Action Segmentation via Transcript-Aware Union-of-Subspaces Learning

Zijia Lu
Northeastern Univeristy
lu.zij@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

Abstract

We address the problem of learning to segment actions from narrated videos, i.e., videos accompanied by narrations that can be parsed into transcripts (ordered list of actions). We propose a framework in which we model actions with a union of low-dimensional subspaces, learn the subspaces using transcripts and refine video features that lend themselves to action subspaces. To do so, we design an architecture consisting of a Union-of-Subspaces Network, which is an ensemble of autoencoders, each modeling a low-dimensional action subspace to capture variations of an action. For learning, at each iteration, we generate positive and negative soft alignment matrices using a constrained alignment algorithm, which we use for discriminative training of our model. Our experiments on two datasets show that our method improves the state-of-the-art action segmentation and alignment.

** A complete version of this workshop paper is currently under review.*

1. Introduction

Localization and classification of human actions in long uncurated videos has been a major challenge in video understanding. Since gathering framewise annotations of videos is costly, there has been an increasing interest in learning from unsupervised videos accompanied with narrations that explain the actions performed in the videos. Although the exact location of each action cannot be inferred from the narration (transcript), it provides useful information about the ordering of actions in a video [9, 1, 17]. Since the narrated videos are less costly to gather, it has motivated a variety of interesting approaches that learn to localize and classify actions using transcripts [17, 7, 15, 16, 5, 19, 10, 3, 11].

Challenges. Despite tremendous advances, existing works on transcript-based action learning still face major challenges. In fact, a successful class of recent methods focuses on alternating between segmenting the videos with transcripts and retraining models with the one obtained segmentations [16, 10]. However, it ignores and discourages other likely segmentations and propagates the initial segmentation errors. Meanwhile, the existing methods often ignore the underlying low-dimensional structures of videos. In fact, it is well known that high-dimensional visual data lie in low-dimensional subspaces [18, 6, 14, 2, 13, 12, 4]. Yet, leveraging such low-dimensional subspaces in the transcript-based setting has been mainly unexplored.

Contributions. In this paper, we address the action segmentation by developing a Transcript-aware Action Subspace Learning (TASL) framework that models actions with a union of low-dimensional subspaces, learns the subspaces using transcripts parsed from narrations and refines video features that lend themselves to action subspaces. To do so, we design a new Union-of-Subspaces Network (USN), which is an ensemble of autoencoders, each modeling a low-dimensional action subspace, that captures variations of each action.

For learning, we alternate between segmenting training videos using transcripts and learning models from segmentations. However, instead of learning a model to reproduce an obtained segmentation, we generate positive and negative soft alignment matrices using the optimal segmentation, which we will use for discriminative learning of subspaces. Our experiments on two datasets show that our method improves the state-of-the-art.

2. Transcript-Aware Multi-Subspace Learning

Problem Statement. Assume we have V videos and their action transcripts $\{(\mathcal{X}^v, \mathcal{T}^v)\}_{v=1}^{V}$, where $\mathcal{X}^v = (\boldsymbol{x}_1^v, \ldots, \boldsymbol{x}_{N_v}^v)$ denotes the collection of framewise unsupervised features for the video v, which has N_v frames. $\mathcal{T}^v = (a_1^v, \ldots, a_{n_v}^v)$ denotes its transcript parsed from the narration, which is the ordered list of n_v actions in the video. We have $a_i^v \in \{1, 2, \ldots, A\}$, where A denotes the total number of actions across videos. The goal of transcript-based action learning is to learn an segmentation model only using the transcripts. Depending on the information provided for a test video, inference can be divided into action alignment, where the video's narration (transcript) is known, and action segmentation, where the narration (transcript) is unknown. For simplicity, we drop the superscript and subscript v in notations (referring to video v), as it would be clear from the context.

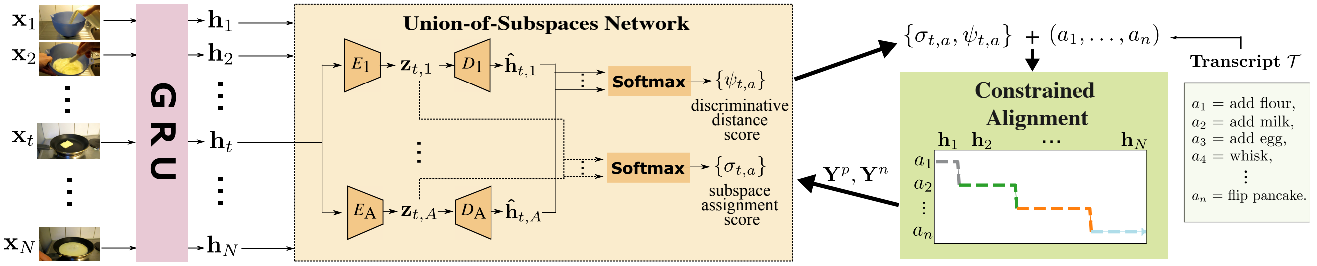

Figure 1: We propose a framework, referred to as Transcript-aware Action Subspace Learning (TASL), for transcript-based segmentation of videos. The framework consists of a Union-of-Subspaces Network (USN), which learns to embed actions into discriminative low-dimensional subspaces, and an efficient constrained video alignment algorithm that generates positive and negative soft alignments, which will be used for parameter learning.

Proposed Framework. As shown in figure 1, our TASL alternates between 1) training the proposed model using the video segmentations, which consists of a GRU for feature learning and a Union-of-Subspaces Network (USN) to learn low-dimensional subspaces of actions; 2) generating valid/invalid segmentations using the constrained alignment algorithm based on the transcripts and the model outputs.

2.1. Discriminative USN Training

In this section, we introduce the designed network architecture and efficient discriminative loss for learning features and low-dimensional subspaces corresponding to actions.

Proposed Architecture. First, we use a recurrent network (here GRUs) as the feature learning module, $(\boldsymbol{h}_1, \dots, \boldsymbol{h}_N) = \text{GRU}\big((\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)\big)$, that captures temporal dependencies between framewise unsupervised features and transforms them into more discriminative features lying in low-dimensional subspaces corresponding to actions. To achieve such low-dimensional embeddings, we design a Union-of-Subspaces Network (USN) that consists of an ensemble of A autoencoders for A actions. The autoencoder a encodes the input feature vector $\boldsymbol{h}_t \in \mathbb{R}^p$ into a low-dimensional embedding vector $\boldsymbol{z}_{t,a} \in \mathbb{R}^d$ ($d \ll p$), which will be decoded to $\hat{\boldsymbol{h}}_{t,a} \in \mathbb{R}^p$. More specifically,

$$\boldsymbol{z}_{t,a} = \boldsymbol{W}_a^e \boldsymbol{h}_t + \mathrm{b}_a^e \in \mathbb{R}^d, \;\; \hat{\boldsymbol{h}}_{t,a} = \boldsymbol{W}_a^d \boldsymbol{z}_{t,a} + \mathrm{b}_a^d \in \mathbb{R}^p, \tag{1}$$

where $\{\boldsymbol{W}_a^e, \boldsymbol{W}_a^d, \mathrm{b}_a^e, \mathrm{b}_a^d\}$ are the learnable weights of the encoder and decoder of the action a, respectively. Here, $\{\boldsymbol{z}_{t,a}\}$ represent the d-dimension embeddings of $\{\boldsymbol{h}_t\}$ on the subspace of action a. With the linear decoder, $\{\hat{\boldsymbol{h}}_{t,a}\}$ are affine transformations of $\{\boldsymbol{z}_{t,a}\}$ using the same combination weights $\boldsymbol{W}_a^d$, thus, lie on the d-dimension subspace. Therefore, subspaces are described by the column spaces of $\{\boldsymbol{W}_a^d\}$. A feature $\boldsymbol{h}_t$ being close to $\hat{\boldsymbol{h}}_{t,a}$ implies that the frame t is close to the subspace a.

Proposed Discriminative Training. Given framewise labels from the alignments, instead of naively learn a subspace for each action a by minimizing the distance $\|\hat{\boldsymbol{h}}_{t,a} - \boldsymbol{h}_t\|_2$ over all frames assigned to it, we develop a method that uses two complementary scores for discriminative training. Since $\|\boldsymbol{z}_{t,a}\|_2$ corresponds to the embedding norm of $\boldsymbol{h}_t$ onto the subspace a, we compute the *subspace assignment score* of $\boldsymbol{h}_t$ to subspace a by

$$\sigma_{t,a} = \frac{e^{\|\boldsymbol{Q}_a \boldsymbol{z}_{t,a}\|_2^2}}{\sum_{a'} e^{\|\boldsymbol{Q}_{a'} \boldsymbol{z}_{t,a'}\|_2^2}} \in [0, 1]. \tag{2}$$

Since not every direction in a subspace is necessarily useful for recognition of the underlying action, e.g., directions corresponding to intersection with other subspaces, we use $\boldsymbol{Q}_a \in \mathbb{R}^{d' \times d}$ ($d' \leq d$) to allow learning discriminative features within each subspace a.

Given $\|\hat{\boldsymbol{h}}_{t,a} - \boldsymbol{h}_t\|_2$ as the distance between $\boldsymbol{h}_t$ and the subspace a, we define the *discriminative distance score*,

$$\psi_{t,a} = \frac{e^{-\|\hat{\boldsymbol{h}}_{t,a} - \boldsymbol{h}_t\|_2^2}}{\sum_{a'} e^{-\|\hat{\boldsymbol{h}}_{t,a'} - \boldsymbol{h}_t\|_2^2}} \in [0, 1], \tag{3}$$

whose maximization for a subspace a enforces that $\boldsymbol{h}_t$ must be close to it and far from other subspaces.

Based on network outputs, our alignment algorithm will produce two soft label matrices (see the next subsection for details): 1) positive soft alignments $\boldsymbol{Y}^p \in [0, 1]^{N \times A}$, whose each row is the probability distribution of frame t belonging to each action, and is computed based on optimal alignment of the video with its transcript; 2) negative soft alignments $\boldsymbol{Y}^n \in [0, 1]^{N \times A}$, whose each row is the probability distribution of frame t to undesired actions. Thus, to learn the GRU and USN, for each video, we define the loss

$$\begin{aligned} \mathcal{L}_{\text{video}} \triangleq \sum_{t=1}^{N} \sum_{a=1}^{A} \Big[& - y_{t,a}^p \big(\log(\sigma_{t,a}) + \rho \log(\psi_{t,a}) \big) \\ & + y_{t,a}^n \big(\log(\sigma_{t,a}) + \rho \log(\psi_{t,a}) \big) \Big], \end{aligned} \tag{4}$$

and minimize the average of this loss over all training videos with respect to the network parameters. Here, ρ controls the trade-off between the subspace assignment score σ and discriminative distance score ψ (here, $y_{t,a}^p$ and $y_{t,a}^n$ are the (t, a)-th elements of $\boldsymbol{Y}^p$ and $\boldsymbol{Y}^n$, respectively). The loss function aims to maximize the embedding norms and the closeness between $\boldsymbol{h}_t$ and the associated subspace based on the positive alignment $\boldsymbol{Y}^p$ while minimizing those to the incorrect subspaces based on $\boldsymbol{Y}^n$.

2.2. Constrained Alignment Algorithm

In this section, we discuss our alignment algorithm that first find the optimal alignment of a video then form positive and negative soft alignments for network training.

Finding Optimal Transcript Alignment. Given the transcript $\mathcal{T} = (a_1, \dots, a_n)$ of a video, our goal is to find the best alignment that assigns each frame to one action in the transcript in order. Notice that an alignment can be fully determined by finding the lengths of actions in the transcript. Let l_i denote the length of action a_i, where we must have $\sum_i l_i = N$. To find the optimal alignment, we obtain the subspace assignment scores $\sigma_{t,a}$ in (2) and search for $\{l_i\}_{i=1}^n$ that give the best total assignment score over the video via an optimization algorithm, i.e., we solve

$$\min_{\{l_i\}, \sum_i l_i = N} \sum_{i=1}^{n} \left[\gamma \mathcal{L}_{reg}(l_1, ..., l_n) + \sum_{t=L_i+1}^{L_i+l_i} -\log(\sigma_{t,a_i}) \right]. \tag{5}$$

Here, $L_i \triangleq \sum_{j=1}^{i-1} l_j$ is the total length of actions prior to a_i (we set $L_1 = 0$) and γ sets a trade-off between negative likelihood and regularization.[1] $\mathcal{L}_{reg}$ is a regularization term. Specifically, let $p(a)$ denote the estimated frequency of action a to occur and $p_a(l)$ denote the probability of action a having length l. We define

$$\mathcal{L}_{reg} = \underbrace{\sum_{i=1}^{n} l_i \log\big(p(a_i)\big)}_{\triangleq \mathcal{L}_{reg}^1} + \underbrace{\sum_{i=1}^{n} -\log\big(p_{a_i}(l_i)\big)}_{\triangleq \mathcal{L}_{reg}^2}, \tag{6}$$

where $\mathcal{L}_{reg}^1$ penalizes imbalanced segmentation within a video by incurring a large cost when most frames are assigned to a frequent action. On the other hand, $\mathcal{L}_{reg}^2$ ensures each action has similar lengths across videos. We model $p_a(l) = \lambda_a^l \exp(-\lambda_a)/l!$ by a Poisson distribution [16] with a parameter λ_a denoting the mean action length. We will estimate both $p(a)$'s and λ_a's from the obtained segmentations. This objective function can be efficiently solved using constrained Viterbi decoding algorithm [16, 11].

Constructing Positive and Negative Soft Alignments. To allow our model to explore multiple candidate alignments and to better distinguish between valid/invalid alignments, we generate candidate valid alignments $\{\boldsymbol{R}_k^p\}$ and invalid alignments $\{\boldsymbol{R}_k^n\}$ using [10] based on the optimal alignment $\{l_i^*\}$. $\boldsymbol{R}_k^p \in \{0,1\}^{N \times A}$ is a discrete label matrix encoding k-th alignment and similarly for $\boldsymbol{R}_k^n$. Its (t,a) entry equals to 1 if frame t is assigned action a(each row has only one 1). To further incorporate the candidate alignments' likelihood, we propose to measure the likelihood score by computing the inner product

$$s(\boldsymbol{R}_k^p) \triangleq \langle \boldsymbol{R}_k^p, \boldsymbol{\Delta} \rangle, \quad \boldsymbol{\Delta} \triangleq \big[\log(\sigma_{t,a})\big] \in \mathbb{R}_-^{N \times A}, \tag{7}$$

which measures the likelihood of the k-th alignment path according to the learned subspace assignment scores $\sigma_{t,a}$. We then form the positive soft alignment matrix by computing the weighted average

$$\boldsymbol{Y}^p \triangleq \sum_k \alpha_k \boldsymbol{R}_k^p, \quad \alpha_k \triangleq \frac{\exp(s(\boldsymbol{R}_k^p))}{\sum_j \exp(s(\boldsymbol{R}_j^p))}. \tag{8}$$

Similarly, we compute the score of the k-th negative alignment $s(\boldsymbol{R}_k^n) \triangleq \langle \boldsymbol{R}_k^n, \boldsymbol{\Delta} \rangle$ and the negative soft alignment matrix $\boldsymbol{Y}^n$ as the weighted average of $\{\boldsymbol{R}_k^n\}$. We will use these two matrices to train our network via (4).

3. Experiments

We evaluate the performance of our proposed TASL method, against state-of-the-art transcript-based action segmentation algorithms, NNV [16] and CDFL [10], on the Breakfast [8] and CrossTask [19] datasets for both *action segmentation* and *action alignment*.

3.1. Experimental Setup

Datasets. We perform experiments on three large datasets. The *Breakfast* [8] dataset consists of 1,712 videos of people performing 10 different cooking activities. It has 48 different actions, including a 'background' class to denote non-action frames. The *CrossTask* [19] dataset contains videos from 18 primary tasks. We use the 14 cooking-related tasks, which include 2,552 videos and 80 different actions. Each video has 14.4 actions on average, while 74.8% of frames correspond to background. Both datasets have released the cleaned transcripts of the videos.

Evaluation Metrics. To evaluate the performance, we use *1) Mean-over-frame (Mof)*, which is the percentage of frames for which the predicted action labels are correct. *2) Intersection over Union (IoU)*, defined as $\frac{1}{A}\sum_a |GT_a \cap D_a|/|GT_a \cup D_a|$, where GT_a is the set of frames belonging to action a and D_a is the set of frames classified as action a. *3) IoU-bg*, which is the same as IoU but excluding the background class.

Implementation Details. For TASL, we initialize $p(a) = 1/A$ and $\lambda_a = 1$. At each iteration, we randomly sample one video and run the model and the constrained alignment algorithm with its transcript. We use the obtained optimal alignment $\{l_i^*\}$ from the current and previous iterations to update the estimation of $p(a)$ and λ_a. We set $\rho = 0.35$, $d_a = 3$ and $\boldsymbol{Q}_a$ in (2) to identity. While we found this is the best setting for the two datasets, it is possible to use a non-identity $\boldsymbol{Q}_a$ to learn a linear combination of projections on each subspace. Due to the alternating nature of learning subspaces and segmenting videos, the performance of all methods changes for different initializations. For a fair

[1] It is also possible to include $\log(\psi_{t,a_i})$ in (5), yet we found excluding it yields better performance.

	Action Segmentation			Action Alignment		
Breakfast	MoF	IoU	IoU-bg	Mof	IoU	IoU-bg
Best						
NNV [16]	42.9	32.2	29.1	59.5	47.0	47.7
CDFL [10]	**50.8**	35.7	33.6	**67.6**	50.5	51.3
TASL(ours)	49.9	**36.6**	**34.3**	65.8	**51.0**	**51.9**
Average						
NNV [16]	40.2	31.2	27.7	55.9	45.2	45.6
CDFL [10]	47.2	34.1	31.3	62.1	47.8	48.4
TASL(ours)	**47.8**	**35.2**	**32.6**	**64.1**	**49.9**	**50.7**
CrossTask	MoF	IoU	IoU-bg	Mof	IoU	IoU-bg
Best						
NNV [16]	27.0	11.0	8.5	34.6	15.3	11.4
CDFL [10]	32.5	11.8	7.7	46.7	17.2	11.5
TASL(ours)	**42.7**	**14.9**	**9.2**	**57.1**	**19.1**	**11.7**
Average						
NNV [16]	26.5	10.7	7.9	34.3	15.1	11.3
CDFL [10]	31.9	11.5	7.5	43.4	17.0	11.3
TASL(ours)	**40.7**	**14.5**	**8.9**	**54.6**	**18.8**	**11.5**

Table 1: Action Segmentation and Alignment Performance.

comparison, we run all methods using their codes for 3 different initializations and report the best run results as '*Best*' performance as well as the averaged results over runs as the '*Average*' performance in the tables.

Experimental Results Table 1 shows the performance of different methods for action segmentation and alignment. Notice that TASL achieves state-of-the-art performance on the two datasets for both action segmentation and alignment tasks, demonstrating that USN effectively learns discriminative action subspaces. For the more difficult task of action segmentation, TASL exceeds CDFL by 0.6% and 1.2% at MoF and IoU respectively on Breakfast for 'Average'. On the more challenging CrossTask dataset, TASL outperforms CDFL by 8.8% and 3% for MoF and IoU.

4. Conclusions

We developed a framework for learning to segment actions in videos using transcripts. We modeled actions by low-dimensional subspaces using an ensemble of autoencoders and proposed an efficient alignment algorithm by generating soft positive and negative alignments and introducing a regularization to prevent unbalanced segmentations. By experiments on Breakfast and CrossTask, we showed our method improves the state-of-the-art.

References

[1] J. B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis. *Journal of the ACM*, 58(1):1–37, 2010. 1

[3] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[4] Anoop Cherian, Basura Fernando, Mehrtash Harandi, and Stephen Gould. Generalized rank pooling for activity recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[5] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[7] D. A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. *European Conference on Computer Vision*, 2016. 1

[8] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1

[10] J. Li, P. Lei, and S. Todorovic. Weakly supervised energy-based learning for action segmentation. *International Conference on Computer Vision*, 2019. 1, 3, 4

[11] Jun Li and Sinisa Todorovic. Set-constrained viterbi for set-supervised action segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3

[12] Sheng Li, Kang Li, and Yun Fu. Temporal subspace clustering for human motion segmentation. *IEEE International Conference on Computer Vision*, 2015. 1

[13] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 1

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *International Conference on Machine Learning*, 2009. 1

[15] A. Richard, H. Kuehne, and J. Gall. Action sets: Weakly supervised action segmentation without ordering constraints. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[16] A. Richard, H. Kuehne, A. Iqbal, and J. Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 4

[17] Y. Shen, L. Wang, and E. Elhamifar. Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb. 2009. 1

[19] D. Zhukov, J. B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic. Cross-task weakly supervised learning from instructional videos. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3