Multi-Stream Dynamic Video Summarization Supplementary Materials

Mohamed Elfeki[1,2], Liqiang Wang[2], and Ali Borji

[1]Microsoft, [2]University of Central Florida
melfeki@microsoft.com, lwang@cs.ucf.edu, aliborji@gmail.com

1 Introduction

As been explained in the original text, our major contributions can be summarized in four points:

(i) Addressing the generalized problem of multi-stream dynamic video summarization where the input consists of multi-view sequences recorded by dynamically moving cameras that intermittently share their field-of-views. The output is a holistic summary that represents important and representative events across all of the views.
(ii) Since we are the first to address this problem, no testing data is applicable. Thus, we introducing a new dataset that includes a variety of real-life activities in uncontrolled environments, while altering people, actions, and places.
(iii) Collecting user annotations for the dataset, and running a comprehensive analysis to ensure reliability and consistency.
(iv) Proposing a new generic approach that operates on multi-view videos by identifying the important events across all the views as well as selecting representative view(s) that can contribute to illustration of the events in the summary. Our method can be used in a supervised or unsupervised setting and applicable to fixed-cameras or dynamic-cameras footage.

This supplementary materials is organized as follows. In section 2, we present more details about the dataset presented in the main text. In section 3, we elaborate on the annotation procedure and illustrate some qualitative results. In section 4, we conduct further analysis on the collected annotation to help us understand the behavior of human annotators, that ultimately should help us for a better understanding of the problem. Finally, in section 5 we provide additional implementation details of our framework to help reproducing our results.

We note that we will be sharing our dataset and materials to facilitate future research in this area.

2 Dataset Description

We collected a total of 12 hours of videos per view for three people performing different activities. During collecting the videos we varied the environments,

people, and activities to ensure a reasonable diversity that occurs in a day-to-day activities. Figure 1 shows sample frames from the sequences covering all the collections. As shown by the samples, the views are dynamically changing their perspective and often don not share the same field-of-view. Also, using egocentric cameras causes the videos to contain rapid changes in illumination, unpredictable camera motion, unusual composition, and often complex hand-object manipulations.

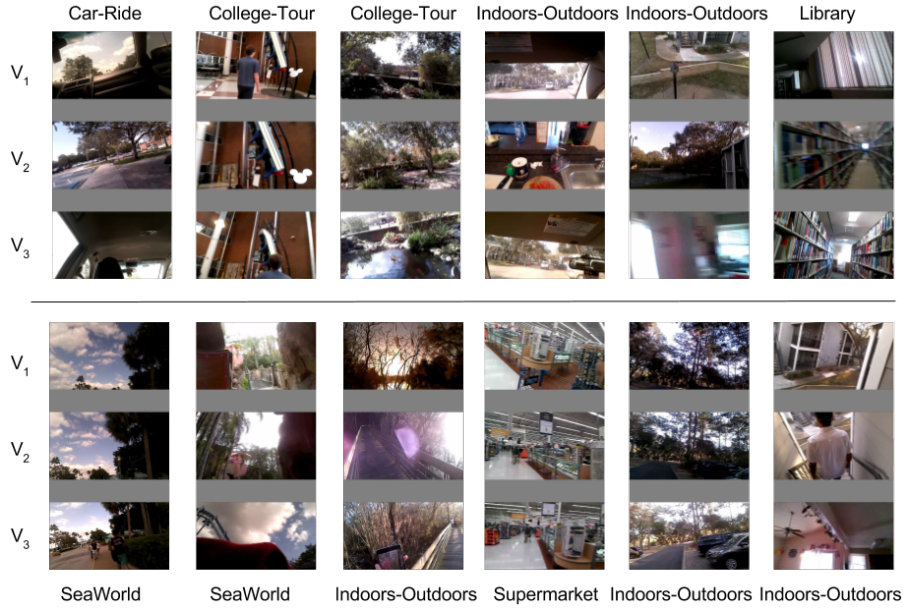

Fig. 1. Sample frames from the dataset.

From all the videos, we extracted 41 sequences that contain the highest diversity in terms of actions and environments. Each sequence contains three views that are temporally aligned across the views (i.e., all the views have the same starting and ending points). The sequences varied in length from 3 to 7 minutes and have been recorded using cameras that have frame rate of 30fps. To facilitate training and evaluating the sequences, we grouped the sequences into 6 collections that represent different scenarios:

- **Car-Ride:** 1 driver and 3 passengers are involved, the driver and two passengers hold the cameras. They all ride a car. The driver drops the passengers off separately in different locations, then comes back and picks them up again. This process is repeated several times in various locations and different orders. At the end, they park the car in a garage and enter a building. The videos were recorded in daytime.

- **Library:** 3 students go to a library. They walk around the library, search for books for the most part. Occasionally, they stop to take a glance at an interesting book, or for a brief talk when they run into each other.
- **Supermarket:** 3 people walk around a supermarket for grocery shopping. Sometimes they stop, examine or pick interesting products and goods. Occasionally they stop and talk to other people.
- **College-Tour:** 1 tour guide and 3 visitors are involved. The guide and two visitors hold the cameras. The guide walks them through the campus explaining most of the locations in daytime. Occasionally they stop and the guide gives them brief details about an attraction. They walk through the buildings in the campus and outside attractions.
- **Indoors-Outdoors:** Several family members, three of them hold cameras. They perform different activities such as cooking, playing cards, walking outside the house, learning driving the car, going to a park, playing Xbox games, etc.
- **SeaWorld:** 3 friends go to a sea-world show in daytime. The sequences includes activities such as driving, walking in the sea-world, checking in through the gate, watching the show, taking photos of each others, etc.

Table 1 shows the number of sequences per collection and the number of frames per sequence; after down-sampling the frame rate to 1 fps.

Collection Name	**Number of sequences**	**Number of frames per sequence**
Car-Ride	4	360, 360, 360, 360
College-Tour	11	190, 300, 270, 300, 240, 280, 260, 210, 300, 240, 240
Library	5	225, 300, 300, 300
Supermarket	9	270, 270, 230, 225, 210, 315, 300, 270, 210
Indoors-Outdoors	8	300, 240, 300, 300, 345, 300, 225, 270
SeaWorld	4	300, 300, 260, 300
Total	**41**	**11,135 x 3 views = 33,405**

Table 1. Statistics of the Dataset

3 Annotation Procedure

As mentioned in section 2, the original frame rate of the cameras used in recording is 30 fps. To generate a human-accessible data, we uniformly subsampled the frames to 1 fps. For each view, we generated non-overlapping shots from every consecutive 3 frames in the data. We chose to include 3 frames per shot empirically such that each shot contains a consistent action per view. Figure 2 shows sample shots (3-consecutive frames) from the sequences for all the views.

Fig. 2. Sample Shots (3-Consecutive Frames) from the datasets.

We asked five human users (4 undergraduate students, and 1 high-school student) to create annotations for all the sequences. Even though, subsampling the frames resulted in 180 to 360 frames per view in each sequence, but it is still a considerable number to show to human annotators. This may cause subjects to forget the details within a view or across the views. To remedy this, we displayed the shots to the users and asked them to select the minimal number of shots that best represent the videos as a summary.

Since the story unfolding dependencies are complex within the views, as well as across the views, we decided to use *annotating in stages* procedure. In *stage one*, the users were asked to track the story unfolding within each view and summarize them independently regardless of the correlation among the views. In *stage two*, the users were asked to track the story unfolding within a pair of views and select the most important events across all the views, and in each important event to pick the most representative view for such event. Similarly in *stage three*, the users were asked to summarize all the views simultaneously such that all the important events from all the views as well as the most representative view(s) for each event are considered.

Using the "annotating in stages" technique helps the users understand the dependencies within each view independently in stage 1, then in stages 2 & 3 they start to develop better understanding and capturing the dependencies across the views. It is worth noting that the annotators were not limited to choose only one representative view for stages 2& 3 of a certain shot. They could choose as many

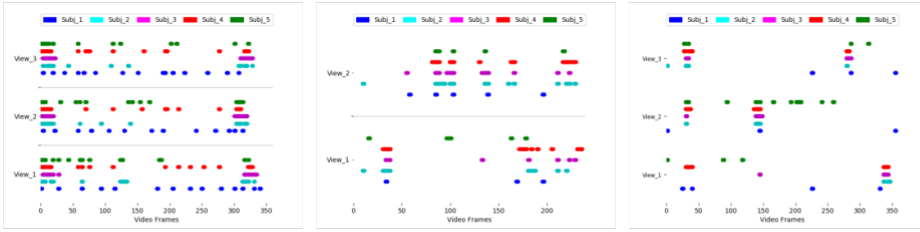

Fig. 3. Visualizations sample of users summaries in the three stages.

views they deem representative for an event as long it constitutes a minimal length summary.

After collecting the annotations, we ran a consensus analysis on the annotations to ensure a reliable and consistent set of annotations for all the stages. As reported in the main text, we computed the average pairwise f1-measures as well as the selection ratio metrics. We find that there is a substantial consensus between the users. We also plotted a visualizations for frames selected by all the annotators to further show a qualitative verification of the consensus. Figure 3 shows sample visualization of the annotations in the three stages.

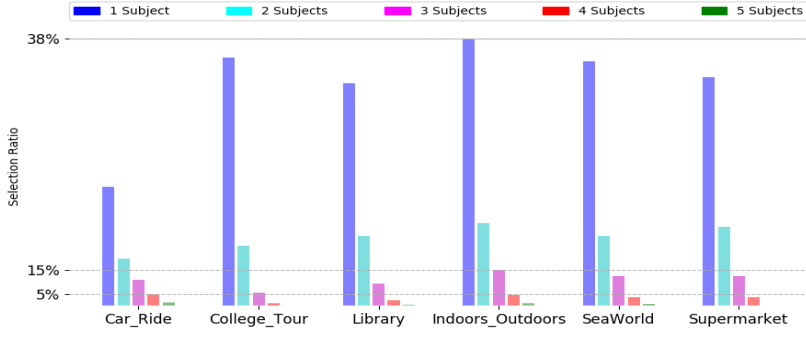

Fig. 4. Percentage of frames selected by at least 1, 2, 3, 4, 5 subjects for the annotations. In every collection, at least 3 annotators agree on $5 - 15\%$ which represents the summary.

For further annotation quality assessment, we used the selection ratio metric. According to [1–3], the usual summary length should be 5-15% of the total length of the sequence. Any frame that is a part of the final summary should be selected by at least three out of the five annotators. Figure 4 shows the ratio of the frames (with respect to sequence length) that have been chosen by at least 1, 2, 3, 4, and 5 subjects, respectively for each collection in stage three. For all the collections, the ratio of the frames chosen by at least three users is within the 5-15% range.

4 Additional Analysis

Can we explicitly see instances where annotators choose frames from one view over the other consistently? To answer this question we ran the following experiment. First, we identified the *conflict shots* which are the shots selected from different views at the same time step in 1-view summaries (i.e., present similar information). Then we calculated the frequency of each view for those time-steps in the 2-view oracle summary which are shown for the oracle summary(i.e., GT) and our method's summary in figures 5 and 6 respectively. Evidently our method's summary resembles the distribution of the ground-truth. Figure 8 shows a qualitative example of the output of our method when applied to a 3-view sequence.

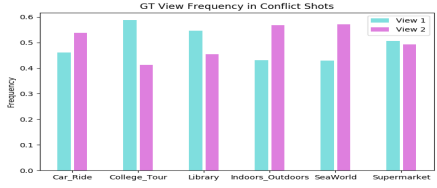

Fig. 5. Conflict shots frequency in GT.

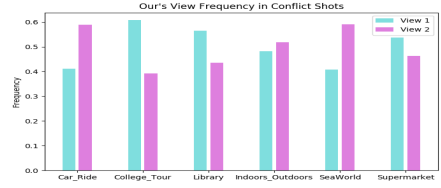

Fig. 6. Conflict shots frequency in Ours

Additionally, we address a shortcoming of the common evaluation metrics that present itself in our setting. Consider the case of two or more views having nearly identical visual content at the same time-step, which happens due to the dynamic overlap of fields-of-view. When annotating the sequences, the user will only include one of the views in the ground-truth summary at important events. However, if the prediction model selects any of the other views, it should not be penalized since the views are visually similar. To address this case, we evaluate the F1-score at several levels of similarity thresholds. That is, if the Euclidean distance of the normalized CNN features between two views at the same time-step is less than a threshold (0%, 10%, 20%, 30%), we do not penalize the prediction model if it selects any of the views instead of the other. We recompute the F1-scores for all unsupervised models at different threshold values. As shown in Fig. 7, our method continues to obtain the highest F1 at all threshold levels.

Finally, we compare our method with multi-video summarization frameworks on a standard benchmark dataset Tour-20. Table 4 shows the comparison with multi-video summarization methods on Tour 20 dataset. We used [4]'s experimental setting, by evaluating using the F-measure at 10% summary length. We report the mean F-measure across all tourist attractions on the Tour-20 dataset.

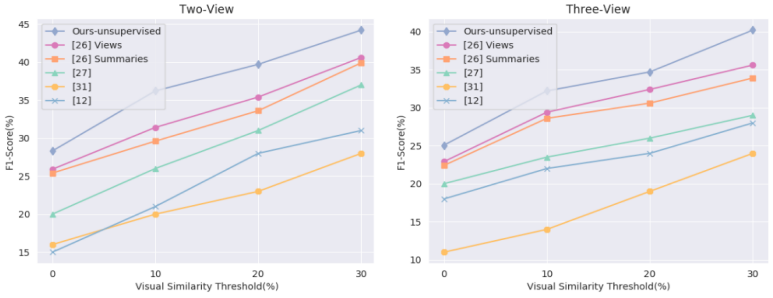

Fig. 7. F1-score computed whereas unsupervised prediction models are not penalized if mistakenly chose a view that is similar to GT view within various threshold levels.

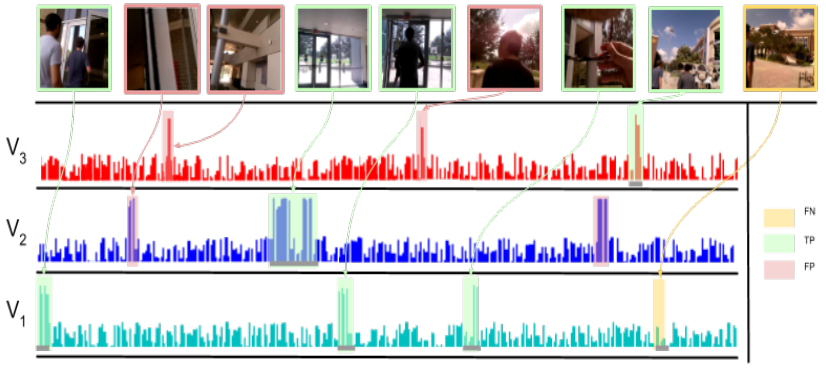

Fig. 8. Our supervised method's summary of a 3-view example: College-Tour Sequence-6

Method	MultiVideoContent [5]	MultiVideoMMR [6]	DiMS [4]	Ours (Unsupervised)	**Ours (Supervised)**
Mean F-measure	0.506	0.517	0.613	0.64	**0.68**

Table 2. Comparison with Multi-video summarization methods on Tour20 Dataset. Numbers show mean F-measure at 10% summary length.

5 Implementation Details

We used Tensor-Flow framework for our system implementation. Using the oracle summary as a ground-truth (Section 3.3 in the main text), we construct two types of labels to match our two network outputs. Those labels are used to train the network in the supervised setting. The first type is the view-selection importance used to train the quality score q_n^m, discussed in Section 4.3 of the original text. The second type of labels is the time-step importance. Both types are used to compute the MLE of Multi-DPP criterion and the normalized cross-entropy loss of the view-selection MLP which is equivalent to the oracle summary labels.

As explained in section 4.4 of the main text, the network is designed to have a number of trainable parameters that is invariant to the number of views in the data. Therefore, we are sharing the LSTM weights among all the units across all the views. We also share the weights of the view-selection MLPs across all the views. Max-pooling is applied to extend the joint features from all the views. As illustrated in section 4.3 of original text, We apply two MLPs to utilize DPP quality-diversity decomposition: Q_v and Φ. We use 256 hidden units for all the LSTM and MLP units. The view-classifier as well as the DPP diversity decomposition MLP contain two hidden layers.

We use a tanh activation layer for the LSTM units and as hidden activations for the MLPs. Additionally, we use a sigmoid activation for the view-selection classifier and a linear activation for the time-step feature MLP Φ. View-selector MLP outputs a scalar value for each view at each time-step. However, Φ MLP outputs a joint feature vector of size 256 at each time-step.

For evaluation, we used the code provided by [1, 7] that is also used by [8–10]. We modified the code to match our multi-view dataset and sat the threshold of the summary length to be 15% of the single-view length.

References

1. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5179–5187
2. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 27. Curran Associates, Inc. (2014) 2069–2077
3. Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. arXiv preprint arXiv:1707.04960 (2017)

4. Panda, R., Mithun, N.C., Roy-Chowdhury, A.K.: Diversity-aware multi-video summarization. IEEE Transactions on Image Processing **26**(10) (2017) 4712–4724
5. Li, Y., Merialdo, B.: Multi-video summarization based on video-mmr. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, IEEE (2010) 1–4
6. Wang, F., Merialdo, B.: Multi-document video summarization. In: 2009 IEEE International Conference on Multimedia and Expo, IEEE (2009) 1326–1329
7. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European conference on computer vision, Springer (2014) 505–520
8. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European Conference on Computer Vision, Springer (2016) 766–782
9. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 1–10
10. Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. arXiv preprint arXiv:1708.09545 (2017)