Human Activity Recognition In Videos

Suvam Bag, Department of Computer Engineering
Rochester Institute of Technology
Rochester, NY, 14623
Sourabh Kulhare, Department of Computer Engineering
Rochester Institute of Technology
Rochester, NY, 14623

I. Introduction

Scene understanding forms one of the most difficult and researched topics in the field of Computer Vision. Action recognition obviously forms an integral part in this task. However the procedures and the complexity of that is not non-trivial or easy in any way. Classification of videos based on different actions performed in them is very difficult and important in the applications of Computer Vision in the real world. For example, if different videos taken from a surveillance camera at different instances of time were to be classified in order to derive some important information out of them, the ideology of scene understanding would play an important role in it. Moving on from normal databases of 2-dimensional images, this paper introduces some innovative approaches to apply the same concepts but instead on 3-dimensional videos. Since any video is actually a combination of a number of frames taken over a fixed period of time, the third dimension is time. We used a specific database called KTH from Royal Institute of Technology which has six classes of human motions - walking, jogging, running, boxing, hand moving, hand clapping. Since some of these actions are very closely related, the importance of good classification is very important from this aspect. The key reason behind the success of this paper is the use of Harris Corner detector in detecting the interest points for the 3-dimensional (3D) Scale Invariant Feature Descriptor (SIFT). Extending the concept of normal 2D SIFT descriptor to 3 dimensional space, the 3D SIFT descriptor works on each frame of the video. Hence time is the third dimension here. These descriptor points were converted to a bag of words approach with the help of k-means clustering. Finally the segmented results forms a histogram which can be used for classification. We used a multiclass Support Vector Machine (SVM) as our classifier. To improve our results on the existing research papers, a majority voting technique was also applied based on a user defined threshold for true positives to improve the accuracy results. Before presenting the details of the methods discussed, we would like to summarize the novel contribution of our paper. They are -

· Harris Corner Detector for detecting interest points
· Forming the 3D SIFT descriptor on these points
· Create bag of words
· Implement SVM classifier

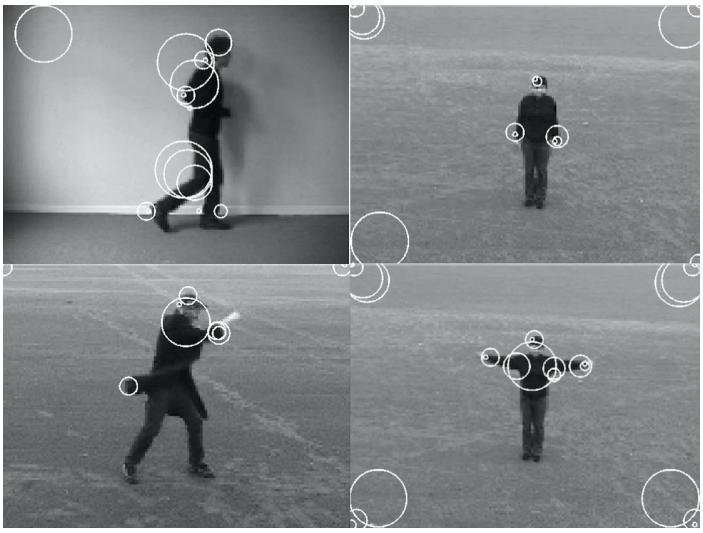

Fig. 1. Interest Points detection

II. Related Work

Human actions often convey the essential descriptive information in videos. There has been plenty of work done in this area but still recognizing human actions in constrained videos and specially in real time videos is one of the challenging problems in Computer Vision. In [10], an unconventional approach was used. Author adds motion characteristics explicitly and categories motion into two parts: dominant motion and residual motion. Author also uses motion descriptor DCS(Divergence, Curl, Shear) for representation of videos. Paper [3] extends the notion of spatial interest points into the spacial temporal domain and shows how the resulting features often reflect interesting events that can be used as informative representation of videos. [4] Uses SURF descriptor and dense optical flow for action recognition. Paper [9] uses a global video descriptor approach. Global descriptor is computed by applying 3-D spatio-temporal filters on the frequency spectrum of a video sequence. In [7], an 'action bank' approach has been applied. Action bank is fundamentally a set of bases of high dimensional action space combined with a simple linear classifier. In this project we used [2] 3-D SIFT approach. SIFT [11] is one of the most accurate and robust descriptor. Its scale invariant nature makes it more suitable for videos. For this

project we used extension of SIFT in third dimension, which is time.

III. Method

Representation of video is just an extension of another dimension in image representation. Here we used the same principle to represent videos. We converted every video files into MAT-files and further attached 3-D matrix representation of video with it. This approach is fast enough to access and process videos. Progressively we calculated interest points on selected frames with the help of Harris interest points detector [12] as shown in Fig. 1. It gives the interest points in the spacial domain within every frame. We accumulated these points and then considered these points our interest points where we want to calculate 3D SIFT [2] descriptors.

$$m_{2D}(x,y) = \sqrt{L_x^2 + L_y^2}, \quad \theta(x,y) = \tan^{-1}(\frac{L_y}{L_x}). \quad (1)$$

SIFT first computes the overall orientation of the neighbourhood and then it creates the sub-histograms. SIFT gradient magnitude and orientation is defined in equation 1, where Lx and Ly are computed using finite difference approximation. Similarly in 3D these computations are done with considering time domain also, extended equations are as follows:

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2}, \quad (2)$$

$$\theta(x,y,t) = \tan^{-1}(L_y/L_x), \quad (3)$$

$$\phi(x,y,t) = \tan^{-1}(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}). \quad (4)$$

It was observed from fig. 2 that 3D computation includes one extra angle. Consequently each pixel has now two values representing the direction of gradients. Combination of these angles also increases the final number of dimensions of descriptor.

Interest points for each video and 3-D matrix representation of video are primary inputs of 3D SIFT . It gives 50 descriptors for every input video. Dimension of each descriptor is 640. After calculating these descriptors for every videos from every class we accumulated them as number of samples and process these samples for Kmeans clustering process. Kmeans clustering provides number of clusters into which we can map these descriptors. Sequential collection of 50 descriptors represents one video. We created cluster histogram for each video based on the values of cluster index of 50 descriptors of video. Each bar in histogram represents number of descriptors belong to that cluster. These histograms is another representation of video in cluster domain. Each histogram is called a "Signature histogram" of video. Finally all the histograms were used for classification. We used SVM classifier. We used one vs all classification concept with binary SVM to do the multi-class classification. We also used cross validation with our data set to get better results.

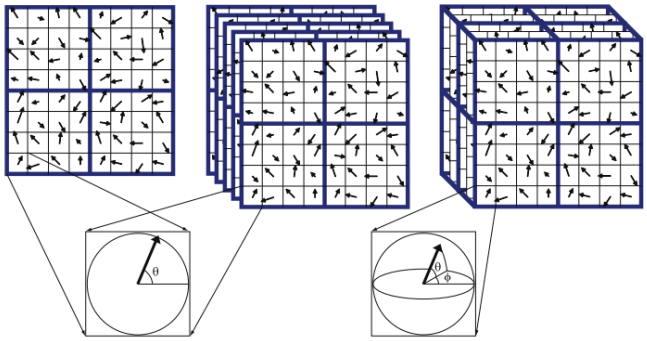

Fig. 2. Extension of SIFT in 3D [2]

IV. Experimental Setup

Target dataset is KTH [1] action database. Database contains 2391 video sequences with length of four seconds in average. It contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping). Every activity is being performed into four different scenarios: outdoors, outdoors with scale variation, outdoor with different clothes and indoors. All sequences were taken over homogeneous backgrounds with a static camera with 25 fps frame rate. Resolution of sequences is 160x120 pixels.

We used MATLAB platform for this paper. Interest point calculation is computationally intense specially with videos. So instead of calculating interest points on each frames, we calculated interest points from every 15th frame to save the computation and to include the full flow of video. Then the best ten interest points from each frame were collected to create interest point collection. As 3D SIFT provides descriptors we pass all of these descriptors for clustering. After many experiments we got a good accuracy for 200 clusters. Following this, classification was done by SVM. Implementation of SVM is done by one vs all scheme. We also used cross validation for better accuracy. We partitioned the data set into 80 % training

Fig. 3. Six classes of human actions [1]

and 20 % testing and changed this partition for every validation turn. Number of descriptors, number of clusters, validation partitioning these are the design parameters those can affect the accuracy.

V. Result

Our main objective was to classify human actions in KTH [1] data set. Actions involve boxing, clapping, walking, running, jogging and hand waiving. Some of the actions in this database are very closely related actions like jogging, running, waiving and clapping. This data set is a very popular benchmark used in many action recognition papers. The confusion matrix for this experiment is shown in fig. 4.

	Boxing	Clapping	Jogging	Running	Waiving	Walking
Boxing	0.90					0.10
Calpping		0.70			0.20	0.10
Jogging			0.60	0.30		0.10
Running			0.22	0.666		0.111
Waiving					0.80	0.20
Walking			0.10			0.90

Fig. 4. Confusion matrix with KTH dataset.

It was observed that better results were obtained with most of the actions like boxing, walking and waiving. Jogging and running are quite similar actions and correlate with each others. Better accuracy can be achieved with the co-occurrence based fusion of bag of words. This approach is described in discussion section. Use of cross validation significantly improved the accuracy of classification.

VI. Discussion

As discussed in the previous section, these accuracies are sensitive to design parameters. During the progression of this paper we experimented with all design parameters. We used only 50 descriptors per videos for less computation, more number of descriptors would have definitely improved our accuracy. Our existing method to select interest points is also very useful and more descriptive interest points. According to our experiments number of clusters also play a significant role in accurate classification. In our case, number of clusters less than 200 and more than 2000 always came up with unstable results. We also experimented with PCA (principle component analysis); we went to the higher number of clusters and then experimented with PCA to reduce the number of clusters but it didn't create valuable impact. In [2], good results are achieved with considering the co-occurrence nature of cluster words. First spatio-temporal word vocabulary was created by K-means clustering process. 3-D SIFT descriptors from the videos are matched to each vocabulary words to create signature histogram representation of videos. [2] then did feature groupings of histograms based on the co-occurrence nature of vocabulary words. A co-occurrence matrix is created to signify that how many times a particular word occur with all other words. Row vectors of this matrix represent contextual distribution of word in terms of other words of vocabulary. Similar contextual distribution vector of any two words signifies that these two words occur more often together. Measure of similarity is done by correlation, if the correlation of two vectors are above than a particular threshold author joins those two words and add their corresponding frequency counts from initial histograms and thus creates new histograms for videos. This approach can be very useful to deal with bag of words model. It optimally comes up with number of words those represent the descriptors more accurately. In future we will try to implement this approach with larger number of descriptors. We will also try to implement this with convolution neural network.

References

[1] I. Laptev, C. Schuldt and B. Caputo *Recognition of human actions* Proc.ICPR'04, Cambridge, UK

[2] P. Sconanner, S. Ali and M. Shah *A 3D Dimentional SIFT Descriptor and its application*, Computer Vision Lab, UCF.ACM MM 2007

[3] I. Laptev, *On time-space interest points* IJCV, 64(2/3): 107-123, 2005.

[4] Wang, A. Klaser, C. Schmid, and C. -L.Liu *Action recognition by dense trajectories.* In CVPR, 2011

[5] A. Klaser, M. Marszalek and C. Schmid. *A spatio-temporal descriptor based on 3D-gradients.* In BMVC, 2008

[6] M.A. Naiel, M.M. Abdelwahab, M. Elsaban, and W.B Mikhael, *Simultaneous Human detection and action recognition employing 2-DPCA-HOG.* 2011

[7] S. Sadanand, J. Corso *Action Bank: A high-level representation of activity in video.* In CVPR, 2012

[8] C. Sch , L. Barbara *Recognizing Human Actions-A Local SVM Approach.* In ICPR, 2004

[9] B. Solmaz, S. Assari and M. Shah *Classifying web videos using a global video descriptor.* MVAP-D-12-00244

[10] M. Jain, H. Jegou, P. Bouthemy, *Better exploiting motion for better action recognition.* In CVPR, 2013.

[11] D.G. Lowe, *Distinctive Image Features from Scale Invarient Keypoints.* In IJCV, 2004

[12] K. Mikolajczyk, and C. Schmid *Scale Affine Invariant Interest Point Detectors.* In ICPV 2004