Human Action Recognition in a Video based on Spatio-Temporal Features

Sushmitha. V C[1] Rashmi. S. R[2]
[1]M. Tech. Student [2]Professor
[1,2]Department of Computer Science & Engineering
[1,2]Dayananda Sagar College of Engineering Bangalore-78, India

Abstract— Recognizing human actions in complex scenes is a challenging problem due to background clutters, camera motion, occlusions, and illumination variations. There are several realistic scenarios where human action recognition (HAR) highlights its importance for security purpose. Challenges in HAR are Application Domain, Variations in Inter and Intra class, Background and Recording settings, Human variation, Action variation…etc. In this paper spatio temporal interest points (STIP) are located in region that shows a high variation of image intensity in all three direction (x,y,t). Spatio temporal corners are located at spatio corners such that they invert motion in two consecutive frame. To classify the frames in video we are using K-nearest neighbor (KNN) by loading the trained features. Experimental results shows the effectiveness of our model.

Key words: Human Action Recognition (HAR), STIP, KNN Classifier

I. INTRODUCTION

Identifying the actions being accomplished by the human in the video sequence automatically and tagging their actions is the prime functionality of intelligent video systems. Recognizing action from videos has not been addressed extensively, primarily due to the tremendous variations that result from camera motion, background clutter, and changes in object appearance, and scale the main challenge is how to extract reliable and informative features from the videos. Automatically recognizing human actions is receiving increasing attention due to its wide range of applications such as video indexing and retrieval, human-computer interaction, and activity monitoring. Although a large amount of research has been reported on action categorization, recognizing actions from realistic video still remains a quite challenging problem due to the significant intra-class variations, occlusion, and background clutter.

II. CHALLENGES IN HUMAN ACTIVITY RECOGNITION

The major applications of Human Action Recognition system is discussed below.

A. Application Domain

The activities of interest and the importance of fine details will vary based on the application domain. For example, for a surveillance system, the main interest is typically in finding the unusual behavior (e.g. falling down, jumping over a fence, etc..,)

B. Variations in Inter and Intra class

The performance of the system depends on the large variations in activity class. For example the activity walking and jogging will vary by only a small degree. A good human action recognition should be able to differentiate the activities of one class with another class.

C. Learning Paradigm Usage

A learning based approach is used to recognize different human activities. The main advantage is robustness to intra class variations. The learning paradigm usage can be either supervised or unsupervised based on the type of training data available.

D. Occlusion

Occluded parts either self or due to objects have a major effect in recognition of actions. As the features from the Occluded body parts are lost often, the system can end up in recognizing an action wrong and yet the system would be correct as the features from the occluded parts have a little effect in the output.

E. Background and Recording settings

Identifying the activities of human with a cluttered or dynamic background is difficult. The quality of a video is also a prime factor in deciding the performance of the system. An efficient activity recognition system should recognize the human even in the varying quality of the video and the cluttered background.

F. Human Variation

Different people will have different body size, since size of body varies detection of object of interest will be challenging.

G. Action Variation

Different people will perform action in different way thus to predict an action system may find difficulties.

III. PROPOSED FRAMEWORK

In this section, we present our framework for HAR. Fig.2 gives an overview of the proposed recognition framework. Given an input video, we first adopt interest point detector to detect STIPs. Then, salient region detection is applied to suppress unwanted background STIPs and obtain valid STIPs with high probability, where valid STIPs mean these STIPs represent local features of object these features refers to a pattern or distinct structure found in an image for example corners, edges.....etc., Based on the local feature KNN classifier is used for the classification of video.

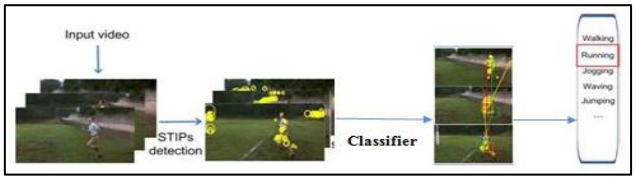

Fig. 1: Overview of proposed recognition frame work

A. Video to Frames

Video that is fed as input to the system is converted into the frames for future processing.

1) STIPS Generation

Most useful and effective approach is to extract local features at STIP and encode the temporal information directly into local feature. Videos are considered as volumes of pixels. For 2d case, interest point structure are searched for that stable under rotation view point, scale and illumination change.

Salient Region Detection: A lot of unwanted background STIPs are detected from complex videos due to background clutters and camera motion. Thus, we need generate huge instances to hit the valid instance. That will make our parameters learning inefficient. In order to prompt the efficiency of parameters learning, a saliency region detection method based on spatial prior is utilized to acquire saliency map that is used to suppress some unwanted background STIPs quickly.

We apply spatial prior based salient region detection to detect salient region in statistic image. For an input video frame, we first segment it into several regions R = {r1, . . . , rc}. Then, the color histogram is built for each region to describe the appearance of region.

- Region Contrast: We use spatially weighted region contrast which incorporates spatial information by introducing a spatial weighting term to increase the effects of closer regions and decrease the effects of farther regions.
- Spatial Prior: spatial prior information is to aid saliency region detection since we observe that actions are more likely to be spatially located in center of video frames, especially for movie videos or videos that track objects.

2) STIPs Suppression

Due to background clutters and camera motion, many unwanted background STIPs are extracted from videos. Then, we need suppress STIPs to generate valid multiple instances. We suppress unwanted background STIPs using saliency maps S as follows: the STIP centered at (x, y) is selected as the final STIP if the saliency value at (x, y) is above the threshold Th.

B. Classifier

The k-nearest neighbor's algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.

1) Algorithm 1: Pseudo-Code Learning Algorithm

- Input: (ΥBi = {Υ(I^{i}_{l} , s)}L l=1,YBi), i = 1, . . . , N
- Output: w,b

1) for j=1:m
2) do Random choose one instance Υ(I i l , s) ∈ ΥBi and compute QP solution w and b for dataset with positive examples (ΥBi , YBi)
3) repeat
4) compute outputs f = w · Υ(I^{i}_{l} , s) for all Υ(I , s)in positive bags, and set Υ(I^{i}_{l} , s) = Υ(I^{i}, s), where ki = argmaxI I ∈B(w · Υ(I^{i}_{l} , s))
5) until no ki has changed during iteration;
6) end.

2) Algorithm 2: Pseudo-Code for Parameters Learning

- Input: (ΥBi = {Υ(I^{i}_{l} , s)}L_{l}=1,YBi), i = 1, . . . , N

1) Initialize: s = k means(x,K)
2) Repeat
3) optimize w and b using Algorithm 1 with ΥBi ;
4) infer latent variables s for each instance in each Bi using Viterbi algorithm;
5) update ΥBi with the inferred s.
6) until no s has changed during iteration;

C. Datasets

- KTH Actions Dataset: KTH actions dataset contains six different actions. These video clips are acted under four different environment. All sequences are divided into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). There are 2391 video clips in total, 1528 video clips for training and 863 video clips for testing. We use training set and validation set to train the model and present recognition accuracy on the test set.
- UCF Sports Dataset: UCF sports dataset [34] contains 10 action categories. This dataset consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN.
- Youtube Actions Dataset: Youtube actions dataset contains 11 action categories: basketball shooting, biking/ cycling, diving, golf swinging, horse riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is very challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc.

D. Saliency Region Detection

Fig.2 gives the saliency maps on sample frames from KTH actions dataset. Because of the clear background, global contrast based salient region detection gives the clear contour of human body. Fine saliency maps which clearly distinguish foreground and background. This is because of the large color histogram contrast between foreground and background. Although global contrast based salient region detection may fail to detect excellent saliency maps it reserves most STIPs due to our spatial prior for saliency region detection.

Fig. 2: Saliency maps for sample frames from KTH actions

1) Spatio-Temporal Interest Points Suppression

Figure 3 gives comparisons between original STIPs and STIPs after suppression. A lot of unwanted background STIPs are detected from unconstraint videos especially when camera moves. After suppression, many unwanted background STIPs are suppressed via saliency maps. It gives high probability of valid STIPs selection for instance generation where these valid STIPs represent the movement of objects, not background.

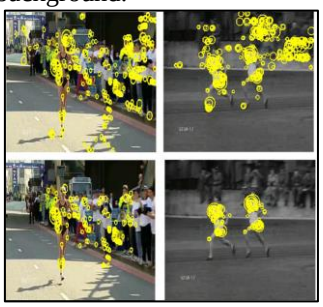

Fig. 3: Comparisons between original STIPs and STIPs after suppression.

IV. EXPERIMENTAL RESULTS

KTH Actions Dataset*:* We found that STIPs after suppression make no difference with original STIP on account of the clear background. Consequently, we use original STIPs to fairly compare with other methods on KTH actions dataset. Table 1 gives the average accuracy of various methods.

Brief Description	Accuracy
SVM, local space time features	71.7%
SVM, spatio-temporal features	91.8%
SVM, dense spatio-temporal features	92.1%
SVM, learned hierarchical invariant spatio-temporal features	93.9%
Incremental learning, pyramid histograms of oriented gradient features	96.1%
Sparse modeling, dictionary learning, motion imagery	97.9%
Sparse representation, covariance manifolds, optical flow	97.4%
Action bank, linear SVM, spatiotemporal orientations	98.2%
Discriminative HMM, "cuboid" features	91.2%
Discriminative Semi-Markov Model, cutting plane learning, "cuboid" features	94.7%
BOW model, SVM, "cuboid" features	85.1%
Discriminative Semi-Markov Model, bundle method, "cuboid" features	95.0%
BOW model, linear SVM, spatio-temporal features	91.4%
Discriminative HMM, spatio-temporal features	92.3%
Discriminative Multiple-instance Markov Model, "cuboid" features	98.0%
Discriminative Multiple-instance Markov Model, spatio-temporal features	96.7%

Table 1: Average accuracy on the KTH actions dataset

Precision and recall values are taken are taken for performance analysis. Precision and Recall is calculated based on the following relation.

$$\text{Precision} = (TP)/(TP+FP)$$

$$\text{Recall} = (TP)/(TP+FN)$$

Where, TP is True Positive which indicates the activities correctly detected and classified by the algorithm, FP is False Positive which designates the activities that are recognized and categorized by the algorithm but does not exist, FN is False Negative which point out the activities that exist but are not noticed and listed by the algorithm and TN is True Negative which stipulate the activities that do not exist and are not detected and tabulated by the algorithm.

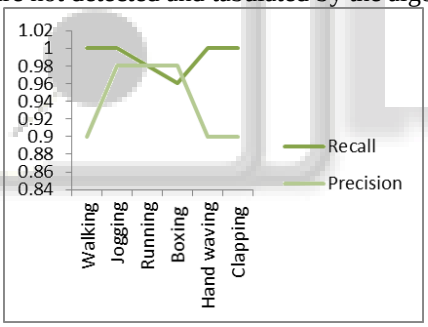

Fig. 4: precision and recall values for different actions

Figure 4 shows the precision and recall values for actions like walking, jogging, running, boxing, handwaving, clapping.

V. CONCLUSION

The proposed method is able to assign activity label to the various activities in video scene with an accuracy of 98%.The values obtained from this method has the high degree accuracy. The proposed method would be effective in analyzing activities in wide area surveillance.

In future, there are different directions in which this work can evolve. Though we have created a standalone application for identifying actions in single known videos, an integrated framework can be created which can identify interesting activity regions as well as recognize them using contextual information. The method can also be extended to be able to correct missed detections and false positives using the contextual information available.

REFERENCES

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8.
[2] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proc. Brit. Mach. Vis. Conf., 2009, pp. 124.1–124.11.
[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 3361–3368.
[4] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1234–1241.
[5] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Action recognition using context-constrained linear coding," IEEE Signal Process. Lett., vol. 19, no. 7, pp. 439–442, Jul. 2012.
[6] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 3337–3344.
[7] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," IEEE Trans. Inf. Forensics Security, vol. 8, no. 10, pp. 1600–1609, Oct. 2013.
[8] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1242–1249.
[9] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 12, pp. 1325–1337, Dec. 1997.
[10] Q. Shi, L. Cheng, L. Wang, and A. Smola, "Human action segmentation and recognition using discriminative semi-Markov models," Int. J. Comput. Vis., vol. 93, no. 1, pp. 22–32, 2011.
[11] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 831–843, Aug. 2000.
[12] P. Natarajan and R. Nevatia, "Coupled hidden semi Markov models for activity recognition," in Proc. WMVC, Feb. 2007, p. 10.
[13] L. Shao and R. Mattivi, "Feature detector and descriptor evaluation in human action recognition," in Proc. ACM Int. Conf. Image Video Retr., 2010, pp. 477–484.
[14] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," Pattern Recognit. Lett., vol. 33, no. 4, pp. 438–445, Mar. 2012.
[15] X. Zhen, L. Shao, D. Tao, and X. Li, "Embedding motion and structure features for human action recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 7, pp. 1182–1190, Jul. 2013.
[16] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in Proc. 11th Eur. Conf. Comput. Vis., Sep. 2010, pp. 577–590.

[17] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, "Cross-viewaction recognition via a continuous virtual path," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2690–2697.
[18] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep./Oct. 2009, pp. 1593–1600.
[19] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1250–1257.
[20] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in Proc. 10th IEEE Int. Conf. Comput. Vis., vol. 2. Oct. 2005, pp. 1808–1815.
[21] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu,"Global contrast based salient region detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2011, pp. 409–416.
[22] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles,"
[23] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in Proc. 15th ICML, vol. 15. 1998, pp. 341–349.
[24] Z. H. Zhou and M. L. Zhang, "Multi-instance multi-label learning with application to scene classification," in Proc. Adv. Neural Inf. Process. Syst., vol. 19. 2007, p. 1609.
[25] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 983–990.
[26] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. 18th Int. Conf. Mach. Learn., 2011, pp. 282–289.
[27] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in Proc. J. Mach. Learn. Res., 2005, pp. 1453–1484.
[28] A. L. Yuille, A. Rangarajan, and A. Yuille, "The concave-convex procedure (CCCP)," in Proc. Adv. Neural Inf. Process. Syst., vol. 2. 2002, pp. 1033–1040.
[29] M. Grant and S. Boyd. (2014, Mar.). CVX: Matlab Software for no. 2, pp. 107–123, Sep. 2005

