Spatio-Temporal Interest Points Chain (STIPC) for Activity Recognition

Fei YUAN*, Gui-Song Xia†, Hichem Sahbi†, Veronique Prinet*
*National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, P.R.C
Email: {fyuan,vprinet}@nlpr.ia.ac.cn
†LTCI-CNRS, TELECOM ParisTech, Paris, France
Email: {xia,sahbi}@telecom-paristech.fr

***Abstract*—We present a novel feature, named Spatio-Temporal Interest Points Chain (STIPC), for activity representation and recognition. This new feature consists of a set of trackable spatio-temporal interest points, which correspond to a series of discontinuous motion among a long-term motion of an object or its part. By this chain feature, we can not only capture the discriminative motion information which space-time interest point-like feature try to pursue, but also build the connection between them. Specifically, we first extract the point trajectories from the image sequences, then partition the points on each trajectory into two kinds of different yet close related points: discontinuous motion points and continuous motion points. We extract local space-time features around discontinuous motion points and use a chain model to represent them. Furthermore, we introduce a chain descriptor to encode the temporal relationships between these interdependent local space-time features. The experimental results on challenging datasets show that our STIPC features improves local space-time features and achieve state-of-the-art results.**

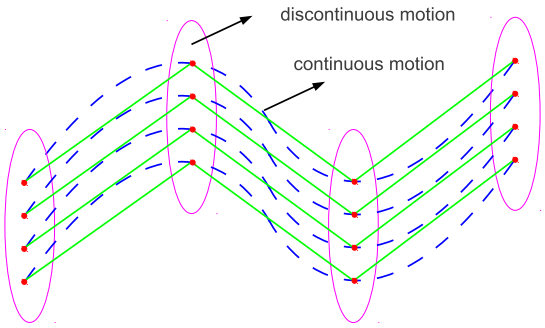


Fig. 1. Continuous motion (in blue color) and discontinuous motion (in red color) used for action recognition. Local space-time features, such as Harris3D [1], focus on the information around the locations having discontinuous motion. In our work, we use both continuous and discontinuous motion for activity recognition.

I. Introduction

Local space-time interest points features [1]–[3] have been widely employed for action recognition. They can be used to form sparse representations of actions and be effectively integrated into a machine learning framework. Impressive results have been reported in both synthetic and realistic scenarios, see [1]–[6]. One possible reason is that they capture the local discontinuous motion information of actions, which is extreme discriminative to certain actions. See Fig. 1 for a graphical illustration of discontinuous motion. When used to represent complicated activities with long-range motions or multiple interactive persons, however, the limitation of such low-level space-time interest points features blows up, since they describe only the local information of activities in a spatio-temporal volume, and the representation based on them, *e.g.*, bag-of-features (BOF), usually discards the geometrical and temporal relationships among features.

In order to exploit the dynamic property of action/activity, trajectory-based methods have been employed [7]–[11]. In trajectory-based methods, trajectory is used as the basic feature unit. Sun *et al.* [7] extracted trajectories by matching SIFT descriptors between consecutive frames, then used an average SIFT descriptor to describe the appearance information of a trajectory, and used a trajectory transition descriptor to encode the motion information on the whole trajectory. Messing *et al.* [8] used the velocity history feature of tracked key-points as basic features. One observation of current trajectory-based features is that the motion/velocity/displacement of points on a trajectory contains lots of redundancy, since most of the motion along a trajectory is similar (see the blue point lines in Fig. 1), thus could be approximated by several motion states (see the green lines in Fig. 1). Furthermore, the most discriminative information around discontinuous points (see the red points in Fig. 1) is not well exploited.

Want *et al.* [11] suggested to compute rich descriptors, such as HOG (Histograms of oriented gradients), HOF (Histograms of oriented optical flow) and MBH (motion boundary histogram), within a size-fixed space-time volume around the trajectory. The volume is then subdivided into a spatio-temporal grid to embed structure information. However, as previous trajectory descriptors, the space-time volume may contain too much similar information and the most discriminative information around discontinuous motion points is suppressed.

In this paper, we propose a new Spatio-Temporal Interest Points Chain (STIPC) feature. The chain feature aims to capture not only the space-time information around discontinuous points on a long-term motion process, but also the continuous motion information between them. These two kinds of information are encoded by a chain model, which makes it possible to exploit the temporal relationships between them. The chain feature is easily extracted: we first extract the point trajectories from the image sequences, then partition the points

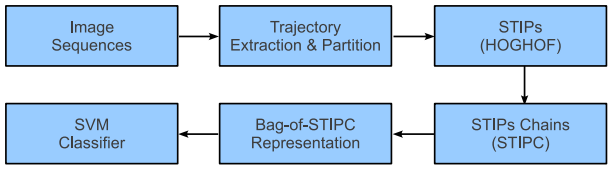


Fig. 2. A graphical illustration of our framework. Trajectories are first extracted from an image sequence and then are partitioned into two sets of points: discontinuous motion points and continuous motion points. Trackable spatio-temporal interest points are detected at discontinuous motion points. Each set of trackable spatio-temporal interest points is represented by a spatio-temporal interest points chain (STIPC), which encodes the temporal revolution of STIPs.

on each trajectory into two kinds of different yet close related points: discontinuous motion points and continuous motion points. Local space-time features around discontinuous motion points, such as HOG and HOF, are then extracted. Finally, a chain model is used to represent them.

Furthermore, to encode the temporal relationships between local space-time features, as well as the continuous motion features, on a chain, we introduce a chain descriptor. This descriptor allows to encode both the short-term relationships and long-term relationships between features on a chain and gives a compact histogram descriptor, which makes it convenient to compute the similarity between chain features.

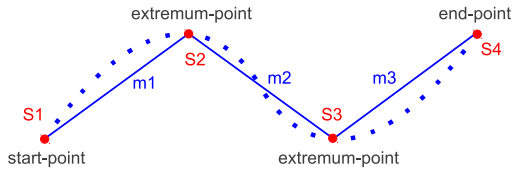
Different from local space-time features, our chain feature contains a set of space-time features with temporal association, which suggests a way to model the relationships between local space-time features. Our chain feature is also different from previous trajectory-based features in the following aspects: it is more discriminative, since discriminative space-time information is extracted around discontinuous motion points; redundant motion information is removed by approximating the continuous motion segments with single motion states.

The rest of this paper is organized as follows. Sec. II gives a detailed description of our approach for extracting trackable spatio-temporal interest points. In Sec. III, we present the chain model for activity representation. We illustrate and interpret the experimental results in Sec. IV, and finally conclude the paper in Sec. V.

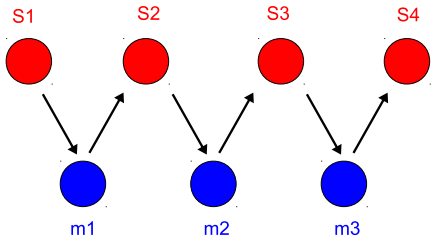
II. Extraction of Trackable Spatio-Temporal Interest Points

Current local space-time features, such as Harris3D [1] and Cuboid [2], are not trackable, even though extracted from a complete motion process of the same objects. One of key reasons is that the local space-time features, as shown in Fig. 1, are often extracted at places where motion changes, thus contain different spatio-temporal motion information.

In this Section, we present a novel method to extract sets of trackable spatio-temporal interest points (TSTIPs). Each set of these trackable spatio-temporal interest points are extracted from a complete motion process of an object or its parts. A graphical illustration of our framework is shown in Fig. 2.



(a) Trajectory partition by computing extrema of the spatio-temporal curvature. The red points stand for discontinuous motion points and blue points stand for continuous motion points.



(b) Spatio-Temporal Interest Points Chain (STIPC). The red points stand for STIPs in a chain, and the blue points stand for the continuous motion between STIPs.

Fig. 3. Spatio-Temporal Interest Points Chain (STIPC) is extracted by partitioning the trajectory.

A. Trajectory Extraction and Partition

Point trajectory could be extracted from an image sequence by either tracking 2-dimension spatially salient points (such as corners or SIFT) frame by frame [7], [8] or by tracking optical flow [11], [12]. In this work, we compute point trajectory using Sundaram' GPU-accelerated method [12]. It's a fast parallel implementation of large displacement optical flow estimation, which runs very fast, thus is suitable for our case.

Given the extracted point trajectories, the next step is to partition each trajectory into two kinds of different points: *discontinuous motion points* and *continuous motion points*. Suppose that a trajectory with the length of Δ lies on a curve $\tau = (p_1, p_2, ..., p_\Delta)$ in the 3-Dimensional spatio-temporal space, where $p_i = (x_i, y_i, t_i)$ is a 3-Dimensional points. It could be divided by computing the extrema of the spatio-temporal curvature as in [13]. Specifically, the spatio-temporal curvature is first smoothed by anisotropic diffusion [14], then its curvature is computed, finally the extrema of the spatio-temporal curvature are extracted with non-maxima suppression. The extrema of the spatio-temporal curvature together with the start-point and end-point of the trajectory, which capture both the changes of speed and the changes of the direction, are noted as discontinuous motion points, and the rest points are taken as continuous motion points. See Fig. 3(a) for an illustration.

B. Trackable Spatio-Temporal Interest Points (TSTIPs)

We propose to extract spatio-temporal interest points around the discontinuous motion points. As said in Sec. II-A, the discontinuous motion points are the points where the motion changes discontinuously, *e.g.*, beginning, changing direction, accelerating, stopping and so on. The information around

discontinuous motion points is reported to be discriminative for representing an activity, which is also the target that spatio-temporal interest points-like features pursue.

We name the spatio-temporal interest points on a trajectory as trackable spatio-temporal interest points (TSTIPs). "Trackable" means that these spatio-temporal interest points correspond to a series of discontinuous motion among a long-term motion of an object or its part, thus they have close interconnection. The interconnection between these spatio-temporal interest points depicts the dynamic property of an activity, therefore is an important cue for activity recognition.

HOGHOF [5], which is reported as an excellent descriptor, is used to describe the TSTIPs. HOG (Histograms of oriented gradients) exploits the appearance information, whereas HOF (Histograms of oriented optical flow) characterizes the motion information. In practice, we use Laptev's implementation of HOGHOF [5].

III. Spatio-Temporal Interest Points Chain (STIPC) for Activity Representation

In this Section, our goal is to model the temporal relationships among TSTIPs. For this purpose, we propose a Spatio-Temporal Interest Points Chain (STIPC) model, where the discontinuous motion, *i.e.*, STIPs, and continuous motion among STIPs are considered together.

A. Spatio-Temporal Interest Points Chain (STIPC)

A graphical illustration of our Spatio-Temporal Interest Points Chain (STIPC) model is shown in Fig. 3(b). Our model exploits the discontinuous motion, *i.e.*, TSTIPs shown with the red nodes in Fig. 3(b) and the continuous motion, shown with blue nodes in Fig. 3(b), in one coherent framework. These two sets of nodes, which stand for two kinds of different, yet complementary and close related features, compose one complete chain.

For the STIPs variables in a STIPC, we quantize each STIP feature into one of finite discretized states in the following way: a global STIPs codebook is first generated by clustering all the STIPs features extracted from the training sequences, where the k-means algorithm is used for clustering and the size of codebook is empirically set to be 300. Then each STIP feature is quantized into the corresponding state of the STIPs codebook. For the continuous motions between STIPs, we also quantize them into finite discretized states. *e.g.*, for continuous motion $m1$ in Fig. 3(a), we first connect the extrema points $S1$ and $S2$, then use the line segment l_{S1S2} to approximately represent the continuous motion between $S1$ and $S2$, finally the orientation of line segments l_{S1S2} are quantized into one of $S = 12$ states.

B. Temporal Modeling on STIPC

We introduce a chain descriptor which can flexibly model the short-term temporal relationships and long-term temporal relationships among the features in a chain under a first-order Markov assumption. Our chain descriptor is inspired from Ling's Proximity distribution descriptor [15], in which both

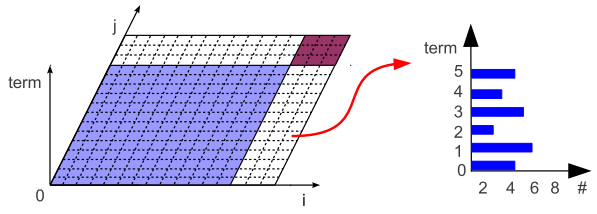


Fig. 4. State transition distribution descriptor. The grid in violet color stands for state transition between STIPs features, the grid in bordeaux color stands for state transition between approximated continuous motion features, while the grid in white color stands for state transition between STIPs and approximated continuous motion features

of local photometric information and local geometric information are combined in a proximity distribution descriptor for category recognition.

Specifically, we construct a two-dimensional array of one-dimension n-term state transition distribution, see Fig. 4 for an illustration. The entry of the two-dimensional array correspond to bins of STIP codebook and discretized motion states. In the third dimension, the term measures the scope of temporal relationships. For example, given a STIPC $S1 \rightarrow m1 \rightarrow S2 \rightarrow m2 \rightarrow S3 \rightarrow m3 \rightarrow S4$ (see Fig. 3(b)), for $term = 1$, the state transitions of $S1 \rightarrow m1, m1 \rightarrow S2, S2 \rightarrow m2, m2 \rightarrow S3, S3 \rightarrow m3, m3 \rightarrow S4$ are considered, for $term = 2$, the state transitions of $S1 \rightarrow S2, m1 \rightarrow m2, S2 \rightarrow S3, m2 \rightarrow m3, S3 \rightarrow S4$ are considered. $term = 0$ means that the temporal relations are not considered.

In total, we obtain a chain descriptor H_r with the size of $C \times C \times N$, where $C = 300 + 12$ stands for the number of quantized states of features and $N = 5$ stands for the n-term state transition we considered.

The similarity between chain descriptor H_r^1 and H_r^2 can be computed in a similar way as in [15]:

$$s(c^1, c^2) = \sum_{i,j=1}^{C} \sum_{r=1}^{N} min(H_r^1(i,j), H_r^2(i,j)) \qquad (1)$$

As suggested in [15], second-order or high-order statistics could be extended to encode more complicated temporal dependency of different features. Furthermore, due to the sparseness of the chain descriptor, computation of the similarity between chain descriptor could be very efficient.

IV. Experimental Results

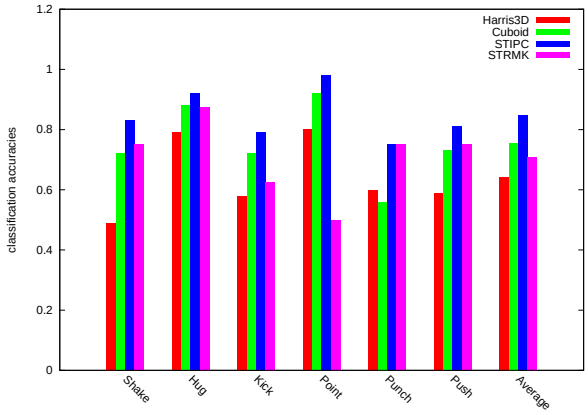
We evaluate the performance of our chain feature and compare to state-of-the-art methods on the recent UT-Interaction dataset [16]. It contains 6 classes of human-human interactions: *shake-hands, point, hug, push, kick* and *punch*. Each class contains 20 video sequences which are divided into two different sets: (1) the first 10 video sequences, Set 1, are taken on a parking lot with slightly different zoom rate, and the backgrounds of the video are mostly static with little camera jitter; (2) the other 10 video sequences, Set 2, are taken on a lawn in a windy day. The backgrounds of the video are moving

slightly, *e.g.* tree moves, and the videos contain more camera jitters. Observe that the background, scale, and illumination of the videos in each set are different.

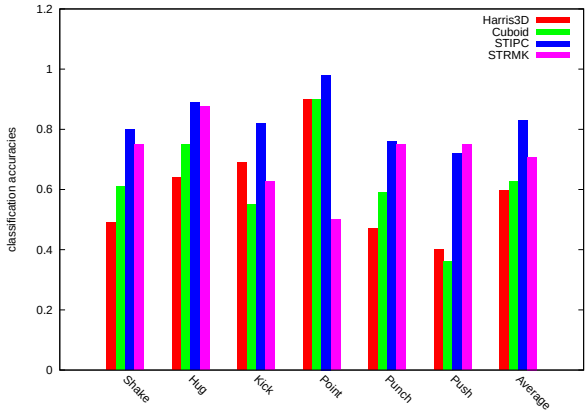
In order to evaluate and compare the performances, we duplicate experimental setting in [16]. Specifically, a 10-fold leave-one-out cross validation (each times, 9 samples are used for training and the left one is used for testing) is implemented on each set. Different methods are compared by using the average recognition rates.

In our experiments, we first generate a codebook of chain features using k-means, then represent activity with the bag-of-features model. We train the model of each class by SVMs with radial basis function kernel.

We compare our chain feature with two popular local space-time features, Harris3D feature [1] and cuboid feature [2]. The results of *Harris3D* and *Cuboid* are taken from the ICPR 2010 Contest on Semantic Description of Human Activities [17]. These results are averaged over 10 different codebooks. Fig. 5 shows the comparison results. We can see that our STIPC feature significantly outperform both of *Harris3D* and *Cuboid*, on both of Set 1 and Set 2. Specifically, on Set 1, the average classification accuracies of *Harris3D* and *Cuboid* are 64.2% and 75.5% respectively, whereas for the proposed STIPC feature, we get the average accuracy of 84.7%, which largely improves the Harris3D feature and Cuboid feature by 20.5% and 9.2% respectively. In set 2, our STIPC feature outperforms STIPs (59.7%) by 23.1% and Cuboid (62.7%) by 20.1% respectively. Our STIPC feature also outperforms the state-of-the-art results in [18], with a improvement of average classification accuracy of 12.9% over [18].



(a) UT-Interaction dataset set 1



(b) UT-Interaction dataset set 2

Fig. 5. Comparison of classification accuracies on UT-Interaction dataset set 1 and set 2.

V. Conclusion

We have presented a novel Spatio-Temporal Interest Points Chain feature (STIPC) for activity representation and recognition, which is designed to encode both of the discriminative discontinuous motion information (described with space-time descriptors) and continuous motion information among a long-term motion of an object or its part. STIPC make it possible to build the connection between local space-time interest points. We have also presented a method to encode the temporal relationships between local space-time features. Experiments on a challenging activity dataset have confirmed that our proposed STIPC outperforms the popular local space-time features.

References

[1] I. Laptev, , and T. Lindeberg, "On space-time interest points," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2003.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. Int. Conference on VS-PETS*, 2005.

[3] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. European Conference on Computer Vision (ECCV)*, 2008.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[5] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] J. Sun, X. Wu, S. Yan, L. F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition." in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2009.

[9] P. Turaga and R. Chellappa, "Locally time-invariant models of human activities using trajectories on the grassmannian," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] P. Siva and T. Xiang, "Action detection in crowd," in *Proc. British Machine Vision Conference (BMVC)*, 2010.

[11] H. Wang, A. Kläser, C. Schmid, and L. Cheng Lin, "Action Recognition by Dense Trajectories," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[12] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," in *Proc. European Conference on Computer Vision (ECCV)*, 2010.

[13] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 203–226, 2002.

[14] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

[15] H. Ling and S. Soatto, "Proximity distribution kernels for geometric context in category recognition," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2007.

[16] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR

contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[17] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (sdha) 2010," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[18] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2009.