Time-slice Prediction of Dyadic Human Activities

Maryam Ziaeefard[1]
maryam.ziaeefard.1@ulaval.ca
Robert Bergevin[1]
robert.bergevin@gel.ulaval.ca
Louis-Philippe Morency[2]
morency@cs.cmu.edu

[1] Computer Vision and Systems Laboratory
Laval University
Quebec, CANADA

[2] Language Technology Institute
Carnegie Mellon University
Pittsburgh, USA

Abstract

Recognizing human activities from video data is being leveraged for surveillance and human-computer interaction applications. In this paper, we introduce the problem of time-slice activity recognition which aims to explore human activity at a smaller temporal granularity. Time-slice recognition is able to infer human behaviors from a short temporal window. It has been shown that the temporal slice analysis is helpful for motion characterization and in general for video content representation. These studies motivate us to consider time-slices for activity recognition. To this intent, we propose a new family of spatio-temporal descriptors which are optimized for early prediction with time-slice action annotations. Our predictive spatio-temporal interest point (Predict-STIP) representation is based on the intuition of temporal contingency between time-slices. Furthermore, we introduce a new dataset which is annotated at multiple short temporal windows, allowing the modeling of the inherent uncertainty in time-slice activity recognition. Our experimental results show performance comparable to human annotations.

1 Introduction

Humans are good at anticipating and correctly predicting the activities of others during social interactions. For example, we do not need to see a full handshake before being able to recognize it. In fact, two people getting closer and lifting hands will most likely shake hands. Humans can naturally model the uncertainty associated with activity recognition. While great progress has been made in computer-based human activity recognition this past decade, computational algorithms are often lacking the predictive capabilities of humans. Also, most recent approaches are expecting a complete video with a large temporal window. Based on intuition from social psychology, we introduce a time-slice approach to human activity recognition which is based on short-term observations. We are interested in improving our understanding of the inherent uncertainty occurring with time-slice observations and building computational algorithms to properly model them. This work has several practical applications, outside the basic research question of better understanding human and computer perception of dyadic actions. It can be beneficial when the whole video stream is not

Pages 167.1-167.13
DOI: https://dx.doi.org/10.5244/C.29.167

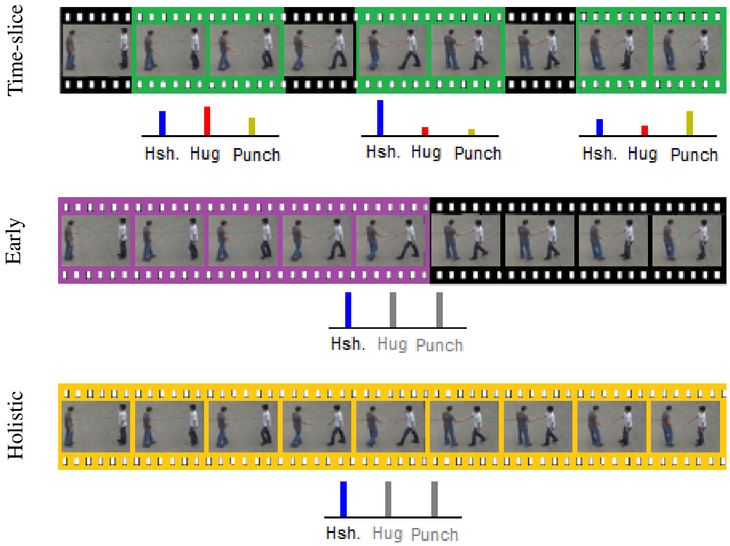


Figure 1: An illustration of human activity recognition problems: The first row illustrates "time-slice" recognition and the labels, i.e., Handshake (Hsh.), Hug, and Punch for different time-slices. The second and third rows show "early" recognition and "holistic" approaches where the label is the same for the whole sequence.

available and activities are not recorded from the start to the end. It can also be useful in video indexing, retrieval, and analysis.

We present in Figure 1 an overview of our approach based on time-slice action prediction and contrast it with the conventional approaches which recognize actions based on either the whole video sequence (referred as "holistic" approach) or the first part of it (early recognition) [30]. Our time-slice approach studies not only the beginning of the action sequence but generalizes this to any short-term observation anywhere in the video sequence. Another key novelty is in the explicit modeling of the uncertainty occurring when predicting actions based on time-slices.

In this paper, we propose a new set of spatio-temporal descriptors using time-slice action annotations for early activity prediction. We show our predictive spatio-temporal interest point (predict-STIP) representation is able to infer time slices of human activities based on discriminative descriptors. We select feature descriptors which are discriminative when an action is clearly occurring during a time-slice and is also visible outside on time-slices with uncertain action. Given their broader temporal range, we hypothesize that these descriptors are better at prediction actions. An overview of our method "Predict-STIP" is illustrated in Figure 2. Our goal is to identify descriptors with broad temporal coverage. The details on how we do it can be described later. Our representation is amenable to early activity recognition. We show the comparison results of this work with the state-of-the-art in early activity recognition.

We introduce a new dataset, named Time-slice Action Prediction (TAP) dataset, to evaluate our proposed feature descriptors and enable future research on this topic. Our dataset could also be used for early activity recognition as well as holistic activity recognition. The dataset was created by extracting time-slices from existing public human action datasets and

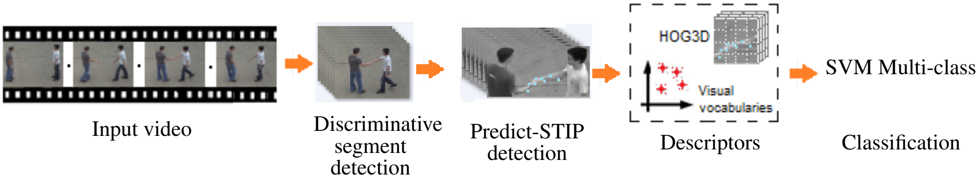


Figure 2: An overview of our method, Predict-STIP. Given an input video sequence, we first extract discriminative segments and then detect Predict-STIPs. HOG3D representation and BoW models are applied to prepare inputs for SVM classifiers.

perform a perception study with multiple annotators giving continuous ratings for each action. The continuous ratings allow to represent the uncertainty in time-slice action prediction.

The outline of the paper is as follows. Section 2 provides an overview of the most relevant works to our paper in activity recognition. We present our new dataset in Section 3. Section 4 explains the methodology of our proposed method. Section 5 shows our experimental results, followed by conclusions in Section 6.

2 Related Work

A number of surveys have been published in activity recognition over the past decade [2, 27, 35]. Given the significant literature review in this area, we focus only on the most relevant works.

Partially observed videos: Very few works have been devoted to recognizing activities from partially observed videos. Ryoo [30] performed the first attempt in early activity recognition and studied how feature distributions change over time. Li and Fu [20] used autoregressive moving average model, ARMA, to model the temporal order of activities for early recognition. Raptis and Sigal [28] trained a model to recognize actions in videos using key-poselets as latent variables for partially observed activity recognition. Yu [38] trained a model using relative locations of space-time points extracted from a video to the center position of that video. A Semantic framework was proposed by Li *et al.* [21] for early recognition of long-duration complex activities by discovering the casual relationships between action units. Early event detection and recognizing human activity from gapped videos have also been studied in [3, 13] which used partially observed videos as input.

Space-time interest points: Recently, space-time interest points (STIPs) have received increasing interest due to their scarcity and reasonable performance for activity recognition. STIP-based methods are invariant to geometric transformations which result in low variation by changes in scale, rotation, and viewpoint. Laptev and Lindeberg [19] proposed the notion of STIP built on the idea of the Harris and Stephens interest point operators [12]. Several other methods have been reported [9, 14, 25] to improve STIP detection for human activity recognition. Chakraborty *et al.* [5] proposed a model for robust Selective STIP detection (S-STIPs) by applying background suppression as well as local and temporal constraints. This method outperforms existing STIP detection techniques and detects more stable and distinctive STIPs. We benefit from the advantages of S-STIPs to extract the initial interest points in our work. For exploring more approaches, we refer readers to a recent comprehensive survey of human action recognition with STIP detector by Das Dawn and Shaikh [8].

Key-components: The use of informative components (frames or time-slices) is in contrast to most research in video-based action recognition which often extracts features from much longer videos. Using a sparse set of frames allows the model to focus on the most discriminative parts of the action which are referred to key-frames in literature review [4, 22, 32, 39]. Key-frames are discrete sets of frames that capture discriminative parts of a video. On the other hand, time-slices are continues sets of frames which represent temporal ordering and dynamic structure of the discriminative part of a video. This paper is the first effort to introduce time-slice for activity recognition.

Trajectory data: Among the local space-time features, tracking interest points through video sequences have been shown to be an efficient representation for action recognition [24, 33, 36]. Shape, appearance, and motion descriptors are extracted from the trajectories of interest points to analyze detailed levels of human movements. Sun *et al.* [33] represented activities using trajectory transition and trajectory proximity descriptors. The trajectory extraction process is based on matching SIFT descriptors between two consecutive frames. The descriptors that are too far apart are discarded. Wang and Schmid [36] proposed a method using improved dense feature trajectories. They estimated the camera motion and removed it from the optical flow to have better motion-based descriptors. In this paper, we track the position of specific spatiotemporal interest points backward and forward in time and extract predictive features based on the persistency of this trajectory data.

3 TAP Dataset

We are interesting in social interactions, more specifically dyadic interactions. We preferred using publicly available datasets so that people can replicate and extend our experiments. We looked for datasets with similar action labels in order to make the time-slice annotation task possible for crowdsourcing.

We have extracted 2132 time-slices from 4 challenging datasets, i.e., UT-Interaction (segmented sets 1 and 2) [29], HMDB [17], TV Interaction [26], and Hollywood [18] datasets. Each time-slice contains one of seven interactions: handshake, high five, hug, kick, kiss, punch, and push. The dataset also contains 204 negative examples, time-slices of full videos that do not have any of the mentioned interactions. We performed a preliminary experiment to validate how many frames were necessary to have good agreement between annotators. We requested some annotators to recognize dyadic interaction examples with 5-, 10- and 15-frame time-slices. We decided to choose 10-frame time-slices for our work since 5-fram time-slices were too short and 15-frame time-slices were not fit to our goal which is studying the inherent uncertainty in activities. Our dataset is available as a public dataset to encourage researchers to continue this line of research. [1]

During our experiments, we grouped together videos from constrained and unconstrained datasets. Constrain, here, refers to the restriction in the settings and activity executing. UT-interaction is our constrained dataset which contain acted interactions with a fixed background and profile viewpoint that are performed for research purpose. On the other hand, unconstrained datasets include activities which are taken in realistic settings, e.g. from TV shows. Unconstrained datasets are more challenging for activity recognition. HMDB, TV Interaction, and Hollywood are our unconstrained datasets. We selected videos of these datasets based on the camera angle ranging from -45 to +45 degree.

[1] http://vision.gel.ulaval.ca/en/Projects/Id$_3 29/Projet.php$

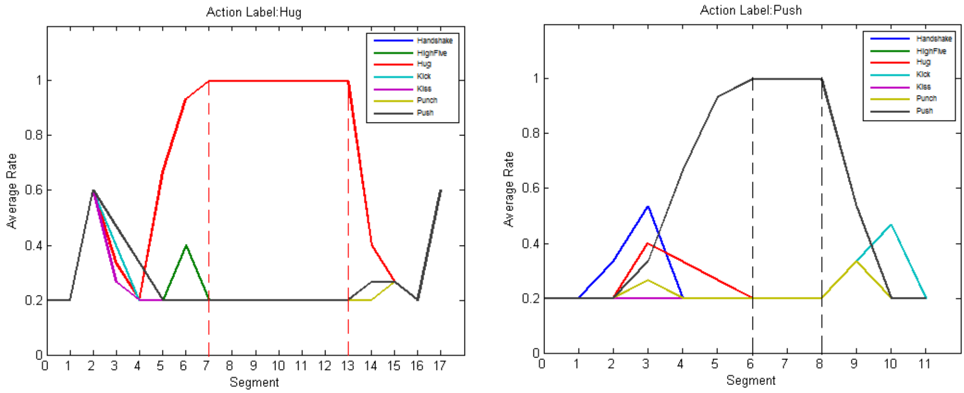


Figure 3: Human annotation: This figure shows the average rate of 3 annotators for two video examples: hug and push. For each possible activity and for each time-slice, the label provided by one annotator is first converted to a number on a linear scale from 0 to 1. The average of those numbers for more than one annotator (we used 3 here) is called the average rate annotation for the time-slice. This average rate will be used to evaluate the performance of our method. Time-slices between dashed lines is the discriminative segment of the interaction.

All time-slices was annotated by multiple online annotators (using the Crowdflower platform [1]). 3 annotators rated each time-slice on how likely a specific action is occurring. For each time-slice and for each action, the annotator was asked to pick one of 5 likelihoods from “Definitely Not Occurring” to “Definitely Occurring”.

Figure 3 illustrates how annotators rated for two example videos. From the figure, we can see the confusion and uncertainty of annotators in first time-slices of videos. As time passes and more information about the activity of interest is observed, they will be better able to recognize the activity.

4 Methodology

Our approach to dyadic human activity recognition consists of three major contributions: i) a new learning approach in which discriminative video segments are used on the basis of human annotations and efficient spatio-temporal features (called predictive) are obtained on the basis of their persistence and ii) a more general definition of the activity recognition problem in which the uncertainty arising from observing a short time-slice from anywhere in the video sequence is explicitly taken into account, and iii) a demonstration that a baseline multi-label classification method can reproduce the features of the human annotation using the proposed learning model for this problem. We introduce discriminative segments since we need feature descriptors which are good when the human agree that an action is clearly occurring. We also require descriptors which have predictive powers when their broader temporal range is considered. We first determine discriminative segments of each video activity based on annotated data where all annotators agreed an interaction of interest is occurring. We then use these segments to select predictive space-time interest points. Each predictive point is described by motion and appearance descriptors to learn the model. In the following subsections, the above steps are explained in more detail.

4.1 Discriminative segments

When analyzing an interaction, we can definitely recognize the ongoing activity from specific time slices such as “two people are shaking each other’s hands” slice in handshaking activity. These slices are referred to discriminative segments in this paper. Discriminative segments, therefore, encode the most relevant slices of video to interested interaction. We define the temporal location of a discriminative segment based on annotated data. In preparing our dataset, we asked 3 users to annotate each time-slice as described in Section 3. To measure the reliability of agreement between annotators, we used Fleiss’ kappa coefficient k [11] that assesses the agreement between more than two raters. This coefficient takes into consideration the agreement occurring by chance as shown in Equation 1. For each interaction video, time-slices where the annotators are in complete agreement, i.e. k=1, on definitely including the interaction of interest, are selected as discriminative segments.

$$k = \frac{\bar{P}_i - P_l}{1 - \bar{P}_i} \tag{1}$$

where $\bar{P}_i$ is the mean value to which annotators agreed for the certain interaction of interest and P_l is the sum of the square of the quantity of all assignments which were to the certain likelihood category. The degree of agreement that is achievable above chance and actually attained above chance are provided by factors $1 - \bar{P}_i$ and $\bar{P}_i - P_l$ respectively.

4.2 Predict-STIP

In this paper, we follow the recent progress in STIP-based recognition strategy. Existing STIP detectors are vulnerable to model the inherent uncertainty in partially observed action recognition and prediction, and therefore, are insufficient for time-slice recognition. To overcome this problem, we introduce a predictive representation which measures how long STIPs are observable in a video. STIPs which are active during the whole video are selected as Predict-STIP (P-STIP). In other words, P-STIPs are the STIPs that exist in first frames of the video and still will appear in upcoming frames.

Given a set of interaction video sequences $\{A_i \mid i = 1 : n\}$ and their associated discriminative segments $\{S_i \mid i = 1 : n\}$, our purpose is to detect P-STIPs P_i of each A_i. Our input variables are sequence of frames $A_i = \{f_i^1, ..., f_i^{e_i}\}$ and $S_i = \{s_i^1, ..., s_i^{N_i}\}$ where e_i and N_i are the length of the full video and the discriminative segment, respectively. To extract P-STIP, we first detect “$stip_{New}$” of s_i^1 as initial landmarks. We then track them backward and forward using Kanade-Lucas-Tomasi (KLT) algorithm [23, 34] to f_i^1 and $f_i^{e_i}$ and check whether or not they have existed during the whole video. We repeat these steps for all frames of S_i. Landmarks that are continuously observable are selected as P-STIPs P_i:

$$P_i = \{p_{(x_{j,t}, y_{j,t})} \in s_i^t, \quad t = 1, ..., N \mid \forall p \quad V_p = 1\} \tag{2}$$

where V is a validity matrix provides a logical array, indicating whether or not each point has existed during the whole video. $(x_{j,t}, y_{j,t})$ is the position of $stip_{New}$ p_j in the frame s_i^t.

To speed up the tracking step and increase the efficiency of our algorithm, we select a new subset of S-STIPs [5], $stip_{New}$, instead of using all densely sampled S-STIPs from S_i. We initialize $stip_{New}$ as S-STIPs extracted from s_i^1 and track them. We then generate putative matches between previously tracked-$stip_{New}$, $stip_T$, and extracted S-STIPs, $stip_E$, of the current frame by finding points that have minimal differences in oriented phase data within

Algorithm 1 Predict-STIP detection from a discriminative segment

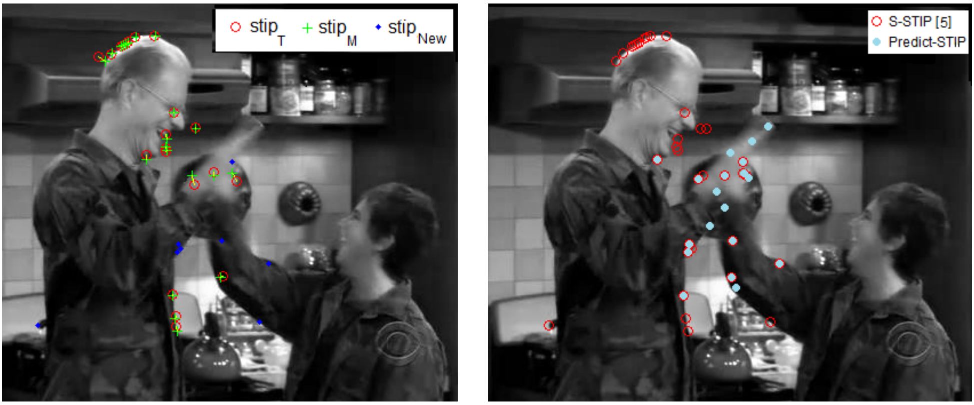
Input: Discriminative segment ($H \times W \times N$): S;
$S = \{s^i \mid i = 1 : N\}$ (contains all frames of a discriminative segment)
Definition:
f_1: The first frame of the full video
f_e: The last frame of the full video
V: Validity matrix provides an M-by-1 logical array, indicating whether or not each point has existed
Ensure: Predict-STIP: $PredectivePoints$

```
$N = size(S,3)$; (Total no. of the discriminative segment's frames)
Initialize $stip_{New}$
Initialize $stip$
for $i = 1 \rightarrow N$ do
    Track $stip_{New}$ backward to $f_1$ and forward to $f_e$ and restore $V$ matrixes
    Let $stip_T$ be $stip$ tracked from $s^{i-1}$
    Let $stip_E$ be S-STIPs extracted from $s^i$
    Match $stip_T$ with $stip_E$ and set $stip_M = stip_T \cap stip_E$
    Update $stip_{New}$ via $stip_E \notin stip_M$
    Update $stip$ via $stip_N \cup stip_T$
end for
Check $V$ matrixes
Find points where $V_{points}$ are always equal to 1 and set as $PredectivePoints$
Return ($PredectivePoints$)
```

windows surrounding each point [16]. Only points that correlate most strongly with each other in both directions are returned as matched points, $stip_M$. Oriented phase data matcher performs better compared to normalized grayscale correlation. We also set a maximum search radius threshold for matching points to improve speed and accuracy since we do not want to match points, e.g. from an arm with points extracted from a leg. Consequently, only points whose Euclidean distance is below the threshold are considered for matching. Afterward, we employ RANSAC algorithm [10] to estimate the fundamental matrix from matching point pairs to excludes outliers and identify strong inliers. Figure 4.a illustrates the matching result of a sample frame. Finally, we update $stip_{New}$ via $stip_E$ that does not belong to $stip_M$. Therefore, the $stip_{New}$ is a new subset of S-STIPs that are not tracked from previous frames and appear in each frame. The pseudo code for the full predictive feature detection is described in Algorithm 1. Figure 4.b shows P-STIPs extracted by our approach and S-STIPs resulted from [5].

4.3 Descriptors and vocabulary building

Several local and global descriptors have been proposed in the past few years for STIP-based methods [6, 7, 15, 31, 37]. In this paper, we use HOG3D descriptors [15] to represent each interaction video. The HOG3D descriptor is based on histograms of 3D gradient orientations, where mean gradient vectors are computed using integral videos. With integral videos, 3D gradients can be efficiently calculated for any arbitrary points in a video. Given P-STIPs



(a)

(b)

Figure 4: **Predict-STIP detection. The matching result of a sample "high five" action is shown in the left figure. The right figure displays S-STIP [5] and our predict-STIP extracted from the example.**

of each interaction video, we construct the HOG3D representation. Local regions are determined first by extracted P-STIPs and then histograms of gradient orientations are computed over a set of gradient vectors from the cuboid neighborhood (4x4x4) around the P-STIPs. All histograms are concatenated to one descriptor vector for each video.

We compute the basic Bag-of-words model and quantize the descriptor vectors, HOG3D extracted at P-STIPs, into 1000 bins associated with visual words using K-means clustering. BoW features are normalized so their L1 norm is 1.

4.4 Learning

The goal of our Predict-STIP method is to determine the interaction category of time-slices of video X among a set of classes $\{1,...,K\}$. Therefore, our purpose is to learn a mapping $f(O) \rightarrow \{1,...,K\}$ where $O \subset X$ refers to the time-slice observations and may occur at any time in the video. We present the videos with BoW descriptors obtained from P-STIPs. For each class of interaction, we learn a model with the corresponded BoW descriptors using multi-class SVM framework in the training phase.

At test time, a query video v_i which is a time-slice of a longer video matched to the models according to the learned appearance and motion predictive features. To this intent, we extract S-STIPs [5] from v_i and match them to the pool of trained P-STIPs. S-STIPs of v_i that matched to P-STIPs are selected as P-STIPs of v_i (lookup table technique). Then BoW descriptors of v_i are extracted. Classification is made based on the score of interaction class-specific models applied on BoW descriptors.

5 Evaluation of predictive model

We present experimental results on two scenarios of our TAP dataset: constrained and unconstrained sets.

5.1 Constrained set

Samples in constrained set are time-slices of 5 interactions (handshake, hug, kick, punch, and push) collected from UT-Interaction dataset. To extract Predict-STIPs, we use a matching function with two adjustable parameters: matching window size and maximum search radius. We set the matching window size to 11 empirically and maximum search radius to 10 according to the resolution of images in the dataset. The number of P-STIPs is different from one action to another action and varies between 15-30.

We evaluate the time-slice recognition performance by using the standard "leave-one-out" method, one video is out each time, and fit the recognition problem in the context of multi-class classification. The average precision for all interactions (compared to human annotation) is given in the second column of Table 1. In order to visualize the performance of our method, we draw its average precision on a per time-slice basis and compare it to the average rate of human annotators (see Figure 5). Because the number of time-slices is not the same for all videos, we compute the averages using video examples with the same number of time-slices. From the figure, we can see in some time-slices our approach outputs higher values than human annotators, e.g. time-slices 11 and 12 for the Hug action. In those cases, our method is thus better in recognizing the action from some limited time-slices.

Interestingly, since Predict-STIP is sufficient for holistic and early activity recognition, we can also compare it with the state-of-the-art on UT-Interaction dataset for those two different recognition context problems. Table 2 shows that our predictive representation outperforms all the state-of-the-art methods.

5.2 Unconstrained set

The unconstrained set is more challenging than constrained set in terms of background clutter, the number of people in the scene, the number of interactions, camera motion, and changes of viewpoints. This set includes time-slices of 6 realistic human interactions (handshake, high five, hug, kick, kiss, and punch) collected from HMDB, TV Interaction, and Hollywood TV show datasets.

The experimental setting of this set is similar to constrained set. The performance of Predict-STIP on this set is also reported in Table 1. The results are obtained based on the number of correctly labeled time-slices compared to the human annotation. From the table, we can see that the results on constrained set are better than unconstrained set because unconstrained set is more challenging. We can also see that the best results are obtained in handshake interaction for both datasets. High five interaction, meantime, has the minimum accuracy rate among 7 listed interactions.

6 Conclusions

In this paper, we have introduced a predictive representation for a new problem of time-slice activity recognition. Time-slice activity recognition aims at exploring and recognizing an activity using a portion of the whole activity. We represented each video based on spatio-temporal descriptors of predictive features extracted from discriminative video segments. We also showed the effectiveness of our approach in a new dataset and compared it to the state-of-the-art. This dataset is available as a public dataset to encourage widespread researchers to explore human activities at a smaller temporal granularity.

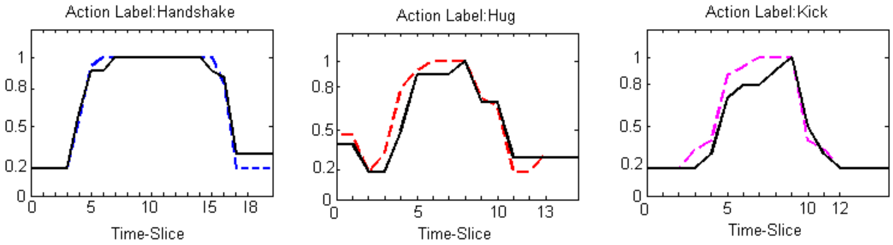


Figure 5: Comparison results of our method with the human annotation. Dashed and black lines show average rate by annotators and average precision of our method at test time, respectively.

	constrained set	unconstrained set
handshake	82%	76.3%
high five	–	61.4%
hug	81%	71%
kick	78%	73.7%
kiss	–	74%
punch	80%	76.2%
push	75%	–

Table 1: The average precision of Predict-STIP on constrained (UT-interaction dataset) and unconstrained sets (selected videos from HMDB, TV Interaction, and Hollywood TV show datasets).

Acknowledgements: This work is supported by a Discovery Grant to Professor Robert Bergevin from the Natural Sciences and Engineering Research Council of Canada and the FRQNT International Internship Grant to Maryam Ziaeefard. This material is also partially based upon work supported by the National Science Foundation under Grant No. IIS-1118018. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government.

Method	Accuracy with half observation	Accuracy with full observation
Our Model	**83%**	**95%**
Raptis and Sigal [28]	73.3%	93.3%
Yu *et al.* [38]	80%	91.7%
Ryoo (Best) [30]	70%	85%
Ryoo and Aggarwal (Best) [29]	31.7%	85%

Table 2: Performance comparison on the UT-Interaction Dataset. Early recognition and holistic recognition results are reported on the second and third columns, respectively.

References

[1] URL www.crowdflower.com/.

[2] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428 – 440, 1999.

[3] Yu Cao, D. Barrett, A. Barbu, S. Narayanaswamy, Haonan Yu, A. Michaux, Yuewei Lin, S. Dickinson, J.M. Siskind, and Song Wang. Recognize human activities from partially observed videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2658–2665, June 2013.

[4] Stefan Carlsson and Josephine Sullivan. Action recognition by shape matching to key frames. In *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, 2001.

[5] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, and Jordi Gonzalez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396 – 410, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[7] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Computer Vision ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 428–441. Springer Berlin Heidelberg, 2006.

[8] Debapratim Das Dawn and SoharabHossain Shaikh. A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, pages 1–18, 2015.

[9] P. Dollar, V. Rabaud, G. Cottrell, and S. 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*, pages 65–72, Oct 2005.

[10] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[11] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[12] Chris Harris and Mike Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[13] M. Hoai and F. De la Torre. Max-margin early event detectors. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2863–2870, June 2012.

[14] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[15] Alexander Kläser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

[16] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[19] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003.

[20] Kang Li and Yun Fu. Arma-hmm: A new approach for early recognition of human activity. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1779–1782, Nov 2012.

[21] Kang Li and Yun Fu. Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8): 1644–1657, Aug 2014.

[22] Li Liu, Ling Shao, and Peter Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 46(7):1810 – 1818, 2013.

[23] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, pages 674–679, 1981.

[24] R. Messing, Chris Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 104–111, Sept 2009.

[25] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(3):710–719, June 2005.

[26] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *Proceedings of the British Machine Vision Conference*, pages 50.1–50.11. BMVA Press, 2010.

[27] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[28] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2650–2657, 2013.

[29] 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.

[30] M.S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1036–1043, Nov 2011.

[31] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, pages 357–360, 2007.

[32] Yasaman S. Sefidgar, Arash Vahdat, Stephen Se, and Greg Mori. Discriminative key-component models for interaction detection and recognition. *Computer Vision and Image Understanding*, 135(0):16 – 30, 2015.

[33] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2004–2011, June 2009.

[34] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, International Journal of Computer Vision, 1991.

[35] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, Nov 2008.

[36] Heng Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558, Dec 2013.

[37] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision ECCV 2008*, volume 5303, pages 650–663. Springer Berlin Heidelberg, 2008.

[38] Gang Yu, Junsong Yuan, and Zicheng Liu. Predicting human activities using spatio-temporal structure of interest points. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1049–1052, 2012.

[39] Zhipeng Zhao and Ahmed M. Elgammal. Information theoretic key frame selection for action recognition. In *BMVC*, 2008.