Human Action Recognition: A Literature Review

Archamol P S
Department of Computer Science
College of Engineering Kidangoor
Kottayam, Kerala India

Nisha C A
Department of Information Technology
College of Engineering Kidangoor
Kottayam, Kerala India

***Abstract*— Understanding human action in videos has received significant research attention in the field of video analysis. Most applications are in summarization, video content retrieval, and human-computer interfaces. Most existing method require manual annotation of relevant portion of action of interest. Human action recognition can be made more reliable without manual annotation of relevant portion of action of interest. . This paper presents not only an update extending previous related surveys, but also focuses on a joint learning framework that identify the temporal and spatial extent of action in videos. Dense trajectories are used as local features to represent the human action .It is more fine grained .Action localization is made by learning the temporal and spatial extents of video. Split and merge algorithm allows the segmentation which is followed by training the video.**

Keywords*— Human action Recognition,Dense Trajectory,Spatial and Temporal extent.

I. Introduction

Human action detection in videos are emerging topics in computer vision since understanding the human action helps in management, summarization and retrieval of videos .In the past two decades significant progress has been made with the invention of local invariant features and bag of features representation. The task is challenging due to variations in action performance, background settings and inter-personal differences. To understand the action in the video ,there are two things to be noted- Action recognition and action localization. One is 'what action' is performed in the video and the other is 'where the action' of interest is taken place.

The problem of assigning videos into several predefined action classes is known as action recognition and action localization is finding the spatio-temporal content of the video. In action recognition we are training the videos into several classes. Training includes both positive and negative samples. After the training phase we can test the videos. In the existing systems, in the training phase we have to manually annotate the relevant part of the action of interest in the video. Manual annotation is tedious, time consuming process and error prone process. So here we introduce an automatic method for finding the relevant portion of the action of interest in the video without human intervention. . Different from previous approaches it does not require reliable human detection and tracking as input .There are also such action detection methods that identify which video contents are occupied by the performer of action. introduces a method to identify the temporal extent of the video. But it ignored the spatial context. Ignoring one domain may produce irrelevant content from that domain.

The proposed method introduces a joint learning framework for finding the spatial and temporal extent of the action. So that it is easy to find the relevant portion of action in the video and thereby easily recognizing the video. The person location is inferred as latent variables. Temporal smoothness is also enforced along with learning the spatial model. Trajectories are extracted from the video to represent the action. Using dense trajectories for representing the video is more fine grained because it is at pixel-level accuracy than single media based solutions.

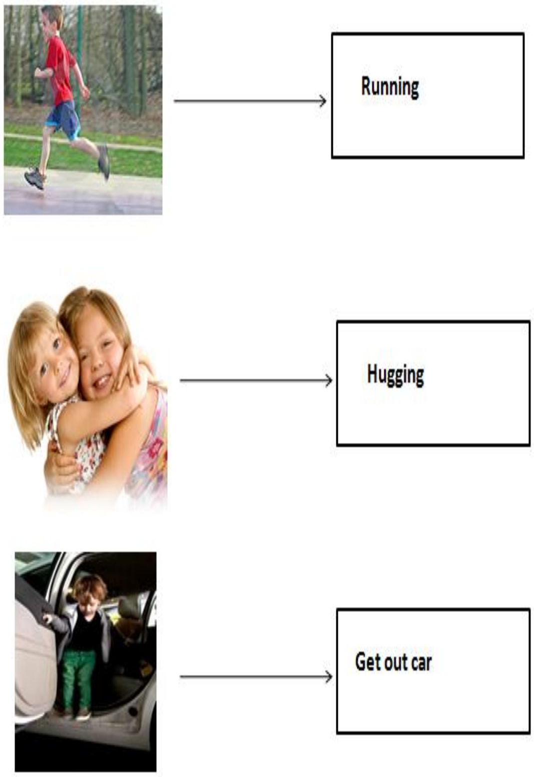

Fig 1 : action detection results. The second column shows the name of the action detected from the frames of the videos.

II. Literature Review

For many years human action recognition has been studied well. Most of the action recognition methods require to manually annotate the relevant portion of the action of interest in the video. In recent years it has been studied that the relevant portion of action of interest can be found out automatically and recognize the action. We can review the action recognition methods

A. Action Recognition

For representing video, feature trajectories have shown efficiency. But the quality and quantity of these trajectories were not sufficient. As the use of dense sampling came popular for image classification Wang *et al.*[1] proposed to use dense trajectories for representing videos. Dense points from each frame are sampled and traced them based on displacement information. For improving the performance Wang *et al.*[2] takes into account the camera motion. The camera motion is estimated by matching feature points between the frames by using SURF descriptors and dense optical flow. Another approach [3] aimed at modeling the motion relationship. The approach operates on top of visual codewords derived from local patch trajectories, and therefore does not require accurate foreground-background separation .Dorr *et al.*[4] proposed another method for finding the informative regions. They used saliency mapping algorithms. As a new method this paper proposes using a joint learning framework for learning spatial and temporal extents of action of interest

B. Action Detection

Recognition was performed using the Mahalanobis distance between the moment description of the input and each of the known actions. Recent popular methods which employ machine learning techniques such as SVMs and AdaBoost, provide one possibility for incorporating the information contained in a set of training examples.[4] introduces the Action MACH filter, a template-based method for action recognition which is capable of capturing intra-class variability by synthesizing a single Action.

Another method is proposed in [5], multiple-instance learning framework, named SMILE-SVM (Simulated annealing Multiple Instance Learning Support Vector Machines), is presented for learning human action detector based on imprecise action location. Wang *et al.* [6]used a figure-centric visual word representation. In that localization is treated as latent variable so as to recognize the action. A spatio-temporal model is learned .During the training[7] model parameters is estimated and the relevant portion is identified.[8] proposed an independent motion evidence feature for distinguishing human actions from background motion.

Most of the methods require that the relevant portion of the video has to be annotated with bounding boxes. Human intervention was tedious. So to overcome the bounding box Brendel *et al.*[9] divides the video into a number of subgroups and then a model was generated that identify the relevant subgroup.

This paper introduces a method that learns both spatial and temporal extents for detection improvement. Dense trajectory is used here as local features to represent the human action.

III. Generating dense trajectories

We are having a set of training videos. Our aim is to find the action performed and to find where the corresponding action is in the video .Our method automatically identifies the relevant portion. We have to only annotate whether it is a positive training video or negative training video.

The video representation will be generated based on local patch trajectories. Wang et al., who generated dense trajectories ,has showed that it performed better than the KLT tracking of sparse local features. So in this paper we adopt dense trajectory approach.

As Large Displacement Optical Flow Tracker can track the object with fast motion and large displacement ,we adopt LDOF tracker to extract trajectories for representing the video. We use a grid step size of 5.For each trajectory we compute four types of descriptors, trajectory, MBH,HOG and HOF. While finding the three types of descriptors it is possible to estimate the camera motion between two consecutive frames. The second frame is warped with camera motion. So it is easy to compute the optical flow between the frames and the descriptors. Here the boundaries of the foreground moving objects is used to estimate the camera motion. To encode the local features standard BOW approach is used. First by applying k-means approach followed by quantizing the descriptors of trajectories in a video. Quantizing is done by assigning the closest vocabulary word based on Eucledian distance. Finally codeword occurrences are normalized by RootSIFT approach and their concatenation is the descriptor of the video.

IV. Generating Object level segmentation

Feature points are densely sampled on a grid, with uniform spacing, and tracked at multiple spatial scales to obtain dense trajectories .Each point is tracked between consecutive frames using median filtering in a dense optical flowfield. We assume that the input video has F frames, and trajectories of P feature points are extracted. Cluster T(p),p=1…P , into several groups such that same group have high similarity. Construct a matrix from T(p) and decompose into M and N. Applying DCT we minimize N(p) and the error.

Calculate the affinity matrix for N(p)[9].Then partition into foreground and background ..Background is the one with lower dimensional space. After several iterations, we can get N(p),p element of foreground. The final step in our split stage is to cluster foreground trajectories into an unknown number of groups. After several iterations we get a foreground moving object. Thus we have generated motion consistent clusters each associated with a unique foreground moving object in the scene.

V. Training And Testing The videos

A. *Learning Videos*

For a given action, we assume we are given M_1 positive training videos, , and M_2 negative training videos, .We use our trajectory split-and-merge algorithm to process the trajectories extracted from each positive training video. In this way, object-level segmentation of the content of a positive training video can be obtained. Subsequently, our goal is to use the training set to train a SVM model, and to learn the spatial and temporal extents of the given action in positive examples.

$V_1^+, V_2^+, .. V_{M1}^+$ are the M_1 positive training videos and $V_1^-, V_2^-, .. V_{M2}^-$ are the M_2 negative training videos. Suppose V_i^+ has N_i foreground moving objects.s_i denote the start frame and e_i denote the end frame and let l_i specify the performer of the given action.l_i can be an element of the foreground moving object. These variables represent the spatial and temporal extent.

From the standard BOW approach we get the descriptors of the video. It is denoted by $d(V_i^+, l_i, s_i, e_i)$. It is then fit to a SVM model.

$\text{Score}(V_i^+, l_i, s_i, e_i) = K(w, d(V_i^+, l_i, s_i, e_i)) + b$

An iterative learning is performed to solve the problem. So after the learning ,we will be getting a specified *w*,b for a specific action. With this we proceed to the testing phase.

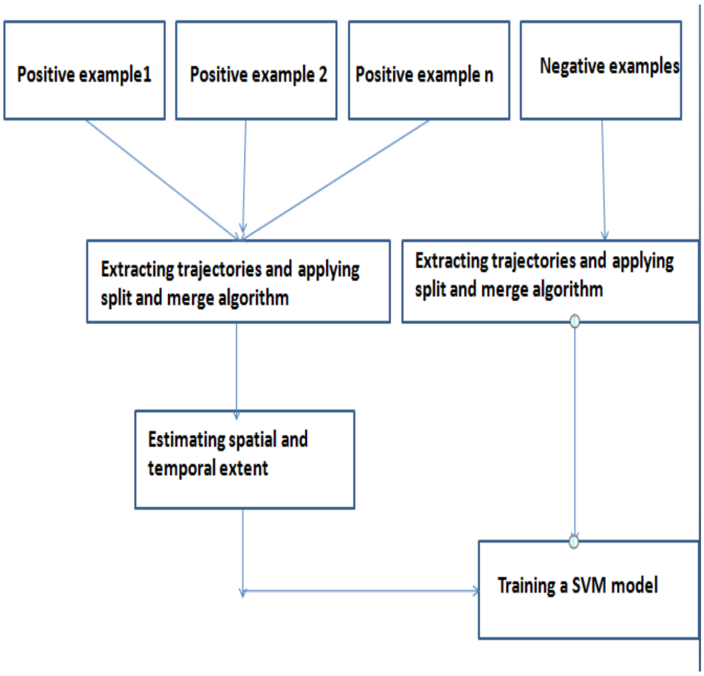
Positive example1
Positive example 2
Positive example n
Negative examples
Extracting trajectories and applying split and merge algorithm
Extracting trajectories and applying split and merge algorithm
Estimating spatial and temporal extent
Training a SVM model

Fig 2: Block diagram of training the video

B. *Testing the action*

When a action video is given, have to find where the action is in the video and what the action is.For the test video, we process split and merge algorithm on the trajectories. After this process, object level segmentation of the video will be obtained. We will be able to identify different foreground moving objects. To construct the descriptors for each foreground moving object standard BOW approach is used. Descriptor of the test video is denoted by $d(V_t, i, s_{ti}, e_{ti})$,for i= 1…N ,where N is the number of foreground moving objects. If a foreground object is detected then multiple descriptors are generated. The descriptor with highest matching score is used to determine the temporal extent of the action.

VI. Conclusion

Human action recognition is made more reliable without manual annotation of relevant portion of action of interest. Dense trajectories are used as local features to represent the human action .It is more fine grained .Action localization is made by learning the temporal and spatial extents of video. This method of action recognition outperforms several state-of-art methods.

Acknowledgment

Sincerely thankful to the guide and other faculties for supporting and helping to do this work. We would also thank the colleagues and the reviewers for helpful comments

References

[1] H.Wang,A .Klaser,C.Schmid and C-L.Liu, "Action recognition by dense trajectories ," in *Proc. IEEE Conf. Comput. Vis.Pattern Recog.*,Jun. 2011,pp 3169-3176.

[2] H.Wang and C Schmiid , "Action recognition with improved trajectories," in *Proc.IEEE Int. Conf.Comput. Vis.*, Dec 2013,,pp 3551-3558.

[3] Y-G Jiang,Q.Dai,X.Xue,W.Liu and C-W Ngo. "Trajectory-based modeling of human actions with motion reference points," in*Proc. Eur.Conf .Comput.Vis.*,Oct 2012,Vol 7576,pp.425-438.

[4] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatiotemporal maximum average correlation height filter for action recognition,"*in Proc. IEEE Conf. Comput. Vis. Pattern Recog*., Jun. 2008,pp. 1–8.

[5] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detectionin complex scenes with spatial and temporal ambiguities," in

Proc. IEEE Conf. Comput. Vis. Pattern Recog., Sep.–Oct. 2009, pp.128–135.

[6] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in Proc. *IEEE Int. Conf. Comput. Vis*., Nov. 2011, pp. 2003–210.

[7] M.Raptis, I.Kokkinos and S.Soatto," Discovering discriminative action parts from mid-level video representations" ,in *Proc ,IEEE Conf.Comput.Vis,.Pattern Recog.,*Jun 2012,pp.1242-1249

[8] W. Brendel and S. Todorovic, "Learning spatiotemporal graphs of human activities," *in Proc. IEEE Int. Conf. Comput. Vis.,* Nov. 2011,

[9] M.Jain,J.van Genmert,H.Jegou ,P.Bouthemy and C.Snoek ,"Action localization with tubelets from motion" ,in *Proc IEEE Conf, Comput.Vis.Pattern Recog*. Jun 2014 pp 740-747