Time-slice Prediction of Dyadic Human Activities

Maryam Ziaeefard
http://vision.gel.ulaval.ca/en/people/ld_842/index.php
Department of Electrical Engineering, Laval University

Robert Bergevin
http://vision.gel.ulaval.ca/en/people/ld_18/index.php
Department of Electrical Engineering, Laval University

Louis-Philippe Morency
http://www.cs.cmu.edu/~morency/
School of Computer Science, Carnegie Mellon University

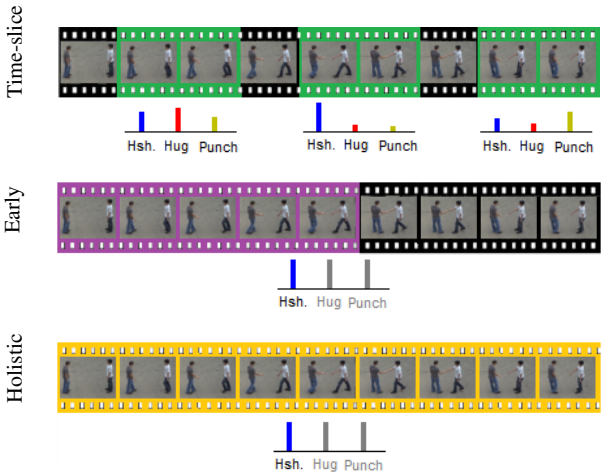

Figure 1: An illustration of human activity recognition problems: The first row illustrates "time-slice" recognition and the labels, i.e., Handshake (Hsh.), Hug, and Punch for different time-slices. The second and third rows show "early" recognition and "holistic" approaches where the label is the same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

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.

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). 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.

TAP Dataset: 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. The dataset was created by extracting time-slices from existing public human action datasets (UT-Interaction, HMDB, TV Interaction, and Hollywood datasets) and 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. 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.

Methodology: **Stage 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. To extract discriminative segments from our dataset, we used Fleiss' kappa coefficient k [2] to measure the reliability of agreement between annotators. 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.

Stage 2- Predict-STIP: Existing STIP detectors are vulnerable to model the inherent uncertainty in partially observed action recognition

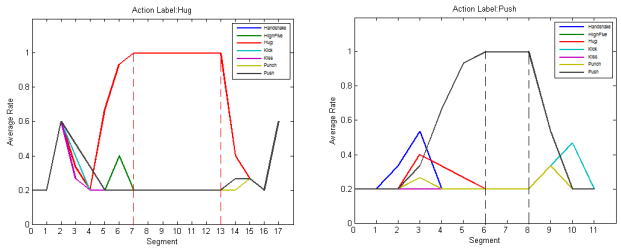

Figure 2: Human annotation: This figure shows the average rate of 3 annotators for two video examples: hug and push. The label provided by one annotator is converted to a number on a linear scale from 0 to 1 called the average rate. This average rate will be used to evaluate the performance of our method. Time-slices between dashed lines is the discriminative segment of the interaction.

	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).

and prediction, and therefore, are insufficient for time-slice recognition. We introduce Predict-STIPs which are active during the whole video. 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 $\{S_i \mid i = 1:n\}$, we first detect a new subset of S-STIPs [1]. We then track them backward and forward to the first and last frames of the video and check whether or not they have existed during the whole video. We repeat these steps for all frames of a discriminative segment. Landmarks that are continuously observable are selected as P-STIPs

Stage 3- Descriptors and vocabulary building: Given P-STIPs of each interaction video, we construct the descriptor vectors HOG3D 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 into 1000 bins associated with visual words using K-means clustering. BoW features are normalized so their L1 norm is 1.

Results: At test time, a query video v_i which is a time-slice of a longer video matched to the models. To this intent, we extract S-STIPs [1] 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. The average precision for all interactions (compared to human annotation) is given in the Table 1.

[1] 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.

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