

NeuralNetwork-Viterbi: A Framework for Weakly Supervised Video Learning

Alexander Richard, Hilde Kuehne, Ahsan Iqbal, Juergen Gall
 University of Bonn, Germany

{richard,kuehne,iqbalm,gall}@iai.uni-bonn.de

1. Introduction

Given platforms like YouTube or Vimeo, the availability of video data has largely increased over the recent years. While approaches for action classification on pre-segmented video clips already perform convincingly well [1], processing of untrimmed videos is still lacking performance. Especially detecting and segmenting fine-grained actions within such videos remains an open challenge in most cases.

One major problem in training such systems is the availability of suitable training data: Manually annotating ground truth actions for a sufficiently large amount of video data is expensive and extremely time consuming. Recent research therefore focused on weakly supervised learning, where no frame-level annotation is required but only an ordered sequence of actions that occur in the video [2, 6]. Such action transcripts can be annotated more quickly and can sometimes even be inferred from subtitles.

In order to learn a model for temporal action segmentation with such weak supervision, CNNs or RNNs have been combined with hidden Markov models (HMMs) and stochastic grammars [6, 3]. While these approaches are particularly suited for videos that contain complex actions and have a huge number of distinct classes, they come with the major problem that their training requires some heuristic ground truth. They rely on a two-step approach that is iterated several times. It consists of first generating a segmentation for each training video using the Viterbi algorithm and then training the neural network using the generated segmentation as pseudo ground-truth. Consequently, the two-step approach is sensitive to the initialization of the pseudo ground-truth and the accuracy tends to oscillate between the iterations [6].

To overcome those problems, we propose a novel learning algorithm that allows for direct learning from the input videos and ordered action classes only. The approach includes a Viterbi-decoding as part of the loss function to train the neural network and does not need any kind of pseudo ground-truth of framewise labeling as initialization. Moreover, it does not suffer from oscillation effects.

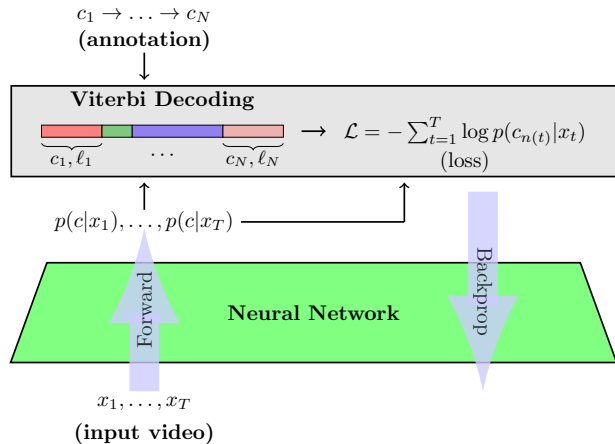


Figure 1. The input video \mathbf{x}_1^T is forwarded through the network and the Viterbi decoding is run on the output probabilities. The frame labels generated by the Viterbi algorithm are then used to compute a framewise cross-entropy loss based on which the network gradient is computed.

2. Temporal Action Segmentation

We address the problem of temporally localizing activities in a video $\mathbf{x}_1^T = (x_1, \dots, x_T)$ with T frames. The task is to find a segmentation of a video into an unknown number of N segments and to output class labels $\mathbf{c}_1^N = (c_1, \dots, c_N)$ and lengths $\mathbf{l}_1^N = (\ell_1, \dots, \ell_N)$ for each of the N segments. Using a background class for uninteresting frames, each frame can be assigned to a segment. For terms of simplicity, we refer to the label assigned to frame x_t as $c_{n(t)}$, where $n(t)$ is the number of the segment frame t belongs to.

Putting the task in a probabilistic setting, we aim to find the most likely video labeling given the video frames, *i.e.*

$$\begin{aligned}
 (\hat{\mathbf{c}}_1^N, \hat{\mathbf{l}}_1^N) &= \arg \max_{\mathbf{c}_1^N, \mathbf{l}_1^N} \{p(\mathbf{c}_1^N, \mathbf{l}_1^N | \mathbf{x}_1^T)\} \\
 &= \arg \max_{\mathbf{c}_1^N, \mathbf{l}_1^N} \left\{ \prod_{t=1}^T p(x_t | c_{n(t)}) \cdot \prod_{n=1}^N p(\ell_n | c_n) \cdot p(c_n | \mathbf{c}_1^{n-1}) \right\},
 \end{aligned} \tag{1}$$

where conditional independence of the frames is assumed for the factorization. We refer to $p(x_t | c_{n(t)})$ as *visual model* (the neural network in our case), to $p(\ell_n | c_n)$ as *length model*

(a Poisson model in this work), and to $p(c_n | c_1^{n-1})$ as *context model* (a finite grammar of all possible training transcripts).

The factorization used in Eq. (1) or a similar factorization is widely used in recent works [5, 6, 3]. The arg max can be efficiently computed using the Viterbi algorithm. Recall that in our weakly supervised setting, I_1^N and accordingly the frame-level annotation of the data is unknown during training.

2.1. Viterbi-based Network Training

Our proposed training procedure is illustrated in Figure 1. During training we randomly draw a sequence \mathbf{x}_1^T and its annotation \mathbf{c}_1^N from the training set. The sequence is then forwarded through a neural network. Note that there are no constraints on the network architecture, all commonly used feed-forward networks, CNNs, and recurrent networks can be used. The optimal segmentation by means of Equation (1) is then computed by application of a Viterbi decoding on the network output. Since \mathbf{c}_1^N is provided as annotation, only I_1^N needs to be inferred during training. We switch notation and write the Viterbi segmentation (\mathbf{c}_1^N, I_1^N) as framewise labels $c_{n(1)}, \dots, c_{n(T)}$, with which the cross-entropy loss over all aligned frames is accumulated:

$$\mathcal{L} = - \sum_{t=1}^T \log p(c_{n(t)} | x_t). \quad (2)$$

Based on the sequence loss \mathcal{L} , the network parameters are updated using stochastic gradient descent with the gradient $\nabla \mathcal{L}$ of the loss. We would like to emphasize that the algorithm operates in an online fashion, *i.e.* in each iteration, the loss \mathcal{L} is computed with respect to a single randomly drawn training sequence $(\mathbf{x}_1^T, \mathbf{c}_1^N)$ only.

Since videos are frequently several thousand frames long but contain only a few of the overall possible actions, the loss in Eq. 2 tends to push the model strongly towards a small subset of actions and the specific appearance of the video frames. In order to avoid this effect and enhance the robustness of our algorithm, we propose to use a buffer \mathcal{B} and store recently processed sequences and their inferred frame labels. In order to make the gradient in each iteration more robust, K frames from the buffer are sampled and added to the loss function,

$$\mathcal{L} = - \left[\sum_{t=1}^T \log p(c_{n(t)} | x_t) + \sum_{k=1}^K \log p(c_k | x_k) \right]. \quad (3)$$

3. Experiments

We evaluate our method on two widely used datasets, the Breakfast dataset [4] and 50Salads [7]. For comparability with other approaches, we use Fisher vectors of improved dense trajectories as input features. The neural network is a

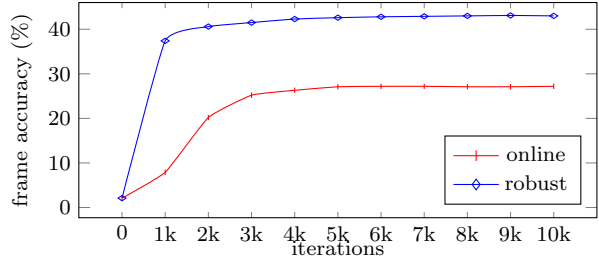


Figure 2. Convergence behaviour of our algorithm on Breakfast.

	Breakfast	50 Salads
CTC [2]	21.8	11.9
HTK [5]	25.9	24.7
ECTC [2]	27.7	—
HMM/RNN [6]	33.3	45.5
NN-Viterbi	43.0	49.4

Table 1. Comparison of our method to several state-of-the-art methods for the task of temporal action segmentation. Results are reported as frame accuracy (%).

recurrent network with a hidden layer of 128 gated recurrent units and a softmax output.

Figure 2 shows the convergence behaviour of our algorithm as a pure online learning approach (loss from Eq. (2)) and with the robustness enhancements (loss from Eq. (3)). While both variants of our algorithm start to converge after 2,000 to 3,000 iterations, the robustness enhancement is particularly advantageous at the beginning of training, adding a huge margin in terms of frame accuracy compared to the pure online variant.

Compared to current state of the art, our method shows a significant improvement compared to the best published results of [6], see Table 1.

References

- [1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [2] D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *ECCV*, 2016.
- [3] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *CVPR*, 2017.
- [4] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.
- [5] H. Kuehne, A. Richard, and J. Gall. Weakly supervised learning of actions from transcripts. *CVIU*, 2017.
- [6] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with RNN based fine-to-coarse modeling. In *CVPR*, 2017.
- [7] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UBICOMP*, pages 729–738, 2013.

This work has been accepted for CVPR 2018 under the same title.