

Online segmentation and classification of modeled actions performed in the context of unmodeled ones

Dimitrios I. Kosmopoulos^{1,3}
dkosmo@ics.forth.gr

Konstantinos Papoutsakis^{1,2}
papouts@ics.forth.gr

Antonis A. Argyros^{1,2}
argyros@ics.forth.gr

¹Institute of Computer Science
FORTH, Greece

²Computer Science Department
University of Crete, Greece

³Dept. of Informatics Engineering
Technological Educational Institute
Crete, Greece

In this work we deal with the problem of online segmentation and classification of visually observable actions, i.e., we have to provide labels given the fact that the visual observations arrive stream-wise on a sequential fashion and we need to decide on the label shortly after they are received, without having available the full sequence.

The video segmentation has been traditionally treated separately from the classification step, however, these two problems are correlated and can be better handled considering simultaneously the low level cues and the high level models representing the candidate classes. Generative models have been used extensively given their ability to build probabilistic models of actions and provide the posterior of assigning labels to observations. Alternatively, discriminative models better predict the conditional probability of the states given the observed features. As a result, several researchers have investigated the use of discriminative models of actions such as CRFs, SVMs [2] or random forests [1]. However, the discriminative models are not without problems, since they cannot easily handle unknown actions, since they were not part of their optimisation process.

In this paper, we show how we seek to mitigate that limitation, by employing a discriminative framework for online simultaneous segmentation and classification of visual actions, which deals effectively with unknown sequences that may interrupt the known sequential patterns. Our framework comprises of two main components: (a) a Hough transform to vote in a 3D space for the begin and end points and the label of the segmented part of the input stream. An SVM is used to model each class and to suggest putative labeled segments on the timeline. (b) A dynamic programming algorithm to identify the most plausible segments among the putative ones, by maximising an objective function for label assignment in linear time.

Hypotheses generation via discriminative voting. In the proposed discriminative voting framework we seek to identify simultaneously (a) the instances of classes C of sub-sequences in time series data, (b) the location \mathbf{x} of the class-specific subsequence, in other words the begin and the end time point in the data. It is inspired by the framework presented in [3], which dealt with Hough transform based object detection.

Let \mathbf{f}_t denote the feature vector observed at time instance t , while $S(C, \mathbf{x})$ denotes the score of class C at a location \mathbf{x} . The implicit model framework obtains the overall score $S(C, \mathbf{x})$ by adding up the individual probabilities $p(C, \mathbf{x}, \mathbf{f}_t, l_t)$ over all observations within a time window l_t .

We define M action primitives, which result from clustering of the visual observation vectors \mathbf{f}_t , using GMMs to represent the distributions of the observation vectors. Let P_i denote the i -th action primitive. Assuming a uniform prior over features and time locations and marginalizing over the primitive entries, we derive:

$$\begin{aligned} S(C, \mathbf{x}) &= \sum_t p(C|P_i) \sum_t p(P_i|\mathbf{f}_t) p(\mathbf{x}|C, P_i, l_t) \\ &= \sum_t w_i \times a_i(\mathbf{x}) = W_c^T A(\mathbf{x}) \end{aligned} \quad (1)$$

We can use maximum margin optimisation, if we observe that the score $S(C, \mathbf{x})$ is a linear function of $p(C|P_i)$, where $A^T = [a_1 a_2 \dots a_M]$, is noted as the activation vector and a_i is given by:

$$a_i(\mathbf{x}) = \sum_t p(\mathbf{x}|C, P_i, l_t) p(P_i|\mathbf{f}_t) \quad (2)$$

The weights W_c^T are class-specific and we notice that they can be optimised in a discriminative fashion to maximise the score for correct segmentations and labels. Given the labels $S(C, \mathbf{x}_i)$ and the respective $A(\mathbf{x}_i)$ we calculate the weights W_c using multiple one-versus-all binary SVM settings. In *testing* we vote in the 3D space using Eq.(1) and then we apply the SVMs in a sliding time window to get the putative segments,

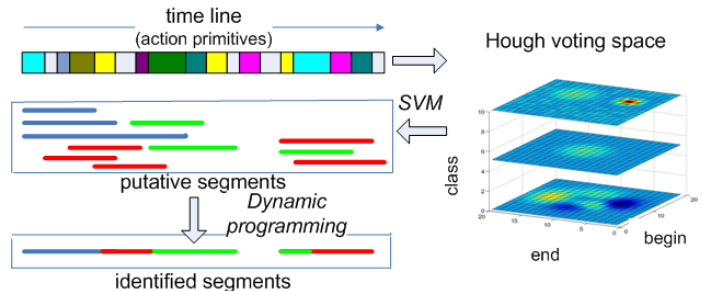


Figure 1: Overview of the proposed method: The action primitives span vote in a 3D Hough voting space (begin-end-class). The SVM receives the votes and suggests the putative segments. The final solution is computed by maximising an objective function via dynamic programming.

considering only the segments that collected enough votes. An additional evaluation step is normally applied to eliminate some false positives using a likelihood-based objective function. An illustrative example of the proposed hypotheses generation process is shown in Fig.1

Hypotheses evaluation via dynamic programming. We merge the proposed K putative segments that may overlap and have the same label. Assuming only one label for each time slot, we propose a variation of the Viterbi algorithm for linear-cost label assignment with regard to the number of input frames based on the likelihood δ_t , which is calculated after the optimal assignment of time instances to classes. The optimal sequence of classes for a time segment $t=1..T$, which contains overlapping candidate segments of different labels is given by the path $\psi_t = C_1, C_2, \dots, C_t$, which is calculated based on dynamic programming.

Experimental Evaluation. The performance of the proposed method was evaluated on synthetic as well as on real data for action recognition (Weizmann and Berkeley MHAD). Actions were provided as segments. For the purpose of identifying actions in continuous data we concatenated those videos. We compare our method against two state of the art methods, [2] and [4], that do online segmentation like our method does. The proposed approach is of comparable accuracy to the state of the art for online stream segmentation and classification and performs considerably better in the presence of previously unseen actions.

Conclusions. Our work proposed a new framework for simultaneous segmentation and classification of sequential data interrupted by unknown actions and we have applied it on synthetic and visual action streams. Under a "closed world" assumption, our method performed similarly or better than the competing discriminative methods. When the actions of interest were interrupted by previously unseen actions our method was still able to classify them and detect the unknown ones. To knowledge, our discriminative method is the first one for online simultaneous segmentation and classification having this property.

- [1] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. on PAMI*, 33(11):2188–2202, November 2011.
- [2] M. Hoai, Z.Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *IEEE CVPR*, 2011.
- [3] S. Maji and J. Malik. Object detection using a max-margin hough transform. In *IEEE CVPR*, 2009.
- [4] Q. Shi, Li Wang, Li Cheng, and A. Smola. Discriminative human action segmentation and recognition using semi-markov model. In *IEEE CVPR*, 2008.