# Sequential Summarization for Instructional Video Understanding

Ehsan Elhamifar
Assistant Professor, College of Computer and Information Science
Northeastern University, Boston, MA
eelhami@ccs.neu.edu

M. Clara De Paolis Kaluza
PhD Candidate, College of Computer and Information Science
Northeastern University, Boston, MA
depaoliskaluza.m@husky.neu.edu

## Abstract

*This submission is a synopsis of our paper, entitled "Subset Selection and Summarization in Sequential Data", which was presented in NIPS 2017 [1].*

People learn how to perform tasks such as assembling a device or cooking a recipe, by watching instructional videos for which there often exists a large amount of videos on the internet. Summarization of instructional videos helps to learn the grammars of tasks in terms of key activities or procedures that need to be performed in order to do a certain task. On the other hand, there is a logical way in which the key actions or procedures are connected together, hence, emphasizing the importance of using the dynamic model of data when performing summarization.

Subset selection, which is the task of finding a small subset of most informative items from a ground set, has become an indispensable too for summarization of image and video and speech data. On the other hand, instructional video and text data contain important structural relationships among segments or sentences, often imposed by underlying dynamic models, that should play a vital role in the selection of key steps. For example, there exists a logical way in which key segments of a video or key sentences of a document are connected together and treating segments/sentences as a bag of randomly permutable items results in losing the semantic content of the video/document. However, existing subset selection methods ignore these relationships and treat items independent from each other.

We have develop a new framework for sequential subset selection that incorporates the dynamic model of sequential data into subset selection. We have develop a new class of objective functions that promotes to select not only high-quality and diverse items, but also a sequence of representatives that are compatible with the dynamic model of data. To do so, we have propose a dynamic subset selection framework, where we equip items with transition probabilities and design objective functions to select representatives that well capture the data distribution with a high overall transition probability in the sequence of representatives, see Figure 1. Our formulation generalizes the facility location objective [2, 3] to sequential data, by incorporating transition dynamics among facilities. Since our proposed integer binary optimization is non-convex, we have develop a max-sum message passing framework to solve the problem efficiently. Please refer to [1] for the mathematical derivations of our optimization and the derivations of our message passing algorithm.

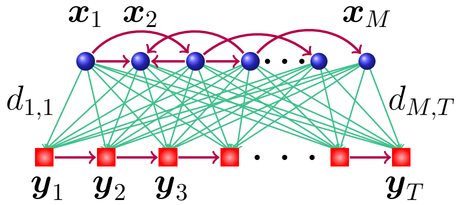

Figure 1: We propose a framework, based on a generalization of the facility location problem, for the summarization of sequential data. Given a source set of items $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ with a dynamic transition model and a target set of sequential items $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, we propose a framework to find a sequence of representatives from the source set that has a high global transition probability and well encodes the target set.

We have applied our SeqFL to the task of summarization of instructional videos to automatically learn the sequence of key actions to perform a task. We use videos from the instructional video dataset [4], which consists of 30 instructional videos for each of five activities. The dataset also provides labels for frames which contain the main steps re-

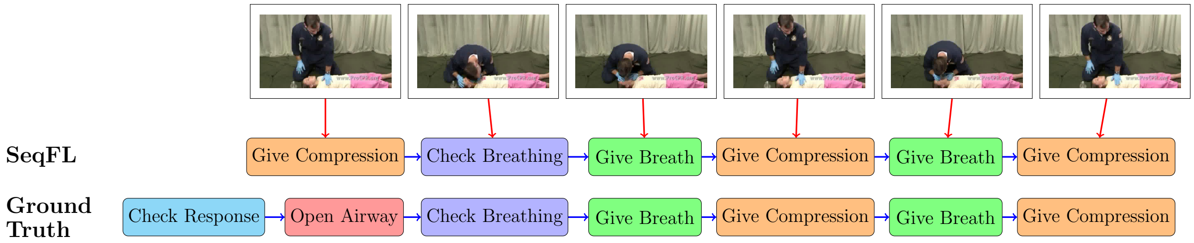

Figure 2: Ground-truth and the automatic summarization result of our method (SeqFL) for the task 'CPR'.

| Task | | kDPP | M-kDPP | Seq-kDPP | DS3 | SeqFL |
|---|---|---|---|---|---|---|
| Change tire | (P, R) | (0.56, 0.50) | (0.55, 0.60) | (0.44, 0.40) | (0.56, 0.50) | (0.60, 0.60) |
| | F-score | 0.53 | 0.57 | 0.42 | 0.53 | **0.60** |
| Make coffee | (P, R) | (0.38, 0.33) | (0.50, 0.44) | (0.63, 0.56) | (0.50, 0.56) | (0.50, 0.56) |
| | F-score | 0.35 | 0.47 | **0.59** | 0.53 | 0.53 |
| CPR | (P, R) | (0.71, 0.71) | (0.71, 0.71) | (0.71, 0.71) | (0.71, 0.71) | (0.83, 0.71) |
| | F-score | 0.71 | 0.71 | 0.71 | 0.71 | **0.77** |
| Jump car | (P, R) | (0.50, 0.50) | (0.56, 0.50) | (0.56, 0.50) | (0.50, 0.50) | (0.60, 0.60) |
| | F-score | 0.50 | 0.53 | 0.53 | 0.50 | **0.60** |
| Repot plant | (P, R) | (0.57, 0.67) | (0.60, 0.50) | (0.57, 0.67) | (0.57, 0.67) | (0.80, 0.67) |
| | F-score | 0.62 | 0.55 | 0.62 | 0.62 | **0.73** |
| All tasks | (P, R) | (0.54, 0.54) | (0.58, 0.55) | (0.58, 0.57) | (0.57, 0.59) | (0.67, 0.63) |
| | F-score | 0.54 | 0.57 | 0.57 | 0.58 | **0.65** |

Table 1: Precision (P), Recall (R) and F-score for the summarization of instructional videos for five tasks.

quired to perform that task. We preprocess the videos by segmenting each video into superframes [5] and obtain features using a deep neural network that we have constructed for feature extraction for summarization tasks. We use 60% of the videos from each task as the training set to build an HMM model whose states form the source set, $\mathbb{X}$. For each of the 40% remaining videos, we set $\mathbb{Y}$ to be the sequence of features extracted from the superframes of the video. Using the learned dynamic model, we apply our method to summarize each of these remaining videos. The summary for each video is the set of representative elements of $\mathbb{X}$, i.e., selected states from the HMM model. The assignments of representatives to superframes gives the ordering of representatives, i.e., the ordering of performing key actions.

Given ground-truth summaries, we compute the precision, recall and the F-score of various methods (see our NIPS'17 paper [1] for details). Table 1 shows the results. Notice that existing methods, which do not incorporate the dynamic of data for summarization, perform similar to each other for most tasks. In particular, the results show that the sequential diversity promoted by Seq-kDPP and M-kDPP is not sufficient for capturing the important steps of tasks. On the other hand, for most tasks and over the entire dataset, our method (SeqFL) significantly outperforms other algorithms, better producing the sequence of important steps to perform a task, thanks to the ability of our framework to incorporate the underlying dynamics of the data. Figure 2 shows the ground-truth and the summaries produced by our method for the task of 'CPR'. Notice that SeqFL sufficiently well captures the main steps and the sequence of steps to perform these tasks. However, SeqFL does not capture two of the ground-truth steps. We believe this can be overcome using larger datasets and more effective feature extraction methods for summarization.

## References

[1] E. Elhamifar and M. C. De Paolis Kaluza, "Subset selection and summarization in sequential data," in *Neural Information Processing Systems*, 2017. 1, 2

[2] P. B. Mirchandani and R. L. Francis, *Discrete Location Theory*. Wiley, 1990. 1

[3] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol. 14, 1978. 1

[4] J.-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *European Conference on Computer Vision*, 2014. 2