# Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization

Rameswar Panda and Amit K. Roy-Chowdhury, *Senior Member, IEEE*

*Abstract*—**Most traditional video summarization methods are designed to generate effective summaries for single-view videos, and thus they cannot fully exploit the complicated intra and inter-view correlations in summarizing multi-view videos in a camera network. In this paper, with the aim of summarizing multi-view videos, we introduce a novel unsupervised framework via joint embedding and sparse representative selection. The objective function is two-fold. The first is to capture the multi-view correlations via an embedding, which helps in extracting a diverse set of representatives. The second is to use a $\ell_{2,1}$-norm to model the sparsity while selecting representative shots for the summary. We propose to jointly optimize both of the objectives, such that embedding can not only characterize the correlations, but also indicate the requirements of sparse representative selection. We present an efficient alternating algorithm based on half-quadratic minimization to solve the proposed non-smooth and non-convex objective with convergence analysis. A key advantage of the proposed approach with respect to the state-of-the-art is that it can summarize multi-view videos without assuming any prior correspondences/alignment between them, e.g., uncalibrated camera networks. Rigorous experiments on several multi-view datasets demonstrate that our approach clearly outperforms the state-of-the-art methods.**

*Index Terms*—**Video summarization; Camera Network; Sparse optimization; Multi-view video.**

## I. INTRODUCTION

**N**ETWORK of surveillance cameras are everywhere nowadays. The volume of data collected by such network of vision sensors deployed in many settings ranging from security needs to environmental monitoring clearly meets the requirements of big data [24], [55]. The difficulties in analyzing and processing such big video data is apparent whenever there is an incident that requires foraging through vast video archives to identify events of interest. As a result, *video summarization*, that automatically extract a brief yet informative summary of these videos has attracted intense attention in the recent years.

Although video summarization has been extensively studied during the past few years, many previous methods mainly focused on developing a variety of ways to summarize *single-view* videos in form of a key-frame sequence or a video skim [53], [11], [70], [9], [27], [26], [29]. However, another important problem and rarely addressed in this context is to find an informative summary from *multi-view* videos [14], [33], [43], [28], [44]. *Multi-view video summarization refers to the problem of summarization that seeks to take a set of input videos captured from different cameras focusing on roughly the same fields-of-view (fov) from different viewpoints and produce a video synopsis or key-frame sequence that presents the most*
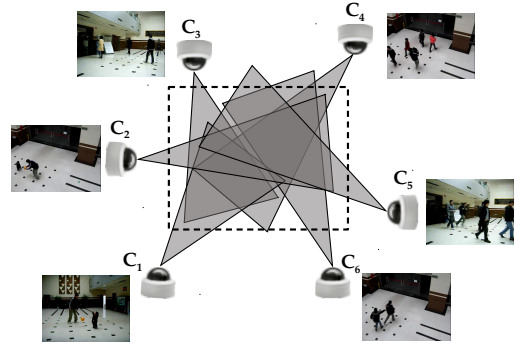
• Rameswar Panda and Amit K. Roy-Chowdhury are with the Department of Electrical and Computer Engineering, University of California, Riverside, CA 925024, USA. E-mails: (rpand002@ucr.edu, amitrc@ece.ucr.edu)



Fig. 1. An illustration of a multi-view camera network where six cameras $C_1$, $C_2$, ..., $C_6$ are observing an area (black rectangle) from different viewpoints. Since the views are roughly overlapping, information correlations across multiple views along with correlations in each view should be taken into account for generating a concise multi-view summary.

*important portions of the inputs within a short duration (See Fig. 1).* In this paper, given a set of videos and its shots, we focus on developing an unsupervised approach for selecting a subset of shots that constitute the multi-view summary. Such a summary can be very beneficial in many surveillance systems equipped in offices, banks, factories, and crossroads of cities, for obtaining significant information in short time.

Multi-view video summarization is different from single-video summarization in two important ways. First, although the amount of multi-view data is immensely challenging, there is a certain structure underlying it. Specifically, there is large amount of correlations in the data due to the locations and fields of view of the cameras. So, content correlations as well as discrepancies among different videos need to be properly modeled for obtaining an informative summary. Second, these videos are captured with different view angles, and depth of fields, for the same scenery, resulting in a number of unaligned videos. Hence, difference in illumination, pose, view angle and synchronization issues pose a great challenge in summarizing these videos. So, methods that attempt to extract summary from single-view videos usually do not produce an optimal set of representatives while summarizing multi-view videos.

To address the challenges encountered in a camera network, we propose a novel multi-view video summarization method, which has the following advantages.

• First, to better characterize the multi-view structure, we project the data points into a latent embedding which is able to preserve both intra and inter-view correlations without assuming any prior correspondences/alignment between the multi-view videos, e.g., uncalibrated camera networks. Our underlying idea hinges upon the basic concept of subspace learning [6], [41], which typically aims to obtain a latent

subspace shared by multiple views by assuming that these views are generated from this subspace.

- Second, we propose a sparse representative selection method over the learned embedding to summarize the multi-view videos. Specifically, we formulate the task of finding summaries as a sparse coding problem where the dictionary is constrained to have a fixed basis (dictionary to be the matrix of same data points) and the nonzero rows of sparse coefficient matrix represent the multi-view summaries.

- Finally, to better leverage the multi-view embedding and selection mechanism, we learn the embedding and optimal representatives jointly. Specifically, instead of simply using the embedding to characterize multi-view correlations and then selection method, we propose to adaptively change the embedding with respect to the representative selection mechanism and unify these two objectives in forming a joint optimization problem. With joint embedding and sparse representative selection, our final objective function is both non-smooth and non-convex. We present an efficient optimization algorithm based on half-quadratic function theory to solve the final objective function.

## II. RELATED WORK

There is a rich body of literature in multimedia and computer vision on summarizing videos in form of a key frame sequence or a video skim (see [40], [62] for reviews).

*Single-view Video Summarization.* Much progress has been made in developing a variety of ways to summarize a single-view video in an unsupervised manner or developing supervised algorithms. Various strategies have been studied, including clustering [1], [9], [18], [57], attention modeling [37], saliency based linear regression model [30], super frame segmentation [20], kernel temporal segmentation [53], crowd-sourcing [26], energy minimization [54], [13], storyline graphs [27], submodular maximization [19], determinantal point process [17], [69], archetypal analysis [60], long short-term memory [68] and maximal biclique finding [7].

Recently, there has been a growing interest in using sparse coding (SC) to solve the problem of video summarization [11], [70], [8], [39], [10], [38] since the sparsity and reconstruction error term naturally fits into the problem of summarization. In contrast to these prior works that can only summarize a single video, we develop a novel multi-view summarization method that jointly summarizes a set of videos to find a single summary for describing the collection altogether.

*Multi-view Video Summarization.* Generating a summary from multi-view videos is a more challenging problem due to the inevitable thematic diversity and content overlaps within multi-view videos than a single video. To address the challenges encountered in multi-view settings, there have been some specifically designed approaches that use random walk over spatio-temporal graphs [14] and rough sets [33] to summarize multi-view videos. A recent work in [28] uses bipartite matching constrained optimum path forest clustering to solve the problem of multi-view video summarization. An online method can also be found in [43]. However, this method relies on inter-camera frame correspondence, which can be a very difficult problem in uncontrolled settings. The work in [31] and [32] also addresses a similar problem of summarization

in non-overlapping camera networks. Learning from multiple information sources such as video tags [64], topic-related web videos [47], [48] and non-visual data [71], [65] is also a recent trend in multiple web video summarization.

This paper has significant differences with our previous work in [45]. First, in [45], we proposed a static multi-view summarization method that extracts a set of key frames to present most important portions of the input videos in form of story-boards. While key frames are a helpful way of indexing videos, they are limited in that all motion information is lost. This limits their use in many surveillance applications where video skimming i.e., replacing all the videos by a shorter compilation of its fragments/shots, seems better suited for obtaining significant information in short time. In this work, we focus on dynamic shot-based video summarization, which not only reduces computational cost but also provides more flexible way of representing videos by considering temporal aspects of activities typically shown in videos. Towards this, we propose a video representation scheme based on spatio-temporal C3D features which have recently shown promising results in several video recognition tasks [61], [47]. Second, in [45], we adopt a two step approach i.e., both embedding and representative selection are performed independently while summarizing multi-view videos. By contrast, in this work, we jointly optimize both of the objectives, such that the embedding can not only characterizes the multi-view structural correlations, but also indicates the requirements of sparse representative selection. Experiments show that joint optimization indeed improves the summarization performance by generating more informative multi-view summaries. Third, we conduct rigorous experiments on three additional multi-view datasets including one large scale dataset captured with 19 surveillance cameras in an indoor setting [43]. We also perform a subjective user study to validate the effectiveness of our approach in generating high quality summaries for a more efficient and engaging viewing experience. New experimentation with spatio-temporal video representation, and the joint optimization well demonstrate the performance improvements in the current framework for summarizing multi-view videos.

## III. PROPOSED METHODOLOGY

In this section, we start by giving main notations and definition of the multi-view summarization problem and then present our detailed approach to summarize multi-view videos.

*Notation.* We use uppercase letters to denote matrices and lowercase letters to denote vectors. For matrix $A$, its $i$-th row and $j$-th column are denoted by $a^i$ and $a_j$ respectively. $||A||_F$ is Frobenius norm of $A$ and $tr(A)$ denote the trace of $A$. The $\ell_p$-norm of the vector $a \in \mathbb{R}^n$ is defined as $||a||_p = (\sum_{i=1}^{n} |a_i|^p)^{1/p}$ and $\ell_0$-norm is defined as $||a||_0 = \sum_{i=1}^{n} |a_i|^0$. The $\ell_{2,1}$-norm can be generalized to $\ell_{r,p}$-norm which is defined as $||A||_{r,p} = (\sum_{i=1}^{n} ||a^i||_r^p)^{1/p}$. The operator $diag(.)$ puts a vector on the main diagonal of a matrix.

*Multi-View Video Summarization.* Given a set of videos captured with considerable overlapping fields-of-view across multiple cameras, the goal of multi-view video summarization is to compactly depict the input videos, distilling its most informative events into a short watchable synopsis. Specifically, it is

composed of several video shots that represent most important portions of the input video collection within a short duration.

Our approach can be roughly described as the set of three main tasks, namely (i) video representation, (ii) joint embedding and representative selection, and (iii) summary generation. In particular, our approach works as follows. First, we segment each video into multiple non-uniform shots using an existing temporal segmentation algorithm and represent each shot by a feature vector using a mean pooling scheme over the extracted C3D features (Section III-A). Then, we develop a novel scheme for joint embedding and representative selection by exploiting the multi-view correlations without assuming any prior correspondence between the videos (Sections III-B, III-C, III-D). Specifically, we formulate the task of finding summaries as an $\ell_{2,1}$ sparse optimization where the nonzero rows of 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 V).

### A. Video Representation

Video representation is a crucial step in summarization for maintaining visual coherence, which in turn affects the overall quality of a summary. It basically consists of two main steps, namely, (i) temporal segmentation, and, (ii) feature representation. We describe these steps in the following sections.

*Temporal Segmentation.* Our approach starts with segmenting videos using an existing algorithm [7]. We segment each video into multiple shots by measuring the amount of changes between two consecutive frames in the RGB and HSV color spaces. A shot boundary is determined at a certain frame when the portion of total change is greater than 75% [7]. We added an additional constraint to the algorithm 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 summary.

*Feature Representation.* Recent advancement in deep feature learning has revealed that features extracted from upper or intermediate layers of a CNN are generic features that have good transfer learning capabilities across different domains [58], [25]. An advantage of using deep learning features is that there exist accurate, large-scale datasets such as Imagenet [56] and Sports-1M [25] from which they can be extracted. For videos, C3D features [61] have recently shown better performance compared to the features extracted using each frame separately [61], [67]. 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 [61]. 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 in our experiments.

### B. Multi-view Video Embedding

Consider a set of $K$ different videos captured from different cameras, where $X^{(k)} = \{x_i^{(k)} \in \mathbb{R}^D, i = 1, \cdots, N_k\}, k =$ $1, \cdots, K$. Each $x_i$ represents the feature descriptor of a video shot in $D$-dimensional feature space. We represent each shot by extracting the shot-level C3D features as described above. As the videos are captured non-synchronously, the number of shots in each video might be different and hence there is no optimal one-to-one correspondence that can be assumed. We use $N_k$ to denote the number of shots in $k$-th video and $N$ to denote the total number of shots in all videos.

Given the multi-view videos, our goal is to find an embedding for all the shots into a joint latent space while satisfying some constraints. Specifically, we are seeking a set of embedded coordinates $Y^{(k)} = \{y_i^{(k)} \in \mathbb{R}^d, i = 1, \cdots, N_k\}, k = 1, \cdots, K$, where, $d$ ($<< D$) is the dimensionality of the embedding space, with the following two constraints: (1) *Intra-view correlations.* The content correlations between shots of a video should be preserved in the embedding space. (2) *Inter-view correlations.* The shots from different videos with high feature similarity should be close to each other in the resulting embedding space as long as they do not violate the intra-view correlations present in an individual view.

*Modeling Multi-view Correlations.* To achieve an embedding that preserves the above two constraints, we need to consider feature similarities between two shots in an individual video as well as across two different videos.

Inspired by the recent success of sparse representation coefficient based methods to compute data similarities in subspace clustering [12], we adopt such coefficients in modeling multi-view correlations. Our proposed approach has two nice properties: (1) the similarities computed via sparse coefficients are robust against noise and outliers since the value not only depends on the two shots, but also depends on other shots that belong to the same subspace, and (2) it simultaneously carries out the adjacency construction and similarity calculation within one step unlike kernel based methods that usually handle these tasks independently with optimal choice of several parameters.

*Intra-view Similarities.* Intra-view similarity should reflect spatial arrangement of feature descriptors in each view. Based on the *self-expressiveness property* [12] of an individual view, each shot can be sparsely represented by a small subset of shots that are highly correlated in the dataset. Mathematically, for $k$-th view, it can be represented as

$$x_i^{(k)} = X^{(k)} c_i^{(k)}, \quad c_{ii}^{(k)} = 0, \quad (1)$$

where $c_i^{(k)} = [c_{i1}^{(k)}, c_{i2}^{(k)}, ..., c_{iN_k}^{(k)}]^T$, and the constraint $c_{ii}^{(k)} = 0$ eliminates the trivial solution of representing a shot with itself. The coefficient vector $c_i^{(k)}$ should have nonzero entries for a few shots that are correlated and zeros for the rest. However, in (1), the representation of $x_i$ in the dictionary $X$ is not unique in general. Since we are interested in efficiently finding a nontrivial sparse representation of $x_i$, we consider the tightest convex relaxation of the $\ell_0$ norm, i.e.,

$$\min \ ||c_i^{(k)}||_1 \ \text{s.t.} \ x_i^{(k)} = X^{(k)} c_i^{(k)}, \ c_{ii}^{(k)} = 0, \quad (2)$$

It can be rewritten in matrix form for all shots in a view as

$$\min \ ||C^{(k)}||_1 \ \text{s.t.} \ X^{(k)} = X^{(k)} C^{(k)}, \ \text{diag}(C^{(k)}) = 0, \quad (3)$$

where $C^{(k)} = [c_1^{(k)}, c_2^{(k)}, ..., c_{N_k}^{(k)}]$ is the sparse coefficient matrix whose $i$-th column corresponds to the sparse representation

of the shot $x_i^{(k)}$. The coefficient matrix obtained from the above $\ell_1$ sparse optimization essentially characterizes the shot correlations and thus it is natural to utilize as intra-view similarities. This provides an immediate choice of the intra-view similarity matrix as $C_{intra}^{(k)} = |C^{(k)}|^T$ where $i$-th row of matrix $C_{intra}^{(k)}$ represents the similarities between the $i$-th shot to all other shots in the view.

*Inter-view Similarities.* Since all cameras are focusing on roughly the same fovs from different viewpoints, all views have apparently a single underlying structure. Following this assumption in a multi-view setting, we find the correlated shots across two views on solving a similar $\ell_1$ sparse optimization like in intra-view similarities. Specifically, we calculate the pairwise similarity between $m$-th and $n$-th view by solving the following optimization problem:

$$\min \ ||C^{(m,n)}||_1 \quad \text{s.t.} \quad X^{(m)} = X^{(n)}C^{(m,n)}, \tag{4}$$

where $C^{(m,n)} \in \mathbb{R}^{N_n \times N_m}$ is the sparse coefficient matrix whose $i$-th column corresponds to the sparse representation of the shot $x_i^{(m)}$ using the dictionary $X$. Ideally, after solving the proposed optimization problem in (4), we obtain a sparse representation for a shot in $m$-th view whose nonzero elements correspond to shots from $n$-th view that belong to the same subspace. Finally, the inter-view similarity matrix between $m$-th and $n$-th view can be represented as $C_{inter}^{(m,n)} = |C^{(m,n)}|^T$ where $i$-th row of matrix $C_{inter}^{(m,n)}$ represent similarities between $i$-th shot of $m$-th view and all other shots in the $n$-th view.

*Objective Function.* The aim of embedding is to correctly match the proximity score between two shots $x_i$ and $x_j$ to the score between corresponding embedded points $y_i$ and $y_j$ respectively. Motivated by this observation, we reach the following objective function on the embedded points $Y$.

$$\mathcal{J}(Y^{(1)}, ..., Y^{(K)}) = \sum_k \mathcal{J}_{\text{intra}}(Y^{(k)}) + \sum_{\substack{m,n \\ m \neq n}} \mathcal{J}_{\text{inter}}(Y^{(m)}, Y^{(n)})$$

$$= \sum_k \sum_{i,j} ||y_i^{(k)} - y_j^{(k)}||^2 C_{intra}^{(k)}(i,j) +$$

$$\sum_{\substack{m,n \\ m \neq n}} \sum_{i,j} ||y_i^{(m)} - y_j^{(n)}||^2 C_{inter}^{(m,n)}(i,j) \tag{5}$$

where $k$, $m$ and $n = 1, \cdots, K$. $\mathcal{J}_{\text{intra}}(Y^{(k)})$ is the cost of preserving local correlations within $X^{(k)}$ and $\mathcal{J}_{\text{inter}}(Y^{(m)}, Y^{(n)})$ is the cost of preserving correlations between $X^{(m)}$ and $X^{(n)}$. The first term says that if two shots $(x_i^{(k)}, x_j^{(k)})$ of a view are similar, which happens when $C_{intra}^{(k)}(i,j)$ is larger, their locations in the embedded space, $y_i^{(k)}$ and $y_j^{(k)}$ should be close to each other. Similarly, the second term tries to preserve the inter-view correlations by bringing embedded points $y_i^{(m)}$ and $y_i^{(n)}$ close to each other if the pairwise proximity score $C_{inter}^{(m,n)}(i,j)$ is high. Problem (5) can be rewritten using one similarity matrix defined over the whole set of video shots as

$$\mathcal{J}(Y) = \sum_{m,n} \sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 C_{total}^{(m,n)}(i,j) \tag{6}$$

where the total similarity matrix is defined as

$$C_{total}^{(m,n)}(i,j) = \begin{cases} C_{intra}^{(k)}(i,j) & \text{if } m = n = k \\ C_{inter}^{(m,n)}(i,j) & \text{otherwise} \end{cases} \tag{7}$$

This construction defines a $N \times N$ similarity matrix where the diagonal blocks represent the intra-view similarities and off-diagonal blocks represent inter-view similarities. Note that an interesting fact about our total similarity matrix construction in (7) is that since each $\ell_1$ optimization is solved individually, a fast parallel computing strategy can be easily adopted for efficiency. However, the matrix in (7) is not symmetric since in $\ell_1$ optimization (2, 4), a shot $x_i$ can be represented as a linear combination of some shots including $x_j$, but $x_i$ may not be present in the sparse representation of $x_j$. But, ideally, a similarity matrix should be symmetric in which shots belonging to the same subspace should be connected to each other. Hence, we reformulate (6) with a symmetric similarity matrix $W = C_{total} + C_{total}^T$ as

$$\mathcal{F}(Y) = \sum_{m,n} \sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 W^{(m,n)}(i,j) \tag{8}$$

With the above formulation, we make sure that two shots $x_i$ and $x_j$ get connected to each other either $x_i$ and $x_j$ is in the sparse representation of the other. Furthermore, we normalize $W$ as $w_i \leftarrow w_i / ||w_i||_\infty$ to make sure the weights in the similarity matrix are of same scale.

Given this construction, problem (8) reduces to the Laplacian embedding [2] of shots defined by the similarity matrix $W$. So, the optimization problem can be written as

$$Y^* = \underset{Y, YY^T = I}{\operatorname{argmin}} \ tr(YLY^T) \tag{9}$$

where $L$ is the graph Laplacian matrix of $W$ and $I$ is an identity matrix. Minimizing (9) is a generalized eigenvector problem and the optimal solution can be obtained by the bottom $d$ nonzero eigenvectors. Note that our approach is agnostic to the choice of embedding algorithms. Our method is based on graph Laplacian because it is one of the state-of-the-art methods in characterizing the manifold structure and performs satisfactorily well in several vision and multimedia applications [15], [36], [42].

### C. Sparse Representative Selection

Once the embedding is obtained, our next goal is to find an optimal subset of all the embedded shots, such that each shot can be described as weighted linear combination of a few of the shots from the subset. The subset is then referred as the informative summary of the multi-view videos. In particular, we are trying to represent the multi-view videos by selecting only a few representative shots. Therefore, our natural goal is to establish a shot level sparsity which can be induced by performing $\ell_1$ regularization on rows of the sparse coefficient matrix [8], [11]. By introducing row sparsity regularizer, the summarization problem can now be succinctly formulated as

$$\min_{Z \in \mathbb{R}^{N \times N}} \ ||Z||_{2,1} \quad \text{s.t.} \quad Y = YZ \tag{10}$$

where $||Z||_{2,1} \triangleq \sum_{i=1}^N ||z^i||_2$ is the row sparsity regularizer i.e., sum of $l_2$ norms of the rows of $Z$. The self-expressiveness constraint $(Y = YZ)$ in summarization is logical as the representatives for summary should come from the original frame set. Using Lagrange multipliers, (10) can be written as

$$\min_Z \ ||Y - YZ||_F^2 + \lambda ||Z||_{2,1} \tag{11}$$

where $\lambda$ is a regularization parameter that balances the weight of the two terms. Once (11) is solved, the representative shots are selected as the points whose corresponding $||z^i||_2 \neq 0$.

*Remark 1.* Notice that both sparse optimizations in (3) and (10) look similar; however, the nature of sparse regularizer in both formulations are completely different. In (3), the objective of $\ell_1$ regularizer is to induce element wise sparsity in a column whereas in (10), the objective of $\ell_{2,1}$ regularizer is to induce row level sparsity in a matrix.

*Remark 2.* Given non-uniform length of shots, (11) can be modified to a weighted $\ell_{2,1}$-norm based objective to consider length of video shots while selecting representatives as

$$\min_Z ||Y - YZ||_F^2 + \lambda||QZ||_{2,1} \qquad (12)$$

where $Q = [diag(q)]$ and $q \in \mathbb{R}^N$ represent the temporal length of each video shot. It is easy to see that problem (12) favors selection of shorter video shots by assigning a lower score via $Q$. In other words, problem (12) tries to minimize the number of shots by considering the temporal length of video shots, such that the overall objective turns to minimizing the length of the final summary.

### D. Joint Embedding and Sparse Representative Selection

We now discuss our proposed method to jointly optimize the multi-view video embedding and sparse representation to select a diverse set of representative shots. Specifically, the performance of sparse representative selection is largely determined by the effectiveness of graph Laplacian in embedding learning. Hence, it is a natural choice to adaptively change the graph Laplacian with respect to the following sparse representative selection, such that the embedding can not only characterizes the manifold structure, but also indicates the requirements of sparse representative selection.

By combining the objective functions (9) and (11), the joint objective function becomes:

$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha \big(||Y - YZ||_F^2 + \lambda||Z||_{2,1}\big) \qquad (13)$$

where $\alpha > 0$ is a trade-off parameter between the two objectives. The first term of the cost function projects the input data into a latent embedding by capturing the meaningful structure of data, whereas the second term helps in selecting a robust set of representatives by minimizing the reconstruction error and the sparsity. Note that the proposed method is also computationally efficient as the sparse representative selection is done in the low-dimensional space by discarding the irrelevant part of a data point represented by a high-dimensional feature, which can derail the representative selection process.

## IV. OPTIMIZATION

The optimization problem in (13) is non-smooth and non-convex. Solving it is thus more difficult due to the non-smooth $\ell_{2,1}$ norm and the additional embedding variable $Y$. Half-quadratic optimization techniques [21], [22] have shown to be effective in solving these sparse optimizations in several vision and multimedia applications [63], [66], [50], [35]. Motivated by such methods, we devise an iterative algorithm to efficiently solve (13) by minimizing its augmented function

alternatively[1]. Specifically, if we define $\phi(x) = \sqrt{x^2 + \epsilon}$ with $\epsilon$ being a constant, we can transform $||Z||_{2,1}$ to $\sum_{i=1}^n \sqrt{||z^i||_2^2 + \epsilon}$, according to the analysis of $\ell_{2,1}$-norm in [21], [35]. With this transformation, we can optimize (13) efficiently in an alternative way as follows.

According to the half-quadratic theory [21], [22], [16], the augmented cost-function of (13) can be written as

$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha \big(||Y - YZ||_F^2 + \lambda tr(Z^T PZ)\big) \qquad (14)$$

where $P \in \mathbb{R}^{N \times N}$ is a diagonal matrix, and the corresponding $i$-th element is defined as

$$P_{i,i} = \frac{1}{2\sqrt{||z^i||_2^2 + \epsilon}} \qquad (15)$$

where $\epsilon$ is a smoothing term, which is usually set to be a small constant value. With this transformation, note that the problem (14) is convex separately with respect to $Y, Z$, and $P$. Hence, we can solve (14) alternatively with the following three steps with respect to $Z, Y$, and $P$, respectively.

(1) *Solving for $Z$:* For a given $P$ and $Y$, solve the following objective to estimate $Z$:

$$\min_Z \alpha \big(tr((Y - YZ)(Y - YZ)^T) + \lambda tr(Z^T PZ)\big) \qquad (16)$$

By setting derivative of (16) with respect to $Z$ to zero, the optimal solution can be computed by solving the following linear system.

$$(Y^T Y + \lambda P)Z = Y^T Y \qquad (17)$$

(2) *Solving for $Y$:* For a given $P$, and $Z$, solve the following objective to estimate $Y$:

$$\min_{Y,YY^T=I} tr(YLY^T) + \alpha tr((Y - YZ)(Y - YZ)^T)$$
$$= \min_{Y,YY^T=I} tr(Y(L + \alpha(I - 2Z + ZZ^T))Y^T) \qquad (18)$$

Eq. 18 can be solved by eigen-decomposition of the matrix $(L + \alpha(I - 2Z + ZZ^T))$. We pick up the eigenvectors corresponding to the $d$ smallest eigenvalues.

(3) *Solving for $P$:* When $Z$ is fixed, we can update $P$ by employing the formulation in Eq. 15 directly.

We continue to alternately solve for $Z, Y$, and $P$ until a maximum number of iterations is reached or a predefined threshold is reached. Since the alternating minimization can stuck in a local minimum, it is important to have a sensible initialization. We initialize $Y$ by solving (9) using an Eigen decomposition and $P$ by an identity matrix. Experiments show that the alternating minimization converges fast by using this kind of initialization. In practice, we monitor the convergence within less than 25 iterations. Therefore, the proposed method can be applied to large scale problems in practice.

---

[1] We solve all the sparse optimization problems using Half-quadratic optimization techniques [21], [22]. Due to space limitation, we only present the optimization procedure to solve (13). However, the same procedure can be easily extended to solve other sparse optimizations (3, 4).

TABLE I
DATASET STATISTICS

| Datasets | # Views | Total Durations (Mins.) | Settings | Camera Type |
|---|---|---|---|---|
| Office | 4 | 46:19 | Indoor | Fixed |
| Campus | 4 | 56:43 | Outdoor | Non-fixed |
| Lobby | 3 | 24:42 | Indoor | Fixed |
| Road | 3 | 22:46 | Outdoor | Non-fixed |
| Badminton | 3 | 15:07 | Indoor | Fixed |
| BL-7F | 19 | 136:10 | Indoor | Fixed |

## V. SUMMARY GENERATION

Above, we described how we compute the optimal sparse coefficient matrix $Z$ by jointly optimizing the multi-view embedding learning and sparse representative selection. We follow the following rules to extract a multi-view summary:

(i) We first generate a weight curve using $\ell_2$ norms of the rows in $Z$ since it provides information about the relative importance of the representatives for describing the whole videos. More specifically, a video shot with higher importance takes part in the reconstruction of many other video shots, hence its corresponding row in $Z$ has many nonzero elements with large values. On the other hand, a shot with lower importance takes part in reconstruction of fewer shots in the whole videos, hence, its corresponding row in $Z$ has a few nonzero elements with smaller values. Thus, we can generate a weight curve, where the weight measures the confidence of the video shot to be included in the final video summary.

(ii) We detect local maxima from the weight curve, then extract an optimal summary of specified length from the local maximums constrained by the weight value and full sequence coverage assumption. Note that the shots with low or zero weights cannot be inserted into final video summary. Furthermore, the weigh curve in our framework allows users to choose different number of shots in summary without incurring additional computational cost. In contrast, many other multi-view video summarization methods need to preset the number of video shots that should be included in the final summary and any change will result in a re-calculation. Therefore, the proposed approach is scalable in generating summaries of different lengths and hence provides more flexibility for practical applications. More details on the summary length and scalability are included in experiments.

## VI. EXPERIMENTS

In this section, we present various experiments and comparisons to validate the effectiveness and efficiency of our proposed algorithm in summarizing multi-view videos.

### A. Datasets.

We conduct rigorous experiments using 6 multi-view datasets with 36 videos in total, which are from [14], [43] (See Tab. I). The datasets are captured in both indoor and outdoor environments with overall 360 degree coverage of the scene, making it more difficult to be summarized. All these datasets are standard in multi-view video summarization and have been used by the prior works [14], [28], [33]. It is important to note that experiments in our prior work [45] was limited to only 3 datasets, whereas in the current work, we conduct experiments on 6 datasets including BL-7F which is one of the largest publicly available dataset for multi-view video summarization.

### B. Performance Measures.

To provide an objective comparison, we compare all the approaches using three quantitative measures, including Precision, Recall and F-measure ($\frac{2 \times Precision \times Recall}{Precision + Recall}$) [14], [28]. For all these metrics, the higher value indicates better summarization quality. We set the same summary length as in [14] to generate our summaries and employ the ground truth of important events reported in [14] to compute the performance measures. More specifically, the ground truth annotations contain a list of events with corresponding start and end frame for each dataset. We took an event as correctly detected if our framework produces a video shot between the start and end of the event. We follow the prior works [14], [28], [43] and consider an event to be redundant if we detect the event simultaneously from more than one camera. Such an evaluation setting gives a fair comparison with the previous state-of-the-art methods [14], [33], [28], [44], [45].

### C. Experimental Settings.

We maintain the following conventions during all our experiments. (i) All our experiments are based on unoptimized MATLAB codes on a desktop PC with Intel(R) core(TM) i7-4790 processor with 16 GB of DDR3 memory. We used a NVIDIA Tesla K40 GPUs to extract the C3D features. (ii) Each feature descriptor is $L_2$-nominalized. (iii) Determining the intrinsic dimensionality of the embedding is an open problem in the field of manifold learning. One common way is to determine it by grid search. We determine it as in most traditional approaches, such as [3]. (iv) The sparsity regularization parameter $\lambda$ is computed as $\lambda_0/\rho$ and $\lambda_0$ is analytically computed from the embedded points [11], (v) We empirically set $\alpha$ to 0.05 and kept fixed for all results.

### D. Comparison with State-of-the-art Multi-view Methods.

*Goal.* This experiment aims at evaluating our approach compared to the state-of-the-art multi-view summarization methods presented in the literature.

*Compared Methods.* We contrast our approach with several state-of-the-art methods which are specifically designed for multi-view video summarization as follows.

• RandomWalk [14]. The method first create a spatio-temporal shot graph and then use random walk as a clustering algorithm over the graph to extract multi-view summaries.

• RoughSets [33]. The method first adopt a SVM classifier as the key frame abstraction process and then applies rough set to remove similar frames.

• BipartiteOPF [28]. This method first uses a bipartite graph matching to model the inter-view correlations and then applies optimum path forest clustering on the refined adjacency matrix to generate multi-view summary.

• GMM [43]. An online Gaussian mixture model clustering is first applied on each view independently and then a distributed view selection algorithm is adopted to remove the content redundancy in the inter-view stage.

*Implementation Details.* To report existing methods results, we use prior published numbers when possible. In particular, for the multi-view summarization methods (RandomWalk, BipartiteOPF and GMM), we report the available results from the corresponding papers and implement RoughSets

TABLE II
PERFORMANCE COMPARISON WITH SEVERAL BASELINES INCLUDING BOTH SINGLE AND MULTI-VIEW METHODS APPLIED ON THE THREE MULTI-VIEW DATASETS. **P**: PRECISION IN PERCENTAGE, **R**: RECALL IN PERCENTAGE AND **F**: F-MEASURE. OURS PERFORM THE BEST.

| Methods | Office | | | Campus | | | Lobby | | | Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** | |
| `Attention-Concate` | 100 | 46 | 63.01 | 40 | 28 | 32.66 | 100 | 70 | 82.21 | TMM2005 [37] |
| `Sparse-Concate` | 100 | 50 | 66.67 | 56 | 55 | 55.70 | 91 | 70 | 78.95 | TMM2012 [8] |
| `Concate-Attention` | 100 | 38 | 55.07 | 56 | 48 | 51.86 | 95 | 72 | 81.98 | TMM2005 [37] |
| `Concate-Sparse` | 93 | 58 | 71.30 | 56 | 62 | 58.63 | 86 | 70 | 77.18 | TMM2012 [8] |
| `Graph` | 100 | 26 | 41.26 | 50 | 48 | 49.13 | 100 | 58 | 73.41 | TCSVT2006 [51] |
| `RandomWalk` | 100 | 61 | 75.77 | 70 | 55 | 61.56 | 100 | 77 | 86.81 | TMM2010 [14] |
| `RoughSets` | 100 | 61 | 75.77 | 69 | 57 | 62.14 | 97 | 74 | 84.17 | ICIP2011 [33] |
| `BipartiteOPF` | 100 | 69 | 81.79 | 75 | 69 | 71.82 | 100 | 79 | 88.26 | TMM2015 [28] |
| `Ours` | **100** | **81** | **89.36** | **84** | **72** | **77.78** | **100** | **86** | **92.52** | Proposed |
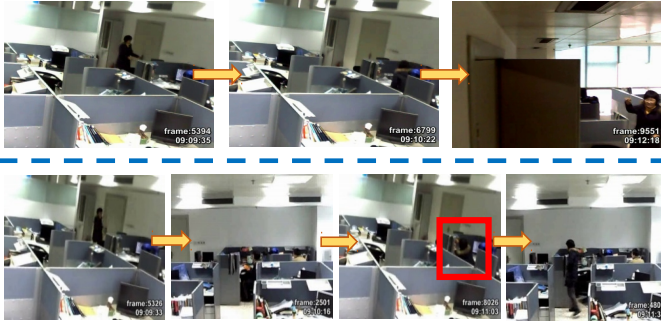


Fig. 2. Sequence of events detected related to activities of a member ($A_0$) inside the Office dataset. Top row: Summary produced by method [14], and Bottom row: Summary produced by our approach. Sequence of events detected in top row: 1st: $A_0$ enters the room, 2nd: $A_0$ sits in cubicle 1, 3rd: $A_0$ leaves the room. Sequence of events detected in bottom row: 1st: $A_0$ enters the room, 2nd: $A_0$ sits in cubicle 1, 3rd: $A_0$ is looking for a thick book to read (as per the ground truth in [14]), and 4th: $A_0$ leaves the room. The event of looking for a thick book to read (as per the ground truth in [14]) is missing in the summary produced by method [14] where as it is correctly detected by our approach (3rd frame: bottom row). This indicates our method captures video semantics in a more informative way compared to [14].

ourselves using the same video representation as the proposed one and tune their parameters to have the best performance.

*Results.* Table II shows the results on three multi-view datasets, namely Office, Campus and Lobby datasets. We have the following key observations from Table II: (i) Our approach produces summaries with same precision as `RandomWalk` and `BipartiteOPF` for both Office and Lobby datasets. However, the improvement in recall value indicates the ability of our method in keeping more important information in the summary compared to both of the approaches. As an illustration, in Office dataset, the event of looking for a thick book by a member while present in the cubicle is absent in the summary produced by `RandomWalk` whereas it is correctly detected by our proposed method. Fig. 2 in this connection explains the whole sequence of events detected using our approach as compared to `RandomWalk`. (ii) For all methods, including `Ours`, performance on Campus dataset is not that good as compared to the other datasets. This is obvious since the Campus dataset contains many trivial events as it was captured in an outdoor environment, thus making the summarization more difficult. Nevertheless, for this challenging dataset, F-measure of our approach is about 6% better than that of the recent `BipartiteOPF`. (iii) Table II also reveals that for all three datasets, recall is generally low compared to precision because users usually prefer to select more extensive summaries in ground truth, which can be verified from the ground truth events from [14]. As a result, number of events

TABLE III
PERFORMANCE COMPARISON WITH GMM BASELINE ON BL-7F DATASET

| Methods | Precision(%) | Recall(%) | F-measure(%) | Reference |
|---|---|---|---|---|
| `GMM` | 58 | 61 | 60.00 | JSTSP2015 [43] |
| `Ours` | 73 | 70 | 71.29 | Proposed |

in ground truth increases irrespective of their information content. (iv) Overall, on the three datasets, our approach outperforms all compared methods in terms of F-measure. This corroborates the fact that the proposed approach produces informative multi-view summaries in contrast to the state-of-the-art methods (See Fig. 3 for an illustrative example).

Table III shows results of our method on a larger and more complex BL-7F dataset captured with 19 surveillance cameras in the 7th floor of the BarryLam Building in National Taiwan University [43]. From Table III, it is clearly evident that our approach significantly outperforms the recent method GMM in generating more informative multi-view summaries. The F-measure of our method is about 11% better than that of GMM [43]. This indicates that the proposed method is very effective and can be applied to large scale problems in practice. We follow the evaluation strategy of [43] and compute the performance measures in the unit of frames instead of events as in Table II to make a fair comparison with the GMM baseline.

### E. Comparison with Single-view Methods.

*Goal.* The objective of this experiment is to compare our method with some single-view summarization approaches to show their performance on multi-view videos. Specifically, the purpose of comparing with single-view summarization methods is to show that techniques that attempt to find summary from single-view videos usually do not produce an optimal set of representatives while summarizing multiple videos.

*Compared Methods.* We compare our approach with several baseline methods (`Attention-Concate` [37], `Sparse-Concate` [8], `Concate-Attention` [37], `Concate-Sparse` [8], `Graph` [51]) that use single-video summarization approach over multi-view datasets to generate summary. Note that in the first two baselines (`Attention-Concate`, `Sparse-Concate`), a single-video summarization approach is first applied to each view and then resulting summaries are combined to form a single summary, whereas the other three baselines (`Concate-Attention`, `Concate-Sparse`, `Graph`) concatenate all the views into a single video and then apply a single-video approach to summarize multi-view videos. Both `Sparse-Concate` and `Concate-Sparse` baselines use (11) to summarize multi-view videos with out any embedding.

| E1 | E2 | E3 | E4 | E6 | E7 | E8 | E9 | E10 | E11 |
|----|----|----|----|----|----|----|----|-----|-----|

| V1 | V2 | V1 | V2 | V1 | V2 | V4 | V3 | V3 | V1 |
|----|----|----|----|----|----|----|----|----|----|

Fig. 3. Summarized events for the Office dataset. Each event is represented by a key frame and is associated with two numbers, one above and below of the key frame. Numbers above the frame (E1, $\cdots$, E26) represent the event number whereas the numbers below (V1, $\cdots$, V4) indicate the view from which the event is detected. Limited to the space, we only present 10 events arranged in temporal order, as per the ground truth in [14].

The purpose of comparing with these two baseline methods is to explicitly show the advantage of our proposed multi-view embedding in generating informative and diverse summaries while summarizing multi-view surveillance videos.

*Implementation Details.* We implement Sparse-Concate and Concate-Sparse ourselves with the same temporal segmentation and C3D feature representation as the proposed one whereas for rest of the single-view summarization methods, we report the available results from the published papers [14], [28].

*Results.* We have the following key findings from Table II and Fig. 4: (i) The proposed method significantly outperforms all the compared single-view summarization methods by a significant margin on all three datasets. We observe that directly applying these methods to summarize multiple videos produces a lot of redundant shots which deviates from the fact that the optimal summary should be diverse and informative in describing the multi-view concepts. (ii) It is clearly evident from the Fig. 4 that both of the sparse representative selection based single-view summarization methods (Sparse-Concate and Concate-Sparse) produce a lot of redundancies (simultaneous presence of most of the events) while summarizing videos on Lobby dataset. This is expected since both of the approaches fail to exploit the complicated inter-view content correlations present in multi-view videos. (iii) By using our multi-view video summarization method, such redundancy is largely reduced in contrast. Some events are recorded by the most informative summarized shots, while the most important events are reserved in our summaries. The proposed approach generates highly informative and diverse summary in most cases, due to its ability to jointly model multi-view correlations and sparse representative selection.

### F. Scalability in Generating Summaries.

Scalability in generating summaries of different length has shown to be effective while summarizing single videos [23], [46]. However, most of the prior multi-view summarization methods require the number of shots to be specified before generating summaries which is highly undesirable in practical applications. Concretely speaking, the algorithm need to be rerun for each change in the number of representative shots that the user want to see in the summary. By contrast, our approach provides scalability in generating summaries of different length based on user constraints without any further analysis of the input videos (*analyze once, generate many*). This is due to the fact that non-zero rows of the sparse coefficient matrix $Z$ can generate a ranked list of representatives which can be subsequently used to provide a scalable representation in generating summaries of desired length without incurring

any additional cost. Such a scalability property makes our approach more suitable in providing human-machine interface where the summary length is changed as per the user request. Fig. 5 shows the generated summaries of length 3, 5 and 7 most important shots (as determined by the weight curve described in Sec. V) for Office dataset.

### G. Performance Analysis with Shot-level C3D Features.

We investigate the importance and reliability of the proposed video representation based on C3D features by comparing with 2D shot-level deep features, and found that the later produces inferior results, with a F-measure of 84.01% averaged over three datasets (Office, Campus and Lobby) compared to 86.55% by the C3D features. We utilize Pycaffe with the VGG net pretrained model [59] 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. III-A. The spatio-temporal C3D features perform best, as they exploit the temporal aspects of activities typically shown in videos.

### H. Performance Analysis with Video Segmentation.

We examined the performance of our approach by replacing the temporal segmentation algorithm [7] by a naive approach that uniformly divides video into several segments of equal length. We use uniform segments with a length of 2 seconds and kept other components fixed while generating summaries. By using the video segmentation algorithm of [7], the proposed approach achieves a F-measure of 86.55% averaged over three datasets (Office, Campus and Lobby). On the other hand, with the use of uniform length segments, our approach obtains a mean F-measure 85.43%. This shows that our approach is relatively robust with the change in segmentation algorithm. Note that our proposed sparse optimization is highly flexible to incorporate more sophisticated temporal segmentation algorithms, e.g., [52] in generating video summaries—we expect such advanced and complex video segmentation algorithms will only benefit our proposed approach.

### I. Performance Comparison with [45].

We now compare the proposed approach with [45] to explicitly verify the effectiveness of video representation and joint optimization for summarizing multi-view videos. Table IV shows the comparison with [45] on Office, Campus and Lobby datasets. Following are the analysis of the results: (i) The proposed framework consistently outperforms [45] on all three datasets by a margin of about 5% in terms of F-measure (maximum improvement of 8% in terms of precision for the office dataset). (ii) We improve around 3% in terms of F-measure for

Fig. 4. Some summarized events for the Lobby dataset. Top row: summary produced by `Sparse-Concate` [8], Middle row: summary produced by `Concate-Sparse` [8], and Bottom row: summary produced by our approach. It is clearly evident from both top and middle rows that both of the single-view baselines produce a lot of redundant events as per the ground truth [14] while summarizing multi-view videos, however, our approach (bottom row) produces meaningful representatives by exploiting the content correlations via an embedding. Redundant events are marked with same color borders. Note that both `Sparse-Concate` and `Concate-Sparse` summarize multiple videos without any embedding by either applying sparse representative selection to each video separately or concatenating all the videos into a single video. Best viewed in color.
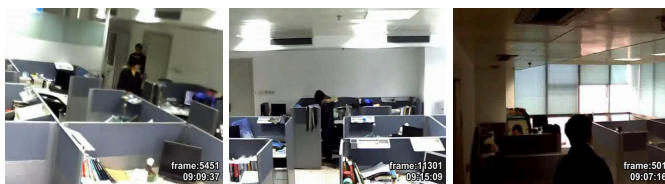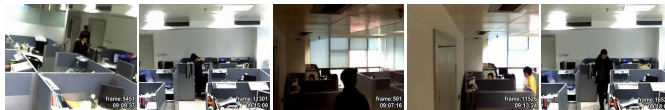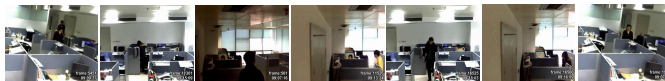


Fig. 5. The figure shows an illustrative example of scalability in generating summaries of different length based on the user constraints for the Office dataset. Each shot is represented by a key frame and are arranged according to the $l_2$ norms of corresponding non-zero rows of the sparse coefficient matrix. (a): Summary for user length request of 3, (b): Summary for user length request of 5 and (c): Summary for user length request of 7.

TABLE IV
F-MEASURE COMPARISON WITH [45]

| Methods | Office | Campus | Lobby | Reference |
|---------|--------|--------|-------|-----------|
| [45] | 84.48 | 75.42 | 88.26 | ICPR2016 [45] |
| Ours | 89.36 | 77.78 | 92.52 | Proposed |

TABLE V
USER STUDY—MEAN EXPERT RATINGS ON A SCALE OF 1 TO 10. OUR APPROACH SIGNIFICANTLY OUTPERFORMS OTHER AUTOMATIC METHODS.

| Methods | Office | Campus | Lobby | Road | Badminton |
|---------|--------|--------|-------|------|-----------|
| RandomWalk | 6.3 | 5.2 | 6.6 | 5.7 | 6.5 |
| BipartiteOPF | 7.1 | 5.8 | 7.4 | 6.0 | 7.2 |
| Ours | 7.6 | 6.5 | 8.2 | 6.7 | 7.9 |

video skims with motion information seems better suited for obtaining significant information in short time.

*J. User Study.*

With 5 study experts, we performed human evaluation of the generated summaries to verify the results obtained from the automatic objective evaluation with F-measure. Our objective is to understand how an user perceive the quality of the summaries according to the visual pleasantness and information content of the system generated summary. Each study expert watched the videos at 3x speed and were then shown 3 sets of summaries constructed using different methods: `RandoWalk`, `BipartiteOPF` and `Ours` for 5 datasets (Office, Campus, Lobby, Road and Badminton). Study experts were asked to rate the overall quality of each summary by assigning a rating from 1 to 10, where 1 corresponded to "The generated summary is not at all informative" and 10 corresponded to "The summary very well describes all the information present in the original videos and also visually pleasant to watch". The summaries were shown in random order without revealing the identity of each method and the audio track was not included to ensure that the subjects chose the rating based solely on visual stimuli. The results are summarized in Table V. Similar to the objective evaluation, our approach significantly outperforms both of the methods (`RandomWalk`, `BipartiteOPF`). This again corroborates the fact that the proposed framework generates a more informative and diverse multi-view summary as compared to the state-of-the-art methods. Furthermore, we note that the relative rank of the different algorithms is largely preserved in the subjective user study as compared to the objective evaluation in Table II.

the more challenging Campus dataset which demonstrates that the current framework is more effective in summarizing videos with outdoor scenes. (iii) We believe the best performance in the proposed framework can be attributed to two factors working in concert: (a) more flexible and powerful video representation via C3D features, and (b) joint embedding learning and sparse representative selection. Moreover, to better understand the contribution of joint optimization, we analyzed the performance of the proposed approach with shot-level C3D features and a 2 step process similar to [45], and found that the mean F-measure on three datasets (Office, Campus and Lobby) decreases from 86.55% to 83.85%. We believe this is because adaptively changing the graph Laplacian with respect to the sparse representative selection helps in better exploiting the multi-view correlations and also indicates the requirement of optimal representative shots to be included in the summary. It also important to note that the approach in [45] is limited to key frame extraction only and hence may not be suitable for many surveillance applications where

### K. Discussions.

*Abnormal Event Detection.* Abnormal event detection and surveillance video summarization are two closely related problem in computer vision and multimedia. In a surveillance setting, where an abnormal event took place, the proposed approach can select shots to represent the abnormal event in the final summary. This is due to the fact that our approach selects representative shots from the mult-view videos such that set of videos should be reconstructed with high accuracy using the extracted summary. Specifically, the proposed approach in (13) favors selecting a set of shots as representatives for constructing the summary which can reconstruct all the events in the input with low reconstruction error. Consider a simple example for an illustration. Let us assume a surveillance setting equipped in a place with only pedestrian traffic. People are walking as usual and suddenly, a car is speeding. In order to reconstruct the part where the car is speeding, our method will choose a few shots from this portion; otherwise the reconstruction error will be high.

*Multi-View Event Capture.* In general, the purpose of overlapping-field of view is to facilitate users to check objects/events from different angles. For an event captured with multiple cameras having a large difference in view angles, the proposed method often selects more than one shot to represent the event in the summary. This is due to the fact that our approach selects representative shots from the multi-view videos such that the whole input can be reconstructed with low error. In our experiments, we have observed a similar situation while summarizing videos on Campus dataset. The summary produced by our approach contains three shots captured with cameras 1, 3, and 4 in an outdoor environment which essentially represent the same event (E23 in the ground truth [14]). However, note that although including shots representing same event from more than one camera in the summary may help an user to check events from different angles, it increases the summary length which often deviates from the fact that length of the summary should be as small as possible. Thus, the objective of our current work is on generating an optimal summary that balances the two main important criteria of a good summary, i.e., maximizing the information content via representativeness and minimizing the length via sparsity.

*Joint Video Segmentation and Summarization.* Note that the proposed approach uses temporal video segmentation as a preprocessing step and then use the shot-level features to extract summaries. Our approach can be modified in two ways to optimize the temporal segmentation for the task of video summarization. First, involving a human in our current approach for giving feedbacks, similar to the concept of relative attributes in visual recognition [49] can help us in adaptively changing the shot boundaries for generating better quality summaries. Second, learning a dynamic agent using Markov decision process (MDP) for moving the shot boundaries (forward or backward with temporal increments) based on the performance of our proposed summarization algorithm is also a possibility in this regard [4]. Developing an efficient framework for joint segmentation and summarization is an interesting practical problem—we leave this as future work, with no existing work, to the best of our knowledge.

## VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we addressed the problem of summarizing multi-view videos via joint embedding learning and $\ell_{2,1}$ sparse optimization. The embedding helps is capturing content correlations in multi-view datasets without assuming any prior correspondence between the individual videos. On the other hand, the sparse representative selection helps in generating multi-view summaries as per user length request without requiring additional computational cost. Performance comparisons on six standard multi-view datasets show marked improvement over some mono-view summarization approaches as well as state-of-the-art multi-view summarization methods.

Moving forward, we would like to improve our method by explicitly incorporating video semantics that may require more complex model with additional techniques such as attention modeling [37] or semantic feature analysis based on user preferences [34]. It is also important to note that unlike single-view videos, understanding semantics in a multi-camera environment is a challenging problem and hence, one may also require data association strategies, e.g, [5] to properly exploit multi-view video semantics for the task of video summarization. Moreover in future, we would like to consider the fact that more than one camera view may be necessary to fully represent an event (e.g., due to occlusion) in a multi-view setting and hence, it may be necessary to include multiple similar shots representing same events from more than one camera for generating a good quality video summary.

## REFERENCES

[1] J. Almeida, N. J. Leite, and R. da S. Torres. VISON: VIdeo Summarization for ONline applications. *PRL*, 2012.
[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.
[3] D. Cai, X. He, and J. Han. Spectral regression for efficient regularized subspace learning. In *ICCV*, 2007.
[4] J. C. Caicedo and S. Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, 2015.
[5] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury. Network consistent data association. *TPAMI*, 2016.
[6] C. Chennubhotla and A. Jepson. Sparse coding in practice. In *Proc. of the Second Int. Workshop on Statistical and Computational Theories of Vision*, 2001.
[7] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.
[8] Y. Cong, J. Yuan, and J. Luo. Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection. *TMM*, 2012.
[9] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de A. AraÃžjo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *PRL*, 2011.
[10] F. Dornaika and I. K. Aldine. Decremental sparse modeling representative selection for prototype selection. *PR*, 2015.
[11] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012.
[12] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI*, 2013.
[13] S. Feng, Z. Lei, and S. Li. Online content-aware video condensation. In *CVPR*, 2012.
[14] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou. Multi View Video Summmarization. *TMM*, 2010.
[15] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely–laplacian sparse coding for image classification. In *CVPR*, 2010.
[16] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *TPAMI*, 1992.

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

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

[19] M. Gygli, H. Grabner, and L. V. Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015.

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

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

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

[23] L. Herranz and J. M. Martinez. A framework for scalable summarization of video. *TCSVT*, 2010.

[24] T. Huang. Surveillance video: the biggest big data. *Computing Now*, 2014.

[25] 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.

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

[27] 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.

[28] S. Kuanar, K. Ranga, and A. Chowdhury. Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. *TMM*, 2015.

[29] S. K. Kuanar, R. Panda, and A. Chowdhury. Video Key frame Extraction through Dynamic Delaunay Clustering with a Structural Constraint. *JVCIR*, 2013.

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

[31] C. D. Leo and B. S. Manjunath. Multicamera video summarization from optimal reconstruction. In *ACCV Workshop*, 2011.

[32] C. D. Leo and B. S. Manjunath. Multicamera Video Summarization and Anomaly Detection from Activity Motifs. *TOSN*, 2014.

[33] P. Li, Y. Guo, and H. Sun. Multi key-frame abstraction from videos. In *ICIP*, 2011.

[34] W.-N. Lie and K.-C. Hsu. Video summarization based on semantic feature analysis and user preference. In *SUTC*, 2008.

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

[36] X. Lu, Y. Yuan, and P. Yan. Image super-resolution via double sparsity regularized manifold learning. *TCSVT*, 2013.

[37] 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.

[38] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng. Video summarization via minimum sparse reconstruction. *PR*, 2015.

[39] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *CVPR*, 2016.

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

[41] H. Nguyen, V. Patel, N. Nasrabadi, and R. Chellappa. Sparse embedding: A framework for sparsity promoting dimensionality reduction. *ECCV*, 2012.

[42] F. Nie, H. Wang, H. Huang, and C. Ding. Unsupervised and semi-supervised learning via $\ell_1$-norm graph. In *ICCV*, 2011.

[43] S.-H. Ou, C.-H. Lee, V. Somayazulu, Y.-K. Chen, and S.-Y. Chien. On-Line Multi-View Video Summarization for Wireless Video Sensor Network. *JSTSP*, 2015.

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

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

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

[47] R. Panda and A. K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017.

[48] R. Panda and A. K. Roy-Chowdhury. Sparse modeling for topic-oriented video summarization. In *ICASSP*, 2017.

[49] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

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

[51] Y. Peng and C.-W. Ngo. Clip-based similarity measure for query-dependent clip retrieval and video summarization. *TCSVT*, 2006.

[52] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric

[54] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis:

videos. In *CVPR*, 2014.

[53] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.
Peeking around the world. In *ICCV*, 2007.

[55] A. K. Roy-Chowdhury and B. Song. Camera networks: The acquisition and analysis of videos over wide areas. *Synthesis Lectures on Computer Vision*, 2012.

[56] 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.

[57] M. M. Salehin and M. Paul. *Fusion of Foreground Object, Spatial and Frequency Domain Motion Information for Video Summarization*. Springer International Publishing, 2016.

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

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

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

[61] 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.

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

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

[64] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua. Event driven web video summarization by tag localization and key-shot identification. *TMM*, 2012.

[65] M. Wang, G. Li, Z. Lu, Y. Gao, and T.-S. Chua. When amazon meets google: Product visualization by exploring multiple web sources. *TOIT*, 2013.

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

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

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

[69] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarizatio. *CVPR*, 2016.

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

[71] X. Zhu, C. C. Loy, and S. Gong. Learning from multiple sources for video summarisation. *IJCV*, 2016.

**Rameswar Panda** received his Bachelors and Masters degree in Electronics and Telecommunication engineering from Biju Patanaik University of Technology, India and Jadavpur University, India in 2011 and 2013 respectively. He is currently pursuing the Ph.D. degree in the department of Electrical and Computer Engineering at University of California, Riverside. His main research interests include computer vision, machine learning, video summarization, person re-identification and video surveillance.

**Amit K. Roy-Chowdhury** received the Bachelors degree in Electrical Engineering from Jadavpur University, Calcutta, India, the Masters degree in systems science and automation from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in Electrical Engineering from the University of Maryland, College Park. He is a Professor of Electrical and Computer Engineering at the University of California, Riverside and a Cooperating Faculty in the department of Computer Science and Engineering. He leads the Video Computing Group at UCR, with research interests in computer vision, image processing, pattern recognition, and statistical signal processing. He is coauthor of the monograph "Camera Networks: The Acquisition and Analysis of Videos over Wide Areas". He has been on the organizing and program committees of multiple conferences and serves on the editorial boards of a number of journals. He is a Fellow of IAPR.