract interest points, build descriptors, and assign codebook as those during training phase. Then SVM and K nearest neighbor(KNN) classifiers are used to classify probable action type according to training sequence model, and calculate accuracy of recognition.

3.1 Interest point Detection:

Interest points are local spatio-temporal features descriptive of the action captured in a video which will be described by variations in intensity .interest point features obtained from local video patches .Interest points providing a rich description and powerful representation of human actions. various interest point detection methods, the Dollar et al. [3] is most widely used for action recognition. The intensity variations in the temporal domain are detected using Gabor filtering. The interest point detection in the spatial domain is based on the detection of corners, such as [8, 9]. Corners are defined as regions where the local gradient vectors point in orthogonal directions.

The response function is as follows,

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (1)$$

Where g(x,y:σ) is the 2D Gaussian smoothing kernel, h_{ev} and h_{od} are quadrature pair of 1D Gabor filter applied temporally.

Where $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and

$$h_{od} = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$$

2D Gaussian smoothing filter is applied on to the spatial dimensions. Interest points of 2D images used for image matching and retrieval, recognition, classification and 3D images is STIP which is usually used for activity or event recognition. The response function is depends upon τ and σ.The space-time interest points are extracted around the local maxima of the response function.

A cuboid contains the spatio-temporally windowed pixel values. the information which contains in cuboid is used to form descriptor for training set. The size of cuboid is near about six times the scale at which they were detected. The extraction of cuboids minimizes preprocessing steps such as foreground subtraction, figure tracking and alignment, etc

3.2 Features descriptor method:

Once spatio temporal interest points are detected, description methods are applied on cuboids to extract information which are present in it. There are various methods of descriptors such as 3D gradient ,transform base descriptors. In this paper discrete wavelet transform is used.

3.2.1 Discrete Wavelet Transform:

DWT decomposes a discrete signal into a set of discrete basis functions (wavelets) and it captures both the frequency information and the space information .among all transform techniques, Wavelet Transform has the best performance by localizing in frequency and also in space. The used of wavelets in detecting local properties of images in content based image retrieval. the main advantage of wavelet transform is the removal of redundancy between neighboring pixels. Discrete Wavelet Transform is a multi-resolution de-compositions that can be used to analyze signals and images.[21]

A set of wavelets can be derived from $\gamma(x)$,

$$\gamma_{a,b}(x) = \frac{1}{\sqrt{a}}\ \gamma\left(\frac{x-b}{a}\right), (a,b \in R, a>0) \quad (2)$$

Where a is dilation parameter and b is translation parameter.

Wavelet transform is its symmetric nature that is both the forward and the inverse transform has the same complexity, building fast compression and decompression routines and very good energy compaction capabilities, high compression ratio etc.

3.3 Codebook formation:

With descriptors PCA is changes original dimensions of descriptors to low space. To create codebook, k means clustering is used to assign low dimensional descriptors to nearest clusters so that to minimize overall distortion. codebook is group of features which required for classify the objects. The size of codebook is an input parameter for K-means clustering.

3.4 Classification method:

The K-NN algorithm is a classification method based on the K (predefined constant), closest training data in the feature space. A vector is classified to one label which is the most frequent label among K nearest training vectors. K-NN classification decision is based on K neighborhood vectors, therefore K-NN can be easily used in multi-modal classification. K-NN is a simple model with few parameters and the computation time for testing phase is independent of the number of classes. Here I used DWT of the small image blocks as feature selection, and apply K-NN as the classifier for human action recognition.

Here used the KNN classifier for multiple actions classifications as Hand Waving, Clapping, Boxing, Running, Walking, Jogging. Which gives much more efficiency than other Classifiers. KNN classifiers are adopted to classify each testing sequence to the most probable action type according to the model built in the training phase, and the correctly classified sequences against all sequences give the more recognition accuracy.

IV. EXPERIMENTAL RESULTS

4.1KTH Dataset:

It contains 600 video sequences and each video has only one action. The dataset consists of six actions (Boxing, Running, Walking, Handclapping, Hand waving, and Jogging) performed by 25 subjects in different scenarios[1]. Categorized into two groups: one is hand motions while the other group is leg motions. Fig. shows the KTH dataset.

Some motions are difficult to recognize such as running and jogging for that we follows training and testing dataset division. For each sequence, only the first 300 frames are selected because actions are performed periodically.

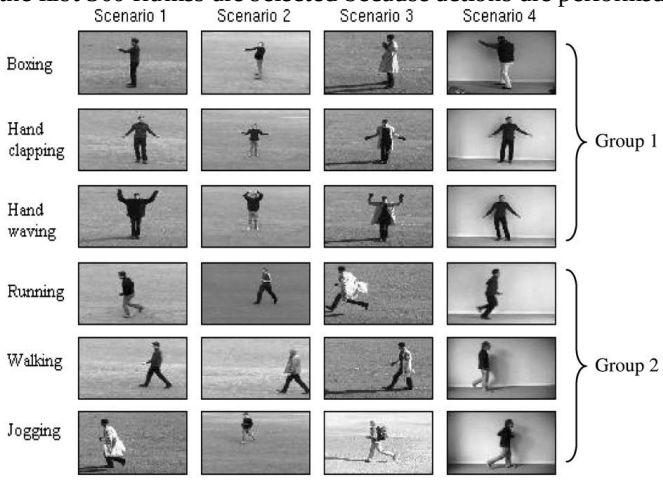

Fig.2 KTH data for each action

The proposed system is implemented using MATLAB 8.2.0. All testing is performed on an Intel (R) core (TM) i3 CPU @2.40 GHz PC with 2GB of RAM running the Windows 7 operating system. Each component of the system is tested individually to accurately analyze.

Firstly the training is done for the different actions by extracting the ST Features with using DWT as feature Descriptor and followed by PCA for dimension reduction. And generated the final codebook Array for the input video stream. And in testing phase the the same procedure as in training is followed and K-NN classifier is used to classify the different actions. As shown in the figures.

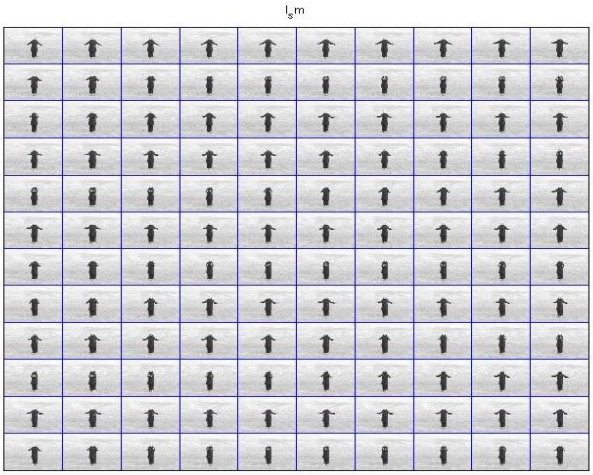

Fig.3 Illustration of the location of the extracted spatial temporal interest points for action *'Hand Waving'*

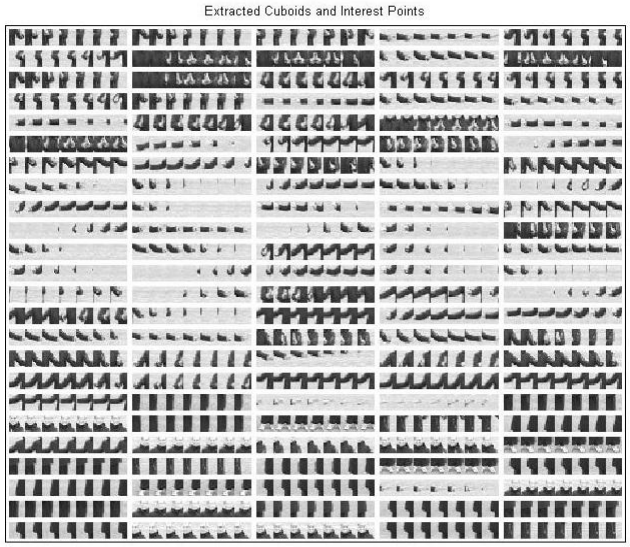

Fig.4 Cuboids extracted for the Hand Waving action

V. CONCLUSION

The experimental results gives more reliability and efficiency in human action recognition, and also indicate the potential of transformation techniques in spatiotemporal video event analysis. Modeling global spatial and temporal distribution of interest points for more accurate and robust action recognition. Our model is able to capture smooth motions, robust to view changes and occlusions, and with a low computation cost. Our experiments on the KTH datasets which contains different actions. One way in raising accuracy is to combine some structural information of each action by means of descriptor fusion. We can improve in the performance of clustering methods and classification algorithms are also desirable.

Acknowledgment:

I am very thankful to Electronics and Telecommunication Department of VIT, Pune, respected H.O.D. Prof. Chopde and all the professors who helped us. It is with great reverence that I wish to express my deep gratitude towards Prof. S.B.Bhorge for him astute guidance, constant motivation and trust, without which this work would never have been possible

REFERENCES

[1] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie. Behavior recognition via sparse spatio-temporal features, in: Proceedings of the Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[2] Ling Shao, Ruoyun Gao, Yan Liu, Hui Zhang, "Transform based spatio-temporal descriptors for human action recognition," Neurocomputing 74 (962–973), 2011

[3] S. Savarese, A. D. Pozo, J. Niebles, and L. Fei-Fei. Spatialtemporal correlations for unsupervised action classification. In IEEE Workshop on Motion and Video Computing, 2008

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In ICCV, pages 726–733, 2003.

[5] J.C. Niebles, H. Wang, Fei-fei Li, Unsupervised learning of human action c-ategories using spatial-temporal words, International Journal of Computer Vision 78 (2008) 299–318.

[6] A. Yilmaz, M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 150–157.

[7] Y. Song, L. Goncalves, P. Perona, Unsupervised learning of human motion, IEEE Transactionson Pattern Analysis and Machine Intelligence 25 (2003) 814–827.

[8] I. Laptev, On space-time interest points, International Journal of Computer Vision 63 (2–3) (2005) 107–123.

[9] Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), 257–267.

[10] Y. Ke, R..Sukthankar, M. Hebert. Efficient visual event detection using volu- metric features, in: Proceedings of the International Conference on Computer Vision, 2005, pp. 166–173.

[11] C. Sch¨uldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In ICPR, volume 3, pages 32–36, 2004.

[12] Matteo Bregonzio, Shaogang Gong and Tao Xiang, Recognising Action as Clouds of Space-Time Interest Points, IEEE, pages 1948-1955, 2009.

[13] C. Rao and M. Shah. View-invariance in action recognition. In CVPR, volume 2, pages 316–322, 2001.

[14] W. Forstner and E. G¨ulch. A fast operator for detection and precise location of distinct points. In Intercommission Conf. on Fast Processing of Photogrammetric Data, pages 281–305, Switzerland, 1987.

[15] J. Liu and M. Shah. Learning human actions via information maximization. In CVPR, 2008.

[16] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. PAMI, 29(12):2247–2253December2007.

[17] M. Vetterli, C. Herley. Wavelets and filter banks: theory and design. *IEEE Transactions on Signal Processing*. Vol. 40, pp. 2207-2232, 1992

[18] S. Wong, T. Kim and R. Cipolla. Learning motion categories using both semantic and structural information. *In Procedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp.1-6, 2007

[19] Z. Xiong and K. Ramchandran, "Wavelet image compression," in *Handbook of Image and Video Processing*, A. Bovik, Ed., 2nd ed. New York: Academic, 2005, ch. 4–5.

[20] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: A new representation for human action recognition. In ECCV, volume 4, pages 817–829, 2008.

[21] T. H. Koornwinder. Wavelets: an elementary treatment of theory and applications. *World Scientific Publishing, New Jersey*. pp. 49-80, 1993.

