Noname manuscript No.
(will be inserted by the editor)

Leveraging Semantic Saliency Maps for Query-Specific Video Summarization

Kemal Cizmeciler · Erkut Erdem[*] · Aykut Erdem

Received: date / Accepted: date

Abstract The immense amount of videos being uploaded to video sharing platforms makes it impossible for a person to watch all the videos understand what happens in them. Hence, machine learning techniques are now deployed to index videos by recognizing key objects, actions and scenes or places. Summarization is another alternative as it offers to extract only important parts while covering the gist of the video content. Ideally, the user may prefer to analyze a certain action or scene by searching a query term within the video. Current summarization methods generally do not take queries into account or require exhaustive data labeling. In this work, we present a weakly supervised *query-focused* video summarization method. Our proposed approach makes use of semantic attributes as an indicator of query relevance and semantic attention maps to locate related regions in the frames and utilizes both within a submodular maximization framework. We conducted experiments on the recently introduced RAD dataset and obtained highly competitive results. Moreover, to better evaluate the performance of our approach on longer videos, we collected a new dataset, which consists of 10 videos from YouTube and annotated with shot-level multiple attributes. Our dataset enables much diverse set of queries that can be used to summarize a video from different perspectives with more degrees of freedom.

Keywords query-specific, video summarization

1 Introduction

Recent advancements in digital imaging technologies and their increasing involvement in people's everyday lives has led to a massive increase in the amount of visual

[*] Corresponding author.

Kemal Cizmeciler · Erkut Erdem
Department of Computer Engineering, Hacettepe University, Ankara, Turkey
E-mail: kemalcizmeci@gmail.com, erkut@cs.hacettepe.edu.tr

Aykut Erdem
Department of Computer Engineering, Koç University, Istanbul, Turkey
E-mail: aerdem@ku.edu.tr

data being uploaded to the Internet. In addition, public and private institutions are now heavily using visual surveillance systems to constantly monitor different areas of cities and buildings thereby further adding to this surplus. Handling this much visual data and providing ways to make them easier to digest are important more than ever, which pose many challenges for computer vision researchers.

Video summarization has gained interest as a prominent research problem which aims at finding the most essential part of a video, eliminating as much redundancy as possible [26]. Video summarization approaches first split a given video into pieces in the form of video frames or video shots, and then extract features from each of these pieces. Then, they select a subset of them by considering notions such as relevance, diversity, and coherency [44]. Most recent studies cast video summarization as a supervised learning problem, and additionally learn to estimate importance of each video piece from a training set containing a number of videos with their groundtruth annotations [6,7].

A key difficulty with these generic video summarization methods lies in their evaluation [34,35,41]. Researchers commonly employ certain evaluation metrics such as F1-score and prediction accuracy to assess the summarization performance, yet these metrics are not highly correlated with the human judgments. The primary reason for this comes from the subjectivity of the summarization process in that each user has certain preferences over the importance of video pieces to be included in the summary, resulting in different summaries by different individuals that are not fully coincide with each other. Hence, designing a summarization approach that can fulfill the preference of each user is almost impossible.

Very recently, the so-called query-specific or query-focused video summarization approaches have been proposed as an attempt to alleviate the aforementioned issue [34,41,35,33,48,9]. These approaches differ from the generic video summarization techniques in one important aspect. The summarization process is carried out by considering a set of preference terms in the form of textual queries. For instance, for a video shot in a restaurant, if the input query is given as '*food and drink*', the parts of the video showing either the food or the drinks get higher importance, instead of close ups of people conversing with each other. As an example, Fig. 1 demonstrates two different summaries obtained from the same video sequence by considering two different query terms. As can be seen, each query term encodes a different concept and the extracted query-specific summary includes only the synopsis of the video relevant to the given query. Hence, this makes the evaluation process much more objective than that of generic video summarization.

Query-specific summarization has its own challenges. First and foremost, it requires a common understanding of given textual queries and the existing visual data. To succeed, the models need to form a bridge between these two different modalities and select the summary shots accordingly by integrating the information extracted from them. Using textual queries allows for a more personalized way of summarizing videos, which widens the range of its real-life applications. As the information considered important could vary from one person to another and from an application domain to a different one, one can use these query-specific summarization models for various reasons such that obtaining snapshots of important and/or interesting events in news media coverage or surveillance videos.

Most existing query-specific summarization methods are supervised approaches and heavily utilize labels associated with video parts together with groundtruth summaries in learning to summarize according to queries [34,41,35,33,48,9]. In

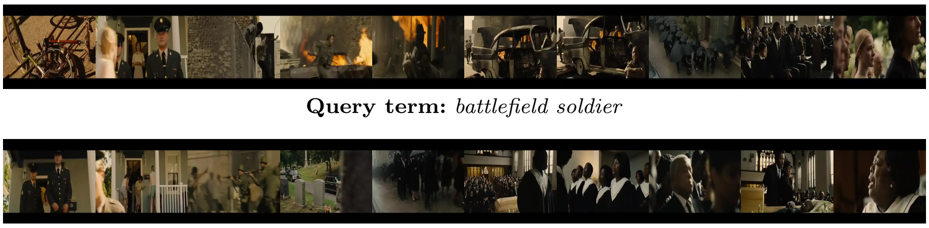

Query term: *battlefield soldier*
Query term: *church audience*

Fig. 1 Two sample human generated summaries of the same video obtained with two different textual queries. The frames included in the summaries are different from each other and mainly reflect the query terms given as the additional input.

this study, we propose a weakly-supervised video summarization approach, which does not require a large set of training videos with specific query terms and the relevant ground truth summaries. We rather assume that some pre-defined visual classifiers are available during summarization for every textual query. We use these visual classifiers to determine the relevance scores between the concepts seen in each video shot and the query terms. Moreover, utilizing these classifiers, we extract class-specific saliency maps that are then used to select the relevant image regions and visual features from video frames. We use these to define novel objective functions that encode different aspects of a good summary, in which we cast the summarization process as a submodular optimization task and employ a greedy search algorithm to select video shots for the summary in relation to given query terms. To our knowledge, we are the first to follow such a strategy for summarizing videos. There are some previous attempts to fuse semantical and visual information through class-specific saliency maps [27], yet they are too limited. They only estimate shot-level importance scores by computing the average saliency scores and then use the topmost important shots as the video summary.

In summary, our contributions can be summarized as follows:

- We propose a new query-specific video summarization approach, which leverage weak supervision in the form of semantic saliency maps obtained from predictions of pretrained attribute/action classifiers. Hence, our approach can be easily extended to novel domains without any training.
- We collect the Activity Related Summaries (ARS) dataset which consists of videos involving group and individual activities. Compared to similar datasets, our ARS dataset involves videos that are significantly much longer and thus we may utilize multiple query terms to obtain fundamentally different summaries of each video.
- We demonstrate the effectiveness of the proposed summarization approach on our ARS dataset as well as on RAD dataset [41], another query-specific video summarization dataset.

Our ARS dataset together with groundtruth attribute annotations and reference summaries are publicly available at the project website[1]. The rest of the paper is organized as follows: In Section 2, we give a brief overview of the existing

[1] https://hucvl.github.io/query-specific-summarization/.

generic and query-specific video summarization approaches along with the semantic saliency models. In Section 3, we introduce our ARS dataset by examining our data collection strategy and providing its statistics. In Section 4, we present the details of our proposed approach for weakly-supervised query-specific video summarization. In Section 5.3, we show the results of our experimental evaluation. Finally, in the last section, we offer some concluding remarks and discuss possible research directions for future work.

2 Related Work

In the following, we first provide a brief overview of generic video summarization methods (Section 2.1). We then review the existing query-specific video summarization approaches in detail (Section 2.2). Finally, we summarize the techniques used for generating semantic saliency maps as a means to interpret predictions of deep models, which we employ in our proposed video summarization framework (Section 2.3).

2.1 Generic Video Summarization

In the broadest sense, video summarization methods can be grouped into two categories based on the form of summaries they generate. While some approaches select a set of keyframes [42,4,11], the second group of studies picks shots (short video clips) [19,6,5]. At a closer look, the methods mainly differ in the strategies they use to represent video frames/clips and in the techniques they use in formalizing the selection process.

As for the features used to encode frames or clips, the early works mostly rely on low and mid-level features, such as color histograms and HOG features [19, 12], motion cues [42,20], simple spatio-temporal [13] or audio-visual features [30], as well as SIFT features [15,14] and Fisher Vectors [29]. On the other hand, the current trend is to exploit high-level features, examples of which include responses of concept detectors [34,35], features extracted from the fully-connected layers of a CNN [19,41], object-specific R-CNN features [44], detected faces [18] or persons [7].

The algorithms used for selecting the frames/shots to be included in the summary mainly formalize two basic properties of a good summary. First, the summary should include representative shots which best describes the entire video (*coverage*). Second, there should be no redundant information that the summary should be composed of a diverse set of video frames or clips (*diversity*). The simplest method one can use is to perform clustering across video shots. [2,39]. Other popular choices include casting the task as a constrained knapsack problem [6, 44], submodular optimization problem [7,41], or employ an energy based [43] or a graph-based formulation [12,19,23]. In [27], a weakly supervised approach is proposed which utilizes video-level labels to learn interestingness scores. It selects the shots which maximum this score internally.

More recent works additionally employ supervised learning strategies for better incorporating high-level semantics into the summarization process. Examples include using deep pretrained CNN features [1] , semantic category knowledge [14, 11,15,29] or attributes [39], interestingness [6] or visual saliency [28] to encode video content. Another interesting recent research direction is to employ audio-visual features [22]. The curious reader can refer to [40] for a more detailed and recent survey.

Due to the subjective nature of the problem, evaluating video summaries is not so straightforward and, indeed, considered as an open research question [25, 10]. In addition to qualitative comparisons, the most common way is to calculate precision, recall and F1-score by measuring the agreement between the generated summary and the ground truth one. The matching between the two summaries can be computed by maximum weight bipartite matching [2,11] where a bipartite graph is formed by considering the summaries as the two opposite ends of the graph and edge weights as the similarity between the frames/shots. Another quantitative metric is VERT [16], which is motivated by the BLEU and Rouge metrics used in machine translation and text summarization, respectively. It depends on counting the number of overlaps between the reference and generated summaries. If there exist multiple reference summaries, once could report either the maximum or the average scores over all reference summaries.

2.2 Query-Specific Video Summarization

As we mentioned before, inherent subjectivity of the generic summarization task led researchers to propose novel summarization problems. For instance, in [29], the authors proposed the task of class-specific summarization that requires learning to summarize a video in regard to its semantic category. In particular, to solve this problem, the authors trained a separate linear SVM classifier for each visual category by taking into consideration a set of training sequences with the user-annotated groundtruth summaries.

Following the aforementioned study, other researchers extended the class-specific summarization and introduced the so-called query-specific summarization as a new summarization, which takes into account additional textual queries to guide the summarization process [34,41,35,33,48,9]. This eliminates the need for training separate summarization models for each category. These textual queries contain one or more keywords and allow the users and consequently the developed methods to focus the video frames relevant to the queries more, which results in a more objective evaluation of the summarization models.

Sharghi et al. developed a probabilistic formulation for query-specific video summarization called Sequential and Hierarchical Determinantal Point Process (SH-DPP) [34]. The proposed algorithm selects a set of key frames from the video by jointly examining each frame in terms of its importance in the context of the whole video and its relevance to the user query. To accomplish this, the authors extended the generic seqDPP [5] framework by introducing a two-layer hierarchy to simultaneously account for the diversity and the relevance of the selected frames. In [35], the authors improved upon their early work by developing a memory-network based personalized summarization model, which allows for joint use of attention schemes defined over user queries and diversity modeling via determinantal point processes. The model in [33] further extends these approaches by additionally accounting for the expected length of the summary and alleviating the exposure bias problem inherent to sequential prediction.

Vasudevan et al. developed another query-specific summarization framework that depends on a linear combination of submodular objective functions [41]. The model learns to summarize a video in regard to a given user query by considering multiple objective functions, encoding the relevance of frames to query terms, the quality and the representativeness of frames and the diversity of the summary elements. Optimal weights of these submodular functions are estimated using a training set of videos by applying an adaptive gradient algorithm.

Zhang et al. introduced a similar kind of personalized summarization method that is based on a conditioned three-player generative adversarial networks model with the user queries being the condition [48]. While the generator tries to fool the discriminator by generating accurate summaries of a given video with the use a common representation of the textual queries and the video, the discriminator tries to discriminate the actual summaries from the generated or the randomly-sampled summaries conditioned on the query terms.

Recently, in [9], Jiang and Han proposed the so-called Hierarchical Variational Network (HVN) model for query-specific summarization. This model learns to select the user diversified summary frames from a given video by taking into account long-range temporal dependencies between the frames, using attention mechanisms to encode the content of the video and the user query and utilizing a multilevel self-attention module along with a variational autoencoder.

The evaluation metrics mentioned in the previous subsection can also be used for evaluating query-specific video summarization methods. However, this requires the availability of reference summaries collected for certain query term(s). In the absence of reference summaries, one can also exploit extra annotations showing the relevance scores of each frame/shot to the existing set of queries. In this case, precision can be predicted as the mean average relevance score whereas the recall corresponds to how many of the relevant clusters are selected [41]. Here, F1-score is calculated as the harmonic mean of the relevance and cluster recall.

As stated in the introduction, the aforementioned query-specific summarization methods are supervised methods in that they are trained over a set of training video sequences by using the query terms associated with video frames and the groundtruth summaries. On the other hand, in our work, we follow a weakly-supervised setting in that we only require a set of pre-defined classifiers for the query terms used during summarization, which allows us to jointly extract the frame-level relevance scores and pixel-level semantic saliency maps.

2.3 Semantic Attention Maps

Explaining the decisions made by deep convolutional neural networks (CNNs) is hard to break down. With too many parameters and nonlinear transformations between the layers, these models are commonly considered as complex black box models. Hence, investigating ways to interpret neural model predictions has become an interesting research topic in its own right.

One of the earliest attempts to interpret CNNs is by Simonyan et al. [37], in which the authors proposed a method to visualize the image regions that provide the most discriminative information about a given class in terms of a class-specific saliency map. In particular, the approach depends on computing the gradient of the class score with respect to the input image. Later, Zeiler and Fergus introduced the

Deconvnet [46] and Springenberg et al. proposed the Guided Backpropagation [38], both of which greatly reduce the noise in the extracted saliency maps and improve their quality by modifying the backpropagating gradients. These techniques give more intuitive visualization regarding the image pixels discriminative with respect to the given class.

Zhang et al. proposed another class-specific image saliency model, referred to as Excitation Backprop [47], which extracts the discriminative image pixels with respect to a given class by taking into consideration a probabilistic winner-takes-all process while modifying the backward signals from the top-most layer. Similarly, Shrikumar et al. suggested the DeepLIFT model [36] that backpropagates the contributions of neurons downwards in the network hierarchy, where each neuron's contribution is estimated by examing the difference between its activation and the neuron's activation on a reference image.

In [49], Zhou et al. combined information from the activations and the gradients of the neurons to extract the image-specific class saliency maps of an image again to understand the decisions of a given network. Their proposed model, which is called Class Activation Mapping (CAM), mainly replaces the intermediate fully-connected layers of a CNN model with convolutional layers and a global average pooling layer, and lets the modified network identify the most informative neurons while predicting a target class, which are then used to estimate a saliency map via weighted averaging of the activations of these neurons.

The Grad-CAM model by Selvaraju et al. [31] can be regarded as a generalization of CAM, which eliminates the need for modifying the CNN model in consideration. It allows for extracting class-specific saliency maps highlighting the informative image directly by inspecting the backprogating gradients of the target class from the final convolutional layer downwards in the network hierarchy. In that sense, it does not require a separate training phase as done in the CAM approach.

In our study, we specifically employ the Grad-CAM model [31] to identify the discriminative image regions with respect to the given query terms by using the pre-trained classifiers. In particular, we use these extracted semantic saliency maps to select the most informative features from the video frames for the query-specific summary. The details are further described in Section 4.

3 Activity Related Summaries Dataset

In this section, we describe our dataset, which we refer to as Activity Related Summaries (ARS) dataset. But before presenting our dataset, we first briefly review existing datasets for query-specific video summarization (Section 3.1). In particular, here we explain how we obtain videos from Youtube (Section 3.2), how we extract semantic attributes (Section 3.3), how we accordingly determine queries (Section 3.4) and how we collect groundtruth summaries (Section 3.5).

3.1 Existing Query-Specific Video Summarization Datasets

There has not been many datasets in the literature for query specific summarization. In the following, we summarize the existing datasets where we highlight their

Table 1 Comparison of datasets used for query-specific summarization. Our Activity Related Summaries dataset mainly differs from the existing ones in terms of how we collect annotations, which affect the query terms and the domain of the videos.

Dataset	Video count	Query count	Length	Annotation	Domain	Unit
RAD [41]	200	200	1-3 min	Relevance	Mixed	Frame
UT Egocentric [14]	4	2415	3-5 hrs	Textual	Egocentric	Shot
TV Episodes [45]	4	1275	45 min	Textual	Episodes	Frame
Activity Related Summaries	10	142	2-10 min	Attribute	Crowds	Shot

properties and key features in comparison with our proposed ARS dataset. Table 1 provides some statistics about these datasets.

RAD dataset [41] consists of 200 videos which are retrieved from Youtube by querying multiple words like *'basketball fights'*. The videos are relatively short (1-3 minutes) and sampled as 1 frame per second. Each frame is annotated with a score between 0-3, showing relevance to the text query. The dataset is not originally created for summarization so it does not contain ground truth summaries. Hence, the evaluation of automatic summaries is not based on intersection or union with human generated summaries, but instead, a summary which selects more related frames gains higher scores.

UT Egocentric dataset [14] contains four videos, each lasting 3-5 hours. The authors first identified hand-crafted 46 two-word queries for each video and asked subjects to tag each frame with each concept as either 1 or 0. In fact, they also eliminated redundant concepts like *'area'*, which has a diverse and ambiguous meaning and does not have much use for summarization purposes. Another key difference is its domain – as its name suggests the dataset consists of egocentric videos which are recorded in free environments, i.e. under no controlled manner or intention. Hence, a big portion of the videos are irrelevant to the queries. Videos are partitioned to 5 second-long shots and evaluation is based on textual annotations.

The TV Episodes dataset [45] consist of four videos, each approximately 45 minutes long. As its name suggests, these videos are episodes of some popular TV shows, hence they involve a third person perspective. The corresponding scenes are professionally shot and thus more related to the main topic. The evaluation is again based on textual annotations.

3.2 Data Collection

In order to collect our video data we performed YouTube searches, like done in the related studies. In doing so, however, we focus on crowd videos as they can be annotated with a rich set of attributes about the observed activities, actors and events [32]. Here we use complex query terms such as *'fight between fans' and 'running and swimming'*. In total, we collected 10 videos, including music and movie clips with 720 × 1280 resolution. The duration of the videos changes between 2-12 minutes. More specifically, there is a video log about a triathlon race, which contains running, swimming and biking activities. There exist three music videos, mainly showing cheering fans in stadiums or streets. There are three movie clips, which respectively show a battlefield scene with a ceremony in a church, dancers in a car and in a club, and finally an orchestra band on a stage with full of dancing

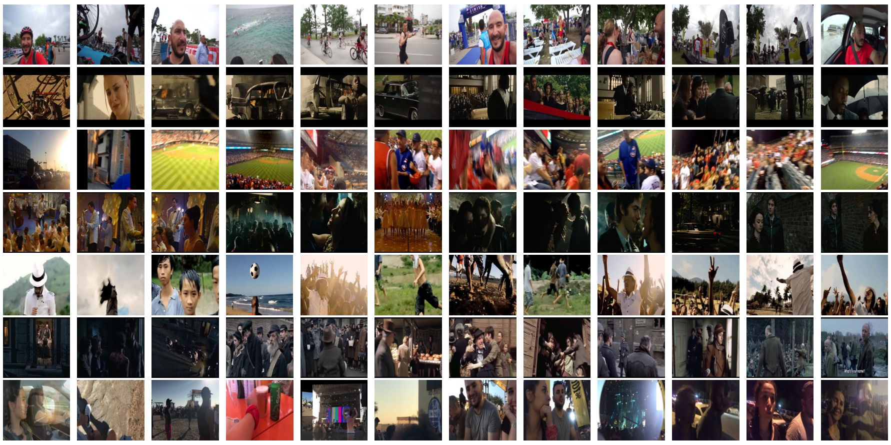


Fig. 2 Example frames from our Activity Related Summaries dataset. Attributes include activities like *war, wave, fight, dance* and places such as *queue, stadium* and actors like *dancer, artist, soldier*.

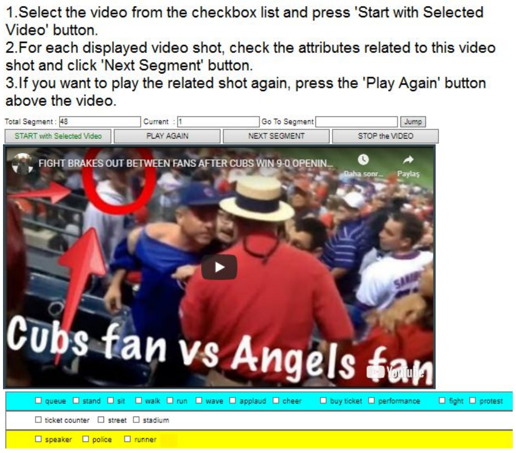
people in front of them. Lastly, there is a rock festival video, which contains singing, resting on a beach and playing activities. Fig. 2 shows some sample frames from our dataset, which highlight the diversity that exists in the video data.

3.3 Tagging of Video Shots with Semantic Attributes

As the videos in our dataset are not very short, we decided to annotate them on shot-level. In order to obtain visually coherent shots, we basically look for shot boundaries that correspond to a change in either semantic content of in visual flow. We determine these shots using the approach of Potapov et al. [29], which utilizes GIST descriptors extracted from each frame [24]. In the end, each shot we extract contains approximately 75 frames on average. To decide the attributes that exist in video shots, we conducted a preliminary user study. We prepared a web-based tool to tag pre-defined shots with attributes that exist in the video sequence. As shown in Fig. 3, a shot is shown to the user and the user selects the relevant attributes observed in the shot. Here we use the exact taxonomy proposed in [32], where the attributes are grouped as activities (*what*), places (*where*), and actors (*who*). In total 10 users participated in our tag collection efforts. To determine the final list of attributes, we take the attribute as tagged if one or more users marked it, i.e. we simply take the union of all user responses. In the end, each video contains approximately 20 attributes and there exist approximately 60 unique attributes in total in all videos (Fig. 4).

3.4 Determining The Queries

Inspired by Vasudevan et al. [41], we selected the attribute pairs for query terms according to four main criteria with different factors: (i) attributes seen always



1.Select the video from the checkbox list and press 'Start with Selected Video' button.
2.For each displayed video shot, check the attributes related to this video shot and click 'Next Segment' button.
3.If you want to play the related shot again, press the 'Play Again' button above the video.

Total Segment : 48 Current : 1 Go To Segment

□ queue □ stand □ sit □ walk □ run □ wave □ applaud □ cheer □ buy ticket □ performance □ fight □ protest

□ ticket counter □ street □ stadium

□ speaker □ police □ runner

Fig. 3 A screenshot of our web interface used in annotating video shots. The attributes correspond to activities, places and actors from [32], and are shown with different colors for clarity.

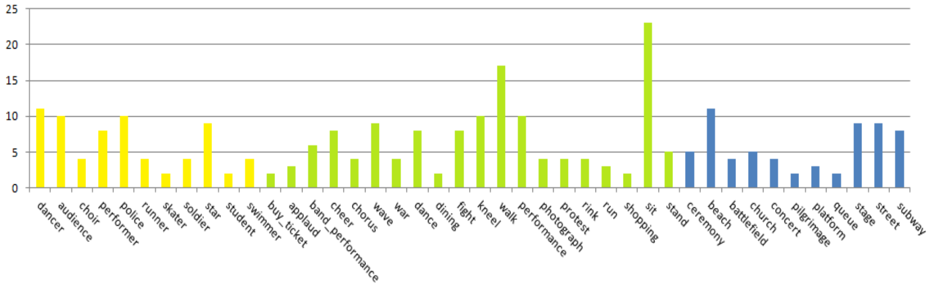
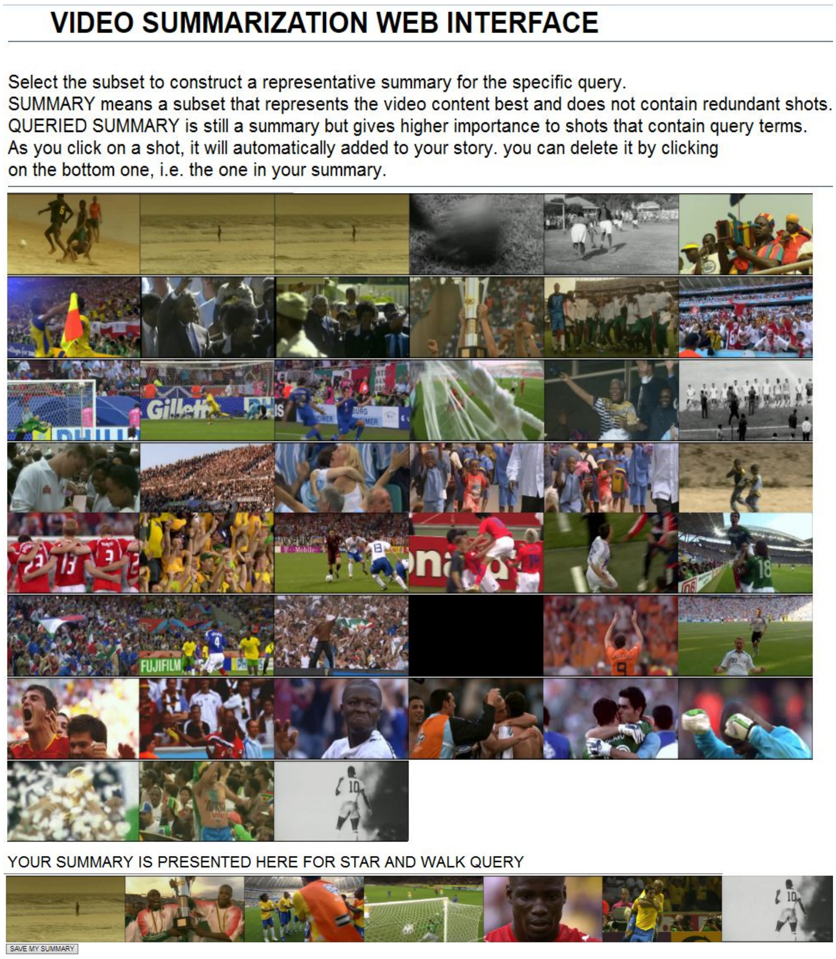


Fig. 4 The frequency of each attribute appearing in more then one query. The colors denote attribute categories of actors (in yellow), activities (in green) and places (in blue).

together in video shots (ii) attributes seen completely separately in video shots (iii) only one of the attributes appear in the video (iv) in some shots the two attributes intersect and jointly appear but not always. The fifth scenario is the non-queried one, i.e. none of the attributes appear in the video and this means the summary is blind to the query. To illustrate our queries, we give one example for each of these scenarios for the Triathlon sequence. The first example is the pair *'pedestrian+stand'* as these two attributes are always observed together. For the second case, it is the pair *'street+swim'* since they could never be in the same shot. An example to the third scenario is *'bazaar+photographer'*, in which the latter exists in the video whilst the former does not. Actually, this corresponds to a one word query. An example to the fourth setting is the type of query we are most interested in investigating. The example is the query *'street+run'*. In the video, both of the attributes exist, in some parts they mutually exist and for some others, they exist in isolation. That is, the runners sometimes run in the streets.



VIDEO SUMMARIZATION WEB INTERFACE

Select the subset to construct a representative summary for the specific query.
SUMMARY means a subset that represents the video content best and does not contain redundant shots.
QUERIED SUMMARY is still a summary but gives higher importance to shots that contain query terms.
As you click on a shot, it will automatically added to your story. you can delete it by clicking
on the bottom one, i.e. the one in your summary.

YOUR SUMMARY IS PRESENTED HERE FOR STAR AND WALK QUERY

Fig. 5 A screenshot of the web interface used in obtaining ground truth query-specific summaries from human subjects. For a specific query for a video, the user selects fixed number of relevant shots as the summary. A sample human generated summary is shown at the bottom row.

3.5 Obtaining The Ground Truth Summaries

In the last stage, we collected human-generated summaries in a query-specific manner. For that purpose, we again designed a web-based user interface, as shown in Fig. 5. On average, we present a user 10 different queries for each video sequence. These queries are either 2-word or single-word queries. When the user selects a query, all of the shots extracted from the video are displayed as animated GIFs and the user is asked to select the shots that constructs the related summary for the given query within a budget constraint of either 7 or 10 shots. In total, we collected ground truth summaries from 5 different users.

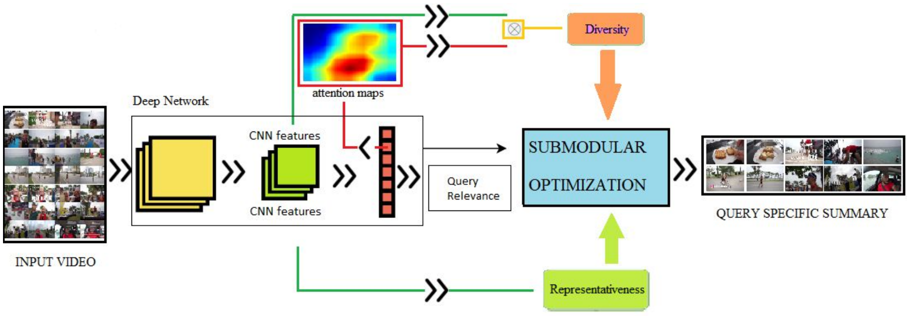


Fig. 6 A system overview of the proposed summarization approach. Given an input video, we first divide it into shots. Then we extract both deep features and high-level semantics of these frames by using a pre-trained deep neural network. We then predict query specific spatial attention maps from each video frame within a shot. We define multiple objectives based on these extracted deep features, high-level semantics and attention maps and determine the video summary using submodular optimization.

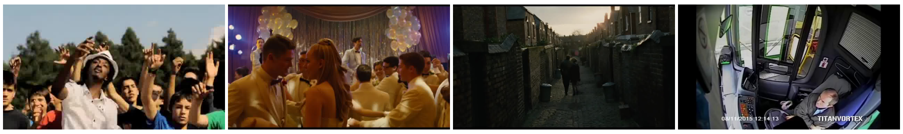
4 Approach

Our query-focused summarization approach depends on submodular optimization [17]. In particular, we cast the summarization process as a subset selection problem where a given video is decomposed into several shots at first and then a number of these shots are selected to form a summary in regard to the input query terms. As mentioned in the Related Work section, selection of these shots is based on several factors. First and foremost, their relevance to the given query terms plays a key role. However, aside from that, diversity and coverage of the selected shots are also important as the extracted personalized summary should be as descriptive as possible while devoid of any redundant information.

In our work, to decide the relevancy of a shot to a given query term, we use a weakly-supervised approach. Specifically, we extract high-level semantics of every shot by using pre-trained visual classifiers of which some of them are related to the given queries (Sec. 4.1). This also allows us to infer query-specific attention maps demonstrating main areas of interest in a spatial manner in regard to the extracted concepts (Sec. 4.2). Then, we employ these semantic features and attention maps during selecting the most *diverse* and the most *representative* set of shots that are at the same time relevant to the given query terms from the whole pool of video shots via an optimization process (Sec. 4.3). This whole process is demonstrated in Fig. 6.

4.1 Predicting High-Level Semantics

To create query-specific summaries, we apply a top-down strategy and discover how relevant a video shot is in regard to the concepts exist in the query by employing some visual classifiers. As will be detailed in the next subsection, this also helps us to select the features related to the query terms via the use of query-specific attention maps.



cheer(0.73) stand(0.89) *dance (0.72) audience(0.61)* *street (0.79) walk(0.63)* *crash (0.62)*

Fig. 7 Sample concepts and their relevancy scores that are extracted by using some deep visual classifiers. As can be seen, the state-of-the-art classifiers are fairly good at estimating high level semantics of the video content even though they were trained on some benchmark datasets.

In our work, we assume that for every query term we have a separate visual classifier that has been pre-trained on some external data and that can be used to detect the concept mentioned in the query. These classifiers could be attribute classifiers or event/activity classifiers, which are selected from a pool that is mainly related to the domain of the videos considered in the summarization process.

In particular, we feed each frame x_i of a video shot to a deep classifier network and apply average pooling along the temporal dimension to obtain the shot representation as well as take the average of class responses. By this way, we not only extract the convolutional features $f(x)$ from the given video frame but also assign a relevancy score $r(x, q_c)$ to the input video shot in regard to a given query term q_c. Fig. 7 shows some illustrative results for the high-level semantics extracted by some visual classifiers on videos from different domains. As these results demonstrate, deep classifiers trained on large-scale benchmark datasets detect the concepts seen in these video shots in a satisfactory manner. We use the extracted deep features and the estimated relevancy scores in our optimization framework, as will described be described in Sec. 4.3.

4.2 Extracting Query-Specific Attention Maps

Using a deep classifier while estimating the relevancy of a video shot to a query term also enables us to extract query-specific attention maps. In particular, with these attention maps, we aim at capturing the most informative image regions in regard to a given query term. Extracting such kind of maps are of interest to our framework as they can be used to perform feature selection. Letting the features encode the relevant (spatial) parts of the video shots while suppressing the background and ignoring the irrelevant parts, results in more robust representations.

In order to obtain the query-specific attention maps, we propose to use Gradient-weighted Class Activation Mapping (Grad-CAM) technique [31]. As mentioned earlier in the related work section, Grad-CAM does not require a separate training procedure. For a given image, it highlights the image regions that are deemed as important by a pre-trained convolutional neural network model while giving its predictions by accumulating the activations in the convolutional layers in a weighted manner.

Assuming that the final convolutional features $f(x) = [f_1(x); f_2(x); \ldots f_K(x)]$ have been extracted via a classifier network (K denoting the number of channels),

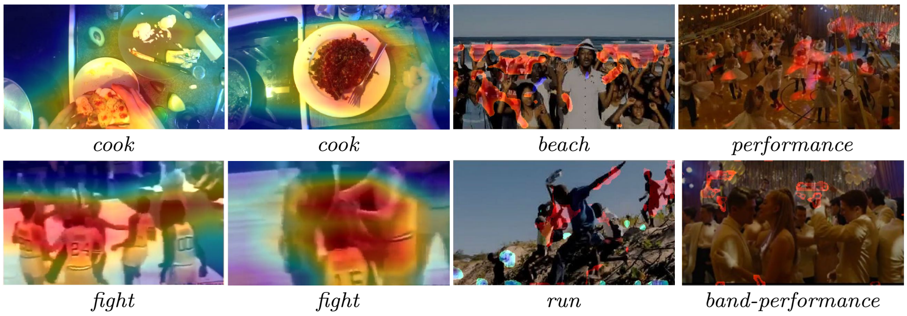


Fig. 8 Sample attention maps extracted from the images by considering some sample concepts. As can be seen, these maps highlights the image regions that are highly relevant to the given query terms.

the attention map $a_c(x)$ for a query term q_c is estimated as follows:

$$a_c(x) = ReLU\left(\sum_{k=1}^{K} \alpha_k^c f_k(x)\right) \tag{1}$$

with α_k^c representing the neuron importance weights for the given query term q_c, which are estimated by performing global-average-pooling to the gradients flowing back to the final convolutional layer of the deep classifier. Since ReLU is used as an activation function, the extracted attention map only highlights the pixels that have a positive contribution on the given visual concept, which makes them suitable for feature selection. That is, the extracted attention map $a_c(x)$ can be used as a soft mask to obtain a more robust feature representation $f'(x)$ as follows:

$$f'(x) = a_c(x) \odot f(x) \tag{2}$$

with $\odot$ denoting the Hadamard product operation. In case there are more than one query terms, we estimate these attention-masked deep features separately for each given query term and concatenate these features to obtain the final query-specific representation of a video shot. In our summarization framework, we use these representations while estimating the diversity of the selected frames, as will described in the next subsection.

In Fig. 8, we show some attention maps that are extracted from sample images considering a number of concepts used as the query terms. As one can see, these maps clearly point out the image regions of interest by providing fairly good localization of the given concept terms.

4.3 Query-Focused Summarization

In order to construct a query-specific summary, we cast summarization as an optimization problem where the shots forming the summary are selected according to an objective function. Formally, let $\mathcal{Y}$ denote the set of all possible solutions $\mathbf{y}$ with a pre-defined cardinality $|\mathbf{y}| = k$ encoding the specified summary length, and $\mathbf{x} = \{x_i\}_{i=1}^{T}$ be the set of mean frames of the video shots with i indicating a shot index and T denoting the total number of shots in the video. Moreover, assume that $\mathbf{f} = \{f(x_i)\}_{i=1:T}$ represent the video features extracted by a deep classifier ψ and $\mathbf{f}'$ denote the attention-masked counterparts of these video features in regard to given query terms. As given below, our objective function consists of three different submodular functions, each capturing a different quality of the extracted summary:

$$\mathbf{y}^* = arg \max_{\mathbf{y} \in \mathcal{Y}} \big(\lambda_1 \phi_{rep}(\mathbf{f}, \mathbf{y}) + \lambda_2 \phi_{div}(\mathbf{f}', \mathbf{y}) + \lambda_3 \phi_{rel}(\mathbf{y}) \big) \tag{3}$$

These qualities are representativeness as encoded by ϕ_{rep}, diversity as encoded by ϕ_{div}, and query relevance as specified by ϕ_{rel}, respectively. In the following, we explain how these qualities contribute to the overall summarization process and how the related submodular functions are defined. In our experiments, we set the weight of each submodular function to 1, assuming that the aforementioned qualities are equally important for a good summary.

Representativeness. Main contribution of this function is to let the extracted summary include the shots representing the input video as a whole. In our formulation, we employ the formulation suggested in [7,41] and measure the representativeness of the selected video shots according to how close they are to the medoids of the shots of the given video as follows:

$$\phi_{rep}(\mathbf{f}, \mathbf{y}) = \sum_{x_i \in \mathbf{x}} \min_{s \in \mathbf{y}} |f(x_i) - f(s)|_2^2 \tag{4}$$

where $f(x)$ represents the mean convolutional features extracted from the frames of the shot x. In particular, these features encode the global characteristics of the video frames, and make the optimization process to favor the shots that look like the best k video shots visually representing the whole video. In that regard, this submodular function does use only the visual information and does not take into account the input query terms.

Diversity. This submodular function is used to ensure that the summary include a set of video shots as diverse as possible. That is, if the summary covers a specific event and/or objects, there is no need to include the video shots to the summary demonstrating similar kind of events and including similar objects. This is achieved by formulating the submodular function for diversity as follows:

$$\phi_{div}(\mathbf{f}', \mathbf{y}) = \sum_{i=1}^{k} \min_{j<i} \left| f'(y_i) - f'(y_j) \right|_2^2 \tag{5}$$

where $f'(y)$ represents the convolutional features encoding the video shot y, which are masked using the query-specific attention maps. In that regard, this formulation integrates the visual information with the context information specified by the given query terms, resulting in a query-dependent formulation of diversity. In particular, diversity is enforced by taking into account the overall distance between the attention-masked features of the video shots in the summary.

Relevance. The aforementioned two qualities, representativeness and diversity, do not explicitly favor including the video shots that are relevant to the given queries in the summary. Hence, we define a third submodular function to achieve this goal. In our formulation, as mentioned before, we consider a weakly-supervised setting in which we assume that a number of visual classifiers are available with each classifier ψ_c returning a probability value of observing a query term c in a given video frame. Since more than once query terms might have been specified for a personalized summary, for such cases, the relevance of a video shot in regard to these query terms is defined as the maximum of those probabilities. Hence, we defined the submodular function for relevance as follows:

$$\phi_{rel}(\mathbf{y}) = \textstyle\sum_{i=1}^{k} \max_{c} \psi_c(y_i) \tag{6}$$

This extra term on relevance encourages to obtain a query-specific summary while the remaining terms respectively enforce the selected shots to represent the whole video and to illustrate diverse events and/or objects as much as possible.

5 Experiments

We perform our experiments on two datasets, our proposed Activity Related Summaries dataset and the RAD dataset [41], respectively. In the following, we provide the details of these experiments. In particular, we first discuss the baseline models considered in our evaluation. Then, for each dataset we describe the experimental setup and present the results of our experiments[2].

5.1 Baseline Approaches

To evaluate the effectiveness of our approach, we compare it against several baseline methods from the literature that have been used for video summarization. In particular, we first use uniform sampling and k-means clustering approaches as weak baselines. These approaches do not take the given query terms into account and just process the given video content, giving generic video summaries. As a strong baseline, in our experiments, we also report the results of a query-specific summarization approach proposed by Vasudevan et al. [41]. This approach, like our method, employs a submodular optimization framework to obtain the summary, however, in contrast to our approach, its objective function lacks an attention-guided diversity term. For all these approaches, we select a fixed and prescribed number of shots as the summary of the input video. In the following, we give the details of these evaluated baseline models.

Uniform Sampling: Most straightforward way to summarize a video is to select shots or frames uniformly. As we work on shot-level summaries, in our first baseline, we select a fixed set of shot in a uniform manner.

k-means Clustering: Another intuitive method in unsupervised video summarization is to perform clustering with k-means. Here, we extract visual features from the frames by using a pre-trained CNN model, and represent each shot by

[2] For additional qualitative results, please refer to the project website at `https://hucvl.github.io/query-specific-summarization`.

the average of these features. Then, we cluster the shots into a set of clusters by k-means method. In our implementation, we employ the squared Euclidean distance metric and initialize the cluster centers by randomly sampling from the data samples. Finally, we generate the summary by selecting the shots closest to each cluster center.

Vis-DSS: This method by Iyer et al [8] is another generic summarization method which employs a subset selection method based on submodular functions for photo album summarization. In particular, we represent each shot with its center frame, and use the deep features extracted from this center frame while selecting the summary shots.

Vasudevan et al. [41]: As a query-specific summarization approach, this method is also based on a submodular optimization framework. It differs from our proposed approach in several aspects. For instance, relevancy between the query text and the video shots is estimated by first learning a common semantic embedding space and then by using the cross-modal distances in this space. Moreover, the weights of the considered submodular functions are learned from ground truth data using simple hyper-parameter search. Of course, our main difference lies in the estimation of diversity term, where we used semantic attention maps.

Panda et al. [27]: As a second strong baseline, this method also exploits weak supervision by utilizing attention maps. This method mainly considers saliency maps extracted by the query terms to calculate importance scores for video shots, and then sorts these shots according to these scores. Finally, it uses the shots having the highest importance scores to form the summary. In that sense, it can be considered as a weakly supervised summarization method, yet it does not take into account representatives or diversity of the video shots in the summary as we do in our submodular optimization scheme.

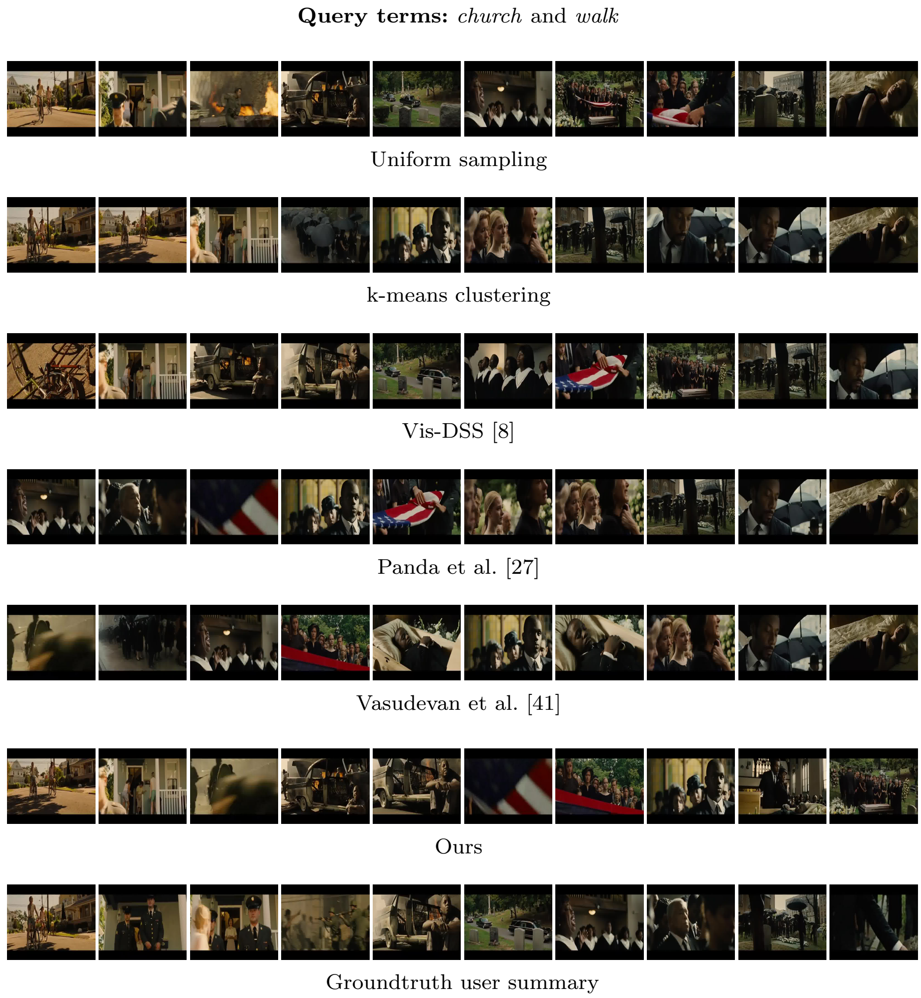
5.2 Experiments on Activity Related Summaries Dataset

Dataset Details and Evaluation Protocol

As mentioned before, the dataset contains summaries for 142 queries from 10 videos. For the quantitative evaluation, we calculate F1-score by comparing summaries with user summaries by maximum weight matching of a bipartite graph, as described in [33]. In this way, not only selecting the exact shots is considered, but also selecting semantically close shot is rewarded. The summaries cover 10-25 percent of the whole duration of the videos. The number of shots varies from 38 to 82, so we decided the total number of shots selected for the summaries as 10 for long videos and 7 for the short ones. This means the total duration of the obtained summaries may be different from one video to another.

Implementation Details.

For our approach, we applied fine-tuning to WWW dataset [32] to compute semantic saliency maps by means of GRADCAM [31] accordingly. Our network architecture is the two-stream model of Shao et al. [32] proposed for attribute prediction. To this network, the inputs are the original frame and the optical flow image denoting motion between two consecutive frames. Since we work on crowd



Query terms: *church* and *walk*

Uniform sampling

k-means clustering

Vis-DSS [8]

Panda et al. [27]

Vasudevan et al. [41]

Ours

Groundtruth user summary

Fig. 9 A sample summary generated by our method along with the reference summary and the summaries obtained with the baseline methods. Our approach selects shots relevant to the input query and overall they are visually more similar to the shots in the groundtruth summary.

videos, we decided to compute optical flow between two consecutive frames to capture activity information better. With motion information, we can infer that if there are two groups getting close to each other, this may refer to a fight or a dance scene. Alternatively, if the motion is small, we can differentiate between the actions of walking and talking. Optical flow images are computed by means of Flownet [3] method. After that, we compute the gradient of the classifier responses in regard to a given textual query with respect to the last convolutional layer (conv5 layer). From these gradients, we obtain the importance of feature maps and extract the final semantic saliency maps accordingly via the weighted combination of class-specific saliency maps.

Table 2 Performance comparison of summarization methods models on our Activity Related Summaries dataset with the best performing model highlighted with a bold typeface. The mean and the standard deviation of the evaluation measures are reported.

	Precision	Recall	F1-score	VERT
Uniform	0.5776 ± 0.1035	0.5986 ± 0.0924	0.5963 ± 0.0938	0.4235 ± 0.1685
K-means	0.5891 ± 0.1129	0.6036 ± 0.1217	0.5935 ± 0.1136	0.4614 ± 0.1287
Vis-DSS [8]	0.6012 ± 0.1779	0.5616 ± 0.1888	0.5947 ± 0.1876	0.4894 ± 0.1969
Panda et al. [27]	**0.6117** ± 0.1619	0.5581 ± 0.1525	0.5814 ± 0.1343	**0.5016** ± 0.1397
Vasudevan et al.[41]	0.5975 ± 0.1332	0.6277 ± 0.1204	0.6178 ± 0.1255	0.4774 ± 0.1441
Ours	0.6087 ± 0.1121	**0.6335** ± 0.1177	**0.6307** ± **0.1149**	0.4942 ± 0.1341

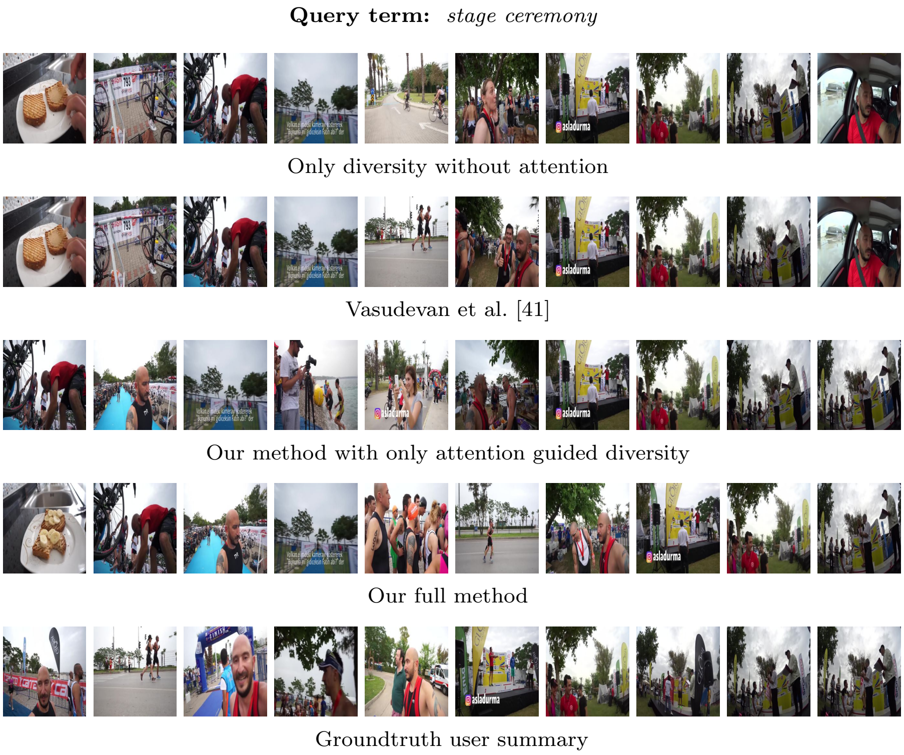
Experimental Results and Discussion.

In Fig. 9, we show a visual comparison of our method against the competing approaches. As can be seen, other methods generally focus on only one aspect at a time. For instance, K-means and Vis-DSS select the representative shots from the given input video but ignores the query terms, giving some shots irrelevant to the query. The approach by Panda et al. [27] selects semantically similar frames as it considers query relevance while constructing the summary. Our method, on the other hand, gives a summary very close to the groundtruth user summary by choosing semantically query-relevant shots that represent the video best. These observations are also validated by the quantitative analysis given in Table 2. Our method outperforms all other methods in terms of F1-score and recall, and gives the second best performance according to precision and VERT, showing the importance of using high-level semantics and query-specific attention maps all together in summarization. Note that the weakly-supervised method by Panda et al. [27], which only uses query-specific attention maps for summarization, performs well in terms of precision, however its recall is among the worst. The same comment can be done for VERT scores as it is also a precision-based metric. Hence, F1-score reflects a better trade-off between precision and recall.

We performed a second set of experiments to analyze the effect of objective functions in isolation and the contribution of attention guided diversity and presented it visually in Fig. 10. We see that if only relevance term is considered in the optimization, our summarization approach already yields a good performance as it encourages to catch semantically close shots. Considering only relevance term, however, is not solely enough as a summary should contain a recap of the input video without any repetition of the same content, attention-guided diversity improves the F1-score compared to the results when it is not taken into account. Table 3 shows the individual effects of submodular functions when utilized in different combinations. Recall that utilizing all the three subfunctions without attention is similar to the setting proposed in [41].

Table 3 Contribution of submodular functions considered in our framework in Activity Related Summaries dataset. 'AG' denotes attention-guidance. We report the mean and the standard deviation of F1-scores.

Representativeness	Diversity	Relevance	F1-score
+			0.5911 ± 0.1149
	+		0.5925 ± 0.1117
		+	0.6105 ± 0.1301
+	+		0.6040 ± 0.1277
+	+	+	0.6178 ± 0.1255
	+ AG		0.5978 ± 0.1213
+	+ AG		0.6122 ± 0.1254
+	+ AG	+	**0.6307** ± 0.1149


Query term: *stage ceremony*

Only diversity without attention

Vasudevan et al. [41]

Our method with only attention guided diversity

Our full method

Groundtruth user summary

Fig. 10 A visual analysis of how the components of the proposed approach affect the extracted summaries. Overall, the full method accurately captures the essence of the query term while selecting shots visually close to the ones in the reference summary.

5.3 Experiments on RAD dataset

Dataset Details and Evaluation Protocol.

We selected 14 videos from RAD dataset which are mainly action categories and most related to sample trained model [21]. Sample frames of the videos that we

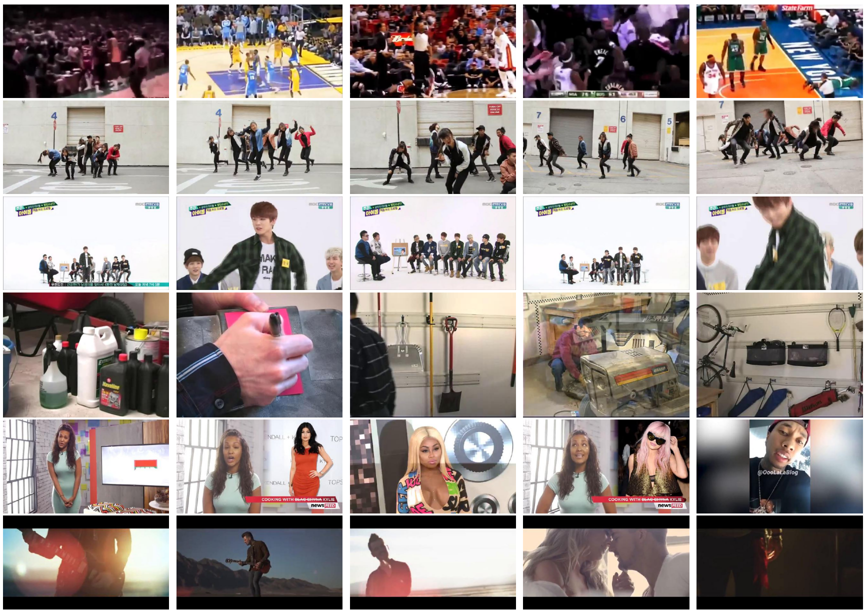


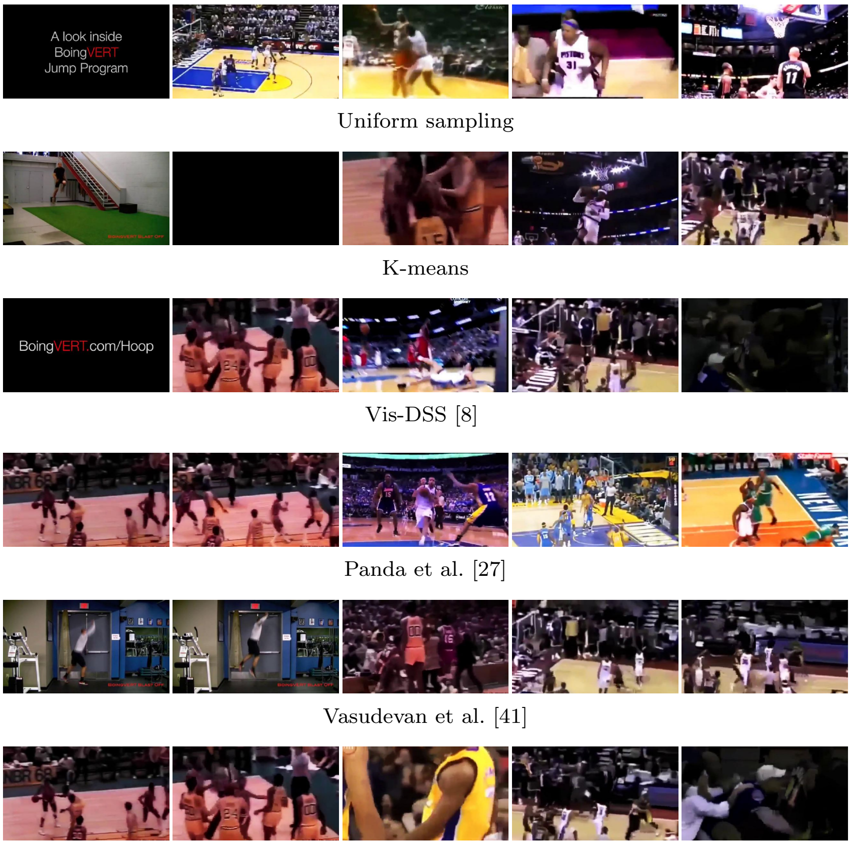
Fig. 11 Example sequences from RAD dataset proposed in [41].

selected from RAD dataset are shown in Fig. 11. Originally, RAD dataset did not aim to measure summary accuracy, so the total relevance of the selected summary is calculated to measure the precision. As explained in [41], the recall is calculated by dividing the number of unique clusters in the selected summary by summary length. In this way, we measure diversity of the summary by expecting to cover more unique clusters. To compare our results to that of Vasudevan et al.'s [41] and previous methods, we fixed our summary length by 5. In test videos, frames are sampled as 1 frame per second. From those sampled set, we aimed to select 5 frames for the summary.

Implementation Details.

For this dataset, we did not apply fine-tuning for attribute predictions, and computed class activation maps and attribute predictions via the sample Moments in Time model offered by Monfort et al. [21]. This model is a product of a large-scale dataset proposed to recognize event and activities in videos. The neural network model is originally based on Resnet50 and was trained with 1M videos from 339 action categories. These categories are generally actions. For a few of them, like barbecuing and drumming, we can infer objects like barbecue and drums. We selected test videos by string matching of video names with targeted actions from Moments in Time dataset [21]. For example, a video name 'bus crash' matches with the '*crash*' action. We matched every video to a query, i.e. crash query for 'bus crash' video. Of course, not all videos we selected are crowd videos and some of them has a few parts containing target actions. This brings a diversity of video genres and allows us to experiment on a variety of video types. We extracted

Query term: *fighting*



Uniform sampling

K-means

Vis-DSS [8]

Panda et al. [27]

Vasudevan et al. [41]

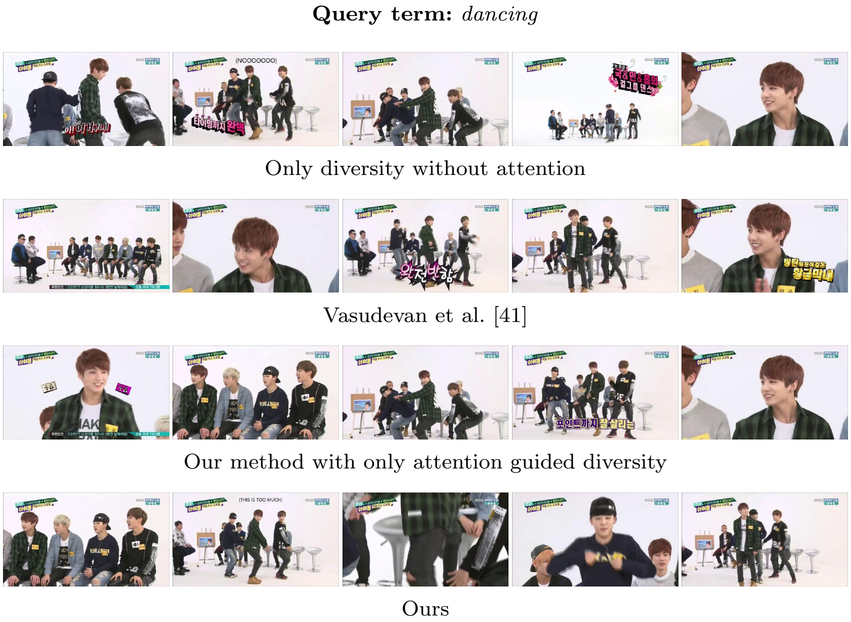
Ours

Fig. 12 Sample summarization results on a sequence from the RAD dataset. Compared to the baseline methods, the summary generated by our approach contains more relevant frames, demonstrating that employing semantic attention maps improve the robustness of the visual features and helps obtaining better results.

convolutional features via this sample model to employ in diversity and representativeness objectives. Moreover, we have 339-dimension probability predictions which is used with queries. For each frame, we first calculated the attribute prediction of the target attribute. For RAD dataset, a shot is represented by the center frame and selected videos contain 64 to 182 shots.

Experimental Results and Discussion

Fig. 12 shows sample summaries of the methods for a query from this dataset. Instead of just focusing on video content or target query attribute or on where people look, we combine these features into an optimization pool and select frames that contain all these features at the same time. In Table 4, we present a quantitative comparison of our approach against the competing approaches. As in our



Query term: *dancing*

Only diversity without attention

Vasudevan et al. [41]

Our method with only attention guided diversity

Ours

Fig. 13 Another visual analysis showing the effects of each component of the proposed summarization approach. Our full model generates a better summary capturing the query term more accurately than the others.

Activity Related Summaries dataset, the results demonstrate the effectiveness our proposed query-focused summarization method. Overall, it outperforms all the other approaches in terms of F1-score while giving highly competitive scores as to all precision-based metrics. The weakly-supervised method by Panda et al. [27] that also employs class-specific attention maps extracted based on the given query term(s) achieves good results when precision and VERT are considered, however, it again gives the worst performance regarding recall and F1-score.

In Table 5, we analyze the effects of submodular functions separately. Representativeness itself contributes most to the overall performance, and combining all functions yields the best performance. Moreover, we can infer that using attention in diversity term (noting attention near the symbol) gives the best results. Again, due to the characteristic of the metric and the dataset, representativeness and diversity play more important roles than relevance since the dataset contains so many frames that contain query concepts and a successful summary should collect different frames from different shots. We can also see in Figure 13 that our method constructs a smooth and coherent summary in terms of semantic video content and emphasizes on query term enough to include query-related frames into the generated summary.

5.4 Runtime Analysis

In this section, we present an analysis regarding the running times of the summarization methods we evaluated in our experiments. As mentioned earlier, these

Table 4 Performance comparison of summarization methods models on RAD dataset with the best performing model highlighted with a bold typeface. The mean and the standard deviation of the evaluation measures are reported.

	Precision	Recall	F1-score	VERT-P
Uniform	0.7541 ± 0.1817	0.4951 ± 0.1925	0.5632 ± 0.1447	0.5973 ± 0.1431
K-means	0.7157 ± 0.1988	0.5113 ± 0.1714	0.5514 ± 0.1873	0.5857 ± 0.1624
Vis-DSS [8]	0.7857 ± 0.2075	0.5343 ± 0.2362	0.5864 ± 0.2104	0.5814 ± 0.2071
Panda et al. [27]	**0.7997** ± 0.1794	0.4571 ± 0.1966	0.5391 ± 0.1927	**0.6193** ± 0.1719
Vasudevan et al.[41]	0.7685 ± 0.1544	0.5527 ± 0.2083	0.6046 ± 0.1817	0.6053 ± 0.1906
Ours	0.7767 ± 0.1834	**0.5771** ± 0.2012	**0.6196** ± **0.2279**	0.6007 ± 0.2025

Table 5 Contribution of submodular functions considered in our framework in RAD dataset [41]. 'AG' denotes attention-guidance. We report the mean and the standard deviation of F1-scores.

Representativeness	Diversity	Relevance	F1-score
+			0.5887 ± 0.1755
	+		0.5574 ± 0.1789
		+	0.5317 ± 0.1724
+	+		0.5912 ± 0.1803
+	+	+	0.6046 ± 0.1817
	+ AG		0.5645 ± 0.1946
+	+ AG		0.5978 ± 0.1874
+	+ AG	+	**0.6196** ± 0.2279

methods differ from each other in the way how they formulate the summarization process – solving it via optimization or by applying greedy strategies, and how the query terms are included, either ignoring them or defining relevance scores or extracting class-specific attention maps. As one expects, these methodological differences affect the overall runtimes of the approaches. In Table 6, we present the total running times of the competing approaches required to summarize a sample video sequence of 228 sec. long that consists of 38 shots (6819 frames in total) using a query term that includes two different keywords. We conduct the experimental analysis on a system with a quad core Intel Core i7 8700k CPU @ 3.7 GHz and 16 GB of memory.

As can be seen from the table, the weakly-supervised model by Panda et al. [27] works much faster than all the other methods under evaluation. Vis-DSS [8], Vasudevan et al. [41] and our approach require additional runtime processing to obtain the summary shots as they all solve an optimization problem based on certain conditions. That said, as discussed in the previous subsections, in general, Panda et al. [27] performs very poorly as compared to other summarization

Table 6 Runtime analysis of the summarization methods. While indicating the timings, we report per-frame runtimes for feature extraction and saliency map extraction steps, and overall runtimes for the optimization step. Saliency map extraction requires a forward pass through the network, and it inherently includes time required for feature extraction.

	Runtime (sec.)		
Method	Feature Extraction (per frame)	Saliency Map Extraction (per frame)	Optimization
Vis-DSS [8]	0.11	-	18.4
Panda et al. [27]	-	0.22	-
Vasudevan et al. [41]	0.13	-	23.7
Ours	-	0.22	23.7

methods in terms of recall and F1-score metrics. Our method extracts and uses class-specific saliency maps during summarization, which introduces a small extra computational cost. However, it is important to note that, since this requires a forward pass through the network, it inherently includes the feature extraction step.

6 Conclusion

We introduced a new method for query-focused video summarization. Our approach is a weakly supervised which integrates saliency maps into a submodular optimization to consider query terms both in capturing relevant shots as well as to represent similarity among shots in a query-adaptive way. We also proposed a new dataset consisting of long videos with rich query terms. We conducted experiments on the proposed dataset called Activity Related Summaries and a subset of the RAD dataset [41]. Experiments showed that our method is superior to simple baseline methods and the previously proposed query-specific summarization methods such as Vasudevan et al. [41] and Panda et al. [27]. Our method works well both in surveillance videos and crowd videos.

Our experiments show that selecting features from the image regions where convolutional networks focuses on while predicting target concepts yields a better summary than not masking out irrelevant features. Moreover, when combining different characteristics in the submodular functions, considering semantic attention maps helps us to extract more diverse summaries. In short, these attention maps provide a generic solution to capture query-related information in a weakly supervised way. Incorporating this attention into the diversity term also leads to a better selection of query relevant shots.

Currently, we assume that we have access to a classifier network trained on a domain similar to that of the query terms. A possible future research direction could be to extend our approach to a zero-shot learning setting. Moreover, it should be noted that our summarization is agnostic to the attribute prediction networks used to compute the semantic attention maps. Hence, additional performance gains may be obtained if one switches to a better performing network architecture. In our current formulation, we use equal weights to denote the importance of representativeness, diversity and query relevance terms. It would be interesting to learn the optimal set of weights through a training stage. Finally, it is important note that query-specific summarization methods suffer from accuracy vs. efficiency tradeoff. To improve the runtime performance, one can use a more efficient CNN architecture or simply consider only the central frame in each shot.

Acknowledgments

This work was supported in part by GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem, BAGEP 2021 Award of the Science Academy to A. Erdem.

References

1. Basavarajaiah, M. and Sharma, P. GVSUM: Generic video summarization using deep visual features. *Multimedia Tools and Applications*, 80:14459–14476, 2021.
2. de Avila, S. E. F., Lopes, A. P. B., da Luz, A., and de Albuquerque Araújo, A. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56 – 68, 2011.
3. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.
4. Goldman, D. B., Curless, B., Salesin, D., and Seitz, S. M. Schematic storyboarding for video visualization and editing. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 862–871. ACM, 2006.
5. Gong, B., Chao, W.-L., Grauman, K., and Sha, F. Diverse sequential subset selection for supervised video summarization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2069–2077, 2014.
6. Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. Creating summaries from user videos. In *Proc. European Conference on Computer Vision (ECCV)*, pages 505–520. Springer, 2014.
7. Gygli, M., Grabner, H., and Van Gool, L. Video summarization by learning submodular mixtures of objectives. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3098, 2015.
8. Iyer, R., Dubal, P., Dargan, K., Kothawade, S., Mahadev, R., and Kaushal, V. Vis-dss: An open-source toolkit for visual data selection and summarization. *arXiv preprint arXiv:1809.08846*, 2018.
9. Jiang, P. and Han, Y. Query-conditioned three-player adversarial network for video summarization. In *Proc. International Conference on Multimedia Retrieval (ICMR)*, 2019.
10. Kaushal, V., Kothawade, S., Tomar, A., Iyer20218, R., and Ramakrishnan, G. How good is a video summary? a new benchmarking dataset and evaluation framework towards realistic video summarization. *arXiv preprint arXiv:2101.10514*, 2021.
11. Khosla, A., Hamid, R., Lin, C.-J., and Sundaresan, N. Large-scale video summarization using web-image priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2705, 2013.
12. Kim, G., Sigal, L., and Xing, E. P. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4225–4232, 2014.
13. Laganière, R., Bacco, R., Hocevar, A., Lambert, P., Païs, G., and Ionescu, B. E. Video summarization from spatio-temporal features. In *Proc. ACM TRECVid Video Summarization Workshop*, pages 144–148. ACM, 2008.
14. Lee, Y. J., Ghosh, J., and Grauman, K. Discovering important people and objects for egocentric video summarization. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1346–1353. IEEE, 2012.
15. Lee, Y. J. and Grauman, K. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.
16. Li, Y. and Merialdo, B. VERT: Automatic evaluation of video summaries. In *Proc. ACM Multimedia*, page 851–854. ACM, 2010.

17. Lin, H. and Bilmes, J. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics, 2011.
18. Liu, W., Mei, T., Zhang, Y., Che, C., and Luo, J. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3707–3715, 2015.
19. Lu, Z. and Grauman, K. Story-driven summarization for egocentric video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2714–2721, 2013.
20. Mendi, E., Clemente, H. B., and Bayrak, C. Sports video summarization based on motion analysis. *Computers and Electrical Engineering*, 39(3):790–796, April 2013.
21. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019.
22. Mundnich, K., Fenster, A., Khare, A., and Sundaram, S. Audiovisual highlight detection in videos. In *Proc. IEEE ICASSP*, 2021.
23. Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):296–305, 2005.
24. Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.
25. Otani, M., Nakashima, Y., Rahtu, E., and Heikkilä, J. Rethinking the evaluation of video summaries. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
26. Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., and Yokoya, N. Video summarization using deep semantic features. In *Asian Conference on Computer Vision*, 2016.
27. Panda, R., Das, A., Wu, Z., Ernst, J., and Roy-Chowdhury, A. K. Weakly supervised summarization of web videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3657–3666, 2017.
28. Pantazis, G., Dimas, G., and Iakovidis, D. K. Salsum: Saliency-based video summarization using generative adversarial networks. *arXiv preprint arXiv:2011.10432*, 2020.
29. Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. Category-specific video summarization. In *Proc. European Conference on Computer Vision (ECCV)*, 2014.
30. Rapantzikos, K., Evangelopoulos, G., Maragos, P., and Avrithis, Y. An audio-visual saliency model for movie summarization. In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 320–323. IEEE, 2007.
31. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
32. Shao, J., Kang, K., Loy, C. C., and Wang, X. Deeply learned attributes for crowded scene understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4657–4666, 2015.
33. Sharghi, A., Borji, A., Li, C., and Yang, T. Improving sequential determinantal point processes for supervised video summarization. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2018.
34. Sharghi, A., Gong, B., and Shah, M. Query-focused extractive video summarization. In *Proc. European Conference on Computer Vision (ECCV)*, pages 3–19. Springer, 2016.
35. Sharghi, A., Laurel, J. S., and Gong, B. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
36. Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proc. International Conference on Machine Learning (ICML)*, 2017.
37. Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proc. International Conference on Learning Representations (ICLR)*, 2014.
38. Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *Proc. International Conference on Learning Representations (ICLR), Workshop Track*, 2015.

39. Sun, K., Zhu, J., Lei, Z., Hou, X., Zhang, Q., Duan, J., and Qiu, G. Learning deep semantic attributes for user video summarization. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 643–648. IEEE, 2017.
40. Tiwari, V. and Bhatnagar, C. A survey of recent work on video summarization: Approaches and techniques. *Multimedia Tools and Applications*, 2021.
41. Vasudevan, A. B., Gygli, M., Volokitin, A., and Van Gool, L. Query-adaptive video summarization via quality-aware relevance estimation. In *Proc. ACM Multimedia*, pages 582–590. ACM, 2017.
42. Wolf, W. Key frame selection by motion analysis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1228–1231. IEEE, 1996.
43. Xiong, B. and Grauman, K. Detecting snap points in egocentric video with a web photo prior. In *Proc. European Conference on Computer Vision (ECCV)*, pages 282–298. Springer, 2014.
44. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J. M., and Singh, V. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2235–2244, 2015.
45. Yeung, S., Fathi, A., and Fei-Fei, L. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824*, 2014.
46. Zeiler, M. D. and Fergus, R. *Visualizing and Understanding Convolutional Networks*, pages 818–833. Springer International Publishing, Cham, 2014.
47. Zhang, J., Zhe, L., Brandt Jonathan, X., Shen, and Stan, S. Top-down neural attention by excitation backprop. In *Proc. European Conference on Computer Vision(ECCV)*, 2016.
48. Zhang, Y., Kampffmeyer, M., Liang, X., Tan, M., and Xing, E. P. Hierarchical variational network for user-diversified i& query-focused video summarization. In *Proc. British Machine Vision Conference (BMVC)*. BMVA, 2018.
49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.