Diversity-aware Multi-Video Summarization (Supplementary Material)

Rameswar Panda, Niluthpol Chowdhury Mithun, Amit K. Roy-Chowdhury

Department of Electrical and Computer Engineering
University of California, Riverside

Table 1: TABLE OF CONTENTS

Page No.	**Comments**
[2]	**Detailed Information on the Experimented Datasets** (a) Descriptive statistics of our Tour20 dataset (b) Information on video data collection (c) Descriptive statistics on the multi-view camera network datasets
[7]	**Detailed Information on Generating Human created Summaries** (a) Task setup (b) Human consistency (c) Some exemplar ground truth summaries
[11]	**Additional Experimental Results** (a) Effectiveness of diversity constraint (b) Exemplar summaries on the Tour20 dataset (c) Summarized events for the Office dataset (d) Summarized events for the Campus dataset (e) Summarized events for the Lobby dataset
[18]	**Algorithmic Details** (a) Detailed derivation of ADMM steps

Detailed Information on the Experimented Dataset

Table 2: **Descriptive Statistics of our Tour20 Dataset.**

Tourist Attractions	**Videos**	**Length**	**Frames**	**Segments**
Angkor Wat, Cambodia	7	26m57s	44,410	803
Machu Picchu, Peru	7	26m15s	43,125	914
Taj Mahal, India	7	22m21s	36,554	705
Basilica of the Sagrada Familia, Spain	6	23m30s	22,641	400
St. Peter's Basilica, Italy	5	14m39s	23,777	406
Milan Cathedral, Italy	10	24m18s	37,749	768
Alcatraz, United States	6	05m22s	09,733	223
Golden Gate Bridge, United States	6	19m21s	33,063	521
Eiffel Tower, Paris	8	16m10s	26,071	495
Notre Dame Cathedral, Paris	8	26m49s	44,583	862
The Alhambra, Spain	6	21m20s	38,087	779
Hagia Sophia Museum, Turkey	6	24m27s	38,608	853
Charles Bridge, Prague	6	27m33s	48,395	769
Great Wall at Mutiantu, Beijing	5	13m16s	22,117	477
Burj Khalifa, Dubai	9	23m21s	40,557	809
Wat Pho, Bangkok	5	11m48s	20,461	382
Chichen Itza, Mexico	8	16m51s	28,737	545
Sydney Opera House, Sydney	10	25m55s	49,735	695
Petronas Twin Towers, Malaysia	9	18m32s	30,009	470
Panama Canal, Panama	6	17m33s	31,625	623
Total	**140**	**6h46m18s**	**669,497**	**12,499**

(Please see Fig. 1, 2 for topic-wise image samples from our Tour20 dataset.)

(b) Video Data Collection:

Topic-oriented video summarization is a relatively unexplored domain; we therefore collected a new dataset, Tour20, that contains 140 videos of total 6 hour 46 minutes duration. We first selected 20 tourist places out of 25 top travel destinations from the Tripadvisor's list. We removed the rest 5 topic places since we did not get any videos while searching on YouTube using the destination name as a search query term. From the search results, we selected only publicly available audio-visual programs associated with Creative Commons license, CC-BY 3.0 and videos of duration less than 15 minutes. We hope the release of our Tour20 dataset will give researchers a new, dynamic tool to evaluate their video summarization algorithms in a repeatable and efficient way.

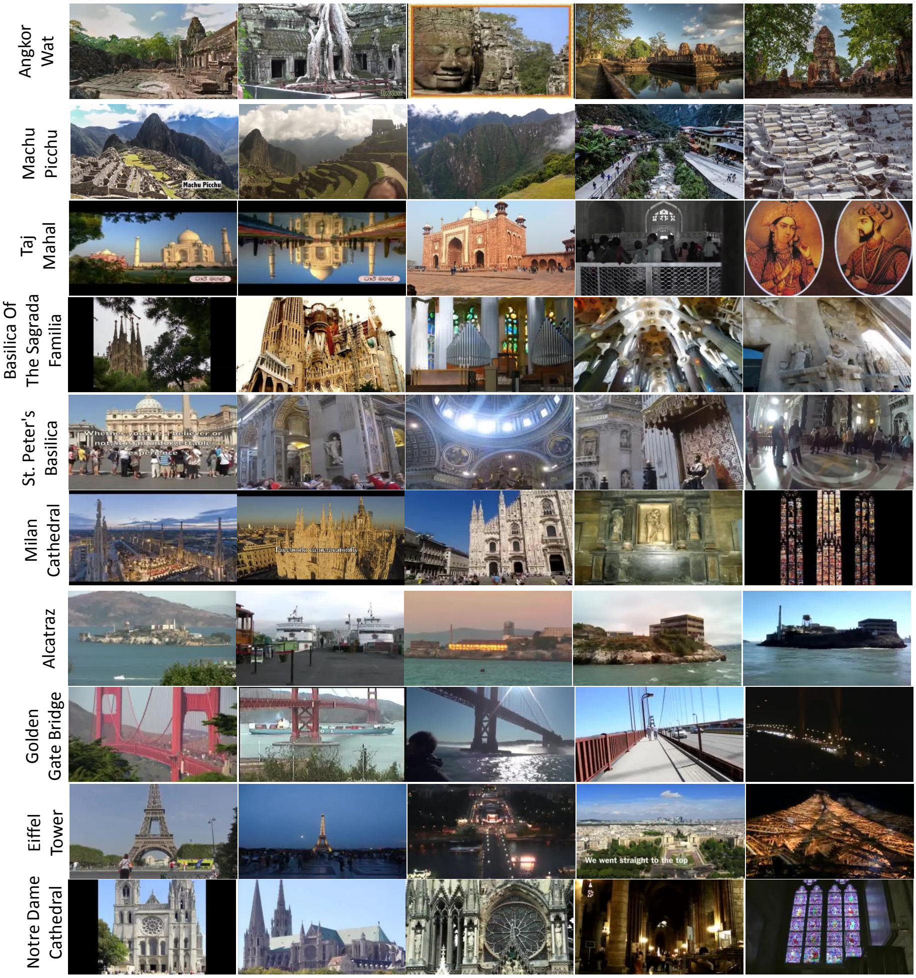


Fig. 1: **Tour20 dataset** contains 140 videos downloaded from YouTube using 20 tourist destination name from the Tripadvisor travelers choice landmarks 2015 list with a duration filter of 15 minutes and an additional constraint such that each video is associated with Creative Commons license, CC-BY 3.0.

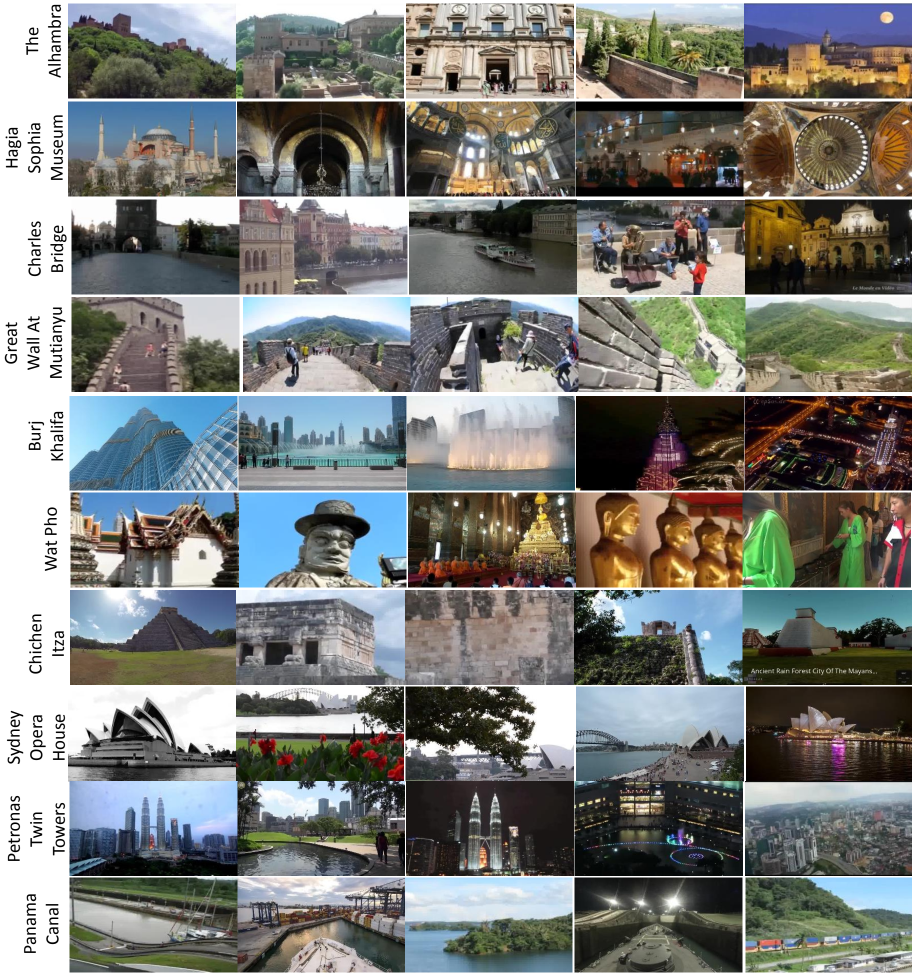


Fig. 2: **Tour20 dataset** contains 140 videos downloaded from YouTube using 20 tourist destination name from the Tripadvisor travelers choice landmarks 2015 list with a duration filter of 15 minutes and an additional constraint such that each video is associated with Creative Commons license, CC-BY 3.0.

(c) Descriptive statistics on the multi-view datasets:

Datasets	Videos	Video Length (Mins.)	Settings
Office	4	11:16/08:43/11:22/14:58	This dataset was captured with 4 stably-held web cameras in an indoor environment (office). The four videos are non-synchronized and also have different frame rates. The change of light conditions across the cameras makes it difficult to generate a good multi-view summary.
Campus	4	15:19/13:51/12:30/15:03	This dataset was taken with 4 handheld ordinary video cameras in an outdoor scene with only 180 degree of coverage. Since the videos are captured by nonspecialists, some of them are unstable and obscure which makes the summarization more challenging.
Lobby	3	08:14/08:14/08:14	This was captured with 3 cameras in a large lobby area. Unlike the previous two datasets, this dataset contains more crowded scenes with richer activities, making it more difficult for summarization.

(Please see Fig. 3 for exemplar frames from the three multi-view datasets.)

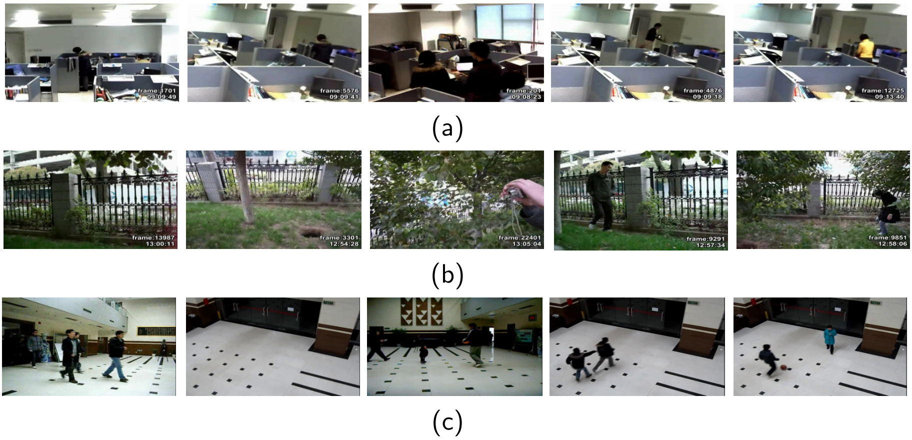


Fig. 3: (a): Office, (b): Campus, and (c): Lobby. As can be seen from figure, the datasets (a) and (c) are captured with fixed cameras in indoor environments whereas the dataset (b) is taken with non-fixed cameras in an outdoor environments.

Generating Human created Summaries

In contrast to single-video summarization, we need a single ground truth summary of all the topic-related videos to evaluate the results in topic-oriented video summarization experiments. We selected three study experts (one graduate student and two undergraduate students) to collect the ground truth summaries in a controlled experiment. Here, we describe the entire task setup that we used collect the ground truth summaries for our Tour20 dataset.

(a) Task setup:

(i) Before the study begins, we first show the name of the tourist place (e.g., Machu Picchu of Peru) to a human and then asked to emulate various concepts related to the place on their mind. They could use the web for this purpose to know the important facts about the place. Our objective is to let the user know about what is important about the place before generating the summaries.

(ii) Given the videos that were pre-segmented into several segments, study experts were asked to select at least 5%, but no more than 15% segments for each video that summarizes most important portions of the video. We set the summary length to be in the range [5%, 15%] of total number of segments to ensure that the input video is indeed summarized rather than being slightly shortened. Furthermore, we asked the users to select a diverse set of segments that can summarize the video collection altogether. We use this diverse set of segments as the ground truth summaries to compare with system generated summaries. They could use a simple interface that allows to watch all the videos of a collection at the same time and select important segments from each video as well as a set of non redundant segments that can describe the collection altogether.

(iii) While audio or embedded text can be used during generating ground truth summaries, we muted the audio to ensure that representative segments are selected based solely on visual stimuli. Moreover, we specify that if something is only mentioned in onscreen text, then it should not be labeled as important. The total user time of the study amounts to over 30 hours.

(b) Human Consistency:

In this section, we analyze the human created summaries in terms of consistency among the study experts. Motivated by [1,2], we compute both pairwise F-measure and the Cronbach's alpha between different ground truth summaries.

F-measure: Given two human created summaries i and j, we compute the pairwise F-measure as follows:

$$F_{ij} = \frac{(1+\beta^2) \times p_{ij} \times r_{ij}}{(\beta^2 \times p_{ij}) + r_{ij}}$$

where

$$p_{ij} = \frac{\#matched\ segments}{\#segments\ in\ ground\ truth\ i}$$

$$r_{ij} = \frac{\#matched\ segments}{\#segments\ in\ ground\ truth\ j}$$

and β balances the relative importance between precision and recall; we set $\beta = 1$. We utilize VSUMM evaluation package [3] for finding matching pair of segments, as in [4]. We compute all the pairwise F-measures of a video collection and report the average measure. The dataset has a mean F-measure of 0.644.

Cronbach's alpha: We additionally computed the Cronbach's alpha which is a standard measure to assess the internal consistency of test.

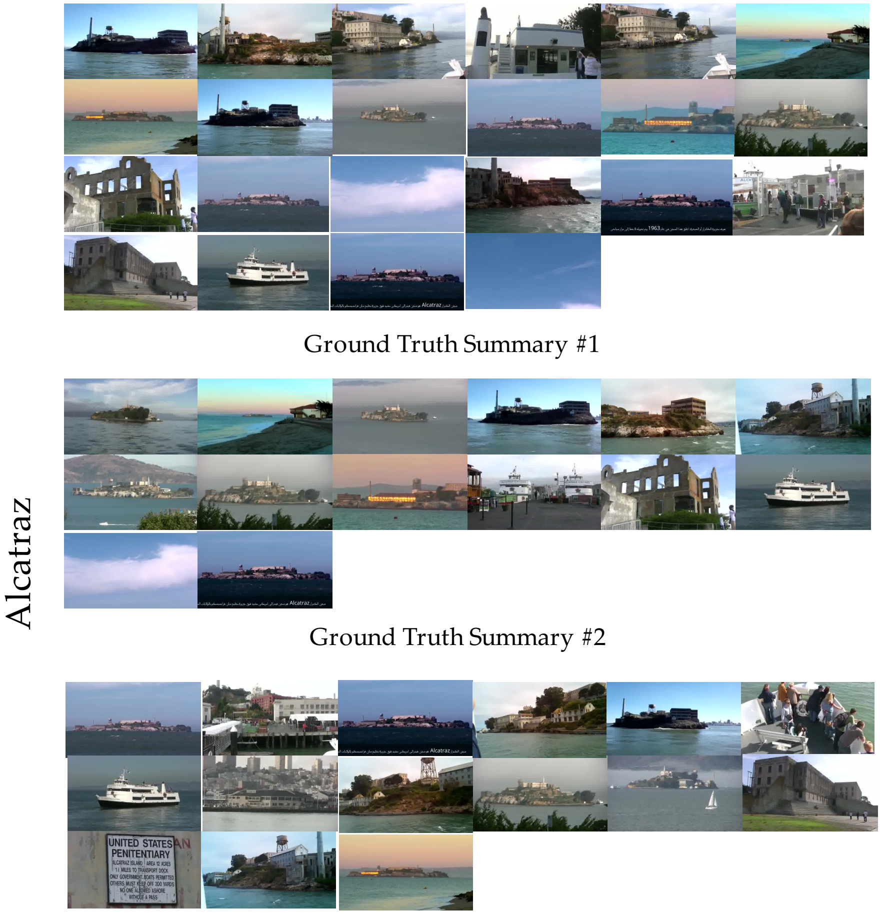
$$\text{Cronobach's \ alpha} = \frac{Nc}{1+(N-1)c}$$

where N is the number of users involved in the test and c is the mean pairwise correlation between all ground truth summaries. The dataset has a mean Cronobach's alpha of 0.944. Ideally alpha is around 0.9 for a good test [5]. See Fig. 4, 5 for some exemplar ground truth summaries of our Tour20 dataset.

Table 3: Human consistency of our Tour20 dataset. We report both pairwise F-measure and Cronobach's alpha. Our dataset has a mean F-measure of 0.643 and mean Cronobach's alpha of 0.944. The analysis of the consistency shows that humans generally agree on what parts of a video are interesting. However, we acknowledge that the individual differences can increase with increase in number of users involved for generating ground truth summaries.

Tourist Attractions	# videos	# segments	**# Selected segments**			**Human Consistency**	
			Human # 1	Human # 2	Human # 3	F-measure	Cronbach's alpha
Angkor Wat	7	803	77	42	53	0.488	0.957
Machu Picchu	7	914	80	58	79	0.485	0.969
Taj Mahal	7	705	53	39	59	0.608	0.942
Basilica of Sagrada Familia	6	400	29	23	16	0.574	0.935
St. Peter's Basilica	5	406	29	28	62	0.588	0.950
Milan Cathedral	10	768	62	43	55	0.578	0.971
Alcatraz	6	223	22	14	15	0.756	0.930
Golden Gate Bridge	6	521	26	26	28	0.799	0.932
Eiffel Tower	8	495	25	25	33	0.810	0.909
Notre Dame Cathedral	8	862	41	59	69	0.557	0.953
The Alhambra	6	779	67	40	81	0.573	0.976
Hagia Sophia Museum	6	853	42	39	38	0.748	0.942
Charles Bridge	6	769	40	30	38	0.797	0.946
Great Wall at Mutiantu	5	477	30	32	38	0.674	0.948
Burj Khalifa	9	809	42	33	30	0.700	0.939
Wat Pho	5	382	39	32	45	0.646	0.938
Chichen Itza	8	545	25	25	32	0.676	0.921
Sydney Opera House	10	695	45	24	60	0.604	0.955
Petronas Twin Towers	9	470	41	29	41	0.589	0.934
Panama Canal	6	623	59	31	45	0.617	0.940
mean						**0.643**	**0.944**

(c) Some Exemplar Ground Truth Summaries:

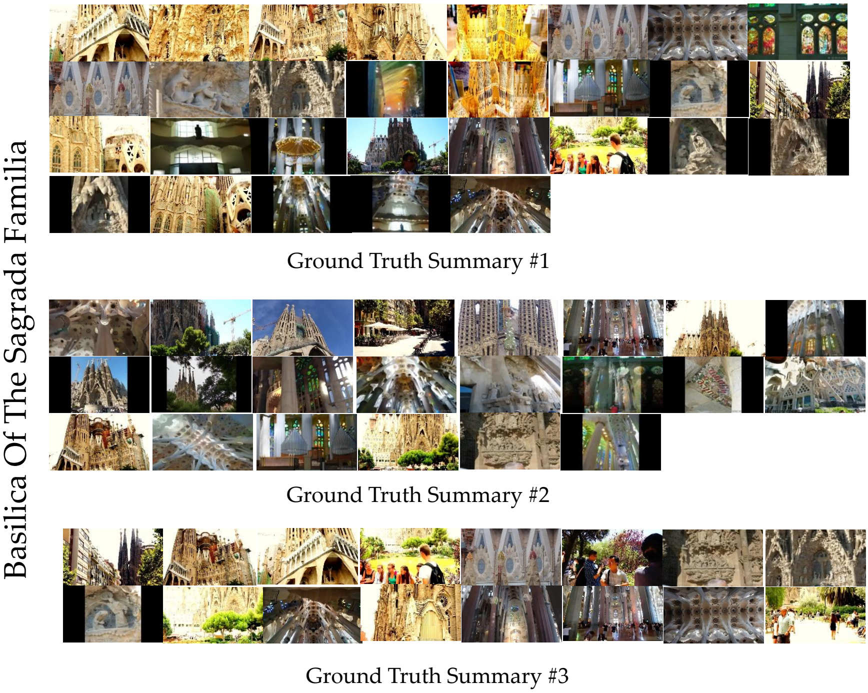


Ground Truth Summary #1

Ground Truth Summary #2

Ground Truth Summary #3

Fig. 4: Ground truth summaries generated for the tourist attraction **Alcatraz**. We represent each segment using the central frame. The video collection contains 6 videos that were segmented into 223 segments. As can be seen from the fig, all three study experts are consistent on finding what parts of the video are interesting. The video collection has a pairwise F-measure of 0.756.



Ground Truth Summary #1

Ground Truth Summary #2

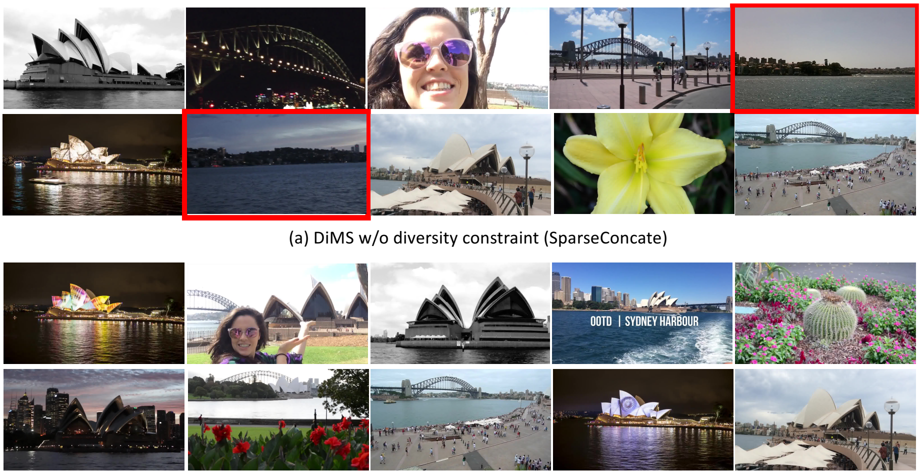
Ground Truth Summary #3

Fig. 5: Ground truth summaries generated for the tourist place **Basilica of the Sagrada Familia**. We represent each segments using the central frame. The video collection contains 6 videos that were segmented into 400 segments. As can be seen from the fig, there exists individual differences between the study experts in selecting representative segments that can visualize different aspects of the catholic church. User #1 mainly focuses on selecting segments that show the interior design of the catholic church and selects 29 representative segments from the video collection, while user #2 and user #3 select 23 and 16 segments respectively. It has a pairwise F-measure of 0.574.

Additional Experimental Results

(a) Example to show the effectiveness of diversity constraint:

In the main paper (Fig. 2), we showed one qualitative example on the effectiveness of diversity constraint. Here, we present one more exemplar summary to further validate our claim on the advantage of the prposed diversity constraint.



(a) DiMS w/o diversity constraint (SparseConcate)

(b) Our approach (DiMS)

Fig. 6: Role of diversity constraint in summarizing videos of **Sydney Opera House**. (a) DiMS w/o diversity constraint (i.e., SparseConcate baseline), and (b) Our approach (DiMS). We show the top 10 segments and represent each segment using the central frame. As can be seen from (a), SparseCocate baseline produces redundant segments (marked with red color boarders) that are selected from different segments but represent the same information. However, our approach, DiMS generate a diverse and informative set of segments by exploring the complementary information present in multiple videos. In this case, our approach achieved the F-measure of 0.614 compared to 0.474 by SparseConcate baseline for a summary of 10% length.

(b) Some exemplar summaries on our Tour20 dataset:

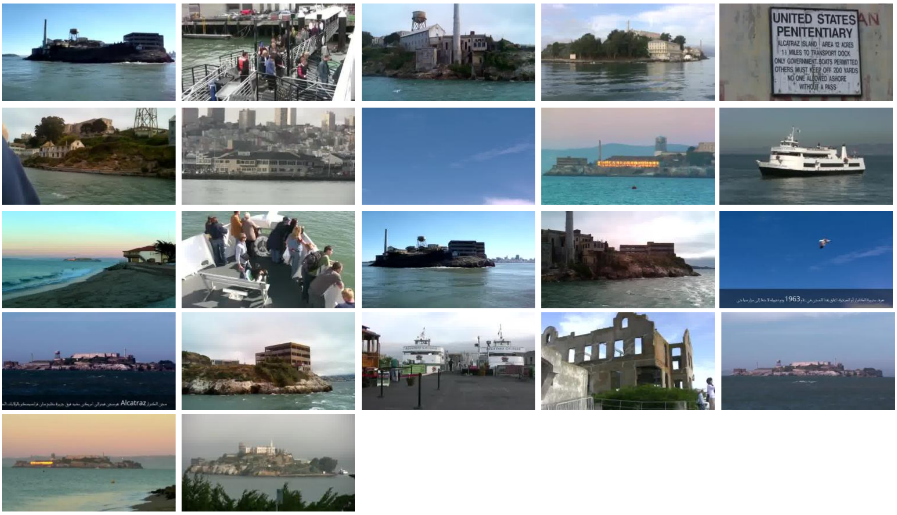

UNITED STATES
PENITENTIARY
ALCATRAZ ISLAND AREA 12 ACRES
1½ MILES TO TRANSPORT DOCK
ONLY GOVERNMENT BOATS PERMITTED
OTHERS MUST KEEP OFF 200 YARDS
NO ONE ALLOWED ASHORE
WITHOUT A PASS
1963
Alcatraz

Fig. 7: Summaries generated by our approach (DiMS) for the tourist attraction **Alcatraz**. We show the summaries at 10% length (*i.e.*, 22 segments out of total 223 segments) and represent each segment using the central frame. As can be seen from the figure, our approach produces informative segments that can describe the whole video collection in few minutes. The F-measure achieved by our approach for this video is the highest (0.755) in our Tour20 dataset.

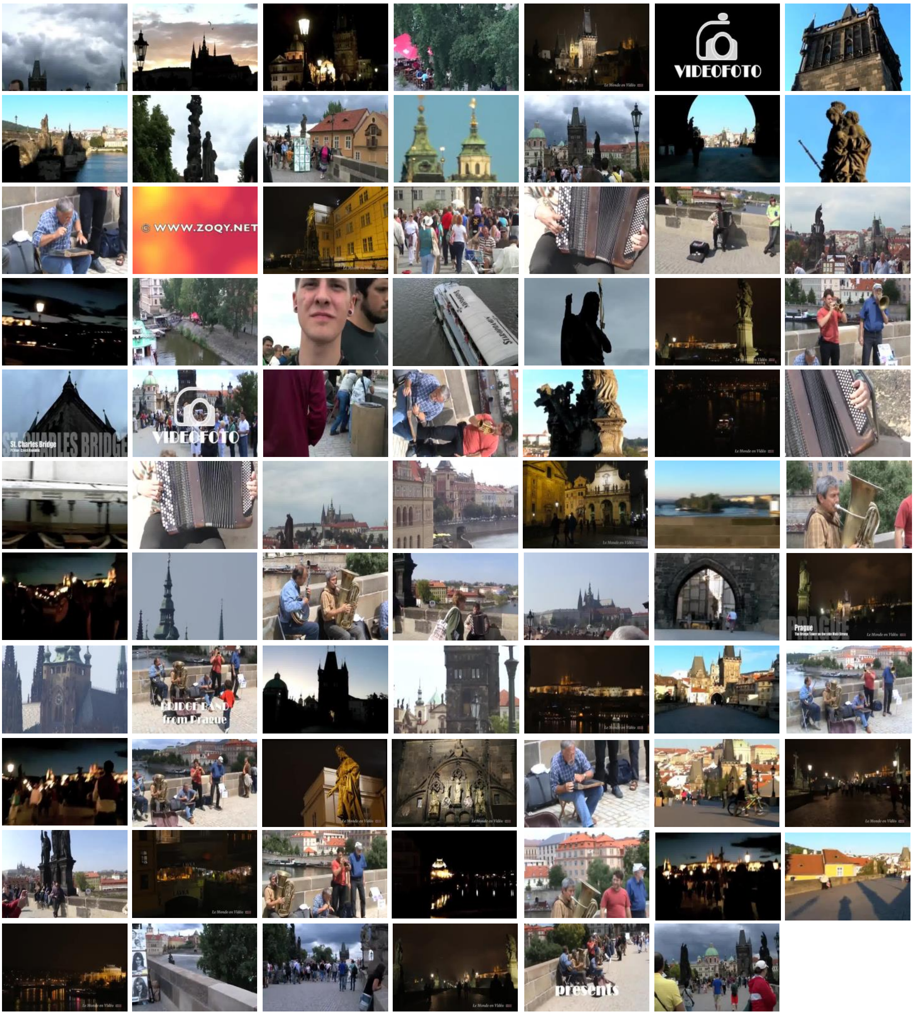


Fig. 8: Representative segments generated by our approach (DiMS) in summarizing videos of **Charles Bridge**. As can be seen, our approach select some redundant segments which reduces the overall quality of the final summary. This shows one of our **failure case**, where performance of our approach is slightly lower than the MultiVideoMMR baseline (0.525 vs 0.534). We believe this is because these videos contain subtle semantics like playing different musical instruments or lighting conditions which are difficult to capture without an additional semantic analysis.

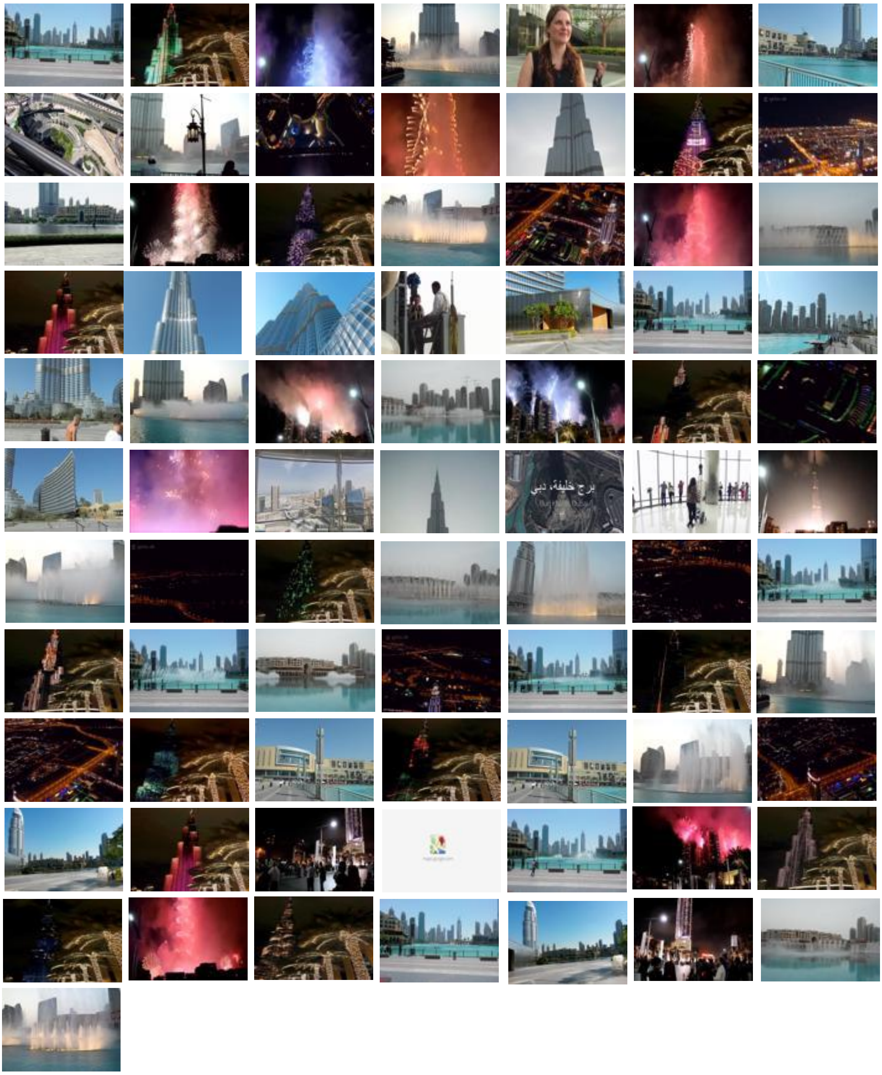


Fig. 9: Representative segments generated by our approach (DiMS) in summarizing videos of **Burj Khalifa**. We show the summaries at 10% length and represent each summarized segments using the corresponding central frame. This shows another **failure case**, where performance of our approach is slightly lower than the MultiVideoContent baseline (0.441 vs 0.450). In fact for this video collection, our approach achieved the lowest F-measure of 0.450 from the whole dataset. These videos contain videos contain fast motion and subtle semantics that define important segments of the video, such as opening the parachute or a nice panning segment from the top of the building. We believe our approach can be improved to handle this situation by including either deep motion features or learning a model for semantic analysis; we leave this as future work.

(c) Summarized events for the Office dataset:

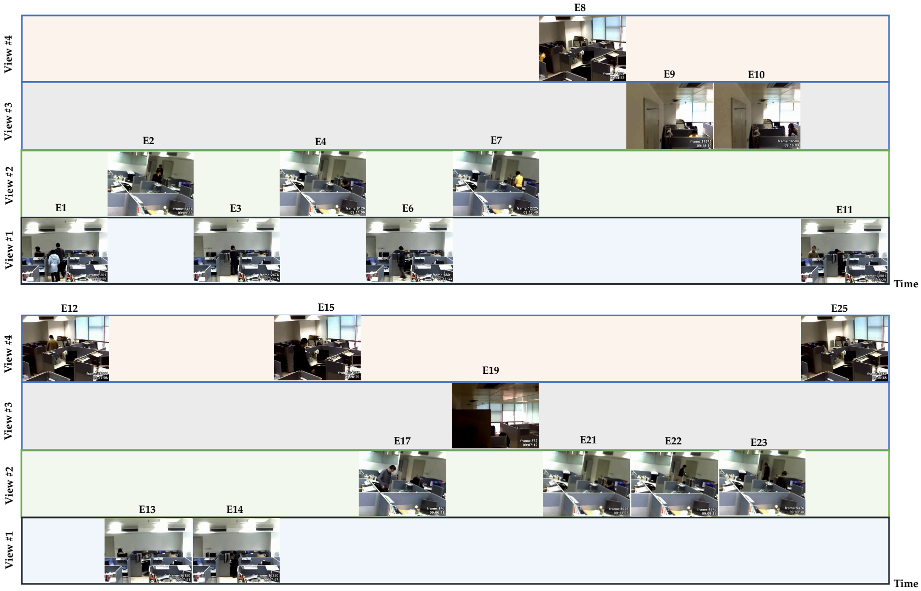


Fig. 10: Summarized events for the Office dataset. X-axis denotes the time line and the Y-axis represent the view (camera) from which the event is detected. Each event is represented by a key frame and an event number. As per the ground truth: A0 represents a girl with a black coat, A1 represents the same girl with a yellow sweater and B0 indicates another girl with a black coat. C and D are two boys. D wears a black topcoat and C wears a dark yellow sweater. E is a old man and F is a young guy about thirty years old. The sequence of events in our summary are: E1: A0, B, and D go out of the room, E2: A0 enters the room, E3: A0 stands in cubicle 1, E4: A0 sits in cubicle 1, E6: A0 leaves the room, E7: A1 enters the room and stands in Cubicle 1, E8: A1 sits in cubicle 1, E9: A1 and C leave the room one after another, E10: B0 enters the room, E11: C enters the room, E12: B0 sits in cubicle 1, E13: B0 goes out of the cubicle, E14: D enters the room and sits down in cubicle 1, E15: D walks to cubicle 2 from cubicle 1, E17: D sits in cubicle 2, E19: F enters the room, E21: A0 is looking for a thick book to read, E22: F leaves the room, E23: E leaves the room and E25: The computer screen in cubicle 2 turns off. As can be seen from the figure, Only 20 events out of 26 events are detected in our summary. It can be noticed that most of the events are detected from the view 1 and 2 as both of the cameras were focused to most activity region in the Fovs which can also be seen from the input videos. (Best viewed in color)

(d) Summarized events for the Campus dataset:

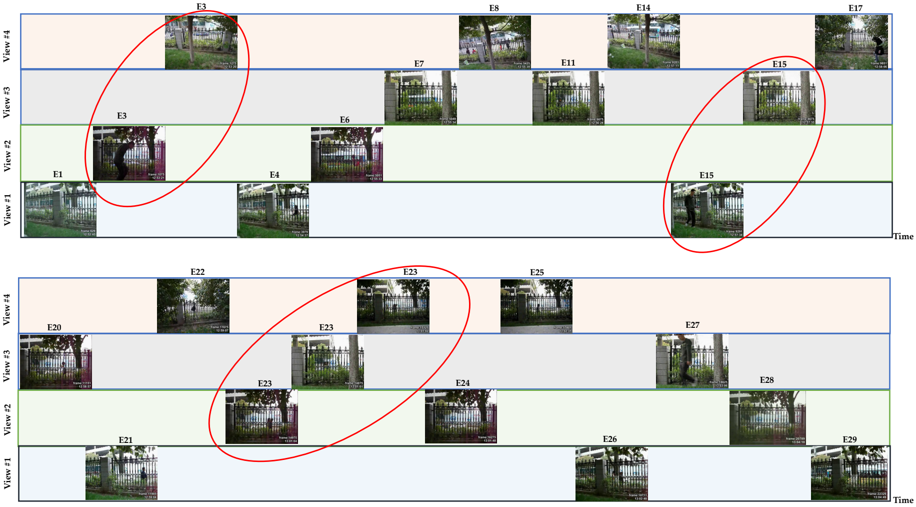


Fig. 11: Summarized events for the Campus dataset. X-axis denotes the time line and the Y-axis represent the view (camera) from which the event is detected. Each event is represented by a key frame and an event number. As per the ground truth: Object nos: A - vehicle, B - bicycle, C - pedestrian, Motion description: A1 - an object moves from left to right, A2 - an object moves from right to left and Position: inside - objects are inside the fence, outside - objects are outside the fence. For notational convenience, we represent an event by a tuple as follows; (C01, A2, outside) indicates a pedestrian moves from right to left inside the fence. The sequence of events in our summary are: E1: (C01, A2, Outside), E3: (C03, A1, inside), E4: (C05, A2, outside), E6: (C07, A2, outside), E7: (A01, A2, outside), E8: (C09, A1, outside), E11: (C12, A1, Outside), E14: (C14, A1, outside), E15: (C15, A1, inside), E17: (C17, A2, inside), E20: (C18, A1, outside), E21: (C19, A1, outside), E22: (C20, A1, outside), E23: (C21, A1, outside), E24: (C23, A1, outside), E25: (C24, A1, outside), E26: (C25, A1, outside), E27: (C27, A2, inside), E28: (C28, A1, outside), and E29: (A06, A1, outside). As can be seen from the figure, there exists some redundancies in our output summary as the video is captured using 4 hand-held cameras in an outdoor environment which makes the summarization difficult. Redundant events are grouped with red color circles. Only 20 out of 24 detected events are unique events in our summary. The dataset contains total 29 events.(Best viewed in color)

(e) Summarized events for the Lobby dataset:

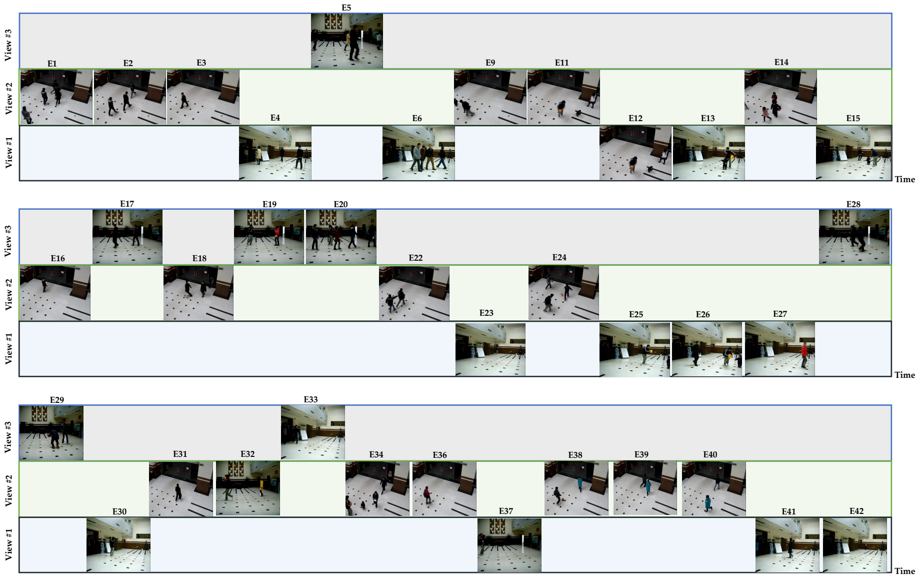


Fig. 12: Summarized events for the *Lobby* dataset. Sequence of events in our summary are: E1: Five persons walk across the lobby towards the gate; a man runs to the gate, E2: Two men walks across the lobby towards the gate, E3: A man run into the lobby from the gate, E4: Four persons walk into the lobby from the gate, E5: A man wlaks across the lobby towards the gate, E6: Three men are walking across the lobby towards the gate, E9: A man plays a ball with a baby, E11: A woman wearing a white coat walks across the lobby towards the gate, E12: A woman with a white coat passes away while a man is playing with a baby, E13: A man throws the ball towards the baby, E14: Two women and a man walk across the lobby from the gate, E15: A man plays a ball with a baby, E16: A man walks across the lobby towards the gate, E17: Two men walk across the lobby towards the gate, E18: Two men wearing black coats walk into the lobby from the gate, E19: A woman wearing a red coat walks across the lobby quickly towards the gate, E20: A man with a baby in his arms, and four other persons walks into the lobby, E22: Two men are running to catch each other, E24: Two men are passing a basketball to each other, E25: A man plays a ball with a baby, E26: A man wearng a black coat walks across the lobby towards the gate, E27: A man wearing a red coat walks across the lobby towards the gate, E28: A man and two women walk acroos the lobby, E29: Two men are playing basketball, E30: A man with a baby in his arms walks walks into the lobby, E31: A man with a brief case taken in hand walks into the lobby, E32: A man with a baby in his arms walks across the lobby, E33: A man with a baby in his arms runs in the lobby, E34: Four men, two women and a man with a baby in his arms walk into the lobby, E36: A man wearing a red coat walks across the lobby, E37: A man wearing a black coat walks across the lobby towards the gate, E38: A man runs quickly with a basketball rolling on the ground, E39: A man runs across the lobby, E40: A woman wearing a blue coat walks into the lobby, E41: A woman wearing a black coat walks into the lobby, E42: A man is wandering in the lobby. As can be seen from the figure, Only 37 events out of 43 events are detected in our summary.

Algorithmic Details

(a) Detailed derivation of ADMM steps:

In this section, we present details about the alternating method of multipliers (ADMM) algorithm to solve the weighted $\ell_{2,1}$-norm minimization problem (Eq. 10 of the main paper).

Specifically, the overall procedure of the ADMM algorithm is as follows:

(1) introduce an auxiliary variable to separate the objective function into separable objectives,

(2) form an augmented Lagrangian with introducing both linear and quadratic error terms through a dual variable,

(3) minimize the augmented Lagrangian iteratively with respect to the primal variable and the dual variable until convergence.

Following the explanations in the main paper, problem (10) is given by:

To facilitate the optimization, we consider an equivalent form of (**??**) by introducing an auxiliary variable U:

$$\min_{Z,U} \|X - XU\|_F^2 + \lambda \|Z\|_{K,2,1} \ s.t. \ \ U = Z \tag{1}$$

The augmented Lagrangian of (1) is:

$$\mathcal{L}_\mu(U, Z, \Lambda) = \frac{1}{2}\|X - XU\|_F^2 + \lambda\|Z\|_{K,2,1} + \langle \Lambda, U - Z\rangle + \frac{\mu}{2}\|U - Z\|_F^2 \tag{2}$$

where $\mathbf{\Lambda}$ is a Lagrangian multiplier, $\langle ., .\rangle$ denote the inner product, $||.||_F$ is the Frobenius norm and $\mu > 0$ is a penalty parameter.

To solve the problem in (2) at each iteration t, ADMM updates the variables in an alternating fashion as

$$U_{t+1} = \arg\min_U \mathcal{L}_\mu(U_t, Z_t, \Lambda_t) \tag{3}$$

$$Z_{t+1} = \arg\min_Z \mathcal{L}_\mu(U_{t+1}, Z_t, \Lambda_t) \tag{4}$$

$$\Lambda_{t+1} = \arg\min_\Lambda \mathcal{L}_\mu(U_{t+1}, Z_{t+1}, \Lambda_t) \tag{5}$$

In the following we present the derivation of specific update rules for (3-5).

Update U when fixing others: The problem (2) becomes:

$$\begin{aligned} &\min_U \ \frac{1}{2}\|X - XU\|_F^2 + \langle \Lambda, U - Z\rangle + \frac{\mu}{2}\|U - Z\|_F^2 \\ \Leftrightarrow &\min_U \ \frac{1}{2}U^T(X^TX + \mu I)U - (X^TX + \mu Z - \Lambda)^T U \end{aligned} \tag{6}$$

Note that it is a convex quadratic problem, hence it reduces to solving the following linear system:

$$(X^TX + \mu I)U = (X^TX + \mu Z - \Lambda) \tag{7}$$

Solving for U yields

$$U = (X^T X + \mu I)^{-1}(X^T X + \mu Z - \Lambda) \tag{8}$$

Update Z when fixing others: By ignoring the variables that are irrelevant to Z, we have

$$\min_{Z} \ \lambda \|Z\|_{K,2,1} + \langle \Lambda, U - Z \rangle + \frac{\mu}{2}\|U - Z\|_F^2 \tag{9}$$

On combining both linear and quadratic error terms into a single term by scaling the dual variable Λ, we get the following form

$$\min_{Z} \ \lambda \|Z\|_{K,2,1} + \frac{\mu}{2}\|U - Z + \Lambda/\mu\|_F^2 \tag{10}$$

Simple manipulation shows that (10) is equivalent to

$$\min_{Z} \ \sum_{i=1}^{n} \Big[\lambda K_{ii} \|Z_i\|_2 + \frac{\mu}{2} \Big\|U_i - Z_i + \frac{1}{\mu}\Lambda_i\Big\|_F^2 \Big] \tag{11}$$

which has a closed form solution by the one-dimensional shrinkage or soft thresholding formula [6]:

$$Z_i = \max\Big\{ \|r_i\|_2 - \frac{\lambda K_{ii}}{\mu}, 0 \Big\} \frac{r_i}{\|r_i\|_2} \tag{12}$$

where

$$r_i := U_i + \frac{1}{\mu}\Lambda_i. \tag{13}$$

We denote the above row-wise soft thresholding operation as follows:

$$Z = Shrink\Big(U + \frac{\Lambda}{\mu}, \frac{\lambda K}{\mu}\Big) = S_{\frac{\lambda K}{\mu}}(U + \Lambda/\mu) \tag{14}$$

Update Λ when fixing others: Having (U_{t+1}, Z_{t+1}) fixed, perform a gradient ascent update with step size of μ on the Lagrange multipliers as

$$\Lambda_{t+1} = \Lambda_t + \delta(U_{t+1} - Z_{t+1}) \tag{15}$$

Computing Platform. We carried out all experiments on a desktop PC with Intel(R) core(TM) i7-4790 processor with 16 GB of DDR3 memory. We used a NVIDIA Tesla K40 GPUs. to extract the C3D features.

Runtime Analysis. Our runtime analysis revealed that the it took on average 3 minutes to extract the C3D features and to compute the segment representations, while the sparse optimization took less than 2 minutes to generate a summary of 10% length from a video collection in topic-oriented video summarization. Similarily in camera network video summarization, our approach is significantly faster in generating summaries as compared to the both RandomWalk and BipartiteOPF. In RandomWalk, the authors first captured the multi-view correlations using a hypergraph based representation and then applied random walk for segment clustering whereas in BipartiteOPF, authors used a bipartite graph matching to model the correlations and then adopted optimum path forest clustering instead of random walk for clustering. We analyzed that both of the approaches are intrinsically complex and slow. Office dataset is a point in the case. For this dataset, we found that the total time needed for our approach is roughly 8.5 min, broken down into 5 min for deep feature extraction and computing correlations, 3.5 min to generate a summary of equal length as RandomWalk. This speed is about 15% faster than that of recently published BipartiteOPF and about 40% faster than that of RandomWalk. Note that these computational times are based on the reported values in the corresponding published papers.

References

1. Gygli, M., Grabner, H., Riemenschneider, H., Gool, L.V.: Creating summaries from user videos. In: ECCV. (2014) 8
2. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: CVPR. (2015) 8
3. de Avila, S.E.F., Lopes, A.P.B., da Luz Jr., A., de A. Arajo, A.: VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. PRL (2011) 8
4. K Zhang, W-L Chao, F.S., Grauman, K.: Summary transfer: exemplar-based subset selection for video summarization. In: CVPR. (2016) 8
5. Kline, P.: The handbook of psychological testing. Psychology Press (2000) 8
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via alternating direction method of multipliers. Foundations and Trends® in Machine Learning (2011) 19