Collaborative Summarization of Topic-Related Videos

Rameswar Panda and Amit K. Roy-Chowdhury
Department of ECE, UC Riverside
rpand002@ucr.edu, amitrc@ece.ucr.edu

Abstract

Large collections of videos are grouped into clusters by a topic keyword, such as "Eiffel Tower" or "Surfing", with many important visual concepts repeating across them. Such a topically close set of videos have mutual influence on each other, which could be used to summarize one of them by exploiting information from others in the set. We build on this intuition to develop a novel approach to extract a summary that simultaneously captures both important particularities arising in the given video, as well as, generalities identified from the set of videos. The topic-related videos provide visual context to identify the important parts of the video being summarized. We achieve this by developing a collaborative sparse optimization method which can be efficiently solved by a half-quadratic minimization algorithm. Our work builds upon the idea of collaborative techniques from information retrieval and natural language processing, which typically use the attributes of other similar objects to predict the attribute of a given object. Experiments on two challenging and diverse datasets well demonstrate the efficacy of our approach over state-of-the-art methods.

1. Introduction

With the recent explosion of "big (video) data" over the Internet, it is becoming increasingly important to automatically extract brief yet informative video summaries in order to enable a more efficient and engaging viewing experience. As a result, *video summarization*, that automates this process, has attracted intense attention in the recent years.

Much progress has been made in developing a variety of ways to summarize videos, by exploring different design criteria (representativeness [25, 11, 66, 8, 49, 6], interestingness [13, 31, 41], importance [17, 60]) in an unsupervised manner, or developing supervised algorithms [27, 18, 15, 40, 50]. However, with the notable exception of [6], one common assumption of existing methods is that videos are independent of each other, and hence the summarization tasks are conducted separately by neglecting relationships that possibly reside across the videos.

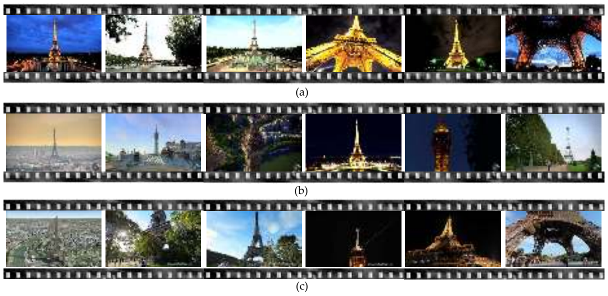

Figure 1. Consider three videos of the topic "*Eiffel Tower*". Each row shows six uniformly sampled shots represented by the middle frame, from the corresponding video. It is clear that all these videos have mutual influence on each other since many visual concepts tend to appear repeatedly across them. We therefore hypothesize that such topically close videos can provide more knowledge and useful clues to extract summary from a given video. We build on this intuition to propose a summarization algorithm that exploits topic-related visual context from video (b) & (c) to automatically extract an informative summary from a given video (a).

Let us consider the video in Fig. 1a. The video is represented by six uniformly sampled shots. Now consider the videos in Fig. 1b and 1c along with the video in Fig. 1a. Are these videos independent of each other or something common exists across them? The answer is clear: all of these videos belong to the same topic "*Eiffel Tower*". As a result, the summaries of these videos will have significant common information with each other. Thus, the context of additional topic-related videos can be beneficial by providing more knowledge and additional clues for extracting a more informative and compact summary from a specified video. We build on this intuition, presenting a new perspective to summarize a video by exploiting the neighborhood knowledge from a set of topic-related videos.

In this paper, we propose a *Collaborative Video Summarization (CVS)* approach that exploits visual context from a set of topic-related videos to extract an informative summary of a given video. Our work builds upon the idea of collaborative techniques [2, 28, 61] from information retrieval (IR) and natural language processing (NLP), which typically use the attributes of other similar objects to predict the attribute of a given object. We achieve this by *finding a sparse set of representative and diverse shots that simulta-*

neously capture both important particularities arising in the given video, as well as, generalities identified from the set of topic-related videos. Our underlying assumption is that a few topically close videos actually have mutual influence on each other since many important visual concepts tend to appear repeatedly across them.

Our approach works as follows. First, we segment each video into multiple non-uniform shots using a temporal segmentation algorithm and represent each shot by a feature vector using a mean pooling scheme over the extracted C3D features (Section 3.1). Then, we develop a novel collaborative sparse representative selection strategy by exploiting visual context from topic-related videos (Section 3.2). Specifically, we formulate the task of finding summaries as an $\ell_{2,1}$-norm sparse optimization problem where the nonzero rows of a sparse coefficient matrix represent the relative importance of the corresponding shots. Finally, the approach outputs a video summary composed of the shots with the highest importance score (Section 3.3). Note that the summary will be of the one video of interest only, while exploiting visual context from additional topic-related videos[1]

The main **contributions** of our work are as follows:

- We propose a novel approach to extract an informative and diverese summary of a specified video by exploiting additional knowledge from topic-related videos. The additional topic-related videos provide visual context to identify what is important in a video.
- We develop a collaborative sparse representative selection strategy by introducing a consensus regularizer that simultaneously captures both important particularities arising in the given video, as well as, generalities identified from the additional topic-related videos.
- We present an efficient optimization algorithm based on half-quadratic function theory to solve the non-smooth objective, where the minimization problem is simplified to two independent linear system problems.
- We demonstrate the effectiveness of our approach in two video summarization tasks—topic-oriented video summarization and multi-video concept visualization. With extensive experiments on both CoSum [6] and TVSum50 [49] video datasets, we show the superiority of our approach over competing methods for both summarization tasks.

2. Related Work

Video summarization has been studied from multiple perspectives [34, 53]. While the approaches might be supervised or unsupervised, the goal of summarization is nevertheless to produce a compact visual summary that encapsulates the most informative parts of a video.

Much work has been proposed to summarize a video using supervised learning. Representative methods use category-specific classifiers for importance scoring [40, 50] or learn how to select informative and diverse video subsets from human-created summaries [18, 15, 45, 65] or learn important facets, like faces, hands, objects [27, 30, 5]. Although these methods have shown impressive results, their performance largely depends on huge amount of labeled examples which are difficult to collect for unconstrained web videos. Our CVS approach, on the other hand, exploits visual context from topic-related videos without requiring any labeled examples, and thus can be easily applied to summarize large scale web videos with diverse content.

Without supervision, summarization methods must rely on low-level visual indices to determine the relevance of parts of a video. Various strategies have been studied, including clustering [1, 9, 16, 38], interest prediction [31, 17], and energy minimization [42, 13]. Leveraging crawled web images is also another recent trend for video summarization [25, 49, 26]. However, all of these methods summarize videos independently by neglecting relationships that possibly reside across them. The use of neighboring topic-related videos to improve summarization still remains as a novel and largely under-addressed problem.

The most relevant work to ours is the video co-summarization approach (CoSum) [6]. It aims to find visually co-occurring shots across videos of the same topic based on the idea of commonality analysis [7]. CoSum also introduced a new benchmark dataset for topic-oriented video summarization. However, CoSum and our approach have significant differences. CoSum constructs weighted bipartite graphs for each pair of videos in order to find the maximal bicliques, which can be computationally inefficient given a large collection of topic-related videos. Our approach, on the other hand, offers a more flexible way to find most representative and diverse video shots through a collaborative sparse optimization framework that can be efficiently solved to handle large number of web videos simultaneously. In addition, CoSum employs a computationally-intensive shot-level feature representation, namely a combination of both observation and interaction features [21], which involves extracting low-level features such as CENTRIST, Dense-SIFT and HSV color moments. By contrast, our approach utilizes generic deep learning features which are more computationally efficient and more accurate in characterizing both appearance and motion.

Our focus on sparse coding as the building block of CVS is largely inspired by its appealing property in modeling sparsity and representativeness in data summarization. In contrast to prior works [8, 11, 66, 36, 37], we develop a novel collaborative sparse optimization that finds shots which are informative about the given video, as well as, the set of topic-related videos.

[1] In this work, we assume that additional topic-related videos are available beforehand. However, in most practical cases, videos retrieved from search engines with topic name as a query may contain outliers and irrelevant videos due to inaccurate query text and polysemy. One feasible choice is to use either clustering [23] or additional meta data to refine the results.

In recent years, collaborative techniques have been successfully applied to several IR and NLP tasks: collaborative recommendation [2, 44], collaborative filtering [61], collaborative ranking [3] and text summarization [56, 54, 55]. The common idea underlying all of these works, including ours, is to make use of the interactions among multiple objects under the assumption that similar objects will have similar behaviors and characteristics.

3. Collaborative Video Summarization

A summary is a condensed synopsis that conveys the most *important* details of the original video. Specifically, it is composed of several shots that represent most important portions of the input video within a short duration. Since, *importance* is a subjective notion, we define a good summary as one that has the following properties.

- **Representative.** The original video should be reconstructed with high accuracy using the extracted summary. We extend this notion of representative as finding a summary that simultaneously minimizes reconstruction error of the given video, as well as the set of topic-related videos.
- **Sparsity.** Although the summary should be representative of the input video, the length should be as small as possible.
- **Diversity.** The summary should be collectively diverse capturing different aspects of the video—otherwise one can remove some of it without losing much information.

The proposed approach, CVS, decomposes into three steps: i) video representation; ii) collaborative sparse representative selection; iii) summary generation.

3.1. Video Representation

Temporal Segmentation. Our approach starts with segmenting videos using an existing algorithm [6]. We segment each video into multiple non-uniform shots with an additional constraint to ensure that the number of frames within each shot lies in the range of [32,96]. The segmented shots serve as the basic units for feature extraction and subsequent processing to extract a video summary.

Feature Representation. Recent advancement in deep learning has revealed that features extracted from upper or intermediate layers of convolutional neural networks (CNNs) are generic features that have good transfer learning capabilities across different domains [46, 67, 24, 43]. In the case of videos, C3D features [52] have recently shown better performance compared to the features extracted using each frame separately [51, 64]. We therefore extract C3D features, by taking sets of 16 input frames, applying 3D convolutional filters, and extracting the responses at layer FC6 as suggested in [52]. This is followed by a temporal mean pooling scheme to maintain the local ordering structure within a shot. Then the pooling result serves as the final feature vector of a shot (4096 dimensional) to be used in the sparse optimization. We will discuss the performance benefits of employing C3D features later in our experiments.

3.2. Collaborative Sparse Representative Selection

We develop a sparse optimization framework that incorporates both information content of the given video and the topic-related videos to extract an informative summary of the specified video. Let v be a video to be summarized and $\tilde{v}$ denote the set of remaining topic-related videos from the video collection. Let the feature matrix of the video v and $\tilde{v}$ are given by $\mathbf{X} \in \mathbb{R}^{d \times n}$and $\tilde{\mathbf{X}} \in \mathbb{R}^{d \times \tilde{n}}$ respectively. d is the dimensionality of the C3D features and n represents the number of shots in the video v. $\tilde{n}$ represents the total number of shots in the remaining topic-related videos $\tilde{v}$.

Formulation. Sparse optimization approaches [8, 11] find the representative shots from a video itself by minimizing the linear reconstruction error as

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 + \lambda_s \|\mathbf{Z}\|_{2,1} \tag{1}$$

where $||\mathbf{Z}||_{2,1} = \sum_{i=1}^{n} ||\mathbf{Z}_i||_2$ and $||\mathbf{Z}_i||_2$ is the ℓ_2-norm of the i-th row of $\mathbf{Z}$. $\lambda_s > 0$ is a regularization parameter that controls the level of sparsity in the reconstruction. Once the problem (1) is solved, the representatives are selected as the points whose corresponding $||\mathbf{Z}_i||_2 \neq 0$.

Clearly, the above formulation summarizes a video neglecting mutual relationships that possibly reside across the videos. Considering the relationships across the topic-related videos, we aim to select a sparse set of representative shots that balances two main objectives: (i) they are informative about the given video, and (ii) they are informative about the complete set of topic-related videos. In other words, we aim to extract a summary that simultaneously minimizes the reconstruction error of the specified video, as well as the set of topic-related videos. Given the above stated goals, we formulate the following objective function,

$$\begin{aligned} \min_{\mathbf{Z}, \tilde{\mathbf{Z}}} \ & \frac{1}{2} \left(\|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 + \alpha \|\tilde{\mathbf{X}} - \mathbf{X}\tilde{\mathbf{Z}}\|_F^2 \right) \\ & + \lambda_s \left(\|\mathbf{Z}\|_{2,1} + \|\tilde{\mathbf{Z}}\|_{2,1} \right) \end{aligned} \tag{2}$$

where parameter $\alpha > 0$ balances the penalty between errors in the reconstruction of video v and errors in the reconstruction of the remaining videos in the collection $\tilde{v}$[2]. The objective function is intuitive: minimization of (2) favors selecting a sparse set of representative shots that simultaneously reconstructs the target video $\mathbf{X}$ via $\mathbf{Z}$, as well as the set of topic-related videos $\tilde{\mathbf{X}}$ via $\tilde{\mathbf{Z}}$, with high accuracy.

Diversity Regularization. The data reconstruction and sparse optimization formulations in (2) tend to select shots that can cover a specified video, as well as the set of topic-related videos. However, there is no explicit tendency to select diverse shots capturing different but also important information described in the set of videos. Prior works [8, 11]

[2] Note that we use a common α to weight the reconstruction term related to the topic-related videos in (2) for simplicity of exposition. However, if we have some prior information on which video is more informative about the topic or close to the specified video, we can assign different αs for different topic-related videos. We leave this problem about the different choice of α as an interesting future work.

handle this issue by manually filtering redundant shots from the extracted summary which can be unreliable while summarizing large scale web videos. Recent works on sparse representative selection [62, 58] also addresses this diversity problem by explicitly adding non-convex regularizers in the objective which makes it difficult to optimize.

Inspired by the recent work on convex formulation for active learning [12] and document compression [63], we introduce two diversity regularization functions, $f_d(\mathbf{Z})$ and $f_d(\tilde{\mathbf{Z}})$ to select a sparse set of representative and diverse shots from the video. Our motivation is that rows in sparse coefficient matrices corresponding to two similar shots are not nonzero at the same time. This is logical since the representative shots should be non-redundant capturing diverse aspects of the input video.

Definition 1. Given the sparse coefficient matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, the diversity regularization functions are defined as:

$$\begin{aligned} f_d(\mathbf{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} Z_{ij} = tr(\mathbf{D}^T\mathbf{Z}), \\ f_d(\tilde{\mathbf{Z}}) &= \sum_{i=1}^{n}\sum_{j=1}^{\tilde{n}} \tilde{d}_{ij} \tilde{Z}_{ij} = tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}) \end{aligned} \tag{3}$$

where $\mathbf{D}$ is the weight matrix measuring the pair-wise similarity of shots in $\mathbf{X}$, and $\tilde{\mathbf{D}}$ measures the similarity between shots in $\mathbf{X}$ and $\tilde{\mathbf{X}}$. There are a lot of ways to construct $\mathbf{D}$ and $\tilde{\mathbf{D}}$. In this paper, we employ the inner product to measure the similarity, since it is simple to implement and it performs well in practice. Minimization of these functions tries to select diverse shots by penalizing the condition that rows of two similar shots are nonzero at the same time.

After adding the diversity regularization functions into problem (2), we have the objective function as follows:

$$\begin{aligned} &\min_{\mathbf{Z},\tilde{\mathbf{Z}}} \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big) \\ &+\lambda_s\big(\|\mathbf{Z}\|_{2,1}+\|\tilde{\mathbf{Z}}\|_{2,1}\big) + \lambda_d\big(tr(\mathbf{D}^T\mathbf{Z}) + tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big) \end{aligned} \tag{4}$$

where λ_d is a trade-off factor associated with the functions.

Consensus Regularization. The objective function (4) favors selecting a sparse set of representative and diverse shots from a target video $\mathbf{X}$ by exploiting visual context from additional topic-related videos $\tilde{\mathbf{X}}$. Specifically, rows in $\mathbf{Z}$ provide information on relative importance of each shot in describing the video $\mathbf{X}$, while rows in $\tilde{\mathbf{Z}}$ give information on relative importance of each shot in $\mathbf{X}$ in describing $\tilde{\mathbf{X}}$. Given the two sparse coefficient matrices, our next goal is to select a unified set of shots that simultaneously cover the important particularities arising in the target video, as well as the generalities arising in the video collection.

To achieve the above goal, we propose to minimize the following objective function:

$$\begin{aligned} &\min_{\mathbf{Z},\tilde{\mathbf{Z}}} \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big) \\ &+\lambda_s\big(\|\mathbf{Z}\|_{2,1}+\|\tilde{\mathbf{Z}}\|_{2,1}\big) + \lambda_d\big(tr(\mathbf{D}^T\mathbf{Z}) + tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big) \\ &+\beta||\mathbf{Z}_c||_{2,1} \quad s.t.\ \mathbf{Z}_c = [\mathbf{Z}|\tilde{\mathbf{Z}}],\ \mathbf{Z}_c \in \mathbb{R}^{n\times(n+\tilde{n})} \end{aligned} \tag{5}$$

where $\ell_{2,1}$-norm on the consensus matrix $\mathbf{Z}_c$ enables $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ to have the similar sparse patterns and share the common components. The joint $\ell_{2,1}$-norm plays the role of consensus regularization as follows. In each round of the optimization algorithm developed later in this paper, the updated sparse coefficient matrices in the former rounds can be used to regularize the current optimization criterion. Thus, it can uncover the shared knowledge of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ by suppressing irrelevant or noisy video shots, which results in an optimal $\mathbf{Z}_c$ for selecting representative video shots.

Optimization. Since problem (5) is non-smooth involving multiple $\ell_{2,1}$-norms, it is difficult to optimize directly. Half-quadratic optimization techniques [19, 20] have shown to be effective in solving these sparse optimizations in several computer vision applications [57, 39, 59, 29, 4]. Motivated by such methods, we devise an iterative algorithm to efficiently solve (5) by minimizing its augmented function alternatively. Specifically, if we define $\phi(x) = \sqrt{x^2+\epsilon}$ with ϵ being a constant, we can transform $\|\mathbf{Z}\|_{2,1}$ to $\sum_{i=1}^{n}\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}$, according to the analysis of $\ell_{2,1}$-norm in [19, 29]. With this transformation, we can optimize (5) efficiently in an alternative way as follows.

According to the half-quadratic theory [19, 20, 14], the augmented cost-function of (5) can be written as follows.

$$\begin{aligned} &\min_{\mathbf{Z},\tilde{\mathbf{Z}}} \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big) \\ &+\lambda_s\big(tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z}) + tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}})\big) + \lambda_d\big(tr(\mathbf{D}^T\mathbf{Z}) + tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big) \\ &+\beta\big(tr(\mathbf{Z}_c^T\mathbf{R}\mathbf{Z}_c)\big) \end{aligned} \tag{6}$$

where $\mathbf{P}, \mathbf{Q}, \mathbf{R} \in \mathbb{R}^{n\times n}$ are three diagonal matrices, and the corresponding i-th element is defined as

$$\begin{aligned} \mathbf{P}_{ii} = \frac{1}{2\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}}, \quad \mathbf{Q}_{ii} = \frac{1}{2\sqrt{||\tilde{\mathbf{Z}}_i||_2^2+\epsilon}}, \\ \mathbf{R}_{ii} = \frac{1}{2\sqrt{||\mathbf{Z}_{ci}||_2^2+\epsilon}} \end{aligned} \tag{7}$$

where ϵ is a smoothing term, which is usually set to be a small constant value. Optimizing (6) over $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ is equivalent to optimizing the following two problems.

$$\begin{aligned} &\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \lambda_d tr(\mathbf{D}^T\mathbf{Z}) \\ &+\lambda_s tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z}) + \beta tr(\mathbf{Z}^T\mathbf{R}\mathbf{Z}) \end{aligned} \tag{8}$$

$$\begin{aligned} &\min_{\tilde{\mathbf{Z}}} \frac{\alpha}{2}\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2 + \lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}) \\ &+\lambda_s tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}}) + \beta tr(\tilde{\mathbf{Z}}^T\mathbf{R}\tilde{\mathbf{Z}}) \end{aligned} \tag{9}$$

Now with fixed $\mathbf{P}, \mathbf{Q}, \mathbf{R}$, the optimal solution of (8) and (9) can be computed by solving the following linear systems:

$$\begin{aligned} (\mathbf{X}^T\mathbf{X} + 2\lambda_s\mathbf{P} + 2\beta\mathbf{R})\mathbf{Z} &= (\mathbf{X}^T\mathbf{X} - \lambda_d\mathbf{D}) \\ (\alpha\mathbf{X}^T\mathbf{X} + 2\lambda_s\mathbf{Q} + 2\beta\mathbf{R})\tilde{\mathbf{Z}} &= (\alpha\mathbf{X}^T\tilde{\mathbf{X}} - \lambda_d\tilde{\mathbf{D}}) \end{aligned} \tag{10}$$

Algo. 1 summarizes the alternative minimization procedure to optimize (5). In step 1, we compute the auxiliary

Algorithm 1 Algorithm for Solving Problem (5)

```
Input: Video feature matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$;
        Parameters $\alpha, \lambda_s, \lambda_d, \beta$, set $t = 0$;
        Construct $\mathbf{D}$ and $\hat{\mathbf{D}}$ using inner product similarity;
        Initialize $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ randomly, set $\mathbf{Z}_c$= [$\mathbf{Z}$, $\tilde{\mathbf{Z}}$] ;
Output: Optimal sparse coefficient matrix Zc.
while not converged do
  1. Compute $\mathbf{P}^t$, $\mathbf{Q}^t$ and $\mathbf{R}^t$ using (7);
  2. Compute $\mathbf{Z}^{t+1}$ and $\tilde{\mathbf{Z}}^{t+1}$ using (10);
  3. Compute $\mathbf{Z}_c^{t+1}$ as: $\mathbf{Z}_c^{t+1}$= [$\mathbf{Z}^{t+1}$ | $\tilde{\mathbf{Z}}^{t+1}$];
  4. $t = t + 1$;
end while
```

matrices **P**, **Q** and **R** which play an important role in representative selection, according to the half-quadratic analysis for $\ell_{2,1}$-norm [19]. In step 2, we find the optimal sparse coefficient matrices **Z** and $\tilde{\mathbf{Z}}$ by solving two linear systems as defined in (10). Step 3 corresponds to the consensus matrix, which is expected to uncover the shared knowledge of **Z** and $\tilde{\mathbf{Z}}$ by enforcing same sparse pattern using a joint $\ell_{2,1}$-norm.

3.3. Summary Generation

Above, we described how we compute the optimal sparse coefficient matrix $\mathbf{Z}_c$ by exploiting visual context from the topic-related videos. To generate a summary, we first sort the shots by decreasing importance according to the ℓ_2 norms of the rows in $\mathbf{Z}_c$ (resolving ties by favoring shorter video shots), and then construct the optimal summary from the top-ranked shots that fit in the length constraint.

4. Experiments

Datasets. We evaluate the performance of our approach using two datasets: (i) the CoSum dataset [6] and (ii) the TVSum50 dataset [49]. To the best of our knowledge, these are the only two publicly available summarization datasets of multiple videos organized into groups with a topic keyword. Both of the datasets are extremely diverse: while CoSum dataset consists of 51 videos covering 10 topics from the SumMe benchmark [17], the TVSum50 dataset contains 50 videos organized into 10 topics from the TRECVid Multimedia Event Detection task [48].

Implementation details. Our results can be reproduced through the following parameters. The regularization parameters λ_s and β are taken as λ_0/γ where $\gamma > 1$ and λ_0 is analytically computed from the data [11]. The other parameters α and λ_d are empirically set to 0.5 and 0.01 respectively and kept fixed for all results.

Compared methods. We compare our approach to the following baselines. For all of the methods, we use what is recommended in the published work.

Clustering (CK and CS): We first clustered the shots using *k*-means (CK) and spectral clustering (CS), with *k* set to 20 [6]. We then generate a summary by selecting shots that are closest to the centroid of top largest clusters.

Sparse Coding (SMRS and LL): We tested two approaches: Sparse Modeling Representative Selection (SMRS) [11] and LiveLight (LL) [66]. SMRS finds the representative shots using the entire video as the dictionary and selecting key shots based on the zero patterns of the coding vector. Note that [8] also uses the same objective function as in [11] for summarizing consumer videos. The only difference lies in the algorithm used to solve the objective function (Proximal vs ADMM). Hence, we compared only with [11]. LL generates a summary over time by measuring the redundancy using a dictionary of shots updated online. We implemented it using SPAMS library [32] with dictionary of size 200 and the threshold $\epsilon_0 = 0.15$, as in [66].

Co-occurrence Statistics (CoC and CoSum): We compared with two baselines that leverage visual co-occurrence across the topic-related videos to generate a summary. Co-clustering (CoC) [10] generates a summary by partitioning the graph into co-clusters such that each cluster contains a subset of shot-pairs with high visual similarity. On the other hand, CoSum finds maximal bicliques from the complete bipartite graph using a block coordinate descent algorithm. We generate a summary by selecting top-ranked shots based on the visual co-occurrence score and set the threshold to select maximal bicliques to 0.3, following [6].

All methods (including the proposed one) use the same C3D feature as described in Sec. 3.1. Such an experimental setting can give a fair comparison for various methods.

4.1. Topic-oriented Video Summarization

Goal: *Given a set of web videos sharing a common topic (e.g., Eiffel Tower), the goal is to provide the users with summaries of each video that are relevant to the topic.*

Solution. The objective function (5) extracts summary of a specified video by exploiting the visual context of topic-related videos. Given a set of videos, our approach can find summaries of each video by exploiting the additional knowledge from the remaining videos. Moreover, one can easily parallelize the computation for more computational efficiency given our alternating minimization in Algo. 1. This provides scalability to our approach in processing large number of web videos simultaneously.

Evaluation. Motivated by [6, 25], we assess the quality of an automatically generated summary by comparing it to human judgment. In particular, given a proposed summary and a set of human selected summaries, we compute the pairwise average precision (AP) and then report the mean value motivated by the fact that there exists not a single ground truth summary, but multiple summaries are possible. Average precision is a function of both precision and change in recall, where precision indicates how well all the representative shots match with the reference summaries and recall indicates how many and how accurately are the representative shots returned in the retrieval result.

For CoSum dataset, we follow [6] and compare each

Table 1. Experimental results on CoSum dataset. Numbers show top-5 AP scores averaged over all the videos of the same topic. We highlight the **best** and second best baseline method. Overall, our approach outperforms all the baseline methods.

Video Topics	Humans			Computational methods						
	Worst	Mean	Best	CK	CS	SMRS	LL	CoC	CoSum	CVS
Base Jumping	0.652	0.831	0.896	0.415	0.463	0.487	0.504	0.561	0.631	**0.658**
Bike Polo	0.661	0.792	0.890	0.391	0.457	0.511	0.492	0.625	0.592	**0.675**
Eiffel Tower	0.697	0.758	0.881	0.398	0.445	0.532	0.556	0.575	0.618	**0.722**
Excavators River Xing	0.705	0.814	0.912	0.432	0.395	0.516	0.525	0.563	0.575	**0.693**
Kids Playing in Leaves	0.679	0.746	0.863	0.408	0.442	0.534	0.521	0.557	0.594	**0.707**
MLB	0.698	0.861	0.914	0.417	0.458	0.518	0.543	0.563	0.624	**0.679**
NFL	0.660	0.775	0.865	0.389	0.425	0.513	0.558	0.587	0.603	**0.674**
Notre Dame Cathedral	0.683	0.825	0.904	0.399	0.397	0.475	0.496	0.617	0.595	**0.702**
Statue of Liberty	0.687	0.874	0.921	0.420	0.464	0.538	0.525	0.551	0.602	**0.715**
Surfing	0.676	0.837	0.879	0.401	0.415	0.501	0.533	0.562	0.594	**0.647**
mean	**0.679**	**0.812**	**0.893**	**0.407**	**0.436**	**0.511**	**0.525**	**0.576**	**0.602**	**0.687**
relative to average human	83%	100%	110%	51%	54%	62%	64%	70%	74%	85%

Table 2. Experimental results on TVSum50 dataset.

Video Topics	Humans			Computational methods						
	Worst	Mean	Best	CK	CS	SMRS	LL	CoC	CoSum	CVS
Changing Vehicle Tire	0.285	0.461	0.589	0.225	0.235	0.287	0.272	**0.336**	0.295	0.328
Getting Vehicle Unstuck	0.392	0.505	0.634	0.248	0.241	0.305	0.324	0.369	0.357	**0.413**
Grooming an Animal	0.402	0.521	0.627	0.206	0.249	0.329	0.331	0.342	0.325	**0.379**
Making Sandwich	0.365	0.507	0.618	0.228	0.302	0.366	0.362	0.375	**0.412**	0.398
ParKour	0.372	0.503	0.622	0.196	0.223	0.311	0.289	0.324	0.318	**0.354**
PaRade	0.359	0.534	0.635	0.179	0.216	0.247	0.276	0.301	0.334	**0.381**
Flash Mob Gathering	0.337	0.484	0.606	0.218	0.252	0.294	0.302	0.318	0.365	**0.365**
Bee Keeping	0.298	0.515	0.591	0.203	0.247	0.278	0.297	0.295	0.313	**0.326**
Attempting Bike Tricks	0.365	0.498	0.602	0.226	0.295	0.318	0.314	0.327	0.365	**0.402**
Dog Show	0.386	0.529	0.614	0.187	0.232	0.284	0.295	0.309	0.357	**0.378**
mean	**0.356**	**0.505**	**0.613**	**0.211**	**0.249**	**0.301**	**0.306**	**0.329**	**0.345**	**0.372**
relative to average human	71%	100%	121%	42%	49%	60%	61%	65%	68%	74%

video summary with five human created summaries[3], whereas for TVSum50 dataset, we compare each summary with twenty ground truth summaries that are created via crowdsourcing. Since the ground truth annotations in TVSum50 dataset contain frame-wise importance scores, we first compute the shot-level importance scores by taking average of the frame importance scores within each shot and then select top 50% shots for each video, as in [6].

Apart from comparing with the baseline methods, we also compute the average precision between human created summaries. We show the worst, average and best scores of the human selections. The worst human score is computed using the summary which is the least similar to the rest of the summaries whereas the best score represent the most similar summary that contain most shots that were selected by many humans. This provides a pseudo-upper bound for this task, and thus we also report normalized AP scores by rescaling the mean AP of human selections to 100%.

Comparison with baseline methods. Tab. 1 shows the AP on top 5 shots included in the summaries for CoSum dataset. We can see that our method significantly outperforms all baseline methods to achieve an average performance of 85%, while the closest published competitor, CoSum, reaches 74%. Moreover, if we compare to the human performance, we can see that our method even outperforms the worst human score of each topic in most cases. This indicates that our method produces summaries comparable to human created summaries. Similarly, for the top-15 results, our approach achieved the highest average score of 83% compared to 69% by the CoSum baseline.

Our approach performed particularly well on videos that have their visual concepts described well by the topic-related videos, e.g., a video of the topic *Eiffel Tower* contains shots that shows the night view of the tower and the remaining videos in the collection also depicts this well (Fig. 1). While our method overall produces better summaries, it has a low performance for certain videos, e.g., videos of the topic *Surfing*. These videos contain fast motion and subtle semantics that define representative shots of the video, such as surfing on the wave or sea swimming. We believe these are difficult to capture without an additional semantic analysis [33]; we leave this as future work.

Tab. 2 shows top-5 AP results for the TVSum50 dataset. Summarization in this dataset is more challenging because of the unconstrained topic keywords. Our approach still outperforms all the alternative methods significantly to achieve an average performance of 74%. Similarly for top-15 results, our approach achieved highest score of 75% compared to 66% by the CoSum baseline.

Test of Statistical Significance. To show statistical significance, we have done t-test of our results and observe that the proposed approach, CVS, statistically significantly outperforms all six compared methods ($p < .01$), except for worst human. To further interpret the not-statistically significant result with respect to worst human, we per-

[3] The original CoSum dataset contains three human created summaries. We have added two more ground truth summaries which are collected using a similar experiment, as in [6].

Table 3. Performance comparison between 2D CNN(VGG) and 3D CNN(C3D) features. Numbers show top-5 AP scores averaged over all the videos of the same topic. * abbreviates topic name for display convenience. See Tab. 1 for full names.

Methods	Base*	Bike*	Eiffel*	Excavators*	Kids*	MLB	NFL	Notre*	Statue*	Surfing	mean
`CVS(Features[6])`	0.580	0.632	0.677	0.614	0.598	0.607	0.575	0.612	0.655	0.623	0.618
`CVS(VGG)`	0.591	0.626	0.724	0.638	0.617	0.642	0.615	0.604	0.721	0.649	0.643
`CVS(C3D)`	0.658	0.675	0.722	0.693	0.707	0.679	0.674	0.702	0.715	0.647	**0.687**

Table 4. Ablation analysis of the proposed approach with different constraints on (5).

Methods	Base*	Bike*	Eiffel*	Excavators*	Kids*	MLB	NFL	Notre*	Statue*	Surfing	mean
`CVS-Neighborhood`	0.552	0.543	0.551	0.583	0.510	0.529	0.534	0.532	0.516	0.527	0.538
`CVS-Diversity`	0.643	0.650	0.678	0.672	0.645	0.653	0.619	0.666	0.688	0.609	0.654
`CVS`	0.658	0.675	0.722	0.693	0.707	0.679	0.674	0.702	0.715	0.647	**0.687**

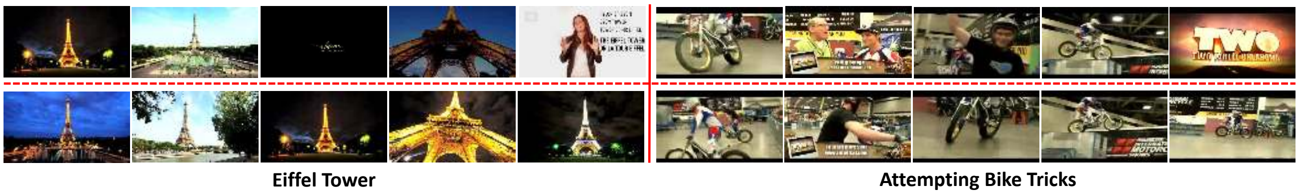

Figure 2. Role of topic-related visual context in summarizing a video. Top row: `CVS` w/o topic-related visual context, and Bottom row: `CVS` w/ topic-related visual context. As can be seen, `CVS` w/o visual context often selects some shots that are irrelevant and not truly related to the topic. `CVS` w/ visual context, on the other hand, automatically selects the maximally informative shots by exploiting the information from additional neighborhood videos. Best viewed in color.

form a statistical power analysis ($\alpha = 0.01$) and see that the power computed for top-5 mAP results on CoSum dataset is 0.279, while on combining with top-15 results, it reaches to 0.877. Similarly, the power reaches 1 for a test that combines both top-5 and top-15 results of both of the datasets. Since, power of a high quality test should usually be > 0.80, we can conclude that our approach statistically outperforms the `worst human` for a large sample size.

Effectiveness of C3D features. We investigate the importance and reliability of C3D features by comparing with 2D shot-level deep features, and found that the later produces inferior results, with a top-5 mAP score of 0.643 on the CoSum dataset (Tab. 3). We utilize Pycaffe [22] with the VGG net pretrained model [47] to extract a 4096-dim feature vector of a frame and then use temporal mean pooling to compute a single shot-level feature vector, similar to C3D features described in Sec. 3.1. We also compare with the shallow feature representation presented in [6] and observe that C3D features performs significantly better over shallow features in summarizing videos (0.618 vs 0.687). We believe this is because C3D features exploit the temporal aspects of activities typically shown in videos.

Performance of the individual components. To better understand the contribution of various components in (5), we analyzed the performance of the proposed approach, by ablating each constraint while setting corresponding regularizer to zero (Tab. 4). With all the components working, the mAP for the CoSum dataset is 0.687. By turning off the neighborhood information from topic-related videos, the mAP decreases to 0.538 (`CVS-Neighborhood`). This corroborates the fact that additional knowledge of topic-related videos help in extracting better summaries, closer to the human selection (see Fig. 2 for qualtitative examples).

Table 5. **User Study—** Average expert ratings in concept visualization experiments. Our approach significantly outperforms other baseline methods in both of the datasets.

Datasets	CK	CS	SMRS	LL	CoC	CoSum	CVS
CoSum	3.70	4.03	5.60	5.63	6.64	7.53	**8.20**
TVSum50	2.46	3.06	4.02	4.20	4.8	5.70	**6.36**

Similarly, by turning off the diversity constraint, the mAP becomes 0.654 (`CVS-Diversity`). We can see that additional knowledge of topic-related videos contributes more than the diversity constraint in summarizing web videos.

4.2. Multi-video Concept Visualization

Goal: *Given a set of topic-related videos, can we generate a single summary that describes the collection altogether?* Specifically, our goal is to generate a single video summary that better estimates human's visual concepts.

Solution. A simple option would be to combine the individual summaries generated from Section. 4.1 and select top ranked shots, regardless of the video, as in the existing existing method [6]. However, such choice will produce a lot of redundant events which eventually reduces the quality of the final summary. We believe this is because, although the individual summaries are informative and diverse, there exists redundancy across the extracted summaries that are relevant to the topic. Our approach can handle this by combining the summaries into a single video, say $\mathbf{X}$ and then extracting a single diverse summary using the final objective function (5) with setting $(\alpha, \beta, \tilde{\mathbf{D}})$ equal to zero.

Evaluation. To evaluate multi-video concept visualization, we need a single ground truth summary of all the topic-related videos that describes the collection altogether. However, since there exists no such ground truth summaries for both of the datasets, we performed human evaluations using 10 experts. Given a video, the study experts were

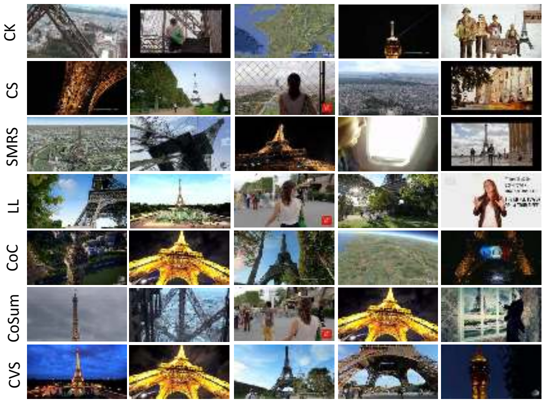

Figure 3. Illustrations of summaries constructed by different methods for the topic *Eiffel Tower*. We show the top-5 results represented by the central frame of each shot. Best viewed in color.

first shown the topic key word (*e.g., Eiffel Tower*) and then shown the summaries constructed using different methods. They were asked to rate the overall quality of each summary by assigning a rating from 1 (worst) to 10 (best).

Results. Tab. 5 shows average expert ratings for both CoSum and TVSum50 datasets. Similar to the results of topic-oriented summarization, our approach significantly outperforms all the baseline methods which indicates that our method generates a more informative summary that describes the video collection altogether. Furthermore, we note that the relative rank of the different approaches are largely preserved as compared to the topic-oriented summarization results. We show a visual comparison between the summaries produced by different methods in Fig. 3. As can be seen, our approach, `CVS`, generates a summary that better estimates human's visual concepts related to the topic.

5. Conclusions

In this work, we present a novel video summarization framework that exploits visual context from a set of topic-related videos to extract an informative summary of a given video. Motivated by the observation that important visual concepts tend to appear repeatedly across videos of the same topic, we develop a collaborative sparse optimization that finds a sparse set of representative and diverse shots by simultaneously capturing both important particularities arising in the given video, as well as, generalities arising across the video collection. We demonstrate the effectiveness of our approach on two standard datasets, significantly outperforming several baseline methods.

Appendix

Since, we have solved (5) using an alternating minimization, we would like to show its convergence behavior. Specifically, the iterative approach in Algo. 1 will monotonically decrease the objective value of (5) in each iteration.

As seen from (6), when we fix $\{\mathbf{P}, \mathbf{Q}, \mathbf{R}\}$ as $\{\mathbf{P}^t, \mathbf{Q}^t, \mathbf{R}^t\}$ in t-th iteration and compute $\mathbf{Z}^{t+1}$, $\tilde{\mathbf{Z}}^{t+1}$, $\mathbf{Z}_c^{t+1}$, the following inequality holds,

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})\\
&\quad+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})+\lambda_s tr((\mathbf{Z}^{t+1})^T\mathbf{P}^t\mathbf{Z}^{t+1})\\
&\quad+\lambda_s tr((\tilde{\mathbf{Z}}^{t+1})^T\mathbf{Q}^t\tilde{\mathbf{Z}}^{t+1}))+\beta\big(tr((\mathbf{Z}_c^{t+1})^T\mathbf{R}^t\mathbf{Z}_c^{t+1})\big)\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})\\
&\quad+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})+\lambda_s tr((\mathbf{Z}^{t})^T\mathbf{P}^t\mathbf{Z}^{t})\\
&\quad+\lambda_s tr((\tilde{\mathbf{Z}}^{t})^T\mathbf{Q}^t\tilde{\mathbf{Z}}^{t}))+\beta\big(tr((\mathbf{Z}_c^{t})^T\mathbf{R}^t\mathbf{Z}_c^{t})\big)
\end{aligned}\tag{11}$$

Adding $\sum_{i=1}^n \frac{\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}$ to both sides of (11), we have

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})\\
&\quad+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})+\lambda_s\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}\\
&\quad+\lambda_s\sum_{i=1}^n\frac{||\tilde{\mathbf{Z}}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\tilde{\mathbf{Z}}_i^t||_2^2+\epsilon}}+\beta\sum_{i=1}^n\frac{||\mathbf{Z}_{ci}^{\,t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_{ci}^{\,t}||_2^2+\epsilon}}\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})\\
&\quad+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})+\lambda_s\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}\\
&\quad+\lambda_s\sum_{i=1}^n\frac{||\tilde{\mathbf{Z}}_i^{t}||_2^2+\epsilon}{2\sqrt{||\tilde{\mathbf{Z}}_i^t||_2^2+\epsilon}}+\beta\sum_{i=1}^n\frac{||\mathbf{Z}_{ci}^{\,t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_{ci}^{\,t}||_2^2+\epsilon}}
\end{aligned}\tag{12}$$

According to the *Lemma* in [35]:

$$\begin{aligned}
&\sum_{i=1}^n\sqrt{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}-\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}\\
&\leq\sum_{i=1}^n\sqrt{||\mathbf{Z}_i^{t}||_2^2+\epsilon}-\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}
\end{aligned}\tag{13}$$

Subtracting Eq. (13) from Eq. (12), we have

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})\\
&+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})+\lambda_s\big(||\mathbf{Z}^{t+1}||_{2,1}+||\tilde{\mathbf{Z}}^{t+1}||_{2,1}\big)+\beta||\mathbf{Z}_c^{\,t+1}||_{2,1}\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})\\
&\quad+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})+\lambda_s\big(||\mathbf{Z}^{t}||_{2,1}+||\tilde{\mathbf{Z}}^{t}||_{2,1}\big)+\beta||\mathbf{Z}_c^{\,t}||_{2,1}
\end{aligned}\tag{14}$$

which establishes that the objective function (5) monotonically decreases in each iteration. Note that the objective function has lower bounds, so it will converge. Empirical results show that the convergence is fast and only a few iterations are needed to converge. Therefore, the proposed method can be applied to large scale problems in practice.

Acknowledgments This work was partially supported by NSF grant IIS-1316934. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] J. Almeida, N. J. Leite, and R. da S. Torres. VISON: VIdeo Summarization for ONline applications. *PRL*, 2012. 2

[2] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 1997. 1, 3

[3] S. Balakrishnan and S. Chopra. Collaborative ranking. In *WSDM*, 2012. 3

[4] R. Bergmann, R. H. Chan, R. Hielscher, J. Persch, and G. Steidl. Restoration of manifold-valued images by half-quadratic minimization. *arXiv preprint arXiv:1505.07029*, 2015. 4

[5] G. K. Bo Xiong and L. Sigal. Storyline representation of egocentric videos with an application to story-based search. In *ICCV*, 2015. 2

[6] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 1, 2, 3, 5, 6, 7

[7] W. S. Chu, F. Zhou, and F. De la Torre. Unsupervised temporal commonality discovery. In *ECCV*, 2012. 2

[8] Y. Cong, J. Yuan, and J. Luo. Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection. *TMM*, 2012. 1, 2, 3, 5

[9] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de Albuquerque Arajo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *PRL*, 2011. 2

[10] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, 2001. 5

[11] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012. 1, 2, 3, 5

[12] E. Elhamifar, G. Sapiro, A. Yang, and S. Sasrty. A convex optimization framework for active learning. In *ICCV*, 2013. 4

[13] S. Feng, Z. Lei, and S. Li. Online content-aware video condensation. In *CVPR*, 2012. 1, 2

[14] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *TPAMI*, 1992. 4

[15] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014. 1, 2

[16] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, and D. D. Feng. A Top-Down Approach for Video Summarization. *TOMCCAP*, 2014. 2

[17] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *ECCV*, 2014. 1, 2, 5

[18] M. Gygli and H. G. L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015. 1, 2

[19] R. He, T. Tan, L. Wang, and W.-S. Zheng. l21 regularized correntropy for robust feature selection. In *CVPR*, 2012. 4, 5

[20] R. He, W.-S. Zheng, T. Tan, and Z. Sun. Half-quadratic-based iterative minimization for robust sparse representation. *TPAMI*, 2014. 4

[21] M. Hoai, Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011. 2

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 7

[23] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: web image search results clustering. In *MM*, 2006. 2

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 3

[25] A. Khosla, R. Hamid, C. J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 1, 2, 5

[26] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014. 2

[27] Y. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1, 2

[28] W. Liu, S. C. Hoi, and J. Liu. Output regularized metric learning with side information. In *ECCV*, 2008. 1

[29] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin. Correntropy induced l2 graph for robust subspace clustering. In *ICCV*, 2013. 4

[30] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[31] Y. F. Ma, X. S. Hua, and H. J. Zhang. A Generic Framework of User Attention Model and Its Application in Video Summarization. *TMM*, 2005. 1, 2

[32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 11:19–60, 2010. 5

[33] T. Mei, L. X. Tang, J. Tang, and X. S. Hua. Near-lossless semantic video summarization and its applications to video analysis. *TOMCCAP*, 2013. 6

[34] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *JVCIR*, 2008. 2

[35] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *NIPS*, 2010. 8

[36] R. Panda, A. Das, and A. K. Roy-Chowdhury. Embedded sparse coding for summarizing multi-view videos. In *ICIP*, 2016. 2

[37] R. Panda, A. Das, and A. K. Roy-Chowdhury. Video summarization in a multi-view camera network. In *ICPR*, 2016. 2

[38] R. Panda, S. K. Kuanar, and A. S. Chowdhury. Scalable video summarization using skeleton graph and random walk. In *ICPR*, 2014. 2

[39] Y. Peng and B.-L. Lu. Robust structured sparse representation via half-quadratic optimization for face recognition. *MTAP*, 2016. 4

[40] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014. 1, 2

[41] Y. Pritch, A. R. Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007. 1

[42] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007. 2

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[44] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, 2001. 3

[45] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 2

[46] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 3

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[48] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, 2006. 5

[49] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 1, 2, 5

[50] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014. 1, 2

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. 2015. 3

[52] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3d: generic features for video analysis. *CoRR, abs/1412.0767*, 2:7, 2014. 3

[53] B. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *TOMCCAP*, 2007. 2

[54] X. Wan and J. Xiao. Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *TOIS*, 2010. 3

[55] X. Wan and J. Yang. Collabsum: exploiting multiple document clustering for collaborative single document summarizations. In *SIGIR*, 2007. 3

[56] X. Wan, J. Yang, and J. Xiao. Single document summarization with document expansion. In *AAAI*, 2007. 3

[57] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2013. 4

[58] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, and H. Yu. Scalable gastroscopic video summarization via similar-inhibition dictionary selection. *Artificial Intelligence in Medicine*, 2016. 4

[59] Y. Wang, C. Pan, S. Xiang, and F. Zhu. Robust hyperspectral unmixing with correntropy-based metric. *TIP*, 2015. 4

[60] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, 2015. 1

[61] G. R. Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen. Scalable collaborative filtering using cluster-based smoothing. In *SIGIR*, 2005. 1, 3

[62] C. Y, J. P, and J. F. Image collection summarization via dictionary learning for sparse representation. In *CVPR*, 2012. 4

[63] J.-g. Yao, X. Wan, and J. Xiao. Compressive document summarization via sparse optimization. In *AAAI*, 2015. 4

[64] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015. 3

[65] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 2

[66] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014. 1, 2, 5

[67] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 3