A Framework towards Domain Specific Video Summarization

Vishal Kaushal
IIT Bombay
vkaushal@cse.iitb.ac.in

Sandeep Subramanian
IIT Bombay
sandeeps94@cse.iitb.ac.in

Suraj Kothawade
IIT Bombay
surajkothawade@cse.iitb.ac.in

Rishabh Iyer
Microsoft Corporation
rishi@microsoft.com

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

Abstract

In the light of exponentially increasing video content, video summarization has attracted a lot of attention recently due to its ability to optimize time and storage. Characteristics of a good summary of a video depend on the particular domain under question. We propose a novel framework for domain specific video summarization. Given a video of a particular domain, our system can produce a summary based on what is important for that domain in addition to possessing other desired characteristics like representativeness, coverage, diversity etc. as suitable to that domain. Past related work has focused either on using supervised approaches for ranking the snippets to produce summary or on using unsupervised approaches of generating the summary as a subset of snippets with the above characteristics. We look at the joint problem of learning domain specific importance of segments as well as the desired summary characteristic for that domain. Our studies show that the more efficient way of incorporating domain specific relevances into a summary is by obtaining ratings of shots as opposed to binary inclusion/exclusion information. We also argue that ratings can be seen as unified representation of all possible ground truth summaries of a video, taking us one step closer in dealing with challenges associated with multiple ground truth summaries of a video. We also propose a novel evaluation measure which is more naturally suited in assessing the quality of video summary for the task at hand than F1 like measures. It leverages the ratings information and is richer in appropriately modeling desirable and undesirable characteristics of a summary. Lastly, we release a gold standard dataset for furthering research in domain specific video summarization, which to our knowledge is the first dataset with long videos across several domains with rating annotations. We conduct extensive experiments to demonstrate the benefits of our proposed solution.

1. Introduction

With the explosion of video data, automatic analysis of videos is increasingly becoming important. Examples of such videos include user videos, sports videos, TV videos or CCTV footages or for that matter videos coming from any other source. One of the popular requirements in this context is the ability to automatically summarize videos. Video summarization finds its uses in a wide variety of applications ranging from security and surveillance to compliance and quality monitoring to user applications aimed at saving time and storage. Broadly speaking, in terms of the kind of video summaries produced video summarization can be of two types - compositional video summarization, which aims at producing spatio-temporal synopsis [26, 28, 25, 27] and extractive video summarization, which aims at selecting key frames (also called key frame extraction, static story boards or static video summarization, eg. [3]) or shots (also called dynamic video summarization or dynamic video skimming, eg. [7]). Past work has tried to address the problem of video summarization using unsupervised as well as supervised techniques. Early unsupervised techniques used attention models [19], MMR [16], clustering [3, 21] etc. and more recently auto encoder based techniques [42] and LSTM based techniques [20] have been used. Use of supervised techniques began with forms of indirect supervision from external sources of information like related images [13], similar videos [2] or title of videos [34]. Gong et al's work [6] was the first work on video summarization which used direct high level supervision in form of user annotations. This was followed by Gygli et al's work [8]. Sharghi, Gong et al took this forward by incorporating a notion of user query in the produced video summaries thereby involving the user in the summary generation process [31, 32]. Theirs became the first work on query focussed video summarization. Then came several deep learning based video summarization techniques. For example, Zhang et al. [45] used LSTMs [10] to model long range dependencies among video key frames to produce summaries. Emphasis was given to sequential structures in videos and their modeling. Mahasseni et al.[20] used LSTM networks in an adversarial setting to generate summaries. Both these methods were domain agnostic. Some researchers viewed video summarization as a subset selection problem [44, 5, 31, 32]. Another key approach was to combine the elements of both deep learning and subset selection in summary generation. Gong et al. achieved this using their seqDPP architecture [6] where an LSTM network was coupled with a DPP (Determinantal Point Process) [15].

In this work we address the challenge of domain specific dynamic video summarization. A good video summary should be a meaningful short abstract of a long video, it must contain *important* events, it should exhibit some continuity and it should be free from redundancy. However, the notion of importance varies with domain. For example a ”six” or a ”wicket” could be considered important in cricket videos while ”entry of birthday girl” and ”cutting of cake” would be considered important in birthday videos. Further the characteristics of a good summary like representativeness, coverage or diversity also depends on the domain. For example, while for surveillance videos a good summary should contain outliers, for a user video coverage and representativeness become more important. We hereby propose a novel domain specific video summarization framework which automatically learns to produce summaries that possess desired characteristics suitable for that domain. Past work on video summarization has focused either on using supervised approaches for ranking the snippets/segments thereby producing a video summary (eg. [24, 19]) or on unsupervised approaches of generating the video summary as a subset of snippets/segments with desired characteristics of representativeness, diversity, coverage, etc. (eg. [16, 3]). Some work has also focused on learning the relative importance of uniformity, interestingness and representativeness for different domains (eg. [8, 30]). None of these works, however, have looked at the joint problem of generating domain specific summaries by automatically learning the concepts that are deemed important for a domain, together with having the desired summary characteristics like diversity, coverage etc. for that domain. Building upon the max margin structured learning framework in [8] we learn a mixture of modular and submodular terms. Modular terms help to capture shots more important to the domain under consideration and submodular terms help to capture characteristics of summary important to that domain. The different weights learnt for different components indicate the varying notion of importance from domain to domain.

Further, having many possible ground truth summaries has been posed as one of the challenges in video summarization [39]. Multiple ground truths are due to the difference in perspectives and the fact that different visual content can have the same semantic meaning. However, one reason is also a lack of any other information about the video. Dealing with domain specific videos, however, allows us to define a notion of importance ratings that are unique and unambiguous for that domain. Such ratings can be seen as a unified representation of all possible ground truth summaries of a video. We establish the importance of ratings in the training dataset in producing good summaries as against binary inclusion/exclusion information. Instead of getting ground truth user summaries from annotators we rather ask them to provide ratings for segments in the entire video. The framework learns to generate more accurate summaries when it is provided more supervision this way. We thus establish that the more efficient way of incorporating domain specific relevances into a summary is to provide a supervision in form of ratings as against multiple ground truths.

We also define a new evaluation measure which is more naturally suited for this task than other standard measures used in literature like F1 score. Our measure evaluates video summaries considering the ratings and not just binary inclusion/exclusion information. It also evaluates summaries not only with respect to what they *should* contain but also with respect to what they *should not* contain and the degree of diversity.

As a part of this work we will also release a gold standard dataset for furthering research in domain specific video summarization. To the best of our knowledge, ours is the first dataset with long videos across several domains with rating annotations.

2. Related Work and Our Contributions

2.1. Domain specific video summarization

One of the earliest works on domain specific video summarization was by Potapov et al.[24] in 2014. They were one of the first to realize the importance of building separate models for summarization for distinct categories of videos. They used an SVM[9] classifier conditioned on video category to produce summaries. The SVM learns to score the segments according to their importance to the domain. The segments having higher scores are then selected greedily and put in temporal order to create the final summary. Sun et al. [36] analyzed edited videos of a particular domain as an indicator of highlights of that domain. After finding pairs of raw and corresponding edited videos, they obtain pair-wise ranking constraints to train their model. Zhang et al. [44] use supervision in the form of human-created summaries to perform automatic keyframe-based video summarization. Their main idea is to nonparametrically transfer summary structures of a particular domain of videos from annotated videos to unseen test videos. By learning a joint model to understand what snippets and what characteristics are important for a domain, we use a more principled approach with a form of supervision which is more efficient.

2.2. Submodular functions for video summarization

Video summarization can be viewed as a subset selection problem subject to certain constraints. Given a set $V = \{1, 2, 3, \cdots, n\}$ of items which we also call the *Ground Set*, define a utility function (set function) $f : 2^V \rightarrow \mathbf{R}$, which measures how good a subset $X \subseteq V$ is. Let $c : 2^V \rightarrow \mathbf{R}$ be a cost function, which describes the cost of the set (for example, the size of the subset). Often the cost c is budget constrained (for example, a fixed set summary) and a natural formulation of this is the following problem:

$$\max\{f(X) \text{ such that } c(X) \leq b\} \qquad (1)$$

The goal is then to have a subset X which maximizes f while simultaneously minimizing the cost function c. It is easy to see that maximizing a generic set function becomes computationally infeasible as V grows.

A special class of set functions, called submodular functions [22], however, makes this optimization easy. A function $f : 2^V \rightarrow R$ is submodular if for every $A \subseteq B \subseteq V$ and $e \in V$ and $e \notin B$ it holds that

$$f(\{e\} \cup A) - f(A) \geq f(\{e\} \cup B) - f(B) \qquad (2)$$

Likewise,a function $f : 2^V \rightarrow R$ is supermodular if for every $A \subseteq B \subseteq V$ and $e \in V$ and $e \notin B$ it holds that

$$f(\{e\} \cup A) - f(A) \leq f(\{e\} \cup B) - f(B) \qquad (3)$$

Submodular functions exhibit a property that intuitively formalizes the idea of "diminishing returns". That is, adding some instance x to the set A provides more gain in terms of the target function than adding x to a larger set A', where $A \subseteq A'$. Informally, since A' is a superset of A and already contains more information, adding x will not help as much. Using a greedy algorithm to optimize a submodular function (for selecting a subset) gives a lower-bound performance guarantee of around 63% of optimal [22] to the above problem, and in practice these greedy solutions are often within 98% of optimal [14]. This makes it advantageous to formulate (or approximate) the objective function for data selection as a submodular function.

This concept has been used in document summarization [18] and in image collection summarization [40]. More recently, Elhamifar et al. [5] used submodular optimization for online video summarization by performing incremental subset selection. One of the first attempts to summarize videos using a submodular mixture of objectives was by Gygli et al. [8]. They, however, did not distinguish between various domains of videos and had a specially crafted video frame interestingness model which played a significant role in the summaries produced. Building upon the approach in [8] we learn a mixture for a domain to produce summary specific to that domain. However, in our work, domain specific importance of snippets/shots is not predicted separately using another model. Rather weighted features are directly used in the mixture as modular terms. These modular components capture the shot level domain importance while the other submodular and supermodular components in the mixture correspond to different desired characteristics of the summary like diversity and coverage.

2.3. Evaluation measures

Different measures have been reported in literature for the purpose of accurately evaluating the quality of the video summary produced. VIPER [4] addresses the problem by defining a specific ground truth format which makes it easy to evaluate a candidate summary. SUPERSEIV [11] is an unsupervised technique to evaluate video summarization algorithms that perform frame ranking. VERT [17] was inspired by BLEU in machine translation and ROUGE in text summarization. More recently approaches by Yeung et al. [43] and Plummer et al. [23] also used text based evaluation methods. De Avila et al. [3] propose a method of evaluation which considers several ground truth summaries. Others, like [13, 7, 5] and [41] more directly use precision and recall type measures. Kannappan et al. [12] propose an approach which is only for static video summaries. The evaluation measure proposed by Potapov et al. [24] is capable of evaluating a summary only against one ground truth. As annotations are done by several users producing several ground truth summaries, they evaluate those annotations against each other to form some kind of an upper bound on performance. Approaches like [46, 35, 8] and [45] combine several ground truths into one before using them for evaluation. This comes at the cost of losing individual opinion. In search for a measure which would work directly on ratings (which is a potential generator of multiple ground truths) and having certain other desired characteristics (as enumerated in the corresponding section below) we developed our own evaluation measure.

2.4. Datasets for video summarization

Different researchers in the past have released different datasets for the purpose of video summarization. Examples include The Video Summarization (SumMe) dataset [7], MED Summaries dataset [24] and Title-based Video Summarization (TVSum) dataset[35]. However, none of these existing datasets were found suitable for our work for following reasons. Firstly, we aim to summarize videos across a large number of domains like surveillance, sports, user etc. in a single framework. For that purpose we need a wide variety of videos summarized uniformly. The various datasets only provide certain subsets of types of videos and they use vastly different methods to annotate those videos. So, it was essential that we had a uniformly annotated, diverse set of videos from diverse domains. Secondly, we wanted to test our method on long videos, as the true benefit of a video summary in real world applications is seen only with respect to long videos. Thirdly, we wanted the annotations to not only capture what is important, but also what is *not* important and what is repetitive. Identifying segments which are relatively long and contain repetitive information (for example, scene of spectators clapping for 5 minutes in a cricket video) and retaining only a fraction of them to be included in the summary, is essential to having a good quality summary.

2.5. Our Contributions

In the following, we summarize the main contributions of this paper.

- We address the problem of Domain specific video summarization, by jointly ranking the most important portions of the video for that domain (for example, a goal in Soccer), while simultaneously capturing diverse and representative shots. We do this by training a joint mixture model with features which capture domain importance along with diversity models.

- We argue how different models capture aspects of the summarization task. For example, diversity is more important in surveillance videos when we want to capture outliers, while representation is more important in personal videos. Similarly importance or relevance plays a critical role in domains like Sports (like a goal in soccer).

- We introduce a novel evaluation criteria which captures these aspects of a summary, and also introduce a large dataset for domain specific summarization. Our dataset comprises of several long videos for different domains (surveillance, personal videos, sports etc.) and to our knowledge is the first domain specific video summarization dataset with long videos.

- We then empirically demonstrate various interesting insights. a) We first show that by jointly modeling diversity, relevance and importance, we can learn substantially superior summaries on all domains compared to just learning any one of these aspects. b) We next show that by learning on the same domain, we can obtain superior results than using learnt mixtures from other domains, thus proving the benefit for domain specific video summarization. c) We then look at the top components learnt for different domains and show how those individual components perform best for that domain if considered in isolation. Moreover, we argue how intuitively it makes sense that these components are important. For example, in surveillance, we see that diversity functions tend to have high ranking compared to other models, while in personal videos (like birthday), we see that representation is important. We moreover also look at the highest ranking snippets based on these components and show how they capture the most important aspects of that domain.

The major contribution of this work is that this is the first systematic study of domain specific video summarization on large videos and we provide several insights into the role of different summarization models for this problem.

3. Methodology

We begin by creating a training dataset comprising of videos from several categories annotated with ratings information. Our method works on ratings and hence better deals with issue of multiple ground truths. Building upon the approach in [8], we create a mixture, but our mixture contains modular terms (to capture the domain specific importance of snippets) and submodular terms (for imparting certain desired characteristics to the summary). For each training video of a domain, the components of the mixture are instantiated and the weights of the complete mixture for that domain are learnt using max margin learning framework. After the training phase, for any given test video of that domain, the weighted mixture is then maximized to produce the desired summary video. Below we describe details of every step in the above methodology.

3.1. Training Data

In this work we focus on videos from five different domains - birthday, cricket, soccer, entry exit and office. The latter two are surveillance videos taken from CCTV cameras installed at various entry/exit locations and offices respectively. We have collected birthday, cricket and soccer videos from internet (existing published datasets / youtube). Due to privacy reasons and to be able to experiment with presence/absence of abnormal events in the surveillance videos, we have collected surveillance videos from our own setup of surveillance cameras. We have 7 Cricket videos (of 276 mins), 9 of Birthday (136 minutes total length) and 21 of Entry Exit (306 minutes).

Next, for each domain, we go over every video and first prepare a table of scenes that occur across different videos in that category and using domain knowledge we assign ratings to those scenes. Negative ratings are assigned to segments which *must not* be included in the summary. Since the ratings are relative (for example, a 2 rated scene is supposed to be more important than a 1 rated scene but less important than a 3 rated scene) it was necessary to gather information about all scenes before starting to rate the scenes in the specific videos. Also going through the extra step of creating scenes document for each category enabled us to come up with a consistent philosophy of ratings and consequently a very high inter annotator correlation. Using this scenes table as annotation guidelines for each category, the annotators were then asked to annotate the videos in each category. Segments that are long and contain repetitive content are explicitly marked repetitive in addition to their rating. For the purpose of annotating, we customized a tool called oTranscribe [1] to make the annotation task easy and to produce the desired annotation JSON. The oTranscribe interface was cleaned up and a lot more keyboard shortcuts were added in for ease of annotation. Shortcuts were added in to mark the beginning and ends of segments, to rate segments, to give a short description of the segment, to mark a segment repetitive and to skip to the end and beginning of the previous segment. Finally, hooks were added in to output the annotation as a JSON file. As a sanity check, we visually verified the annotations thus produced by looking at the annotated videos. The annotated videos were produced by overlaying the labels on top of the original videos.

3.2. Learning framework

The task of video summarization is posed as a discrete optimization problem for finding the best subset representing the summary. Given a video V we split it into a snippets v_i of fixed length. Now we have a set Y_v of all snippets in the video. Our problem reduces to picking $y \subset Y_v$ such that $|y| \leq k$ that maximizes our objective.

$$y^* = \underset{y \subseteq Y_v, |y| \leq k}{\text{argmax}} \; o(x_v, y) \tag{4}$$

y^* is the predicted summary, x_v the feature representation of the video snippets and $o(x_v, y)$ is the weighted mixture of components each capturing some aspect of the domain. Different weights are learnt for different domains.

$$o(x_v, y) = w^T f(x_v, y) \tag{5}$$

where $f(x_v, y) = [f_1(x_v, y), ..., f_n(x_v, y)]$ and $f_i(x_v, y)$ are the various modular, submodular and supermodular components. Given N pairs of a video and a reference summary (V, y_{gt}), we learn the weight vector w by optimizing the following large-margin [38] formulation:

$$\min_{w \geq 0} \frac{1}{N} \sum_{n=1}^{N} L_n(w) + \frac{\lambda_1}{2}||w_1||^2 + \frac{\lambda_2}{2}||w_2||^2 \tag{6}$$

where $L_n(w)$ is the generalized hinge loss of training example n and w_1 and w_2 are the weight vectors for the modular terms and the submodular terms respectively.

$$L_n(w) = \max_{y \subseteq Y_v^n} (w^T f(x_v^n, y) + l_n(y)) - w^T f(x_v^n, y_{gt}^n) \tag{7}$$

This objective is chosen so that each human reference annotation scores higher than any other summary by some margin. For training example n, the margin we chose is denoted by $l_n(y)$. We use $1 - normalizedScore(y)$ as margin where normalized score is computed using min-max normalization of the score generated by our evaluation measure, given the ratings, as described below.

3.3. Components of the mixture

Our mixture contains several hand picked components. Every component serves to impart certain characteristics to the optimal subset (the predicted summary).

Set Cover: For a subset X being scored, the set cover is defined as $f_{sc}(X) = \sum_{u\in U} min\{m_u(X), 1\}$ u is a concept belonging to a set of all concepts U, $m_u(X) = \sum_{x\in X} w_{xu}$ and w_{xu} is the weight of coverage of concept u by element x. This component governs the coverage aspect of the candidate summary and is monotone submodular.

Probabilistic Set Cover: This variant of the set cover function is defined as $f_{psc}(X) = \sum_{u\in U}(1 - \prod_{x\in X}(1 - p_{xu}))$ where p_{xu} is the probability with which concept u is covered by element x. Similar to the set cover function, this function governs the coverage aspect of the candidate summary, viewed stochastically and is also monotone submodular.

Facility Location: The facility location function is defined as $f_{fl}(X) = \sum_{v\in V} \max_{x\in X} sim(v, x)$ where v is an element from the ground set V and $sim(v, x)$ measures the similarity between element v and element x. Facility Location governs the representativeness aspect of the candidate summaries and is monotone submodular.

Saturated Coverage Saturated Coverage is $f_{satc}(X) = \sum_{v\in V} min\{m_v(X), c\}$ where $m_v(X) = \sum_{x\in X} sim(v, x)$ measures the relevance of set X to item $v \in V$ and c is a saturation hyper parameter that controls the level of coverage for each item v by the set X. Saturated Coverage is similar to Facility Location except for the fact that for every category, instead of taking a single representative, it allows for taking potentially multiple representatives. Saturated Coverage is also monotone submodular.

Generalized Graph-Cut Generalized Graph Cut is $f_{gc}(X) = \sum_{i\in V, j\in X} sim(i, j) - \lambda \sum_{i,j\in X} sim(i, j)$ Similar to above two functions, Generalized Graph Cut also models representation. When λ becomes large, it also tries to model diversity in the subset. λ governs the tradeoff between representation and diversity. For $\lambda < 0.5$ it is monotone submodular. For $\lambda > 0.5$ it is non-monotone submodular.

Disparity-min: Denoting the distance measure between snippet/shot i and j by d_{ij}, disparity-min is defined as a set function $f_{disp}(X) = \min_{i,j\in X, i\neq j} d_{ij}$. It is easy to see that maximizing this function involves obtaining a subset with maximal minimum pairwise distance, thereby ensuring a diverse subset of snippets or shots. In principle this is similar to determinantal point processes (DPP), but DPP becomes computationally expensive at inference time. This function, though not submodular, can be efficiently optimized via a greedy algorithm.

Continuity: We work on a set of 2 second snippets as ground set. A summary (subset) could thus may not look continuous enough to give good viewing experience. Thus we add this continuity term in the mixture which would give more score when nearby snippets are chosen - this would ensure a visually more coherent and appealing summary. Essentially it is modeled as a redundancy function (this function is super-modular) within a shot as follows: $f_{cont}(X) = \sum_{s\in S}\sum_{x,x'\in s\cap X} w_{x,x'}$ where S is the set of shots as a result of a shot boundary detection algorithm and $w_{x,x'}$ is the similarity between two snippets which can be defined as how close they are to each other based on their index. That is, the features used here are the indices of the snippets.

Modular components: We use weighted features of snippets (described in the next section) as modular terms in the mixture.

3.4. Features used for instantiating the components

Let the video $\mathcal{V}$ be a set of frames $f_i, i = \{0, 1, 2, \cdots, n\}$. Let us define the ground set $V = \{S_0, S_1, S_2, \cdots, S_k\}$ as a set of *snippets* where each snippet is 2 seconds long. A snippet for a video with frame rate r would thus contain $2r$ consecutive frames. Feature vectors are calculated for each snippet independently by aggregating the feature vectors of the frames/images in that snippet. Different components of the mixture (as above) are instantiated for each video using the features such as VGG [33], GoogleNet [37], YOLO entities and features from Pascal VOC and COCO [29] and Color Histogram features. A comprehensive list of the features used and how they are computed can be found in the extended version of this paper in the supplementary material.

3.5. Evaluation measure

To serve the desired purpose and to be suitable to be used in our framework, we wanted an evaluation measure which would satisfy the following characteristics: 1) The reward for including an r rated snippet must be greater than the reward for including an $r - 1$ rated snippet, 2) A negative rated snippet must be penalized

An r rated segment, no matter how big, should not displace an $r + 1$ rated segment from a budget constrained gold summary, 3) No number of r rated segments should displace an $r + 1$ rated segment from a budget constrained gold summary, 4) In the gold summary, segments marked non-repetitive should not be broken unless it is absolutely necessary (possibly the last one to fit within the boundary) and 5) No reward should be given for picking more than β seconds of a segment marked repetitive. After careful design, we came up with the following formulation. It is not very difficult to see that this formulation satisfies the above characteristics.

The score function for video V, $S_V : y \rightarrow R$, is defined as:

$$\begin{aligned} S_V(y) = &\sum_{x_i\in X_P} |y \cap x_i| * (1 + \frac{|y \cap x_i|}{|x_i|}) * e^{\alpha * rating(x_i)} \\ &+ \sum_{x_i\in X_R} \min(|y \cap x_i|, \beta) * (1 + \frac{\min(|y \cap x_i|, \beta)}{\min(|x_i|, \beta)}) \\ &* e^{\alpha * rating(x_i)} \\ &- \sum_{x_i\in X_N} |y \cap x_i| * k \end{aligned}$$

where, X_P is the set of segments in V marked non-repetitive and rated positive, X_R is the set of segments in V marked repetitive and rated positive, X_R is the set of segments in V rated negative, $\alpha > 0$ is the reward scaling hyper-parameter,

$\beta > 0$ is the repetitiveness cut-off factor and $k > 0$ is the penalty factor.

This function is neither submodular nor supermodular. However, this can be written as a sum of a submodular and a supermodular function (details and proof in supplementary material) and hence the bounds discussed below hold true when this appears as the margin in the discrete optimization of the loss augmented objective (Equation 7).

3.6. Discrete Optimization

Our framework entails two different discrete optimization problems - maximization of the weighted mixture in the loss augmented inference, Equation 7, and during inference to obtain the summary once the mixture is learnt (Equation 4). For efficient optimization with guaranteed bounds, it is important to understand certain characteristics of these components. Note that our mixture of set functions can be written as follows:

$$f(X) = \alpha f^{\text{msub}}(X) + \beta f^{\text{nmsub}}(X) + \gamma f^{\text{sup}}(X) \\ + \delta f^{\text{d}}(X)$$

where $f^{\text{msub}}(X)$ is a monotone submodular function, $f^{\text{nmsub}}(X)$ is a non-monotone submodular function, $f^{\text{sup}}(X)$ is a monotone supermodular function, and $f^{\text{d}}(X)$ is a dispersion function (also called disparity-min) ($f^{\text{d}}(X) = \min_{i,j \in X} d_{ij}$), and $\alpha, \beta, \gamma, \delta \geq 0$. Moreover, we assume that each of the functions above are non-negative (without loss of generality). Note that in the above, we have grouped all monotone submodular, non-monotone submodular, supermodular functions together. The only function which is neither submodular nor supermodular is Disparity Min.

We would like to understand the theoretical guarantees for the following optimization problem:

$$\max\{f(X) \mid X \subseteq V, |X| \leq k\} \tag{8}$$

for various values of $\alpha, \beta, \gamma, \delta$.

Theorem 1. *The following theoretical results hold for solving the optimization problem of Equation 8:*

1. *We obtain an approximation factor of* $1 - 1/e$ *if* $\alpha \geq 0$ *and* $\beta = \gamma = \delta = 0$.
2. *We obtain an approximation factor of* $1/2$ *if* $\delta > 0$ *and* $\alpha = \beta = \gamma = 0$.
3. *We obtain an approximation factor of* $1/e$ *if* $\alpha \geq 0, \beta > 0$ *and* $\gamma = \delta = 0$.
4. *We obtain an approximation factor of* $1/4$ *if* $\alpha > 0, \delta > 0$ *and* $\beta = \gamma = 0$
5. *We obtain an approximation factor of* $1/2e$ *if* $\alpha, \beta, \delta \geq 0$ *and* $\gamma = 0$.
6. *We obtain an approximation factor of* $\frac{(1-e^{(1-\kappa^l)\kappa_k})}{\kappa_k}$ *where* $k(X) = f^{msub}(X)$ *and* $l(X) = f^{sup}(X)$*, if* $\alpha, \gamma \geq 0$ *and* $\beta = \delta = 0$.
7. *We obtain an approximation factor of* $\frac{(1-e^{(1-\kappa^l)\kappa_k})}{2\kappa_k}$ *where* $k(X) = f^{msub}(X)$ *and* $l(X) = f^{sup}(X)$*, if* $\alpha, \gamma, \delta \geq 0$ *and* $\beta = 0$.
8. *The optimization problem of Equation 8 is inapproximable unless P = NP, if* $\beta, \gamma > 0$ *and* $\alpha, \delta \geq 0$.

The proofs for each of the cases enumerated above are included in the extended version of the paper in supplementary material.

3.7. Generating Ground Truth Summaries

The use of ratings allow us to generate multiple ground truth summaries for a video. The total number of possible summaries could be exponential in video duration (a variant of knapsack on duration of segments), so for our experiments we randomly generate up-to 500 ground truth summaries for each video for each budget percentage (5, 15 and 30). Starting from the highest rating, if all segments of that rating can be fit in the budget, they are included in the summary. If all segments of a rating cannot be included in the remaining budget, using a flavor of standard coin exchange problem, maximal combinations of segments are enumerated such that possibly only the last segment gets broken.

4. Experiments and Results

What follows is a description of various experiments performed and results observed using the above framework for domain specific video summarization. For surveillance videos (entry exit and office), since night videos are black and white we do not use the color histogram features based on hue and saturation. For videos in Soccer, Cricket and Birthday domains we additionally perform shot detection to identify distinct shots in the video and create a feature which keeps track of the snippets present in each shot. This is consumed by the continuity component in the mixture. However, we do not use this continuity component in the mixture for the surveillance videos where the notion of a shot is not well defined in those videos. Also unless explicitly stated, for training, during each epoch, a random ground truth summary, out of the many possible summaries, was chosen for each video so that over a large number of epochs, all ground truths get covered. We do a train test split of 70-30 with respect to the number of videos in each domain in the dataset. The hyper-parameters used for the evaluation measure while training were $\alpha = 1$, $\beta = 6$ and $k = 2$. We arrive at best values for λ_1 and λ_2 by testing the models on the held-out validation set.

Sanity of ground truths and behavior of evaluation measure We perform the following sanity checks on the ground truth summaries produced and the behavior of our evaluation measure:

- scores of all ground truth summaries of a particular budget for a video should be same, asserting that the synthesis of ground truths as above is consistent with $S_V(y)$

- scores of ground truth summaries should always be greater than randomly produced summaries - for all lengths, for all videos for all categories

In this experiment we compare the normalized scores of the ground truth summaries against the normalized scores of 1000 random summaries picked exclusively from segments rated highly positive. We do the standard min-max normalization. We plot the minimum, maximum and average scores for both the random summaries and the ground truth summaries. To ensure that the random summaries do not get very low scores (and hence favoring the ground truth summaries during comparison), the random summaries used in this experiment are not truly random. They do not include the negatively rated segments.

The results (included in the extended version) show that all ground truth summaries score higher than any random summary across all domains and all videos. We also verify the ground truth summaries visually by representing them as videos and visually assessing their quality.

Learning experiments For learning weights in the above formulation, we constrain the weights of all submodular components to be always positive. There is no such constraint for modular components in the mixture. We compare AdaGrad and stochastic gradient descent and find AdaGrad to work better for all experiments and hence all results reported use AdaGrad. The reported numbers are losses (i.e. $normalizedGTScore$ - $normalizedSummaryScore$). We compare the following results: a) *All Modular:* Train with only modular terms in the mixture, b) *All Submodular:* Train with only submodular terms in the mixture, c) *Full:* Complete Mixture (all modular terms and all submodular terms).

We compare all these to random summaries, uniform summaries and the best individual component baselines instantiated with different features are used individually to produce summaries. Results are reported in Table 1. We observe that combining both the submodular and modular terms in the mixture (full) provides the best results, as compared to just using submodular and modular terms alone or as compared to any of the baselines. Moreover, the learnt mixtures for only modular and either submodular also outperform Random, Uniform and the average individual submodular functions. We see that the best individual submodular functions also perform better than random or uniform baselines, but not as well as the learnt mixtures, thus proving the benefit of learning for this problem. We also verify the goodness of a summary both quantitatively (using the scores from our evaluation measure) and qualitatively (by visualizing the summary produced). This establishes our hypothesis that joint training can significantly help.

Verification of learning domain specific characteristics To demonstrate that the learnt summaries are domain specific, we test the model learnt on one domain in producing summaries of another domain. The results in Table 2 shows that models learnt on one domain perform poorly on other domains, establishing that the model has indeed learnt domain specific characteristics.

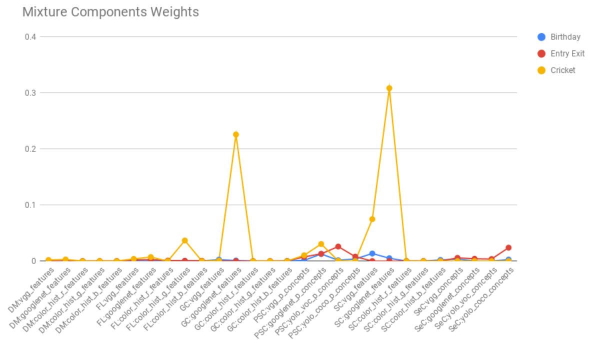
Domain	Method	ScoreLoss
Birthday	All Modular	0.7234
	All Submodular	0.7307
	Full	**0.6625**
	Random	0.7378
	Uniform	0.7569
	Submodular	0.7432
EntryExit	All Modular	0.5967
	All Submodular	0.6306
	Full	**0.5884**
	Random	0.7706
	Uniform	0.7785
	Submodular	0.6306
Cricket	All Modular	0.8140
	All Submodular	0.8275
	Full	**0.7733**
	Random	0.8911
	Uniform	0.8979
	Submodular	0.8275
Office	All Modular	0.3871
	All Submodular	0.4783
	Full	**0.3696**
	Random	0.5743
	Uniform	0.5399
	Submodular	0.5590
Soccer	All Modular	0.8849
	All Submodular	0.7645
	Full	**0.6533**
	Random	0.9217
	Uniform	0.8747
	Submodular	0.9152

Table 1. Learning Experiments comparing all submodular, all modular, full mixture with random, uniform and submodular baselines

Model Trained On	Model Tested On	ScoreLoss
Birthday	Birthday	**0.6625**
	Soccer	0.9753
	Cricket	0.9177
EntryExit	EntryExit	**0.5884**
	Soccer	0.9900
	Cricket	0.9710
	Birthday	0.8009
Cricket	Cricket	**0.7733**
	Soccer	0.8284
	Birthday	0.8103

Table 2. Domain Specific Exp. Results

Next, we look at the weights learnt for the different submodular components. Figure 1 (top) shows magnitude of the learnt weights the different domains. We see that different domains prefer different submodular components and features to produce good summaries. For example, scene features (googlenet_features) are important for Cricket, while object detection features (yolo_voc_p_concepts and yolo_coco_concepts) are more important for Surveillance Videos. For Cricket, the scene of ground or pitch or crowd has a lot of bearing on the importance and for Surveillance, detection of entities assume more importance, given the static scene. Next, we look at the correlation between the components which achieve the best weight in the learnt mixture and the components which achieve the best score when run in isolation. We see in the bottom table of Figure 1 that there is a strong correlation between the two. In particular about 6 to 7 out of the top ten components are the same in both buckets. It is also informative to look at the components themselves. For Birthday and Cricket, we see that Saturated Coverage and Facility Location (i.e. the representative models) are the winners, while in Office (which are surveillance videos), we see a lot of Disparity Min (diversity) functions as winners. A more detailed view of weights learnt for different domains as well as tests demonstrating significance of our results are provided in the extended version of this paper in the supplementary material.



Birthday	**SC:vgg_features** PSC:googlenet_p_concepts **SC:color_hist_r_features** **SC:googlenet_features** PSC:yolo_coco_p_concepts PS SeC:yolo_coco_concepts **GC:vgg_features** **SC:color_hist_b_features** **FL:vgg_concepts**	**SC:vgg_features** **GC:vgg_features** GC:googlenet_features **SC:googlenet_features** FL:googlenet_features **FL:vgg_features** GC:color_hist_b_features **SC:color_hist_b_features** **SC:color_hist_r_features** GC:color_hist_r_features
Cricket	**SC:googlenet_features** **GC:googlenet_features** **SC:vgg_features** **FL:color_hist_g_features** PSC:googlenet_p_concepts PSC:vgg_p_concepts **FL:googlenet_features** **GC:vgg_features** DM:googlenet_features PS	**GC:vgg_features** **SC:vgg_features** **GC:googlenet_features** **SC:googlenet_features** GC:color_hist_r_features SC:color_hist_r_features **FL:color_hist_g_features** SC:color_hist_s_features GC:color_hist_s_features **FL:googlenet_features**
Office	**PSC:yolo_coco_p_concepts** PSC:googlenet_p_concepts **GC:vgg_features** **GC:googlenet_features** SeC:vgg_concepts SeC:yolo_coco_concepts **DM:vgg_features** **DM:googlenet_features** **DM:color_hist_b_features** **DM:color_hist_r_features**	**DM:googlenet_features** **PSC:yolo_coco_p_concepts** **DM:color_hist_b_features** **DM:color_hist_r_features** DM:color_hist_g_features PSC:vgg_p_concepts **DM:vgg_features** **GC:vgg_features** GC:color_hist_g_features **GC:googlenet_features**

Figure 1. (top) Mixture Component Weights and (bottom) the top components based on mixture weights and top components when evaluated in isolation.

Finally, we look at the top ranked frames based on the learnt mixture for each domain. We see that the the frames intuitively capture the most important aspects for that domain. For example, in office footages, the top frames are frames where people are either entering/leaving or meeting, in Birthday videos, we see a shot where people are taking a selfie and a shot where the birthday girl is posing are ranked the best. Similarly in Cricket, shots where the player is about to hit, and where there is a four being hit are selected. Hence we see that the joint training has learnt specific domain specific importance of events along with the right weights for diversity and representation.

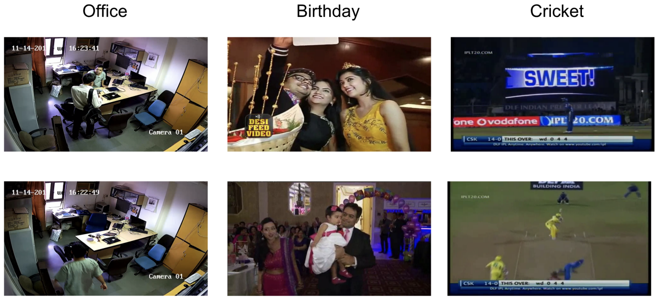


Figure 2. Top Frames based on the Domain specific learnt mixtures

Importance of ratings in generating multiple ground truths To show that ability to generate multiple ground truth from our ratings based annotation is beneficial, we compare the scores obtained by training with a single ground truth summary in every epoch against the scores obtained by training with a random ground truth summary in every epoch.

Birthday	Random GTs	**0.6625**
	Same GT	0.6818
EntryExit	Random GTs	**0.5883**
	Same GT	0.6188

Table 3. Using multiple ground truths gives better accuracy

The results on Table 3 suggests that the model indeed learns better when multiple ground truth summaries are used hence establishing the importance of the system of ratings which allow us to generate many ground truths.

5. Conclusion

Motivated by the fact that what makes a good summary differs across domains, we set out to develop a framework which would automatically learn what is considered important for a domain, both in terms of the kind of snippets to be selected and also in terms of the desired characteristics of the summary produced in terms of representativeness, coverage, diversity, *etc.* We also establish that ratings provide a more efficient way of supervision to impart domain knowledge necessary to create such summaries. Further, we propose a novel evaluation measure well suited for this task. In the absence of any existing dataset which would lend itself well to this particular problem, we created a gold standard dataset and will be making it public as a part of this work. Through several experiments we demonstrated the effectiveness of our solution in producing domain specific summaries which can be seen as a first significant breakthrough in this direction.

References

[1] E. Bentley. otranscribe: Video annotation tool, 2013.

[2] 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.

[3] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[4] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *icpr*, page 4167. IEEE, 2000.

[5] E. Elhamifar and M. C. D. P. Kaluza. Online summarization via submodular and convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1783–1791, 2017.

[6] 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.

[7] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[8] 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.

[9] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] M. Huang, A. B. Mahajan, and D. F. DeMenthon. Automatic performance evaluation for video summarization. Technical report, MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, 2004.

[12] S. Kannappan, Y. Liu, and B. Tiddeman. A pertinent evaluation of automatic video summary. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2240–2245. IEEE, 2016.

[13] 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.

[14] A. Krause. *Optimizing sensing: Theory and applications*. ProQuest, 2008.

[15] A. Kulesza, B. Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[16] Y. Li and B. Merialdo. Multi-video summarization based on video-mmr. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE, 2010.

[17] Y. Li and B. Merialdo. Vert: automatic evaluation of video summaries. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 851–854. ACM, 2010.

[18] H. Lin and J. A. Bilmes. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*, 2012.

[19] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE transactions on multimedia*, 7(5):907–919, 2005.

[20] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail. Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In *Machine Learning and Applications (ICMLA), 2013 12th international conference on*, volume 2, pages 303–308. IEEE, 2013.

[22] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

[23] B. A. Plummer, M. Brown, and S. Lazebnik. Enhancing video summarization via vision-language embedding. In *Computer Vision and Pattern Recognition*, volume 2, 2017.

[24] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014.

[25] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *2009 Advanced Video and Signal Based Surveillance*, pages 195–200. IEEE, 2009.

[26] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[27] Y. Pritch, A. Rav-Acha, and S. Peleg. Nonchronological video synopsis and indexing. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1971–1984, 2008.

[28] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 435–441. IEEE, 2006.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[30] A. Sahoo, V. Kaushal, K. Doctor, S. Shetty, R. Iyer, and G. Ramakrishnan. A unified multi-faceted video summarization system. *arXiv preprint arXiv:1704.01466*, 2017.

[31] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[32] A. Sharghi, J. S. Laurel, and B. Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2127–2136, 2017.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] 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.

[35] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187. IEEE Computer Society, 2015.

[36] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer, 2014.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[38] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.

[39] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3, 2007.

[40] S. Tschiatschek, R. K. Iyer, H. Wei, and J. A. Bilmes. Learning of submodular functions for image collection summarization. In *Advances in neural information processing systems*, pages 1413–1421, 2014.

[41] 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.

[42] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pages 4633–4641, 2015.

[43] S. Yeung, A. Fathi, and L. Fei-Fei. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824*, 2014.

[44] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1059–1067, 2016.

[45] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[46] 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.