

Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach

Aidean Sharghi[†], Jacob S. Laurel[‡], and Boqing Gong[†]

[†]Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816

[‡]Department of Computer Science, University of Alabama at Birmingham, AL 35294

ai.dean.sharghi@knights.ucf.edu, jsl@uab.edu, bgong@crcv.ucf.edu

Abstract

Recent years have witnessed a resurgence of interest in video summarization. However, one of the main obstacles to the research on video summarization is the user subjectivity — users have various preferences over the summaries. The subjectiveness causes at least two problems. First, no single video summarizer fits all users unless it interacts with and adapts to the individual users. Second, it is very challenging to evaluate the performance of a video summarizer.

To tackle the first problem, we explore the recently proposed query-focused video summarization which introduces user preferences in the form of text queries about the video into the summarization process. We propose a memory network parameterized sequential determinantal point process in order to attend the user query onto different video frames and shots. To address the second challenge, we contend that a good evaluation metric for video summarization should focus on the semantic information that humans can perceive rather than the visual features or temporal overlaps. To this end, we collect dense per-video-shot concept annotations, compile a new dataset, and suggest an efficient evaluation method defined upon the concept annotations. We conduct extensive experiments contrasting our video summarizer to existing ones and present detailed analyses about the dataset and the new evaluation method.

1. Introduction

Recent years have witnessed a resurgence of interest in video summarization, probably due to the overwhelming video volumes showing up in our daily life. Indeed, both consumers and professionals have the access to ubiquitous video acquisition devices nowadays. While the video data is a great asset for information extraction and knowledge discovery, due to its size and variability, it is extremely hard for users to monitor or find the occurrences in it.

Jacob S. Laurel contributed to this work while he was an NSF REU student at UCF thanks to the support of NSF CNS #1461121.

Intelligent video summarization algorithms allow us to quickly browse a lengthy video by capturing the essence and removing redundant information. Early video summarization methods were built mainly upon basic visual qualities (e.g., low-level appearance and motion features) [13, 16, 24, 28, 36, 43, 52], while recently more abstract and higher-level cues are leveraged in the summarization frameworks [14, 17, 18, 23, 37, 44, 47, 50].

However, one of the main obstacles to the research on video summarization is the user subjectivity — users have various preferences over the summaries they would like to watch. The subjectiveness causes at least two problems. First, no single video summarizer fits all users unless it interacts with and adapts to the users. Second, it is very challenging to evaluate the performance of a video summarizer.

In an attempt to solve the first problem, we have studied a new video summarization mechanism, query-focused video summarization [37], that introduces user preferences in the form of text queries about the video into the summarization process. While this may be a promising direction to personalize video summarizers, the experimental study in [37] was conducted on the datasets originally collected for the conventional generic video summarization [25, 48]. It remains unclear whether the real users would generate distinct summaries for different queries, and if yes, how much the query-focused summaries differ from each other.

In this paper, we explore more thoroughly the query-focused video summarization and build a new dataset particularly designed for it. While we collect the user annotations, we meet the challenge how to define a good evaluation metric to contrast system generated summaries to user labeled ones — the second problem above-mentioned due to the user subjectivity about the video summaries.

We contend that the pursuit of new algorithms for video summarization has actually left one of the basic problems underexplored, i.e., how to benchmark different video summarizers. User study [26, 29] is too time-consuming to compare different approaches and their variations at large scale. In the prior arts of automating the evaluation pro-

<p>Dense Tags:</p> <p>Face Computer Men Phone Hands Chair Room Desk Hall</p> <p>Caption: I looked at my phone</p>	<p>Dense Tags:</p> <p>Chair Computer Room Desk Office</p> <p>Caption: I walked around my bedroom</p>
<p>Dense Tags:</p> <p>Lady Food Men Drink Hands Hat Computer Market Building Desk</p> <p>Caption: I waited in line with my friend</p>	<p>Dense Tags:</p> <p>Sky Street Building Hands Car Tree Window</p> <p>Caption: I drove the car in traffic</p>

Figure 1: Comparing the semantic information captured by captions in [48] and by the concept tags we collected.

cedure, on one end, a system generated summary has to consist of exactly the same key units (frame or shot) as in the user summaries in order to be counted as a good one [9, 39, 46]. On the other end, pixels and low-level features are used to compare the system and user summaries [14, 17, 18, 50, 52], whereas it is unclear what features and distance metrics match users' criteria. Some works strive to find a balance between the two extremes, e.g., using the temporal overlap between two summaries to define the evaluation metrics [15, 16, 35, 51]. However, all such metrics are derived from either the temporal or visual representations of the videos, without explicitly encoding how humans perceive the information — after all, the system generated summaries are meant to deliver similar information to the users as those directly labeled by the users.

In terms of defining a better measure that closely tracks what humans can perceive from the video summaries, we share the same opinion as Yeung et al.'s [48]: it is key to evaluate how well a system summary is able to retain the semantic information, as opposed to the visual quantities, of the user supplied video summaries. Arguably, the semantic information is best expressed by the concepts that represent the fundamental characteristics of what we see in the video at multiple grains, with the focus on different areas, and from a variety of perspectives (e.g., objects, places, people, actions, and their finer-grained entities, etc.).

Therefore, as our first contribution, we collect dense per-video-shot concept annotations for our dataset. In other words, we represent the semantic information in each video

shot by a binary semantic vector, in which the 1's indicate the presence of corresponding concepts in the shot. We suggest a new evaluation metric for the query-focused (and generic) video summarization based on these semantic vector representations of the video shots¹.

In addition, we propose a memory network [40] parameterized sequential determinantal point process [14] for tackling the query-focused video summarization. Unlike the hierarchical model in [37], our approach does not rely on the costly user supervision about which queried concept appears in which video shot or any pre-trained concept detectors. Instead, we use the memory network to implicitly attend the user query about the video onto different frames within each shot. Extensive experiments verify the effectiveness of our approach.

The rest of the paper is organized as follows. We discuss some related works in Section 2. Section 3 elaborates the process of compiling the dataset, acquiring annotations, as well as a new evaluation metric for video summarization. In section 4 we describe our novel query-focused summarization model, followed by detailed experimental setup and quantitative results in Sections 5.

2. Related Work

We discuss some related works in this section.

This work extends our previous efforts [37] on *personalizing* video summarizers. Both works explore the query-

¹Both the dataset and the code of the new evaluation metric are publicly available at <http://www.ai-dean-sharghi.com/cvpr2017>.

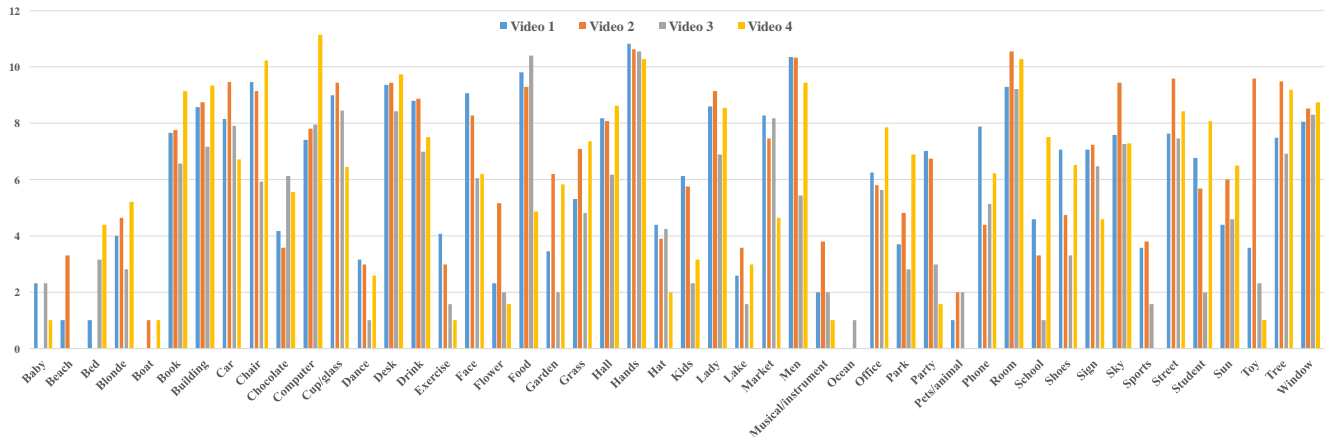


Figure 2: The frequencies of concepts showing up in the video shots, counted for each video separately.

focused video summarization, but we study this problem more thoroughly in this paper through a new dataset with dense per-video-shot tagging of concepts. Our memory network based video summarizer requires less supervision for training than the hierarchical model in [37].

Unlike our user-annotated semantic vectors for the video shots, Yeung et al. asked annotators to caption each video shot using a sentence [48]. A single sentence targets only limited information in a video shot and misses many details. Figure 1 contrasts the concept annotations in our dataset with the captions for a few video shots. The concept annotations clearly provide a more comprehensive coverage about the semantic information in the shots.

Memory networks [4, 40, 41, 42, 45] are versatile in modeling the attention scheme in neural networks. They are widely used to address question answering and visual question answering [3]. The query focusing in our summarization task is analogous to attending questions to the “facts” in the previous works, but the facts in our context are temporal video sequences. Moreover, we lay a sequential determinantal point process [14] on top of the memory network in order to promote diversity in the summaries.

A determinantal point process (DPP) [21] defines a distribution over the power sets of a ground set that encourages diversity among items of the subsets. There have been growing interest in DPP in machine learning and computer vision [1, 2, 5, 7, 11, 12, 19, 20, 22, 27, 30, 31, 38]. Our model in this paper extends DPPs’ modeling capabilities through the memory neural network.

3. Dataset

In this section, we provide the details on compiling a comprehensive dataset for video summarization. We opt to build upon the currently existing UT Egocentric (UTE) dataset [25] mainly for two reasons: 1) the videos are consumer grade, captured in uncontrolled everyday scenarios, and 2) each video is 3–5 hours long and contains a diverse

set of events, making video summarization a naturally desirable yet challenging task. In what follows, we first explain how we define a dictionary of concepts and determine the best queries over all possibilities for the query-focused video summarization. Then we describe the procedure of gathering user summaries for the queries. We also show informative statistics about the collected dataset.

3.1. Concept Dictionary and Queries

We plan to have annotators to transform the semantic information in each video shot to a binary semantic vector (cf. Figures 1 and 3), with 1’s indicating the presence of the corresponding concepts and 0’s the absence. Such annotations serve as the foundation for an efficient and automatic evaluation method for video summarization described in Section 3.2.1. The key is thus to have a dictionary that covers a wide range and multiple levels of concepts, in order to have the right basis to encode the semantic information.

In [37], we have constructed a lexicon of concepts by overlapping nouns in the video shot captions [48] with those in the SentiBank [6]. Those nouns serve as a great starting point for us since they are mostly entry-level [34] words. We prune out the concepts that are weakly related to visual content (e.g., “AREA”, which could be interpreted in various ways and applicable to most situations). Additionally, we merge the redundant concepts such as “CHILDREN” and “KIDS”. We also add some new concepts in order to construct an expressive and comprehensive dictionary. Two strategies are employed to find the new concept candidates. First, after watching the videos, we manually add the concepts that appear for a significant frequency, e.g., “COMPUTER”. Second, we use the publicly available statistics about YouTube and Vine search terms to add the terms that are frequently searched by users, e.g., “PET/ANIMAL”. The final lexicon is a concise and diverse set of 48 concepts (cf. Figure 2) that are deemed to be comprehensive for the UTE videos of daily lives.

User 1: Sky D Lady D Street D Market D Building D Hands D Tree D Car D Window
 User 2: Sky D Lady D Street D Hands D Tree D Car D Hat D Window
 User 3: Sky D Lady D Street D Hands D Tree D Car D Hat D Window

Figure 3: All annotators agree with each other on the prominent concepts in the video shot, while they miss different subtle concepts.



Figure 4: Two summaries generated by the same user for the queries {HAT, PHONE} and {FOOD, DRINK}, respectively. The shots in the two summaries beside the green bars exactly match each others, while the orange bars show the query-specific shots.

We also construct queries, to acquire query-focused user summaries, using two or three concepts as opposed to singletons. Imagine a use case of video search engines. The queries entered by users are often more than one word. For each video, we formalize 46 queries. They cover the following four distinct scenarios: i) all the concepts in the query appear in the same video shots together (15 such queries); ii) all concepts appear in the video but never jointly in a single shot (15 queries), iii) only one of the concepts constituting the query appears in some shots of the video (15 queries), and iv) none of the concepts in the query are present in the video (1 such query). We describe in the Suppl. Materials how we obtain the 46 queries to cover the four scenarios. Such queries and their user annotated summaries challenge an intelligent video summarizer from different aspects and extents.

3.2. Collecting User Annotations

We plan to build a video summarization dataset that offers 1) efficient and automatic evaluation metrics and 2) user summaries in response to different queries about the videos. For the former 1), we collect user annotations about the presence/absence of concepts in each video shot. This is a quite daunting task conditioning on the lengths of the videos and the size of our concept dictionary. We use Amazon Mechanical Turk (MTurk) (<http://www.mturk.com/>) for economy and efficiency considerations. For the latter 2), we hire three student volunteers to have better quality control over the labeled video summaries. We uniformly partition the videos to 5-second-long shots.

3.2.1 Shot Tagging: Visual Content to Semantic Vector

We ask MTurkers to tag each video shot with all the concepts that are present in it. To save the workers' time from watching the shots, we uniformly extract five frames from

each shot. A concept is assumed relevant to the shot as long as it is found in any of the five frames. Figure 3 illustrates the tagging results for the same shot by three different workers. While all the workers captured the prominent concepts like SKY, LADY, STREET, TREE, and CAR, they missed different subtle ones. The union of all their annotations, however, provides a more comprehensive semantic description about the video shot than that of any individual annotator. Hence, we ask three workers to annotate each shot and take their union to obtain the final semantic vector for the shot. On average, we have acquired 4.13, 3.95, 3.18, and 3.62 concepts per shot for the four UTE videos, respectively. In sharp contrast, the automatically derived concepts [37] from the shot captions [48] are far from enough; on average, there are only 0.29, 0.58, 0.23, and 0.26 concepts respectively associated with each shot of the four videos.

Evaluating video summaries. Thanks to the dense concept annotations per video shot, we can conveniently contrast a system generated video summary to user summaries according to the semantic information they entail. We first define a similarity function between any two video shots by intersection-over-union (IOU) of their corresponding concepts. For instance, if one shot is tagged by {CAR, STREET} and another by {STREET, TREE, SIGN}, then the IOU similarity between them is $1/4 = 0.25$.

To find the match between two summaries, it is convenient to execute it by the maximum weight matching of a bipartite graph, where the summaries are on opposite sides of the graph. The number of matched pairs thus enables us to compute precision, recall, and F1 score. Although this procedure has been used in the previous work [17, 10], there the edge weights are calculated by low-level visual features which by no means match the semantic information humans obtain from the videos. In sharp contrast, we use the IOU

Table 1: Inter-user agreement evaluated by F1 score (%) (U1, U2, and U3: the three student volunteers, O: the oracle summary).

U1-U2	U1-U3	U2-U3	U1-O	U2-O	U3-O
55.27	55.85	62.67	64.97	79.75	80.07

similarities defined directly over the user annotated semantic vectors as the edge weights.

3.2.2 Acquiring User Summaries

In addition to the dense per-video-shot concept tagging, we also ask annotators to label query-focused video summaries for the 46 queries described in Section 3.1.

To ensure consistency in the summaries and better quality control over the summarization process, we switch from MTurk to three student volunteers in our university. We meet and train the volunteers in person. They each summarize all four videos by taking queries into account — an annotator receives 4 (videos) \times 46 (queries) summarization tasks in total. We thus obtain three user summaries for each query-video pair.

However, we acknowledge that it is infeasible to have the annotators to summarize all the query-video pairs from scratch — the UTE videos are each 3–5 hours long. To overcome this issue, we expand each temporal video to a set of static key frames. First, we uniformly extract five key frames to represent each shot in the same way as in Section 3.2.1. Second, we pool all the shots corresponding to the three textual summaries [48] as the initial candidate set. Third, for each query, we further include all the shots that are relevant to it into the set. A shot is relevant to the query if the intersection of the concepts associated with it and the query is nonempty. As a result, we have a set of candidate shots for each query that covers the main story in the video as well as those of relevance to the query. The annotators summarize the video by removing redundant shots from the set. There are 2500 to 3600 shots in the candidate sets, and the summaries labeled by the participants contain only 71 shots on average.

Oracle summaries. Supervised video summarization methods [14, 16, 37, 50, 51] often learn from one summary per video, or per query-video pair in query-focused summarization, while we have three user generated summaries per query. We aggregate them into one, called the oracle summary, per query-video pair by a greedy algorithm. The algorithm starts from the common shots in the three user summaries. It then greedily chooses one shot every time such that this shot gives rise to the largest marginal gain over the evaluated F1 score. We leave the details to the Suppl. Materials. The oracle summaries achieve better agreements with the users than the inter-user consensus (cf. Table 1).

Summaries of the same video differ due to queries. Figure 4 shows two summaries labeled by the same user for

Table 2: The average lengths and standard deviations of the summaries for different queries.

	User 1	User 2	User 3	Oracle
Vid1	143.7 \pm 32.5	80.2 \pm 47.1	62.6 \pm 15.7	82.5 \pm 33.9
Vid2	103.0 \pm 45.0	49.9 \pm 25.2	64.4 \pm 11.7	64.1 \pm 11.7
Vid3	97.3 \pm 38.9	50.1 \pm 9.6	58.4 \pm 9.3	59.2 \pm 9.6
Vid4	79.9 \pm 30.3	34.4 \pm 7.3	28.9 \pm 8.7	35.6 \pm 8.5

two distinct queries, {HAT, PHONE} and {FOOD, DRINK}. Note that the summaries both track the main events happening in the video while they differ in the query-specific parts. Besides, table 2 reports the means and standard deviations of the lengths of the summaries per video per user. We can see that the queries highly influence the resulting summaries; the large standard deviations attribute to the queries.

Budgeted summary. For all the summaries thus far, we do not impose any constraints over the total number of shots to be included into the summaries. After we receive the annotations, however, we let the same participants further reduce the lengths of their summaries to respectively 20 shots and 10 shots. We call them *budgeted* summaries and leave them for future research.

4. Approach

We elaborate our approach to the query-focused video summarization in this section. Denote by $V = \{V_t\}_{t=1}^T$ a video that is partitioned to T segments, and by q the query about the video. In our experiments, every segment V_t consists of 10 video shots each of which is 5-second long and is used in Section 3.2 to collect the concept annotations.

4.1. Query Conditioned Sequential DPP

The sequential determinantal point process (DPP) [14] is among the state-of-the-art models for generic video summarization. We condition it on the query q as our overarching video summarization model,

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_T = y_T | V, q) \quad (1)$$

$$= P(Y_1 = y_1 | V_1, q) \prod_{t=2}^T P(Y_t = y_t | V_t, y_{t-1}, q) \quad (2)$$

where the t -th DPP variable Y_t selects subsets from the t -th segment V_t , i.e., $y_t \subseteq V_t$, and the distribution $P(Y_t = y_t | V_t, y_{t-1}, q)$ is specified by a conditional DPP [21],

$$P(Y_t = y_t | V_t, y_{t-1}, q) = \frac{\det[L(q)]_{y_t, y_{t-1}}}{\det[L(q) + I_t]}. \quad (3)$$

The nominator on the right-hand side is the principle minor of the (L-ensemble) kernel matrix $L(q)$ indexed by the subsets $y_t \cup y_{t-1}$. The denominator calculates the determinant

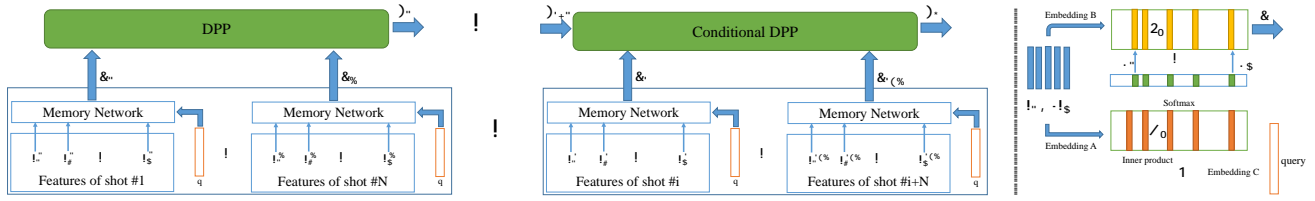


Figure 5: Our query-focused video summarizer: Memory network (right) parameterized sequential determinantal point process (left).

of the sum of the kernel matrix and a corrupted identity matrix whose elements indexed by y_{t-1} are 0's. Readers are referred to the great tutorial [21] on DPP for more details.

Note that the DPP kernel $L(q)$ is parameterized by the query q . We have to carefully devise the way of parameterizing it in order to take account of the following properties. In query-focused video summarization, a user selects a shot to the summary for two possible reasons. One is that the shot is quite related to the query and thus becomes appealing to the user. The other may attribute to the contextual importance of the shot; e.g., the user would probably choose a shot to represent a prominent event in the video even if the event is not quite relevant to the query. To this end, we use a memory network to model the two types of importance (query-related and contextual) of a video shot simultaneously.

4.2. Memory Network to Parameterize DPP Kernels

The memory network [40] offers a neural network architecture to naturally attend a question to “facts” (cf. the rightmost panel of Figure 5). In our work, we shall measure the relevance between the query q and a video shot and incorporate such information into the DPP kernel $L(q)$. Therefore, it is straightforward to substitute the question in memory network by our query, but the “facts” are less obvious.

As discussed in Section 3.1, there could be various scenarios for a query and a shot. All the query concepts may appear in the shot but possibly in different frames; one or two concepts of the query may not be present in the shot; it is also possible that none of the concepts are relevant to any frame in the shot. In other words, the memory network is supposed to screen all the video frames in order to determine the shot’s relevance to the query. Hence, we uniformly sample 8 frames from each shot as the “facts”. The video frames are represented using the same feature as [37] (cf. f_1, \dots, f_K on the rightmost panel of Figure 5).

The memory network takes as input the video frames $\{f_k\}$ of a shot and a query q . The frames are transformed to memory vectors $\{m_k\}$ through an embedding matrix A . Similarly, the query q , represented by a binary indication vector, is mapped to the internal state u using an embedding matrix C . The attention scheme is implemented simply by a dot product followed by a softmax function,

$$p_k = \text{Softmax}(u^T m_k), \quad (4)$$

where p_k carries how much attention the query q incurred over the frame f_k .

Equipped with the attention scores $\{p_k\}$, we assemble another embedding $\{c_k\}$ of the frames, obtained by the mapping matrix B in figure 5, into the video shot representation o :

$$o = \sum_k p_k c_k, \quad (5)$$

which is conditioned on the query q and entails the relevance strength of the shot to the query. As a result, we expect the DPP kernel parameterized by the following

$$[L(q)]_{ij} = o_i^T D^T D o_j \quad (6)$$

is also flexible in modeling the importance of the shots to be selected into the video summary. Here i and j index two shots, and D is another embedding matrix. Note that the contextual importance of a shot can be inferred from a shot’s similarities to the others by the kernel matrix, while the query-related importance is mainly by the attention scheme in the memory network.

4.3. Learning and Inference

We learn the overall video summarizer, including the sequential DPP and the memory network, by maximizing the log-likelihood of the user summaries in the training set. We use stochastic gradient descent with mini-batching to optimize the embedding matrices $\{A, B, C, D\}$. The learning rates and numbers of epochs are chosen using the validation set. At the test stage, we sequentially visit the video segments V_1, \dots, V_T and select shots from them using the learned summarization model.

It is notable that our approach requires less user annotations than the SH-DPP [37]. It learns directly from the user summaries and implicitly attend the queries to the video shots. However, SH-DPP requires very costly annotations about the relevances between video shots and queries. Our new dataset does supply such supervisions, so we shall include SH-DPP as the baseline method in our experiments.

5. Experimental Results

We report experimental setup and results in this section.

Features. We extract the same type of features as used in the existing SH-DPP method [37] in order to have fair

Table 3: Comparison results for query-focused video summarization (%).

	SeqDPP [14]			SH-DPP [37]			Ours		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Vid1	53.43	29.81	36.59	50.56	29.64	35.67	49.86	53.38	48.68
Vid2	44.05	46.65	43.67	42.13	46.81	42.72	33.71	62.09	41.66
Vid3	49.25	17.44	25.26	51.92	29.24	36.51	55.16	62.40	56.47
Vid4	11.14	63.49	18.15	11.51	62.88	18.62	21.39	63.12	29.96
Avg.	39.47	39.35	30.92	39.03	42.14	33.38	40.03	60.25	44.19

comparisons. First, we employ 70 concept detectors from SentiBank [6] and use the detection scores for the features of each key frame (8 key frames per 5-second-long shot). However, it is worth mentioning that our approach is not limited to using concept detection scores and, more importantly unlike SH-DPP, does not rely on the per-shot annotations about the relevance to the query — the per shot user labeled semantic vectors serve for evaluation purpose only. Additionally, we extract a six dimensional contextual feature vector per shot as the mean-correlations of low-level features (including color histogram, GIST [33], LBP [32], Bag-of-Words, as well as an attribute feature [49]) in a temporal window whose size varies from 5 to 15 shots. The six-dimensional contextual features are appended to the key frame features in our experiments.

Data split. We run four rounds of experiments each leaving one video out for testing and one for validation, while keeping the remaining two for training. Since our video summarizer and the baselines are sequential models, the small number (i.e., two) of training videos is not an issue as the videos are extremely long, providing many variations and supervisions at the training stage.

5.1. Comparison Results

Query-focused video summarization. We contrast our video summarizer, the memory-network based sequential determinantal point process, to several closely related methods. We first include SH-DPP [37], the most recent approach to the query-focused video summarization. Our model improves upon SeqDPP [14] by taking the query into account and parameterizing the DPP kernel by the memory network. SeqDPP is thus directly comparable to ours. We concatenate the query features (binary indication vectors) with the shot features and input them to SeqDPP and SH-DPP. We set the same dimensionality for all the embedding spaces in our and the two baseline methods. It turns out the 128D embeddings are chosen due to their performances on the validation videos.

Table 3 compares the performances of the three video summarizers. Each video is taken in turn as the test video and the corresponding results are shown in each row. The average results are included as the last row. Precision, re-

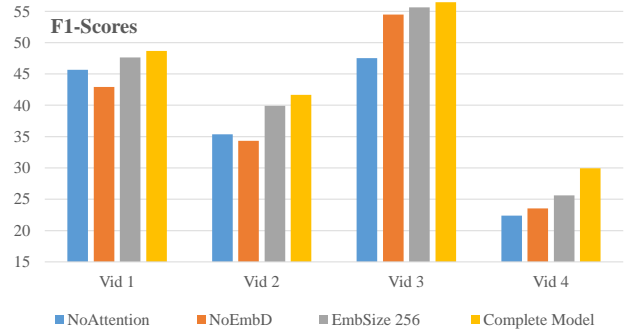


Figure 6: The Effectiveness of various individual components in our proposed video summarizer.

call, and F1 score are reported for all the video summarizers. Our approach outperforms the other two by a large margin (more than 10% F1 score on average). It seems like Video 4 is especially challenging for all the methods. For Video 2, our summarizer generates a little longer summaries than the others do. In the future work, we will explore how to control the summary length in the sequential DPP model.

Component-wise analyses. To investigate how each component in our framework contributes to the final results, we conduct more experiments by either removing or modifying them. Figure 6 shows the corresponding results.

The main benefit from the memory network is the attention mechanism (cf. equation (6)). If we instead use a uniform distribution for the attention scores $\{p_i\}$ and append the query information u directly after the memory network output o , the results become worse on all the four videos (cf. NoAttention in Figure 6). The NoEmbD results are obtained after we remove the last embedding matrix D when we compute the DPP kernels. Finally, EmbSize 256 are the results when we change the 128D embeddings in our approach to 256D. The performance drops from our complete model verify that all the corresponding components are complementary, jointly contributing to the final results.

Generic video summarization. Recall that our queries incur four different scenarios (cf. Section 3.1). When there are no video shots relevant to the query, it reduces to the generic video summarization in some sort. We single out such queries and contrast our summarizer to some existing and recent methods for generic video summarization:

Table 4: Comparison results for generic video summarization, i.e., when no video shots are relevant to the query.

	SubMod [16]			Quasi [52]			Ours		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Vid1	47.86	51.28	49.51	57.37	49.36	53.06	65.88	59.75	62.66
Vid2	56.53	46.50	51.03	46.75	63.34	53.80	35.07	67.31	46.11
Vid3	62.46	66.72	64.52	53.93	46.44	49.91	65.95	53.12	58.85
Vid4	34.49	37.25	35.82	13.00	77.88	22.31	22.29	67.74	33.5
Avg.	50.34	50.44	50.22	42.76	59.25	44.77	47.3	61.98	50.29

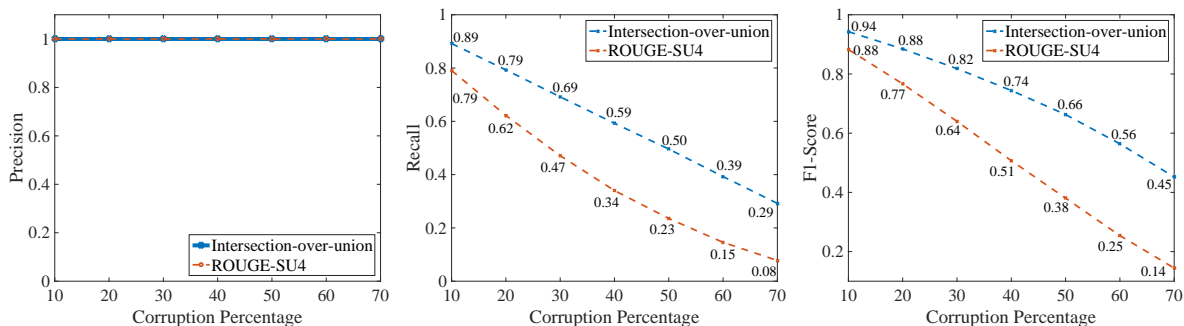


Figure 7: A nice behavior of our evaluation metric. When we randomly remove video shots from the user summaries, the recall between the original user summaries and the corrupted ones decreases almost linearly. The evaluation by ROUGE-SU4 [48] is included for reference.

SubMod [16] which employs submodular functions to encourage diversity and Quasi [52] which is an unsupervised method based on group sparse coding. Unlike the DPP type of summarizers, the baseline methods here are not able to automatically determine the lengths of the summaries. We tune the threshold parameter in Quasi such that the output lengths are no more or less than the oracle summary by 20 shots. For SubMod we set the budget parameter such that it generates summaries that are exactly as long as the oracle summaries. As shown in Table 4, our approach still gives the best overall performance even though we reveal the oracle summaries’ lengths to the baseline methods, probably due to its higher neural network based modeling capacity.

5.2. A Nice Behavior of Our Evaluation Metric

Our evaluation method for video summarization is mainly motivated by Yeung et al. [48]. Particularly, we share the same opinion that the evaluation should focus on the semantic information which humans can perceive, rather than the low-level visual features or temporal overlaps. However, the captions used in [48] are diverse, making the ROUGE-SU4 evaluation unstable and poorly correlated with human judgments [8], and often missing subtle details (cf. Figure 1 for some examples).

We rectify those caveats by instead collecting dense concept annotations. Figure 1 exhibits a few video shots where the concepts we collected provide a better coverage than the captions about the semantics in the shots. Moreover, we conveniently define an evaluation metric based on the IOU similarity function between any two shots (cf. Section 3.2.1)

thanks to the concept annotations.

Our evaluation metric has some nice behaviors. If we randomly remove some video shots from the user summaries and compare the corrupted summaries with the original ones, an accuracy-like metric should give rise to linearly decreasing values. This is indeed what happens to our recall as shown in Figure 7. In contrast, the ROUGE-SU4 recall, taking as input the shot captions, exhibits some nonlinearity. More results on randomly replacing some shots in the user summaries are included in the Suppl. Materials.

6. Conclusion

In this work, our central theme is to study the *subjectiveness* in video summarization. We have analyzed the key challenges caused the subjectiveness and proposed some solutions. In particular, we compiled a dataset that is densely annotated with a comprehensive set of concepts and designed a novel evaluation metric that benefits from the collected annotations. We also devised a new approach to generating personalized summaries by taking user queries into account. We employed memory networks and determinantal point processes in our summarizer, so that our model leverages their attention schemes and diversity modeling capabilities, respectively. Extensive experiments verify the effectiveness of our approach and reveals some nice behaviors of our evaluation metric.

Acknowledgements. This work is supported by NSF IIS #1566511, a gift from Adobe Systems, and a GPU from NVIDIA. We thank Fei Sha, the anonymous reviewers and area chairs, especially R2, for their insightful suggestions.

References

- [1] R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar. Learning the parameters of determinantal point process kernels. In *ICML*, pages 1224–1232, 2014. **3**
- [2] A. Agarwal, A. Choromanska, and K. Choromanski. Notes on using determinantal point processes for clustering with applications to text clustering. *arXiv preprint arXiv:1410.6975*, 2014. **3**
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. **3**
- [4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. **3**
- [5] N. K. Batmanghelich, G. Quon, A. Kulesza, M. Kellis, P. Golland, and L. Bornn. Diversifying sparsity using variational determinantal point processes. *arXiv preprint arXiv:1411.6307*, 2014. **3**
- [6] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 459–460. ACM, 2013. **3, 7**
- [7] W.-L. Chao, B. Gong, K. Grauman, and F. Sha. Large-margin determinantal point processes. UAI, 2015. **3**
- [8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. **8**
- [9] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3584–3592, 2015. **2**
- [10] S. E. F. De Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo. Vsum: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011. **4**
- [11] M. Gartrell, U. Paquet, and N. Koenigstein. Low-rank factorization of determinantal point processes for recommendation. *arXiv preprint arXiv:1602.05436*, 2016. **3**
- [12] J. A. Gillenwater, A. Kulesza, E. Fox, and B. Taskar. Expectation-maximization for learning determinantal point processes. In *Advances in Neural Information Processing Systems*, pages 3149–3157, 2014. **3**
- [13] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz. Schematic storyboarding for video visualization and editing. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 862–871. ACM, 2006. **1**
- [14] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, pages 2069–2077, 2014. **1, 2, 3, 5, 7**
- [15] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. **2**
- [16] M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3090–3098, 2015. **1, 2, 5, 8**
- [17] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2698–2705, 2013. **1, 2, 4**
- [18] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4225–4232, 2014. **1, 2**
- [19] A. Kulesza and B. Taskar. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1193–1200, 2011. **3**
- [20] A. Kulesza and B. Taskar. Learning determinantal point processes. 2011. **3**
- [21] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012. **3, 5, 6**
- [22] J. T. Kwok and R. P. Adams. Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, pages 2996–3004, 2012. **3**
- [23] J. Kwon and K. M. Lee. A unified framework for event summarization and rare event detection. In *CVPR*, pages 1266–1273, 2012. **1**
- [24] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu. Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 144–148. ACM, 2008. **1**
- [25] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, volume 2, page 7, 2012. **1, 3**
- [26] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015. **1**
- [27] C. Li, S. Jegelka, and S. Sra. Fast dpp sampling for nyström with application to kernel methods. *arXiv preprint arXiv:1603.06052*, 2016. **3**
- [28] T. Liu and J. R. Kender. Optimization algorithms for the selection of key frame sequences of variable length. In *European Conference on Computer Vision*, pages 403–417. Springer, 2002. **1**
- [29] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013. **1**
- [30] Z. Mariet and S. Sra. Fixed-point algorithms for learning determinantal point processes. *Advances in Neural Information Systems (NIPS)*, 2015. **3**
- [31] Z. Mariet and S. Sra. Kronecker determinantal point processes. *arXiv preprint arXiv:1605.08374*, 2016. **3**
- [32] T. Ojala, M. Pietikainen, and T. Maenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. **7**

- [33] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 7
- [34] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2768–2775, 2013. 3
- [35] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV 2014 - European Conference on Computer Vision*, 2014. 2
- [36] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 435–441. IEEE, 2006. 1
- [37] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 1, 2, 3, 4, 5, 6, 7
- [38] J. Snoek, R. Zemel, and R. P. Adams. A determinantal point process latent variable model for inhibition in neural spiking data. In *Advances in Neural Information Processing Systems*, pages 1932–1940, 2013. 3
- [39] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, 2015. 2
- [40] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015. 2, 3, 6
- [41] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 3
- [42] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3
- [43] W. Wolf. Key frame selection by motion analysis. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 1228–1231. IEEE, 1996. 1
- [44] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *European Conference on Computer Vision*, pages 282–298. Springer, 2014. 1
- [45] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*, 2016. 3
- [46] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2244, 2015. 2
- [47] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. 1
- [48] S. Yeung, A. Fathi, and L. Fei-Fei. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824*, 2014. 1, 2, 3, 4, 5, 8
- [49] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 771–778, 2013. 7
- [50] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. *arXiv preprint arXiv:1603.03369*, 2016. 1, 2, 5
- [51] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. *arXiv preprint arXiv:1605.08110*, 2016. 2, 5
- [52] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2513–2520, 2014. 1, 2, 8