A Combined SVM/HCRF Model for Activity Recognition based on STIPs Trajectories

Mouna Selmi, Mounim A. El-Yacoubi and Bernadette Dorizzi
Institut Mines-Telecom / Telecom SudPari, Evry, France

Keywords: Human Activity Recognition, Hidden Conditional Random Field, SVM/HCRF Combination, Space-time Interest Points' Trajectories.

Abstract: In this paper, we propose a novel human activity recognition approach based on STIPs' trajectories as local descriptors of video sequences. This representation compares favorably with state of art feature extraction methods. In addition, we investigate the use of SVM/HCRF combination for temporal sequence modeling, where SVM is applied locally on short video segments to produce probability scores, the latter being considered as the input vectors to HCRF. This method constitutes a new contribution to the state of the art on activity recognition task. The obtained results demonstrate that our method is efficient and compares favorably with state of the art methods on human activity recognition.

1 INTRODUCTION

Recognition of human activity from video sequences has a wide range of real-world applications such as video surveillance, assistance for elder care, human robot interaction, video indexing, etc. As a consequence, a large number of approaches have been devoted to human activity recognition. Based on the features used for recognition, these approaches can be classified into two categories: holistic approaches (Blank et al., 2005) and local approaches (Dollar et al., 2005); (Laptev, 2005). The first category uses explicit body representation by extracting features from the whole silhouette and exploits both spatial information and motion trajectory. However, it requires background segmentation and tracking of the body or of the body parts which may be difficult in the case of complex scenes that contain a dynamic background, fast motion and self occlusion. The second category uses local Interest Points. It achieves state-of-the-art performance for motion recognition tasks in complex real-world scenes when combined with a bag-of-word (BOW) representation. The major advantage of these approaches is to provide a concise representation of events while avoiding the pre-processing phases related to foreground/background segmentation and to tracking. These interest points are usually described by histograms of gradients (HOG) and histograms of optical flows (HOF). However, these gradient-based descriptors ignore the spatio-temporal layout of the local features which may be very informative. Addressing this problem, some recent research (Messing et al., 2009); (Matikainen et al., 2009); (Sun et al., 2009); (Wang et al., 2011) exploit the trajectories information of local features. These approaches outperform BOW based approaches in activity recognition tasks in complexes real-world scenes. In (Messing et al., 2009), trajectories of Harris interest points are used for complex daily living activity recognition. These trajectories are tracked with a standard KLT method and encoded as sequences of log-polar quantized velocities. In (Matikainen et al., 2009), Matikainen et al use a quantization of local features trajectory based on a k-means clustering and affine transformation. Sun et al. (Sun et al., 2009) recently used similar techniques to model SIFT-feature trajectories. They considered a fixed-dimensional velocity description using the stationary distribution of a Markov chain velocity model. Wang et al. (Wang et al., 2011) used dense trajectories by sampling dense points from each frame in multiple scales. The trajectories are described by HOG, HOF and motion boundary histogram (MBH) calculated in a volume surrounding each trajectory. However, these methods track local features that were proposed essentially for *image* recognition tasks which are not necessarily adapted to *space-time* data. These image features correspond to points with a significant local variation of intensities in the corresponding frame

Selmi M., A. El-Yacoubi M. and Dorizzi B. (2013).
A Combined SVM/HCRF Model for Activity Recognition based on STIPs Trajectories.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 568-572
DOI: 10.5220/0004267405680572

without any consideration of the temporal context.

In this work, we consider the use of the trajectories of local space-time interest points (STIPs) that correspond to points with significant local variation in both space and time, thus extending the approaches above which are limited to 2D interest points. In fact, STIPs have proven to be a strong feature extraction method that has given impressing results in real-world human action recognition tasks. Our motivation is that STIPs' trajectories can provide rich spatio-temporal information about human activity at the local level. For sequence modeling at the global level, a suitable statistical sequence model is required.

Hidden Markov Models (HMMs) (Rabiner, 1989) have been widely used for temporal sequence recognition. However, HMMs make strong independence assumptions on feature independence that are hardly met in human activity tasks. Furthermore, generative models like HMMs often use a joint model to solve a conditional problem, thus focusing on modeling the observations that at runtime are fixed anyway. To overcome these problems, Lafferty et al. (Lafferty et al., 2001) have proposed powerful discriminative models: Conditional Random Fields (CRF) for sequence text labeling. CRF is a sequence labeling model that has the ability to incorporate a long range dependency among observations. CRF assign to each observation in a sequence a label but it cannot capture intrinsic sub-structures of observations. To deal with this, CRF is augmented with hidden states that can model the latent structures of the input domain with the so called Hidden CRF (HCRF) (Quattoni, 2004). This makes it better suited to modeling temporal and spatial variation in an observation sequence. Such a capability is particularly important as human activities usually consist of a sequence of elementary actions. However, HCRF needs a long time range for the training phase. To overcome this problem we propose to combine HCRF with a discriminative local classifier (e.g SVM). The local classifier predicts confidence of activity labels from input vectors. We use the predicted confidence measurements of different classes from the local discriminative classifier as the input observation to the HCRF model. Assuming, as is the usual case, that the number of classes is significantly lower than feature dimensionality, this will reduce as much the feature space dimensionality during HCRF inference while exploiting the high discriminative aspect of SVM.

To summarize, the first objective of this paper is to investigate the use of STIPs' trajectories as activity descriptors. To the best of our knowledge, such a descriptor has not been addressed before in the state of the art. The second objective is to assess the discriminant power of HCRF-SVM combination on a daily living activities recognition task. This constitutes the second contribution of our work.

The organization of the paper is as follows. Section 2 gives a brief description of local space time features. HCRF and its combination with SVM are reviewed in Section 3. In Section 4, the databases used for experiments are described and results are detailed and compared with the state of the art. Section 5 draws some conclusions and sketches futures directions of this work.

2 LOCAL SPACE-TIME TRAJECTORIES

Local space-time features capture structural and temporal information from a local region in a video sequence. A variety of approaches exist to detect these features (Wang et al., 2009). One of the most popular methods is the one detecting Space Time Interest Points (STIP), proposed by Laptev et al. (Laptev et al., 2001), that extends Harris corner detector to the space- time domain. The main idea is to find points that have a significant change in space and time.

To characterize the detected points, histograms of gradients (HOG) and histograms of optical flows (HOF) are usually calculated inside a volume surrounding the interest point and used as descriptors.

To provide a description at the video action level, one of the most popular methods is to represent each video sequence by a BOW of HOG/HOF STIP's descriptors. However, this representation does not capture the spatio-temporal layout of detected STIPs. To overcome this limitation, a number of recent methods encode the spatio-temporal distribution of interest points. Nevertheless, these methods typically ignore the spatio-temporal evolution of each STIP in the video sequence. As mentioned above, some approaches have attained a good result when using the trajectories of 2D-interest points that are mainly adapted to 2D space domain. In this section, we present our approach of activity representation based on the trajectories of STIPs (Figure 1) which are adapted to video data.

To construct our basic feature, we first extract STIPs from the video sequences. Then we track them with Kanade-Lucas-Tomasi (KLT) tracker

(Lucas and Kanade, 1981) for a fixed number of frames T.

The trajectories description considered in this work is based on the following three temporal sequences:

- horizontal and vertical position trajectories:

$$P_n=<x_n, y_n>,$$

- path-tangent angle:

$$\theta_n = \arctan(\frac{y_n - y_{n-1}}{x_n - x_{n-1}}),$$

- path velocity magnitude:

$$V_n = \sqrt{(x_n - x_{n-1})^2 + (y_n - y_{n-1})^2}.$$

where $n = 1..N$ is the time index and N is the time duration.

As shown in Section 4, STIPs' trajectories outperform 2D-interest points' trajectories and HOG/HOF descriptor.

Figure 1: Example of STIPs trajectories in KTH dataset.

3 SVM-HIDDEN CONDITIONAL RANDOM FIELDS COMBINATION

HCRF is a powerful discriminative model that can be used to predict a label z from an input y. y is a vector of local observations $y = \{y_1, y_2, \ldots, y_m\}$ and each local observation y_i is represented by a feature vector, z is a member of a set Z which represents the class labels.

An HCRF models the conditional probability of a class label given a set of observations by:

$$P(z|y,\theta,\omega) = \sum_h P(z,h|y,\theta,\omega) = \frac{\sum_h e^{\Psi(z,h,y:\theta,\omega)}}{\sum_{z' \in Z, h \in H^m} e^{\Psi(z',h,y:\theta,\omega)}}$$

where the summation is over hidden state sequences **h**, and ω represents a specific window size. The potential function Ψ(z, **h**, y : θ, ω), parameterized by θ and ω measures the compatibility between a label, a set of observations and a configuration of the hidden states. The following objective function is used in estimating the parameters:

$$L(\theta) = \sum_{i=1}^{n} \log P(z_i|y_i,\theta,\omega) - \frac{1}{2\sigma^2}\|\theta\|^2.$$

where n is the total number of training sequences. The first term is the conditional log likelihood of the labels given the data while the second is the log of a Gaussian prior with variance σ^2.

The training phase aims at finding the best parameters, $\theta^* = \arg\min -L(\theta)$. It can be carried out by gradient descent using LBFGS optimization method (Byrd et al., 1995).

Given a new test sequence y and the parameters values $\theta*$ learned from training examples, the label for the sequence is determined by:

$$\tilde{z} = \arg\max_{z \in Z} P(z|y,\omega,\theta^*).$$

As mentioned in the introduction, HCRF suffers from a slow convergence during training. In fact, HCRF uses, for parameter estimation, the forward–backward algorithm that needs to scan the training set many times. The training duration is correlated to the dimension of data. To deal with this problem, we use a SVM-HCRF combination. Using the probability scores given by SVM as input of HCRF will reduce considerably feature's dimension and so will accelerate the training and testing phases. In addition, SVM is a popular method for classification tasks due to its high discriminative and generalization properties. It also has the ability to use high-dimensional feature spaces via kernels. Thus, using SVM as a local classifier can improve the recognition rate while ensuring a significant training speed-up.

To train SVM, first we split the video sequence into segments. Each segment is locally represented by a local Bag of Words (frequency histogram of visual word occurrences): visual words are constructed using k-means based on trajectories descriptors. We train SVM on these local Bag of Words and the generated activity probability scores will be considered as the input of HCRF.

4 EXPRIMENTS

We start this section by an evaluation of STIP's trajectories performance, and then we evaluate the performance of our model in a complex daily living activity recognition task.

4.1 STIPs Trajectories Evaluation

In this section, we evaluate the proposed STIPs' trajectories in action recognition task and compare it

to state-of-the-art methods. We provide detailed analysis below.

4.1.1 Dataset Description

The KTH dataset (Schuldt et al., 2004) is one of the most common datasets in evaluation of action recognition. It consists of six types of human actions: boxing, hand clapping, hand waving, jogging, running and walking; performed by 25 subjects in 4 different scenarios. We divide the samples into test set (9 subjects) and training set (16 subjects) as is usually done.

4.1.2 Evaluation Framework

In this section, we compare STIPs and Speeded Up Robust Features (SURF) (Bay et al., 2006) trajectories. SURF is a scale and rotation invariant detector of distinctive key points from images. The extracted points correspond to corners detected on the integral image. We also compare the performance of STIPs' descriptors: trajectories and HOG/HOF. Theses comparisons are based on the bag of features approach where each video sequence is represented as the frequency histogram over visual words. We begin by extracting STIPs. We track them during 15 frames using KLT tracker. STIPs' trajectories are then quantized into visual words using k-means clustering. In our experiments, we set the number of visual words to 80 which gives the optimal result. Finally, the video descriptor is obtained by assigning every local descriptor to the nearest visual word. This descriptor is the input of a Gaussian kernel SVM.

4.1.3 Results and Interpretation

As Table1 shows, our trajectory descriptor outperforms the HOG/HOF descriptor that contains rich information about texture and local motion. And it also outperforms the SURFs' trajectories' descriptor based on Velocity histories. The accuracy gained is principally due to the fact that STIPs are sparse and correspond to points with significant local variation in both space and time. In addition, most detected SURFs belong to the background.

Table 1: Comparison results on KTH dataset.

Methods	Rates (%)
STIPs trajectories	**84.25**
HOG/HOF [4]	80
Velocity Histories(Messing et al., 2009)	74

4.2 SVM-HCRF Combination

4.2.1 Dataset Description

We evaluate our model on the Rochester dataset (Messing et al., 2009). It consists of 10 complex daily living activities: answering a phone, dialling a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. It was each performed three times by five persons. To ensure appearance variation, the activities are performed by people having different shapes, sizes, genders, and ethnicities.

4.2.2 Evaluation Framework

To evaluate our activity recognition system, we compute recognition accuracy using the leave-one-person-out cross-validation method. Each time, we first leave out all the sequences pertaining to one person. Then, we train the model using all the remaining sequences, and we use the 10 activities of the omitted person as test data. We average out the results from all the persons to obtain the average recognition rate.

As stated above, we first extract STIPs' trajectories for L=15 frames. For each trajectory, we calculate its basic descriptor (P_n, θ_n,V_n) sequence. Then, we construct K visual words using k-means (K is empirically optimized, K=100). Each video is split into smaller segments using a sliding window of length 10. Each segment is locally represented as a frequency histogram of word occurrences (local Bag of Words). Based on these features, we train a Gaussian kernel SVM in order to obtain the activity probability scores for each segment. These probability vectors are the input of our HCRF model which is trained using ten hidden states.

4.2.3 Results and Interpretation

Table 2 compares our results to the state of the art. Our approach significantly outperforms STIP-based Bag-of-word approach based on HOG/HOF descriptor. It also outperforms considerably the recognition rate given by the "Latent Velocity Histories" method (Messing et al., 2009): This method is based on the velocity histories of the SURF trajectories and Markov chain. Note that the authors (Messing et al., 2009) have obtained an important improvement (22%) by adding texture and

color information. However, it is clear that color information is actually irrelevant to activity recognition. Although in (Messing et al., 2009), it brings high improvements, this is due to the fact the activities are performed in the same kitchen consisting of the same objects (refrigerator, plates, knives, etc.) whose color does not change. Had the activities were performed in different locations with different objects, using color would have actually decreased performance.

Table 2: Results comparison on Rochester dataset.

Methods	Rates (%)
Our approach	**80**
BOW of HOG/HOF (Messing et al., 2009)	59
Latent Velocity Histories(Messing et al., 2009)	67
Augmented Velocity Histories(Messing et al., 2009)	89

5 CONCLUSIONS

In this work we have introduced an approach for activity recognition based on STIPs' trajectories and HCRF model. Our STIPs' trajectories are more robust than state-of-the-art interest point descriptors. In fact, this descriptor captures the motion information of efficient sparse interest points, namely "3D Harris points".

We also have described a new approach to train an HCRF for high-dimensional feature sequences. Our approach is based on SVM-HCRF combination. It is faster and more scalable than standard HCRF model.

Experiments show that the proposed method achieves high accuracy. The results obtained using SVM-HCRF based on STIPs' trajectories compare favorably with the state of the art on the same dataset. This is quite promising if we bear in mind that no color information is considered in our approach.

We have shown that trajectory descriptors outperform STIP descriptors by a margin of 5%, while consisting of a radically different feature representation. The former is based on the intrinsic information of the trajectory while the later is based on a rough HOG/HOF description of STIP neighborhood. An interesting consequence is that the two representations are good candidates for feature combination as their orthogonality means a real potential for accuracy improvement. We intend to explore this direction in our future work.

REFERENCE

M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, (2005). Actions as space-time shapes, 2005. In *International Conference on Computer Vision.*

P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65-7.

I. Laptev (2005). On space-time interest points. In. *J. Comput. Vision*.

R. Messing, C. Pal, and H. Kautz, (2009). Activity recognition using the velocity histories of tracked keypoints. In *ICCV*.

P. Matikainen, M. Hebert, and R. Sukthankar, (2009). Trajectons: Action recognition through the motion analysis of tracked features. In *ICCV workshop on Video-oriented Object and Event Classification.*

J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, (2009). Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*.

Heng Wang, Alexander Kl¨aser, Cordelia Schmid, Cheng-Lin Liu, (2011). Action Recognition by Dense Trajectories. In *CVPR*: 3169-3176

Rabiner, L., (1989). A tutorial on HMM and selected applications in speech recognition. *In Proceedings of the IEEE*, 77(2):257–286.

J. Lafferty, A. McCallum, and F. Pereira, (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*.

A. Quattoni, M. Collins, and T. Darrell, (2004). Conditional random fields for object recognition. In *NIPS*.

H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, (2009). Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*.

B. D. Lucas and T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679.

Bay, H., Tuytelaars, T., and Gool, L. J. V., (2006). Surf: Speeded up robust features. In *ECCV (1)'06*, pages 404–417.

R. H. Byrd, P. Lu and J. Nocedal, (1995). A Limited Memory Algorithm for Bound Constrained Optimization, *SIAM Journal on Scientific and Statistical Computing*, 16, 5, pp. 1190–1208.

C. Schuldt, I. Laptev, and B. Caputo, (2004.). Recognizing human actions: A local svm approach. In *ICPR*, pages 32–36.