ACTION RECOGNITION USING INTEREST POINTS CAPTURING DIFFERENTIAL MOTION INFORMATION

Gaurav Kumar Yadav[†], *Prakhar Shukla*[⋆], *Amit Sethi*[†]

[†]Department of Electronics and Electrical Engineering, IIT Guwahati
[⋆]Department of Mathematics, IIT Guwahati, Guwahati

ABSTRACT

Human action recognition has been a challenging task in computer vision because of intra-class variability. State-of-the-art methods have shown good performance for constrained videos but have failed to achieve good results for complex scenes. Reasons for their failing include treating spatial and temporal dimensions without distinction as well as not capturing temporal information in video representation. To address these problems we propose principled changes to an action recognition framework that is based on video interest points (IP) detection with capturing differential motion as the central theme. First, we propose to detect points with high curl of optical flow, which captures relative motion boundaries in a frame. We track these points to form dense trajectories. Second, we discard points on the trajectories that do not represent change in motion of the same object, yielding temporally localized IPs. Third, we propose a video representation based on spatio-temporal arrangement of IPs with respect to their neighboring IPs. The proposed approach yields a compact and information-dense representation without using any local descriptor around the detected IPs. It significantly outperforms state-of-the-art methods on UCF youtube dataset, which has complex action classes, as well as on KTH dataset, which has simple action classes.

Index Terms— Video interest points, action recognition, dense trajectories, Optical flow

1. INTRODUCTION

Human action recognition has been studied very extensively due its potential applications in video surveillance, search, and retrieval. Defining and recognizing a class of actions is fraught with problems such as large variations in motion, posture, and clothing of people, as well as variations in scene illumination and background. A widely applicable solution to deal with all these variations is yet to come. Until the advent of standardized action datasets such as Weizman [1] and KTH [2], comparing techniques was not easy. However, most datasets and techniques were still based on simplifying assumptions such as uncluttered background, isolated actions and static camera until more complex datasets such as UCF11 [3] and techniques such as [4] came about. We compare our methods on both simple [2] and complex [3] datasets.

Most of the methods for human action recognition can be classified broadly into two categories [5]: hierarchical and single-layered approaches. Hierarchical approaches break a complex activity into simple activities or sub-events. Multiple layers of sub-events are constructed for the analysis of complex activities. Such methods however, are more complex and recognition of the high level activities run into problems if the sub-events or low-level activities are not reliably recognized. On the other hand, single-layered approaches tend to be faster and more suitable for real time applications because these recognize actions directly from the video. These are further classified into two categories: Space-time approaches and sequential approaches. Space-time approaches tend to be the fastest because they do not consider temporal order unlike sequential approaches and are further divided into three categories: those based on space-time volumes, trajectories, and interest points. Our method is based on trajectories and interest points while it also partly incorporates elements of sequential approaches to incorporate the advantages of these three groups of techniques.

A lot of action recognition methods have been proposed based on interest points (IPs) such as [6], [7], and [8] which are characterized by their detectors, descriptors and fusion methods. Although many interest point detectors have been proposed for videos such as STIP [9], selective-STIP [4], Cuboid [10], n-SIFT [11], Mo-SIFT [12] , curl of optical flow (COF) [13] etc., except Mo-SIFT and COF all other methods treat temporal dimension in a manner similar to the two spatial dimensions, thus extending 2-d spatial interest point detectors to 3-d. This is not appropriate as shown in [13] because of unique properties of temporal dimension such as object persistence and smoothness. We extend this approach significantly based on three contributions that capture our proposed theme that *differential motion has important information about an action*.

Our first contribution is to use and extend the interest point detector proposed in [13] that was based on unique properties of the time axis, and captured points on relative motion boundaries. The threshold applied to curl of optical flow was fixed in [13], which led to large variations in

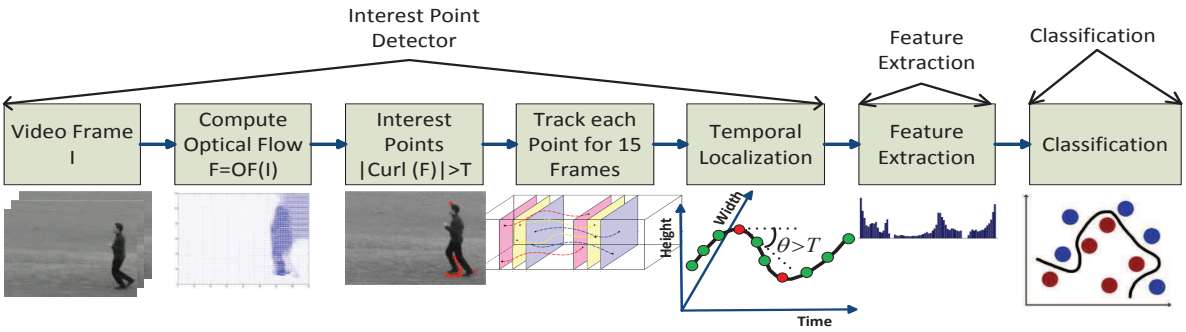


Fig. 1. Block diagram of the proposed action recognition system.

the number of IPs across videos. We introduce an adaptive thresholding which gives a sparse set of meaningful points.

Temporal localization of IPs by discarding uninformative points on trajectories can reduce memory requirements and increase information density of a video representation. Some methods for IP-based action recognition have used temporal localization [9], [11] but they don't mention it as an explicit intermediate goal. Our second contribution is to use and show utility of temporal localization.

To extract IP descriptors a 3D volume around the interest point is considered. Further, to get a fixed-dimensional vector representation of a video, descriptors from a varying number of IPs of a video are fused using different methods such as bag-of-words (BoW) [14] and latent Dirichlet allocation (LDA) [15]. Such fusion methods lose the temporal order of IPs which can contain discriminative information, although some temporal information can be captured in their descriptors. Our third contribution is to propose a video representation that captures temporal information and gives a compact fixed-dimensional representation of a video.

2. PROPOSED ALGORITHM

Video classification using interest points has three major parts – IP detection, video representation, and classification. Differential motion, whether it is between objects or between close-by time instants for the same object, contain important information about action classes. Such information is somewhat lost in local IP descriptor-based methods especially when descriptor fusion methods (such as bag-of words) are also used. The proposed algorithm combines the desirable features of these three parts of video classification algorithms. First, we detect IPs based on a modification of COF[13], which captures only points associated with action. These points are tracked to obtain dense trajectories, capturing temporal information. Then, we perform temporal localization and prune uninformative points by retaining only an informative subset of points on the trajectories. Next, we compute a video representation for classification that is based on spatial and, more importantly, temporal distances between consecutive retained points on the trajectories, thus capturing information important for action recognition. These innovations have yielded substantial improvements compared to state-of-the-art on benchmark datasets. We now describe each module as shown in Fig. 1 in more detail as follows:

Adaptive interest point detection: Optical flow captures object persistence and smoothness in time dimension. The IP detector proposed in [13] was based on curl of optical flow to capture the relative tangential motion between objects such as foreground and background. For the optical flow (velocity) vector $\overrightarrow{V}$ of a point in a frame expressed as a linear combination of unit vectors in $\overrightarrow{i}$ and $\overrightarrow{j}$ in x and y directions respectively as $V_x\,\overrightarrow{i} + V_y\,\overrightarrow{j}$, the curl is defined as:

$$\nabla \times \overrightarrow{V} = \frac{\partial V_x}{\partial y} - \frac{\partial V_y}{\partial x} \tag{1}$$

If $\left|\nabla \times \vec{V}\right|$ is greater than a threshold T at a point, and it is also a spatial local maxima then we consider it as an interest point. In [13], the threshold T was fixed across videos, which caused the following problems due to scene change between videos. While a high threshold yields too sparse IPs for certain videos, a low threshold introduces spurious IPs based on non-zero curl due to background movement, noise, and video compression artifacts that vary from scene to scene. We adapted the threshold for each video according to the base level of curl magnitude that occurs even when there is no activity in that scene. We computed the threshold separately for each video as follows:

$$T_i = \underset{j}{\text{median}}(\max_{k}(\left|\nabla \times \vec{V}_{ijk}\right|)) \tag{2}$$

where $\left|\nabla \times \vec{V}_{ijk}\right|$ denotes the curl magnitude of point k in a video frame j for video i.

Tracking: After detecting points in each frame, tracking was done using KLT tracker [16]. Trajectories tend to drift during tracking because of noise. Hence, we fixed the trajectory length to 15 frames as suggested in [8].

Pruning-based temporal localization: Temporal localiza-

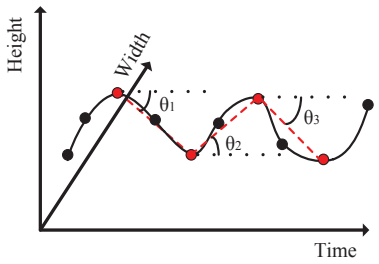


Fig. 2. Illustration shows the trajectory of an interest point. Red dots are the temporally localized points. Change in slope is indicated by the angle.

tion was done based on the change in the direction of trajectory as an indicator of its unpredictability, as shown in Fig. 2. If the angle between the tangents at a point and its predecessor on a trajectory was more than $20°$, then the point was retained otherwise it was pruned.

Video representation: Since we do not use any IP descriptor, we retain spatial and, importantly, temporal information associated with motion of an action by capturing the relation between consecutive temporally localized (unpruned) points on the trajectories. This information is characterized by both distance and angle between these points. Weighted orientation histograms of projections of line segments connecting consecutive retained IPs on to xy, yt and tx planes, were calculated, where x and y denotes the spatial dimensions of the frame and t denotes the time dimension. The weight (contribution) was proportional to the length of the line segment, while the bin was decided by the orientation of its projection on that plane. The histogram had 36 bins dividing the angle range of $0°$ to $180°$ into intervals of $5°$. Histograms for the three planes were concatenated together and normalized to form the video representation.

Classification: After forming a compact video representation, we used it as input for an SVM to classify the actions. This is similar to a lot of state-of-the-art methods.

3. EXPERIMENT AND RESULTS

We compared the performance of our proposed technique with several state-of-the-art and a few pioneering legacy techniques on a simple and a complex video action datasets.

Dataset description: KTH database consists of simple scenarios of six types of human actions - walking, jogging, running, boxing, hand waving and hand clapping [2]. The actions were performed by 25 subjects in different scenarios such as outdoors and indoors, with different scale and clothing variations. UCF 11 is a complex data set with videos taken from the wild, and consists of 11 action categories - basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog [3]. This dataset has variations in camera motion, object appearance and pose, object scale, viewpoint, background clutter, illumination conditions, etc. For each category, the videos are grouped into 25 groups with more than 4 clips in each group. The video clips in the same group share some common features, such as the same actor, similar background or viewpoint. It has been suggested to use entire groups instead of splitting their constituent clips into training or testing sets (specifically, leave-one-group-out) to test robustness of action recognition techniques to unseen scene variations.

Table 1. Comparison of the average accuracy of the proposed method with state-of-art methods on KTH dataset. Methods not based on IPs are denoted by *.

Methods	Accuracy(%)
Proposed method	**98.20%**
*Jhuang et. al. [17]	96.00%
Mo-SIFT [12]	95.80%
Kovashika et. al. [6]	94.53%
Gilbert et. al. [7]	94.50%
Wang et. al. [8]	94.20 %
COF [13]	93.40%
Laptev et al. [9]	91.80%
Cuboid [10]	80.00%

Experimental setup: Classification of actions was done using Libsvm package [20]. Best parameter selection was done using 3-fold cross-validation on testing dataset. KTH dataset was divided into 80% training and 20% testing set randomly. The accuracy was averaged over several random selections of training and testing data. For UCF11 dataset we have followed the leave-one-group-out cross-validation(LOOCV) approach as suggested in [3]. LOOCV scheme leads to 25 cross-validation results for each action class, which were averaged.

Action recognition results: Table 1 and Table 2 show comparison of action recognition performance of the proposed method and state-of-the-art methods. KTH dataset contains simple actions in constrained environment for which state-of-the-art methods have already achieved good results.

Table 2. Comparison of the average accuracy of the proposed method with state-of-art methods on UCF11 dataset. Methods not based on IPs are denoted by *, and methods that don't seem to have used LOOCV scheme for testing are denoted by $^?$.

Methods	Accuracy(%)
Proposed method	**91.30%**
Wang et. al. [8]	84.20 %
$^{*?}$Mota et. al. [18]	75.40 %
*Ikizler-Cinbis et. al. [19]	75.21 %
Liu et. al. [3]	71.20%

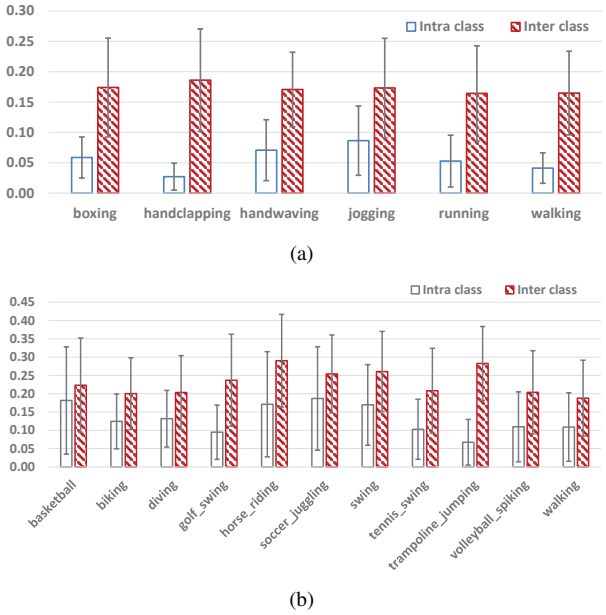


Fig. 3. Inter- and intra-class squared distances for the proposed video representation for (a) KTH and (b) UCF11 datasets.

The proposed method showed better results than other methods that use IPs for KTH dataset and almost as good as more complex methods, as shown in Table 1. Importantly, for UCF11 dataset, the proposed method outperformed all the state-of-the-art methods, including those not based on IPs, as shown in Table 2.

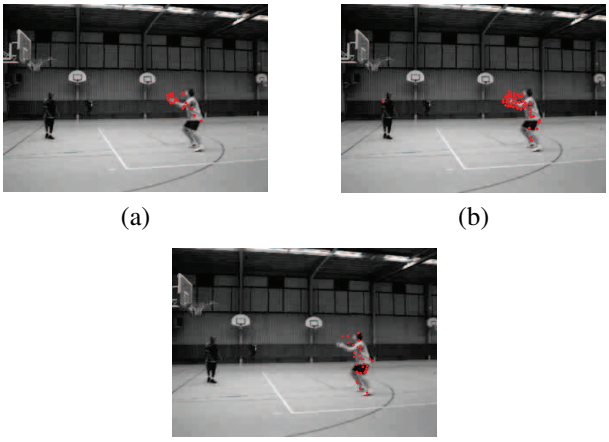
Class separation: We have also shown the discriminative ability of the proposed video representation by comparing inter- and intra-class mean-squared distance in Fig. 3. The graph shows low inter-class and high intra-class variance, which indicates that the distinctness factor of descriptors are high for most classes for both the datasets.

Impact of adaptive threshold: As shown in Fig. 4, the proposed thresholding method improved the quality of IPs detected when compared with the COF [13] and selective-STIP [4], which have shown to yield more meaningful IPs than other techniques. Our interest point are sparse as well as on the boundary of the moving object where relative motion is large whereas in case of selective-STIP and COF the number of IPs is either too large (some of which are on the background) or IPs are not exactly on the boundary of moving object. Motion boundary is important for characterizing an action when the background is not static.

Impact of temporal localization: The proposed method has shown a drastic improvement with temporal localization as shown in Table 3. The reason is that our video representation was based on orientation of connecting temporal localized points on the trajectories, and does not use local descriptors commonly used in other IP-based video representations. These points are action-specific which results in different histogram for different actions. If we consider all the points on the trajectories rather few points then most of the consecutive pairs of points will have small changes in orientation, and will contribute to only a few bins of the histogram. This will make it difficult to distinguish between action classes.

Table 3. Comparison of the average accuracy of the proposed method with and without temporal localization.

Ineterest point extraction	KTH	UCF11
With temporal localization	**98.2%**	**91.3%**
Without temporal localization	57.2%	39.5%



(a) (b)

(c)

Fig. 4. Example of interest points of the proposed method (a) with adaptive threshold, (b) with fixed threshold [13], and (c) Selective-STIP [4]

4. CONCLUSION

We proposed a video representation for action recognition that performs better than state-of-art methods on two widely used benchmark datasets. We detect IPs on motion boundaries to discard moving background, and retain only those IPs that represent new information based on large change in trajectory direction. Thus it is much more compact than using entire trajectories. Additionally, we have shown that temporal localization plays an important role in improving the information content in certain video representations where local IP descriptors are not used. We conjecture that using temporal localization will improve the performance of video representations based on local IP descriptors as well. In future, this work can be extended for testing on large class database such as Hollywood dataset and UCF101 dataset. Scale and view point invariance can also be introduced in the video representation to further improve the performance [21].

5. REFERENCES

[1] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[2] Christian Schüldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[3] Jingen Liu, Jiebo Luo, and Mubarak Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1996–2003.

[4] Bhaskar Chakraborty, Michael B Holte, Thomas B Moeslund, Jordi Gonzalez, and F Xavier Roca, "A selective spatio-temporal interest point detector for human action recognition in complex scenes," in *Computer Vision (ICCV), 2011*. IEEE, 2011, pp. 1776–1783.

[5] Jake K Aggarwal and Michael S Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 16, 2011.

[6] Adriana Kovashka and Kristen Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[7] Andrew Gilbert, John Illingworth, and Richard Bowden, "Action recognition using mined hierarchical compound features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 883–897, 2011.

[8] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[9] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[11] Warren Cheung and Ghassan Hamarneh, "N-sift: N-dimensional scale invariant feature transform for matching medical images," in *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*. IEEE, 2007, pp. 720–723.

[12] Ming-yu Chen and Alexander Hauptmann, "Mosift: Recognizing human actions in surveillance videos," in *Technical Report, Carnegie Mellon University*, 2009.

[13] Gaurav Kumar Yadav and Amit Sethi, "A flow-based interest point detector for action recognition in videos," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2014, p. 41.

[14] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[15] Yang Wang and Greg Mori, "Human action recognition by semilatent topic models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 10, pp. 1762–1774, 2009.

[16] Carlo Tomasi and Takeo Kanade, *Detection and tracking of point features*, School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[17] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio, "A biologically inspired system for action recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. Ieee, 2007, pp. 1–8.

[18] Virginia F Mota, Jessica Souza, Arnaldo de A Araujo, Marcelo Bernardes Vieira, et al., "Combining orientation tensors for human action recognition," in *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*. IEEE, 2013, pp. 328–333.

[19] Nazli Ikizler-Cinbis and Stan Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," *Computer Vision–ECCV 2010*, pp. 494–507, 2010.

[20] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[21] Ashok Veeraraghavan, Anuj Srivastava, Amit K Roy-Chowdhury, and Rama Chellappa, "Rate-invariant recognition of humans and their activities," *Image Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1326–1339, 2009.