Query-focused Video Summarization

Jacob Laurel[1], Aidean Sharghi[2], and Boqing Gong[2]

[1]University of Alabama at Birmingham [2]University of Central Florida

{jslaurel, aidean, bgong}@crcv.ucf.edu

Abstract

We introduce a new approach to video summarization based on a user query. We introduce a novel dataset and explain a baseline algorithm that was used to test our set. Our set has been gathered for 4 different videos with summaries for 60 different query pairs per video. We show that our dataset can be used to train a variety of supervised algorithms for the task of video-summarization.

1. Introduction

With an millions of new videos being created every day, coupled with recent developments in egocentric video technology (i.e.'GoPro' and 'GoogleGlass'), the goal of summarizing video content to a more manageable level is becoming ever more important . Recent trends have focused on summarizing videos for the purpose of detecting objects or generating storylines**,** however the goal of summarizing a video for a given query is unexplored [**1**,**2**]. Being able to customize a video summary for a user, based on a specific query or pair of queries, represents huge advance in the state-of-the-art. Motivated by this problem, we introduce the first Query-focused Video Summarization dataset and propose a novel benchmark algorithm. Our algorithm makes used of a Determintantal Point Process (DPP), a well-known mathematical tool that can be used to enforce diversity and distinctiveness in a set of randomly chosen objects.

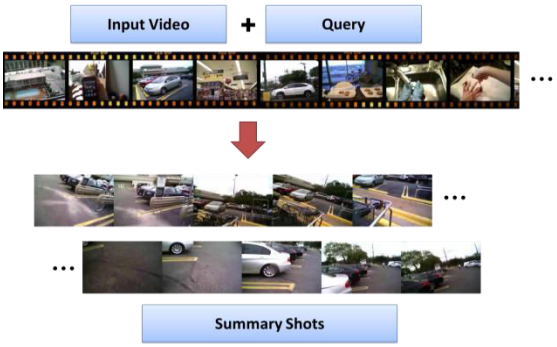


Figure1. Imagine a user wishes to obtain a video summary for "car". The generated summary would contain representative, yet query-dependent frames as shown above

1.1. Relevant Literature

Though query-focused video summarization is a new area, video summarization is an established topic of research, that has seen great strides in recent years[**1,6**], for example Lu and Graumann introduced story-driven video summaries, wherein a summary is automatically generated based on applying methods from text analysis to generate a textual "storyline" that can then be searched by a user [**2**] . Most existing video summarization methods rely on selecting some ideal set of frames based on some metric, whereas our approach selects frames probabilistically based on

enforcing a notion of distinctiveness that follows from the Determinantal Point Process (DPP). For this reason, we seek to build upon existing works that have also applied DPPs towards solving Video Summarization [**3**]. It is also worth noting that making an automatic summarization tailored to a specific query is also a well-known problem in its own right. Though our work is, to the best of our knowledge, the first to apply this idea to video summaries, the idea of generating textual document summaries based on a user's query has been of interest to the machine learning and NLP communities for much of the last decade [**4**].

2. Our Approach

Our approach and contributions are twofold: we gather and introduce a new fully summarized data set and also provide a benchmark algorithm that learns the queries in the data set and then makes use of Determintantal Point Processes (DPP) to generate the video summary. We then applied our algorithm to our data to establish a standard baseline.

2.1 Creating the Query-Focused Dataset:

We explain the process of collecting the query-dependent video summaries for the 4 videos in the UT Egocentric dataset [**1,2**]. The ground truth data was collected from human users through Amazon Mechanical Turk (AMT), an online crowd-sourcing platform that has recently been used for the process of gathering large datasets [**7**]

2.2 Selecting the Queries

The initial and arguably most important step to the data collection was to select the relevant queries to use for the data set (Note we use the term query and concept interchangeably). This consisted of referencing SentiBank's list of queries as well as comparing these against common Youtube and Vine search terms [**5**]. Once the initial set of queries was selected, it was given to 3 users to remove both less important queries and queries that were too specific or uncommon. Thus the final query list was filtered down to 48 concepts, which based on user preferences, was deemed to be fully comprehensive for most video genres.

2.3 Collecting the shots and annotations

We collect the shots from the 4 egocentric videos by first sampling the video at 1 fps. We define a shot to be a set of 5 consecutive frames, which is based on the assumption that over a 5 second period, the video content remains relatively constant. This assumption was later found to hold true based on user-evaluations. We give 3 different AMT workers the same shot and ask them to select any concepts from the 48 possible, which are found in the shot. We apply both an Intersection-over-Union threshold of 0.25, as well as lower limit on the minimum number of annotations which was selectively applied at certain shots, in order to achieve high-quality annotations and remove spurious ones. The union of all annotations provided by the 3 AMT workers was taken to be the ground truth for that specific shot. For each video, the final Intersection-over-Union ratio, as well as the average number of annotations, and the histogram count for each of the 48 concepts was computed in preparation for the Query-Pair selection.

2.4 Query Pair Selection

For each video, we select the set of Query Pairs by partitioning the Pairs into 4 disjoint sets, and then sampling 15 pairs from each of the sets.

2.4.1 Case 1 – Intersection:

The first set, denoted as the ***IntersectionSet***, represents the set of all Pairs such that both query concepts in the pair were found together in the same shot, in at least 30 separate shots. This signifies the scenario where the user wants the first query *AND* the second query. We employ the threshold of $\boldsymbol{T = 30}$ since it was found to be the optimal threshold based on the previously computed concept histogram to ensure that both concepts weren't simply tagged together accidentally by AMT workers, and that instead, the pairing represented an important conceptual relationship for the scene. We then rank the list of ***IntersectionSet*** pairs by how often both concepts occur in the same frame, or rather how large the Intersection count of the two queries is, over the set of all shots. We then randomly sample this ranked list (without replacement) for 15 different query pairs, where the probability of a query pair being selected is proportional to its Intersection count (i.e a pair of queries occurring together in 1000 shots is substantially more likely to be selected than a pair occurring in only 50).

$$Count_{Q_1 \cap Q_2} = \sum_{\{Shot_1 \ldots Shot_N\}} \mathbb{I}(Q_1)\, \mathbb{I}(Q_2) \quad (1)$$

$$P(Query1, Query2) \propto Count_{Q_1 \cap Q_2} \quad (2)$$

2.4.2 Case 2 – Union:

The second set, denoted as ***UnionSet***, represents the set of all pairs such that both query concepts in the pair occur at *least* $\boldsymbol{T}$ *times* in the video, while never actually occurring in the same shot together, thus the intersection of the two queries over the set of all shots is empty even though their union is high. This signifies the scenario where the user wants the first query *OR* the second query. In order to select the most representative pair of queries we again rank the list of queries in ***UnionSet***, however to rank them, we introduce a new metric, which is analogous to parallel resistance in an electrical network. We use the computed histogram to determine how many times each concept in the pair occurs in the video and we then take the product of both counts divided by their respective sum. The purpose of this metric is to ensure that for any given pair, the union is high if and only if both concepts appear several times throughout the video. This metric favors pairs such as "Book + Street" (histogram counts: 246 and 226 respectively), while disfavoring pairs such as "Computer + Grass (histogram counts: 264 and 38 respectively). As before, we then randomly sample this list for 15 different pairs where a given pair's selection probability is proportional to the proposed metric.

$$Count_{Q_1} = \sum_{\{Shot_1 \ldots Shot_N\}} \mathbb{I}(Q_1) \quad (3)$$

$$Count_{Q_2} = \sum_{\{Shot_1 \ldots Shot_N\}} \mathbb{I}(Q_2) \quad (4)$$

$$P(Query1, Query2) \propto \frac{Count_{Q_1} \cdot Count_{Q_2}}{Count_{Q_1} + Count_{Q_2}} \quad (5)$$

2.4.3 Case 3 – One Concept Missing:

The third disjoint set, denoted as ***OneMissingSet*** represents the set of all pairs such that one of the concepts occurs greater than $\boldsymbol{T}$ times, while the other concept occurs *less than* $\boldsymbol{T}$ *times*, and is thus defined as "missing" from the video. We first take the list of all concepts that occur *at least* $\boldsymbol{T}$ *times,* but that were not used in the previous two scenarios. Our reasoning for this is to make sure all significant concepts are included for the purposes of being able to train on them. We then sample from this set (without replacement) if its cardinality is greater than 15, and if it is less than 15 we simply take all of the concepts. In the case where this set's cardinality is less than 15, we obtain the remaining queries by sampling

the set of all remaining concepts occurring *at least* ***T*** times uniformly without replacement. For each of the 15 sampled queries, we then randomly select a concept from the set of all concepts occurring *Less than* ***T*** *times* and then pair the two queries together, thus producing a pair of queries where one concept *is* present while the other concept is defined to be "missing" since it occurs less than our threshold ***T***.

2.4.4 Case 4 –Both Concepts Missing:

The fourth case represents the set of all possible pairs of queries such that both queries occur *less than* ***T*** *times*, meaning both are not truly present in the video. As with the previous case, we uniformly sample without replacement to sample 15 possible pairs from this set.

2.5 Ground Truth Summarization Task

We collect the summaries for each video for each of the 4 aforementioned cases. We first generated a global summary by using the textual summarization annotations provided in the original UT Egocentric dataset to select relevant frames from each annotated region of the video. The collection of the global summary was performed by 2 users who were told to give preference to selecting frames that both closely matched the available textual annotations while also being visually distinct and important.

2.5.1 Case 1 – Intersection:

We collected the "Intersection" summary for each of the chosen 15 query pairs by first taking the set of all shot images where both queries in the given pair occurred (based on previously gathered annotations), and then writing both of the queries on top of the shot image to provide a reminder to the user that the image contains both concepts. Then for each of these 15 different query pair image sets, we take the given set's union with the global summary. We then give each of these 15 sets to 3 AMT workers instructing them to remove redundant, unimportant and incorrectly labeled frames, while also telling them to give preference to shot images that contain the queries

2.5.2 Case 2 – Union:

We collected the Union Summary for each of the 15 concept pairs in a similar fashion, though for each query pair, we first took the union of all images containing either query in the pair (i.e. if the pair was "Men + Flower", all shots having either 'Men' *or* 'Flower' were taken). Then for each respective pair, the previously obtained set of images was then combined with the global summary. As before, we wrote the query on top of the image if it had been previously annotated as such, in order to give the workers a gentle cue. We then gave each of these 15 unique sets to 3 separate AMT workers, just as before.

2.5.3 Case 3 – One Concept Missing:

We collected the summaries for the 15 concept pairs where one concept was missing in a slightly different fashion. For each of the 15 concept pairs we take the set of all image shots where the non-missing concept appears and then take the union of this set with the set of all images from the global summary. As before, we then write the name of the query on any image that contains that query. We then give each of these 15 sets of images to 3 separate AMT workers.

2.5.4 Case 4 –Both Concepts Missing:

We handle the scenario where neither concept is present by simply returning the global summary.

2.6 Algorithm

As previously stated, our algorithm makes use of a Determinantal point process (DPP). We first partition the video into sets of 10 frames sampled at 1 fps. We then compute the 10x10 DPP kernel matrix by taking the cosine distance similarity between SentiBank feature vectors for each of 10 choose 2 combination of frames. We then add the restriction that we only use determinants of 2x2 matrices (thus we are using a k-DPP with a k of 2). We exhaustively compute all the determinants and take the maximum which corresponds to some pair of frames *i* and *j*, which are then included in the final summary. We repeat this procedure over all sets of 10 frames for the entire video.

3. Results

We analyze our data set and compare it with other existing sets, specifically the MED dataset and the Co-Sum dataset in order to show the desirable properties of our set [**8,9**].

3.1 Annotations

The first desirable property of our set is the representativeness of the different annotations. Our set has been annotated for 48 different concepts whereas other video summarization sets have much less, as shown in Table 1. This property allows our dataset to be used for training other algorithms in a much broader range than existing methods

Table 1. Comparison of our method with the MED Category-specific data set as well as the Co-Sum dataset.

	Ours	MED[**8**]	Co-Sum[**9**]
Number of concepts	48	10	10
Annotators per Video	3	1	3
Total Summaries	136	160	10

We also show the frequency of some of the different annotations, taken for all videos:

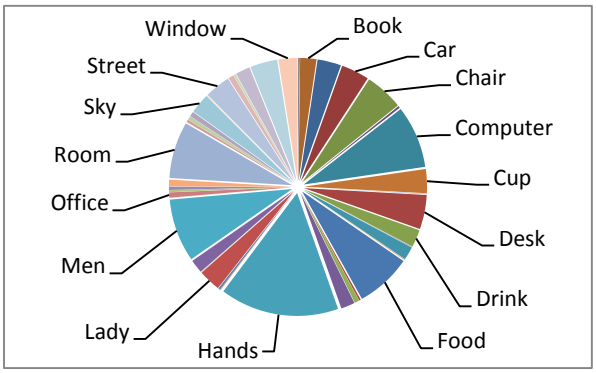


Figure 2. Display of some of the most common annotations that were given by AMT workers.

The results agree with basic intuition that for egocentric video, concepts such as hands and food would be among the most popular categories.

Our dataset was also found to have good inter-user agreement with an average for all videos of roughly 0.33. We believe this value may even be a lower than the true accuracy, given that many of the images had multiple correct annotations, thus a high union, which could artificially skew this metric.

$$\frac{|User1 \cap User2 \cap User3|}{|User1 \cup User2 \cup User3|} \approx 0.33 \quad (6)$$

3.2 Ground Truth Summaries

The second desirable property of our dataset is that there are 136 ground truth summaries for each of the 4 videos. This is because for each video we had 3 users annotate the entire video for 3 different cases (***IntersectionSet***, ***UnionSet***, ***OneMissingSet***) where each case had 15 different pairs where a summary was generated. The final summary for each video was simply the global summary of the entire video which was only gathered once per video, and this summary was also used for the 4th case (where both queried concepts were missing).

3.3 Algorithm Results

We briefly tested our algorithm on parts of the data, though we leave more extensive testing for future work. The results of the algorithm showed good consistency with the ground truth, and can be seen below.

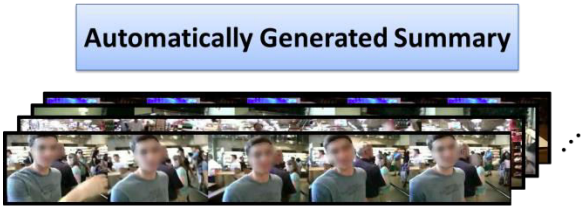


Figure 3. Results of a video summary generated using the k-DPP for the query pair "Men + Computer" where the *intersection* is the desired scenario.

4. Conclusions

We introduce the first query-based summarization dataset and provide it to the Computer Vision community to allow other researchers to reproduce results. Our dataset can be used to train a wide class of algorithms on 45 unique concept pairs per video, for the purpose of automatically generating custom, query-based summaries.

5. Acknowledgements

We wish to thank Dr.Mubarak Shah, Dr. Niels Lobo and Dr. Boqing Gong for their support. This work was supported by the NSF for the UCF REU in Computer Vision.

6. References

1. Discovering Important People and Objects for Egocentric Video Summarization. Y. J. Lee, J. Ghosh, and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 2012.

2. Story-Driven Summarization for Egocentric Video. Z. Lu and K. Grauman. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 2013

3. Diverse Sequential Subset Selection for Supervised Video Summarization. B. Gong, W. Chao, K. Grauman, and F. Sha. In Advances in Neural Information Processing Systems (NIPS), Montreal, Canada, Dec 2014.

4. A. Nenkova and K. Mckeown, "A Survey of Text Summarization Techniques," Mining Text Data, pp. 43–76, 2012.

5. Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel and Shih-Fu Chang. "Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs," ACM Multimedia Conference, Barcelona, Oct 2013

6. A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, “Large-Scale Video Summarization Using Web-Image Priors,” 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

7. C. Vondrick, D. Ramanan, and D. Patterson, “Efficiently Scaling Up Video Annotation with Crowdsourced Marketplaces,” Computer Vision – ECCV 2010 Lecture Notes in Computer Science, pp. 610–623, 2010.

8. D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, “Category-Specific Video Summarization,” Computer Vision – ECCV 2014 Lecture Notes in Computer Science, pp. 540–555, 2014.

9. W.-S. Chu, Y. Song, and A. Jaimes, “Video co-summarization: Video summarization by visual co-occurrence,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.