Diversity-aware Multi-Video Summarization

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

***Abstract*—Most video summarization approaches have focused on extracting a summary from a single video; we propose an unsupervised framework for summarizing a collection of videos. We observe that each video in the collection may contain some information that other videos do not have, and thus exploring the underlying complementarity could be beneficial in creating a diverse informative summary. We develop a novel diversity-aware sparse optimization method for multi-video summarization by exploring the complementarity within the videos. Our approach extracts a multi-video summary which is both interesting and representative in describing the whole video collection. To efficiently solve our optimization problem, we develop an alternating minimization algorithm that minimizes the overall objective function with respect to one video at a time while fixing the other videos. Moreover, we introduce a new benchmark dataset, Tour20, that contains 140 videos with multiple human created summaries, which were acquired in a controlled experiment. Finally, by extensive experiments on the new Tour20 dataset and several other multi-view datasets, we show that the proposed approach clearly outperforms the state-of-the-art methods on the two problems—topic-oriented video summarization and multi-view video summarization in a camera network.**

***Index Terms*—Video summarization; Sparse optimization.**

I. Introduction

With the recent explosion of big video data, it is becoming increasingly important to automatically extract a brief yet informative summary of these videos 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.

Although video summarization has been extensively studied during the past few years, many previous methods mainly focused on summarizing a *single video* by developing a variety of selection criteria (e.g., representativeness [18], [79], [8], interestingness [46], [26]) to prioritize frames/segments for the output summary. Another important problem and rarely addressed in this context is to find an informative summary from *multiple* videos. Similar to single video summarization problem, the *multi-video summarization* approach seeks to take a set of related videos and extracts key frames/video skims that presents the most important portions of the input videos within a short duration. Application areas include any scenarios where the user is confronted with watching or browsing a set of related videos, like videos given by a search [42], [73], [77] or videos captured with multiple video sensors in a camera network [20], [36], [55], [56]. Given that browsing through all the videos is a very time consuming task, we want to explore whether we can automatically create a video summary that can describe the whole video collection within a short duration.

• Rameswar Panda, Niluthpol Chowdhury Mithun 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, nmithun@ece.ucr.edu, amitrc@ece.ucr.edu)

Multi-video summarization is related to the general problem of single-video summarization with two important distinctions. First, these videos are topically related and hence inter-video statistical dependencies need to be properly exploited for obtaining an informative and diverse summary. Second, different environmental factors like difference in illumination, pose and synchronization issues across the multiple topic-related videos also pose a challenge in summarizing such videos. Thus, direct use of methods that attempt to extract summary from single videos may not produce an optimal set of representatives while summarizing multiple topic-related videos.

To address the challenges encountered in a multi-video setting, we propose a Diversity-aware Multi-Video Summarization (DiMS) approach to generate an informative summary by exploring the complementarity between a set of videos. We observe that each video in the set may contain some information that other videos do not have, and thus exploring the underlying complementarity is of great importance for the success of multi-video summarization. We achieve this by developing a novel sparse optimization that jointly summarizes a set of videos to find a single summary that can optimally describe the video collection. Our summarization approach consider two aspects. One, it considers "interestingness" prior in the sparse representative selection to extract summary that is both interesting and representative of the input video. In particular, segments with high interestingness score are more likely to be selected as key video segments compared to the segments with low interestingness score. Second, we introduce a diversity regularizer in the optimization framework to explore the complementarity within multiple videos in extracting a high quality multi-video summary. We finally develop an efficient alternating minimization algorithm to solve our optimization problem. Furthermore, rather than manually evaluating the produced summaries, we introduce a new benchmark dataset with multiple ground truth summaries for each video as well as for the video collection. This data allows to asses the performance of any single-video or multi-video summarization algorithm in a fast and repeatable manner.

Contributions: We address an important, and practical problem in this paper—how to extract an informative yet diverse video summary from a collection of videos. Towards solving this problem, we make the following contributions. (1) we propose an unsupervised approach for multi-video summarization by exploring the complementarity within a set of videos; (2) we develop a novel diversity-aware sparse optimization method that can be efficiently solved by an alternating minimization algorithm; (3) we introduce a new dataset, Tour20, along with clear ground truth summaries to evaluate summarization algorithms in a fast and repeatable manner. To the best of our knowledge, this is the biggest dataset for summarization available. (4) we show the effec-

tiveness of our approach in two tasks—topic-oriented video summarization and multi-view video summarization in a camera network. With extensive experiments on both Tour20 and several standard multi-view datasets, we show the superiority of our approach over competing methods for both of the tasks.

II. Related Work

There is a rich body of literature in image processing and computer vision on summarizing videos in form of a key frame sequence or a video skim. It is beyond the scope of this paper to do a comprehensive review. Interested readers can check [50], [71] for a more comprehensive summary. Roughly, all these summarization methods can be divided into two categories: single-video and multi-video summarization.

Single-Video Summarization: Much progress has been made in developing a variety of ways to summarize a single video in an unsupervised manner or developing supervised algorithms. Representative methods along the direction of supervised algorithms use category-specific classifiers for importance scoring [60], [68] or learn how to select informative and diverse video subsets from human-created summaries [25], [23], [76] or learn important facets, like faces, hands, objects, diversity [37], [45], [2]. Although these supervised techniques have shown impressive results, their performance largely depends on huge amount of labeled examples which are difficult to collect in many cases. Nevertheless, it is generally feasible to have only a limited number of users to annotate training videos, which may lead to a biased summarization model.

Without supervision, summarization methods rely on low-level visual indices to determine the important parts of a video. Various strategies have been studied, including clustering [12], [24], [29], [57], maximal biclique finding [7], interest prediction [46], [26], and energy minimization [61], [19]. Leveraging crawled web images or videos is also another recent trend for video summarization [32], [67], [33], [58].

Recently, there has been a growing interest in using sparse coding (SC) to solve the problem of video summarization [18], [79], [8], [48], [15] since the sparsity and reconstruction error term in SC naturally fits into the problem of summarization. Another recent work [17] finds a subset of the source set to efficiently describe the target set, given pairwise dissimilarities between two sets. In contrast to these prior works that can only summarize a single video, we develop a multi-video summarization method that jointly summarizes a set of videos to find a single summary for describing the collection altogether. Moreover, we consider interestingness of segments along with representativeness in the sparse optimization to extract summaries that are both interesting and representative.

Multi-Video Summarization: Generating a summary from multiple videos is a more technically challenging problem due to the inevitable thematic diversity and content overlaps within multiple videos than a single video. Generally, the applications of multi-video summarization can be roughly divided into two categories. The first category is to summarize a group of topically related web videos given by a search. Some of early works in this category focused on videos of specific genres, such as TV news [42], [73] and generated an automatic summary by frame clustering [74] or leveraging genre specific information, e.g., speech transcripts in news [41], [64]. However, they generally fail to summarize large scale open world web videos since they are unstructured and range over a wide variety of content. A system for rapid browsing of multiple videos are proposed in [9]. A recent approach to the problem of summarizing multiple sensor-rich videos in geo-space can be seen in [78]. A supervised approach to summarize multiple videos captured with hand-held devices is presented in [77]. However, these systems relies on meta-data sensor information or semantics related to a geographical area (e.g., weather and lighting condition) which are mostly unavailable while summarizing unconstrained web videos.

The other category of multi-video summarization is to summarize videos captured with video sensors at the same time with overlapped or partially overlapped field-of-views in a camera network. Representative methods in this category use random walk over spatio-temporal shot graphs [20] and rough sets [40] to summarize multi-view videos. A recent work in [36] uses bipartite matching constrained optimum path forest clustering to solve the problem of summarizing multi-view videos. An online method for summarization can also be found in [53]. In [38], [39], summarization is performed by detecting abnormal events between sensors in a non-overlapping camera network.

Since both of the categories of multi-video summarization are inherently related, we develop, to our best knowledge, the first generalized framework to extract an informative summary by exploring the complementary information within multiple videos. We demonstrate the generalizability of our framework with extensive experiments on several datasets.

III. Diversity-aware Multi-Video Summarization

In this section, we start by giving notations and definitions of the main concepts of our approach, and then present our detailed approach to summarize multiple videos.

Notation: We use uppercase letters to denote matrices and lowercase letters to denote vectors. For matrix $A = (a_{ij})$, 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 ℓ_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 ℓ_0-norm is defined as $||a||_0 = \sum_{i=1}^n |a_i|^0$. The Frobenius norm of $A \in \mathbb{R}^{n\times m}$ is defined as $\sqrt{\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2}$. The $\ell_{2,1}$-norm can be generalized to $\ell_{r,p}$-norm which is defined as $||A||_{r,p} = (\sum_{i=1}^n ||a_i||_r^p)^{1/p}$. When $r \geq 1$ and $p \geq 1$, the $\ell_{r,p}$-norm is a valid norm since it satisfies the three basic conditions of a norm including the triangle inequality $||A||_{r,p} + ||B||_{r,p} \geq ||A + B||_{r,p}$. However, when $r < 1$ or $p < 1$, $\ell_{r,p}$-norm is not valid as well as the ℓ_0, but we also call them norms for convenience. The operator $diag(.)$ puts a vector on the main diagonal of a matrix. 1 denotes a vector whose elements are equal to one.

Video Summary: Given a set of videos, our goal is to find a summary that conveys the most *important* details of the original video collection. Specifically, it is composed of several video segments that represent most important portions of the input video collection within a short duration. Since, *importance* is a subjective notion, we define a good summary as one that has the following properties:

• *Representativeness.* The set of videos should be reconstructed with high accuracy using the extracted summary.
• *Interestingness.* The summary should contain the most interesting parts of the input videos, e.g., in a collection of videos related to *Eiffel Tower*, one does not want to miss a segment that depicts the colorful night view of the tower.
• *Sparsity.* Although the summary should be representative and interesting, the length should be as small as possible.
• *Diversity.* The summary should be diverse as much as possible capturing different aspects of the input video collection. In other words, the amount of content redundancy should be small in the final set of extracted summaries.

We develop a diversity-aware sparse optimization framework to generate a multi-video summary that characterizes all the above desirable properties of an optimal summary. The proposed approach, DiMS, decomposes into three steps: i) video representation; ii) diversity-aware sparse representative selection; iii) summary generation.

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.

1) Temporal Segmentation: Our approach starts with segmenting videos using an existing algorithm [7]. We divide each video into multiple non-uniform segments by measuring the amount of changes between two consecutive frames in the RGB and HSV color spaces [3]. A segment 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 segmentation algorithm to ensure that the number of frames within each segment lies in the range of [32,96]. The video segments serve as the basic units for feature extraction and subsequent processing to extract a video summary.

2) Feature Representation: Deep convolutional neural networks (CNNs) have been successful at large-scale object recognition [35]. Beyond the object recognition task itself, recent advancement in deep 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 [65], [31]. An advantage of using deep learning features is that there exist accurate, large-scale datasets such as Imagenet [62], and Sports-1M [31] from which they can be extracted. In addition, GPU-based extraction of such features are much faster than that for the traditional hand crafted features such as CENTRIST, Dense-SIFT.

In the case where the input is a video clip, C3D features [70] have recently shown better performance compared to the features extracted using each frame separately [70]. We therefore extract C3D features, by taking sets of 16 input frames, applying 3D convolutional filters, and extracting the responses at FC6 layer as suggested in [70]. This is followed by a temporal mean pooling scheme to maintain the local ordering structure within a video segment. Then the pooling result serves as the final feature vector of a video segment (4096 dimensional) to be used in the sparse optimization. We will discuss the performance benefits of employing C3D features later in our experiments.

Note that in our current work, we did not consider the audio information while representing videos. However, we believe that audio (if available) can be used as a potential side information along with visual features to select important segments from a video. One can easily incorporate audio features in our framework by combining both audio and visual features to represent a video segment or following aggregation mechanisms similar to [41], [28]—we leave this as an interesting direction for future research. Our proposed sparse optimization approach as described in next section, is quite flexible in handling multi-modal information while summarizing videos—we expect more sophisticated ones will only benefit our approach.

B. Diversity-aware Sparse Representative Selection

We develop a sparse optimization framework that jointly summarizes a set of videos to extract a summary that describes the collection together. Consider a set of m relevant videos given by a video search or generated from a multi-view camera network, where $X^{(v)} = \{x^i \in \mathbb{R}^d, i = 1, \cdots, n_v\}, v = 1, \cdots, m$. Each x^i represents the feature descriptor of a segment in d-dimensional feature space. We represent each video segment by extracting C3D features as described above.

1) Formulation: Sparse optimization approaches [8], [18] find the representative segments from a single video $X^{(v)}$ by minimizing the linear reconstruction error as

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 \; s.t. \; \|Z^{(v)}\|_{2,0} \leq k, \; {Z^{(v)}}^T 1 = 1 \quad (1)$$

The constraint on $\ell_{2,0}$ norm of $Z^{(v)}$ implies that only k video segments are chosen as the representative whereas the affine constraint ${Z^{(v)}}^T 1 = 1$ makes the selection of representatives invariant with respect to the global translation of the data.

This is an NP-hard problem since it requires searching over every subset of the k columns of $X^{(v)}$. A standard ℓ_1 relaxation to the problem (1) is given by

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 \; s.t. \; \|Z^{(v)}\|_{2,1} \leq \tau, \; {Z^{(v)}}^T 1 = 1 \quad (2)$$

where $\|Z^{(v)}\|_{2,1} = \sum_{i=1}^{n_v} ||z_i^{(v)}||_2$ and $\tau > 0$ controls the level of sparsity in the reconstruction.[1] Using Lagrange multiplier, the optimization problem (2) can be written as,

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda_s^{(v)}\|Z^{(v)}\|_{2,1} \; s.t. \; {Z^{(v)}}^T 1 = 1 \quad (3)$$

where $\lambda_s^{(v)}$ is a regularization parameter. Once problem (3) is solved, the summary is generated by selecting segments whose corresponding $||Z_i^{(v)}||_2 \neq 0$. We keep the constraint ${Z^{(v)}}^T 1 = 1$ since it can be easily handled as we will show later.

Introducing Interestingness of Video Segments: Note that in problem (3), all segments are treated equally without considering the interestingness of some specific segments. Specifically, sparse optimization approaches [8], [18] only characterizes the reconstruction capability and sparsity but does not account for the fact that the selected video segments should be interesting.

[1]Note that we use τ instead of k since $\|Z\|_{2,1}$ is not necessarily bounded by k after the relaxation.

As a result, it may leave out some crucial segment(s) in the summary. A good summarization method can certainly benefit from incorporating such interestingness prior knowledge from application domain or user specifications. To better leverage interestingness along with representativeness, we propose a simple extension to (3) as follows [48]:

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda_s^{(v)}\|Q^{(v)}Z^{(v)}\|_{2,1} \ s.t. \ \ {Z^{(v)}}^T 1 = 1 \tag{4}$$

where $Q^{(v)} = [diag(q^v)]^{-1}$ and $q^v \in \mathbb{R}^{n_v}$ represent the interstingness score of each video segment. It is easy to see that problem (4) favors selection of interesting segments by assigning a lower score via $Q^{(v)}$. Thus, given a video, minimization of (4) selects segments that are both interesting and representative. More details on the video interestingness prior are presented in Sec. IV.

The sparse optimization (4) extracts a good summary from a single video. However, summarizing multiple videos is ubiquitous in video search or in a camera network, hence, extending (3) into multi-video setting is of vital importance for many multimedia applications. One direct way to extend into multi-video setting is to apply (3) to each of the video, and then combine the results to produce a single summary. Mathematically, we have the naïve multi-video summarization approach as follows:

$$\min_{Z^{(1)},Z^{(2)},\cdots,Z^{(m)}} \sum_{v=1}^{m}\|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \sum_{v=1}^{m}\lambda_s^{(v)}\|Q^{(v)}Z^{(v)}\|_{2,1}$$
$$s.t. \ \ {Z^{(v)}}^T 1 = 1, \ Z^{(v)} \in \mathbb{R}^{n_v \times n_v}, \ \forall \ 1 \le v \le m \tag{5}$$

This approach summarizes videos independently without considering complementarity of different videos, hence, produces redundant information in the final summary.

Introducing Complementarity of Multiple Videos: The objective function (5) summarizes multiple videos independently, without any constraint. Considering the presence of complementary information within multiple videos, we introduce a diversity regularization function to select a sparse set of representative and diverse video segments. Specifically, to explore the complementary information, we enforce a regularizer that penalizes the condition that two correlated segments from two distinct videos are present in the summary at the same time. For example, if the i-th segment from v-th video is highly correlated to the j-th segment in w-th video, then we do not need to select both of them simultaneously.

Definition 1. Given the sparse coefficient matrices $Z^{(v)}$ and $Z^{(w)}$, the diversity regularization function is defined as:

$$f_d(Z^{(v)}, Z^{(w)}) = \sum_{i=1}^{n_v}\sum_{j=1}^{n_w}\|z_i^{(v)}\|_2 c_{ij} \|z_j^{(w)}\|_2 = \|W^{(vw)}Z^{(v)}\|_{2,1} \tag{6}$$

where c_{ij} measure the correlation between i-th segment from v-th video and the j-th segment in w-th video. The second equality follows from the simple manipulation as $W_{ii}^{(vw)} = \sum_{j=1}^{n_w} c_{ij}\|z_j^{(w)}\|_{2,1}$. More details on correlation between different video segments are given in Sec. IV.

Minimization of (6) tries to explore the complementarity by penalizing the condition that rows of two similar video segments from two distinct videos are nonzero at the same time. This amounts to enforcing the sparse coefficient matrices of different videos to be of maximum diversity.

Overall Objective Function: After adding the diversity regularization function into problem (5), we have the final objective function as follows:

$$\min_{Z^{(1)},Z^{(2)},\cdots,Z^{(m)}} \sum_{v=1}^{m}\|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda_s \sum_{v=1}^{m}\|Q^{(v)}Z^{(v)}\|_{2,1}$$
$$+\lambda_d \sum_{\substack{1\le v,w\le m \\ v\ne w}} f_d(Z^{(v)}, Z^{(w)})$$
$$s.t. \ \ {Z^{(v)}}^T 1 = 1, \ Z^{(v)} \in \mathbb{R}^{n_v \times n_v}, \ \forall \ 1 \le v \le m \tag{7}$$

where λ_s and λ_d are two trade-offs associated with the sparsity and diversity regularization functions respectively.

2) Optimization: It is difficult to solve the constrained problem (7). In this section, we propose an alternative algorithm to solve this optimization problem efficiently. With the alternating minimizing strategy, we can approximately solve (7) in the manner of minimizing with respect to one video once at a time while fixing the other videos. Specifically, we minimize the following objective function with respect to $Z^{(v)}$ while keeping all others fixed:

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda_s\|Q^{(v)}Z^{(v)}\|_{2,1}$$
$$+\lambda_d \sum_{w=1,v\ne w}^{m}\|W^{(vw)}Z^{(v)}\|_{2,1} \ s.t. \ \ {Z^{(v)}}^T 1 = 1 \tag{8}$$

To reformulate the problem (8), we need the following lemma.

Lemma 1. For any two diagonal positive semidefinite matrices $W^{(1)}, W^{(2)} \in \mathbb{R}^{n\times n}$, the following equality holds for any matrix $Z \in \mathbb{R}^{n\times n}$:

$$\|W^{(1)}Z\|_{2,1} + \|W^{(2)}Z\|_{2,1} = \|WZ\|_{2,1} \tag{9}$$

where $W = W^{(1)} + W^{(2)}$. The proof follows directly from the fact that $\ell_{2,1}$-norm is a valid norm and the equality in triangle inequality holds if both $W^{(1)}$ and $W^{(2)}$ are positive semidefinite matrices. □

From lemma 1, it is easy to reformulate problem (8) as following:

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda_s\|Q^{(v)}Z^{(v)}\|_{2,1} + \lambda_d\|W^{(v)}Z^{(v)}\|_{2,1}$$
$$s.t. \ \ {Z^{(v)}}^T 1 = 1 \tag{10}$$

where $W^{(v)} = \sum_{w=1,v\ne w}^{m} W^{(vw)}$. Note that both second and third term in (10) are functions of the same variable $Z^{(v)}$ with two trade-offs λ_s and λ_d respectively. From lemma 1, we can approximate (10) with one trade-off parameter λ as following:

$$\min_{Z^{(v)}} \|X^{(v)} - X^{(v)}Z^{(v)}\|_F^2 + \lambda\|K^{(v)}Z^{(v)}\|_{2,1} \ s.t. \ \ {Z^{(v)}}^T 1 = 1 \tag{11}$$

where $K^{(v)} = Q^{(v)} + W^{(v)}$. For convenience, ignoring the superscripts, we get

$$\min_{Z} \|X - XZ\|_F^2 + \lambda\|Z\|_{K,2,1} \ s.t. \ \ Z^T 1 = 1 \tag{12}$$

where $\|Z\|_{K,2,1}$ denotes the weighted $\ell_{2,1}$-norm of Z and is defined as $\|Z\|_{K,2,1} = \|KZ\|_{2,1}$. When we replace X with

$[X^T, \alpha * 1]^T$ where α approaches to infinity, (12) is equivalent to the following problem:

$$\min_{Z} \|X - XZ\|_F^2 + \lambda \|Z\|_{K,2,1} \tag{13}$$

We can prove equation (12) is equivalent to (13) by expanding (13) as follows:

$$\|X - XZ\|_F^2 = \|X^* - X^*Z\|_F^2 + \alpha \|1^T - 1^T Z\|_F^2 \tag{14}$$

where X^* is the original X presented in (12). When α approaches to infinity, $Z^T 1$ approaches to 1. Thus, problem (12) is equivalent to (13).

The objective function (13) is a convex weighted $\ell_{2,1}$-norm minimization problem which can be efficiently solved using Alternating Direction Method of Multipliers (ADMM) [4]. The ADMM procedure to solve (13) is summarized in Algo. 1. [2]

The above alternating procedure of DiMS is carried out until convergence, as shown in Algo. 2.

Algorithm 1 An ADMM solver for (13)

Input: Video feature matrix X, K and $\lambda, \mu > 0$
while *not converged* **do**
  $U \leftarrow (X^T X + \mu I)^{-1}(X^T X + \mu Z - \Lambda)$;
  $Z \leftarrow \max\left\{\|U + \Lambda/\mu\|_2 - \frac{\lambda K}{\mu}, 0\right\} \frac{U+\Lambda/\mu}{\|U+\Lambda/\mu\|_2}$ (row-wise);
  $\Lambda \leftarrow \Lambda + \mu(U - Z)$;
end while
Output: Sparse coefficient matrix Z.

Algorithm 2 Algorithm for solving (7)

Input: Video feature matrices $X^{(1)}, X^{(2)}, \cdots, X^{(m)}$
for *each* v **do**
  Initialize $Z^{(v)}$ by solving (5);
end for
while *not converged* **do**
  for *each* v **do**
    Obtain $Z^{(v)}$ by solving (13);
  end for
end while
Output: Coefficient matrices $Z^{(1)}, Z^{(2)}, \cdots, Z^{(m)}$.

C. Summary Generation

Above, we described how we compute the sparse coefficient matrices where the nonzero rows indicate the representatives for the summary. We follow the following rules to generate a summary of specified length: (i) We first sort the representative segments in a video $X^{(v)}$ by decreasing importance according to the ℓ_2 norms of the rows in $Z^{(v)}$ (resolving ties by favoring shorter video segments). (ii) We then sort the videos according to the number of nonzero rows in the corresponding sparse coefficient matrix (informative score) and compute the number of segments that should be selected from each video based on the relative score and user-defined summary length. (iii) Finally, we construct the video summary by placing the selected segments from the most informative video at the beginning and then appending segments from other videos based on the relative informative score.

[2]We provide details about the ADMM in the supplementary material. The supplementary material associated with this paper is available at http://www.ee.ucr.edu/~amitrc/publications.php

IV. Discussions

Interestingness of Video Segments: As existing approaches, we compute the interestingness score of each segment by taking into account the rest of the video segments. Specifically, we first compute the interestingness score of a segment as the sum of scores predicted for each frame that belong to the segment and then take the relative score over the maximum predicted score in a video. We follow [26] to compute the interestingness score of each frame by considering attention, aesthetic quality and presence of landmarks/persons. Note that these forms of interestingness prediction are often used in several vision tasks and are quite flexible [37], [16], [10], [13]. However, one can also learn a regression model to predict an interestingness score of domain relevance [25], [68], [75] or compute with user specifications via human in the loop [27]—we expect more sophisticated ones will only benefit our proposed approach. Concretely speaking, our method is not dependent on a particular definition on interestingness.

Correlation between Video Segments: There are a lot of ways to measure the correlation between two video segments c_{ij}. In this paper, we employ Scott and Longuet-Higgins (SLH) algorithm [63] with Gaussian kernel to measure the correlation, since it is simple to implement and it performs well in several vision tasks [54], [69]. Specifically, given the segment-level feature similarity matrix S, computed via a Gaussian kernel between two videos, SLH algorithm finds an orthonormal matrix C that permutes the rows of S in order to maximize its trace. Mathematically,

$$C = \underset{C^T C = I}{\arg\max}\ tr(C^T S) \tag{15}$$

Maximizing the above function is a singular value decomposition problem and the optimal solution is given by $C^* = UDV^T$, where the SVD decomposition of $S = UEU^T$ and D is obtained by replacing singular values of E by ones. We use the matrix C^* as the correlation matrix after setting the negative values to 0 [69]. We will discuss the performance benefits in employing such correlations compared to the cosine similarity in the experiments. It is also important to mention here that our proposed formulation (7) is highly flexible to incorporate any form of correlations defined between two video segments.

Sparsity Regularization Parameter: The regularization parameter λ in (13) puts a trade-off between two opposing terms: the reconstruction error and number of representative segments. In other words, we obtain a small reconstruction error by selecting more representative segments and vice versa. As indicated by the update equation of Z in Algo. 1, when λ is large enough, e.g., $\lambda \geq \lambda^{max}$, we get $Z = 0$ that means we select no representative segments. Thus, to avoid an empty selection, we let $\lambda \leq \lambda^{max}$ and obtain $\lambda^{max} = \max_{0 \leq i \leq n} ||x_i^T X||_2$, as in [18]. In our experiments, we let $\lambda = \frac{\lambda^{max}}{\alpha}$ and tune α between the interval [2,30] [18].

Intialization in Algo. 2: Since the alternating minimization can make the Algo. 2 stuck in a local minimum, it is important to have a sensible initialization. We initialize the sparse coefficient matrices of $m-1$ videos by solving (5) using Algo. 1, which is a special case (when $\lambda_d = 0$ in (7)) of our method. After the initialization, the following question remain: from which view we should start the alternating minimization? One

possible way is to randomly start with any video and repeat the minimization over all videos until convergence. However, since we have some prior knowledge on which video is more informative in the collection, we can start with initializing and fixing more informative videos, and optimize with respect to the least informative video. More specifically, we start with the specific $Z^{(v)}$ which has more number of nonzero rows after solving (5) since the number of nonzero rows indicate the relative importance of each video in the collection.

Stopping Criteria: In Algo. 1, the stop criteria is set to $||U^{(t)}-Z^{(t)}||_\infty \leq \epsilon$ *or* $t \geq 2000$, where t is the iteration number and ϵ is set to 10^{-7} throughout the experiments. Similarly, in Algo. 2, we set the stop criteria as $\frac{|f^{(t+1)}-f^{(t)}|}{f^{(t)}} < 10^{-2}$, where $f^{(t)}$ is the objective value in the t-th iteration.

Convergence Analysis: We can prove the convergence of the proposed Algo. 2 as follows: we divide the problem (7) into m number of subproblems and each of them is a convex problem with respect to one variable (Algo. 1). The convergence of Algo. 1 is guaranteed by the existing Alternating Direction Method of Multipliers (ADMM) theory [22]. Therefore, by solving the subproblems alternatively, our proposed algorithm will guarantee that we can find the optimal solution to each subproblem and finally, the algorithm will converge to the local solution. In all our experiments, we monitor the convergence is reached within less than 10 iterations.

Time Complexity Analysis: As discussed earlier, our overall problem can be divided into m number of subproblems and each of them can be solved using Algo. 1, we first analyze the computational complexity of Algo. 1 and then present the total complexity of our method. We also show that the proposed approach allows for parallel implementation, which can further reduce the computational time to a large extent.

In Algo. 1, each iteration contains three substeps (See Appendix for details): (i) solving a linear system with respect to U for once and is not repeated for each iteration. Solving this requires at most complexity of $O(n_v^3)$. However, we can solve this via n_v independent smaller linear systems over the n_v columns of U. Thus, with P parallel processing resources, we can reduce the computational time to $O(n_v^3/P)$, (ii) update with respect to Z can be done in $O(n_v^2)$ computational time. However, since the solution correspond to one-dimensional shrinkage and thresholding operation, we can perform the update via n_v indepedent shrinkage operations over the n_v rows of Z. Thus, with P parallel processing resources, this can be reduced to $O(n_v^2/P)$, (iii) similarly, update on Λ can be done in $O(n_v^2/P)$ computational time with P parallel processing resources by performing n_v independent updates over rows or columns. As a result, the computational complexity of Algo. 1 is $O(n_v^3 + 2 * n_v^2) \approx O(n_v^3)$ and it reduces to $O(n_v^3/P)$ with P parallel processing resources. The proposed approach invokes Algo. 1 for each subproblem i.e., with respect to one video alternatively. By adopting the same procedure, the computational complexity of our approach is $O(\sum_{v=1}^{m} n_v^3)$. Note that time complexity for solving a linear system can be reduced from $O(n_v^3)$ to $O(n_v^{2.376})$ using the Coppersmith-Winograd algorithm. Therefore, the time complexity of our approach is $O(\sum_{v=1}^{m} n_v^{2.376})$ and it reduces to $O([\sum_{v=1}^{m} n_v^{2.376}]/P)$ with P parallel processing resources.

V. Experiments

In this section, we present various experiments and comparisons to validate the effectiveness and efficiency of our proposed algorithm in two summarization tasks such as topic-oriented video summarization and multi-view video summarization in a camera network, as explained below.

A. Topic-oriented Video Summarization

Goal: Large collections of web videos contain clusters of videos belonging to a topic with typical visual content and repeating patterns across the videos. *Given a set of topic-related videos generated from a video search, can we generate a single summary that describes the collection altogether?* Specifically, our goal is to generate a single video summary that can describe the whole video collection.

Dataset: To evaluate topic-oriented video summarization, we need a single ground truth summary of all the topic-related videos that can describe the videos altogether. However, since there exists no such publicly available dataset that fits our need, we introduce a new large dataset, Tour20, that allows for the automatic evaluation of summarization methods in a fast and repeatable manner. We selected 20 tourist attractions from the Tripadvisor travelers choice landmarks 2015 list[3], and collected 140 videos from YouTube under the Creative Commons license (See Tab. II for names of the tourist attractions). Such a summary can be a great source of information for prospective tourists when they plan to visit the place and would like to get a preview of its main parts[4]. It is also important to note that all prior works [78], [77], [42], [41] conducted experiments on personal test sets, which are not publicly available, thus making it hard for others to reproduce or to compare the presented results. We hope the release of our Tour20 dataset will give researchers a new, dynamic tool to evaluate their video summarization algorithms in a repeatable and efficient way[5]. To the best of our knowledge, this is the biggest publicly available summarization dataset with 140 videos totaling about 7 hours (669,497 frames and 12,499 segments).

Performance Measures. Motivated by [26], [76], [67], we assess the quality of an automatically generated summary by comparing it to human judgment. Specifically, given a proposed summary and a set of human selected summaries, we compute the pairwise F-measure and then report the mean value motivated by the fact that there exists not a single ground truth summary, but multiple summaries are possible.

Ground truth Summaries. Previous topic-oriented video summarization approaches generated video summaries and then let humans assess their quality by comparing different system generated summaries. Specifically, users are shown different summaries and are asked to select the better one or assign a rating from a predefined scale. While simple and fast, this approach does not scale well because the user study has to be re-run every time a change is made. Another alternative is to let the humans watch the whole video and select some

[3]https://www.tripadvisor.com/TravelersChoice-Landmarks#1

[4]Although we focus on summarizing multiple videos of a tourist attraction as an application area in our experiments, our approach is quite general to summarize any type of videos generated from a search.

[5]The Tour20 dataset along with the complete groundtruth summaries are publicly available to download in http://www.ee.ucr.edu/~amitrc/datasets.php.

TABLE I
COMPARISON WITH SINGLE-VIDEO SUMMARIZATION METHODS ON TOUR20 DATASET. NUMBERS SHOW MEAN F-MEASURES AT 10% SUMMARY LENGTH, *i.e.*, SUMMARY CONTAINING ONLY 10% OF TOTAL VIDEO SEGMENTS. WE HIGHLIGHT THE **BEST** AND SECOND BEST BASELINE METHOD. OUR APPROACH (DiMS) STATISTICALLY OUTPERFORMS ALL BASELINE METHODS BY A SIGNIFICANT MARGIN ($p < .01$).

F-measure	ConcateKmeans	ConcateSpectral	ConcateSparse	KmeansConcate	SpectralConcate	SparseConcate	Graph	DT	SubMod	DiMS(ours)
mean	0.396	0.413	0.450	0.455	0.465	0.503	0.457	0.476	0.512	**0.613**

of the important segments as the summary. This approach has the advantage that, once the ground truth summaries are obtained, experiments can be carried out indefinitely, which is desirable especially for multimedia systems that involve multiple iterations and testing. We take this approach in our work to generate ground truth summaries.

Given the videos that were pre-processed into several segments, we asked three study experts to select at least 5%, but no more than 15% segments for each video as well as a single set of diverse segments that can describe the video collection altogether. We muted audio to ensure that important video segments are selected based solely on visual stimuli. Moreover, we also specify that if some embedded text is only mentioned in on-screen text, then it should not be labeled as important. They could use a simple interface that allows to watch all the videos of a collection at the same time and select important segments from each video. Note that obtaining these ground truth summaries was very time consuming. The study experts are requested to watch the whole video before selecting ground truth segments as whether a segment is important or not is a relative judgment within a video. Since the dataset contains important segments for each video as well as a diverse set of segments to describe the collection altogether, it can be used to evaluate both single-video and multi-video summarization algorithms in an repeatable and efficient way.

To assert the consistency of human created summaries, we compute both pairwise F-measure and the Cronbach's alpha between them, as in [26], [67]. The dataset has a mean F-measure of 0.643 and mean Cronobach's alpha of 0.944. Ideally alpha is around 0.9 for a good test [34]. More details on the dataset consistency and exemplar human created summaries can be found in the supplementary material.

Compared Methods. We compare our approach with several methods that fall into four main categories: (1) classical clustering based methods such as ConcateK means [1], ConcateSpectral [72], ConcateSparse [18], KmeansConcate [1], SpectralConcate [72], SparseConcate [18]) that use single-video summarization approach over multiple videos to generate a summary. The first three baselines (ConcateKmeans, ConcateSpectral, ConcateSparse) concatenate all the videos into a single video and then apply k-means, spectral clustering and sparse coding [18] to the concatenated video respectively, whereas in the other three baselines (KmeansConcate, SpectralConcate, SparseConcate), the corresponding approach is first applied to each video and then the resulting summaries are combined to form a single video summary. (2) graph clustering based methods including Graph [59] and DT [51]. Graph uses normalized cut-based clustering [59] over the graph constructed using the concatenated video [44], whereas DT uses Delaunay triangulation-based graph clustering to automatically extract informative and diverse segments from a video. Specifically, a Delaunay graph is first constructed using the video segments and then all the edges are classified into short edges and separating edges using average and standard deviation of edge lengths at each vertex. More details about the Delaunay graph clustering for summarizing videos can be seen in [51]. We apply Delaunay graph clustering to each video separately and then the resulting summaries are combined to form a single summary. (3) a submodularity based method (SubMod) [6], [43] that uses three selection criteria (Exhaustive, Mutually Exclusive and Interestingness) to extract informative segments from a video. We follow [6] to model the first two selection criteria and follow [25] to model interestingness in summarization. We use the same method [26] to compute the interestingness score of each video segment and use a greedy algorithm proposed by Nemhauser *et.al.* [52] to solve the combined submodular function. Similar to the DT baseline, we apply submodular maximization to each video separately and then the resulting summaries are combined to form a single summary. (4) state-of-the-art methods including MultiVideoContent [73], MultiVideoMMR [42] which are specifically designed for multi-video summarization. MultiVideoContent [73] uses a greedy approach with a content inclusion measure to summarize multiple videos whereas MultiVideoMMR [42] extends the concept of maximal marginal relevance [5] to the video domain for the same purpose.

Note that Eq. (5) represents the SparseConcate baseline that summarizes multiple videos without any diversity constraint. The purpose of comparing with single-video summarization methods is to show that techniques that attempt to find informative summary from a single-video usually do not produce an optimal set of representatives while summarizing multiple videos. Note that the recent two multi-video summarization methods in [77], [78] use meta-data sensor information or semantics related to a geographical area (e.g., weather and lighting condition) and are hence left out for comparison.

Experimental Settings. All methods use the same C3D feature as described in Sec. III-A. For all the compared methods (including ours), we generate a summary at 10% summary length, i.e., summary containing 10% of total segments in a video collection. Such a setting can give a fair comparison for various methods. We follow [76] and utilize VSUMM evaluation package [12] for finding matching pair of segments.

Comparision with Single-Video Baseline Methods. Table I shows the mean F-measure at 10% summary length on Tour20 dataset. While comparing with the single-video baseline methods, we have the following key findings from Table I: (1) The proposed method, DiMS statistically significantly outperforms all the compared single-video summarization methods ($p < .01$). We observe that directly applying these methods to summarize multiple videos produces a lot of redundant segments which deviates from the fact that the optimal summary should be diverse and can describe the multi-video concepts. This is probably because these methods are specific to single-

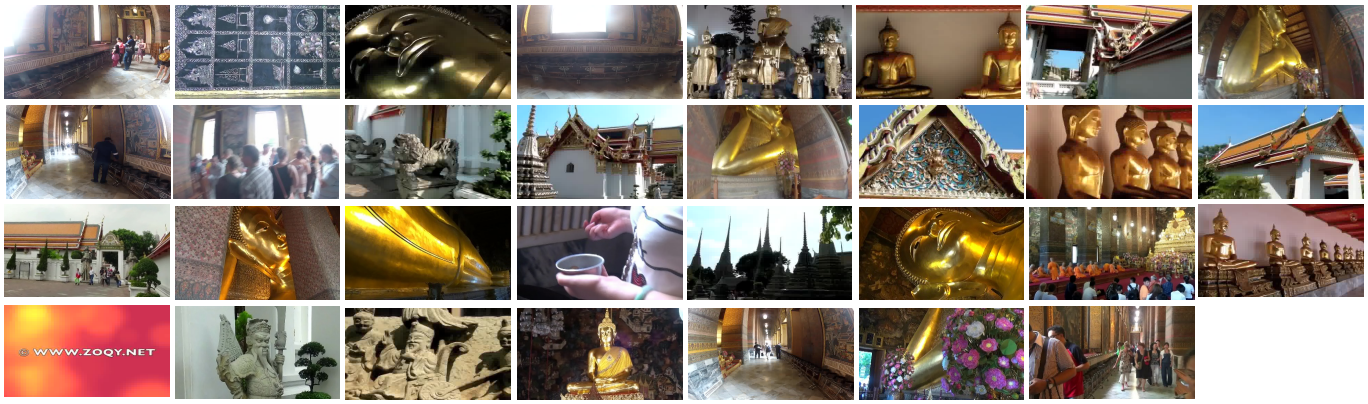

Fig. 1. Representative video segments generated by our approach (DiMS) in summarizing videos of the tourist attraction Wat Pho. We show the summaries at 10% length and represent each summarized segment using the corresponding central frame. As can be seen, our approach generates a summary that visualizes most of the concepts related to Wat Pho. Our approach achieved the highest F-measure of 0.722 compared to 0.625 by the MultiVideoContent baseline.

TABLE II
COMPARISON WITH MULTI-VIDEO SUMMARIZATION METHODS ON TOUR20 DATASET. NUMBERS SHOW MEAN F-MEASURES AT 10% SUMMARY LENGTH. WE HIGHLIGHT THE **BEST** AND SECOND BEST BASELINE METHOD. OVERALL, OUR APPROACH (DiMS) STATISTICALLY SIGNIFICANTLY OUTPERFORMS BOTH METHODS ($p < .01$). NAME OF THE TOURIST PLACES ARE PRESENTED IN THE FORMAT "NAME (# VIDEOS)".

Tourist Attractions	MultiVideoContent	MultiVideoMMR	DiMS(ours)
Angkor Wat (7)	0.431	0.452	**0.567**
Machu Picchu (7)	0.438	0.507	**0.582**
Taj Mahal (7)	0.593	0.533	**0.679**
Basilica of Sagrada Familia (6)	0.488	0.492	**0.597**
St. Peter's Basilica (5)	0.586	0.602	**0.699**
Milan Cathedral (10)	0.481	0.473	**0.571**
Alcatraz (6)	0.652	0.668	**0.755**
Golden Gate Bridge (6)	0.527	0.515	**0.618**
Eiffel Tower (8)	0.436	0.446	**0.562**
Notre Dame Cathedral (8)	0.463	0.473	**0.550**
The Alhambra (6)	0.553	0.582	**0.662**
Hagia Sophia Museum (6)	0.473	0.536	**0.585**
Charles Bridge (6)	0.453	**0.534**	0.525
Great Wall at Mutiantu (5)	0.493	0.507	**0.673**
Burj Khalifa (9)	**0.450**	0.392	0.441
Wat Pho (5)	0.625	0.603	**0.722**
Chichen Itza (8)	0.514	0.492	**0.582**
Sydney Opera House (10)	0.503	0.512	**0.614**
Petronas Twin Towers (9)	0.453	0.486	**0.643**
Panama Canal (6)	0.512	0.544	**0.639**
mean	**0.506**	**0.517**	**0.613**

Fig. 2. Role of diversity constraint in summarizing videos of Taj Mahal. (a) DiMS w/o diversity constraint (i.e., SparseConcate baseline), and (b) Our approach (DiMS). We show the top 10 segments generated using 10% summary length. As can be seen from (a), SparseConcate baseline finds redundant segments (marked with red color boarders) since it does not consider diversity of multiple videos. Our approach DiMS, in contrast, generate a more informative summary capturing different but also important information described in the videos by exploring the complementary information.

video summarization and thus can not take the advantage of the complementary information among multiple videos. (2) Among the alternatives, the SubMod baseline is the most competitive. However, the gap is still significant due to the fact that the proposed optimization approach efficiently explores the complementary information in creating an optimal summary from multiple videos. The mean F-measure performance improvements over SubMod is about 10% (0.613 vs 0.512) on our newly introduced Tour20 dataset. (3) Furthermore, note that our approach DiMS outperforms the naïve approach, SparseConcate, that summarizes multiple videos without any constraint with a clear margin (0.613 vs 0.503). This explicitly corroborates the effectiveness of our proposed diversity regularization (Eq. 6) in creating an informative and compact multi-video summary (See Fig. 2 for an illustrative example). (4) Our approach outperforms both of the graph clustering based methods (Graph, DT) by a significant margin due to its ability to efficiently model multi-video correlations.

Comparision with State-of-the-art Methods. Table II shows the topic-wise mean F-measure performance of our method along with two multi-video summarization methods on Tour20 dataset. Following observations can be made from Table II: (1) Our method achieves the highest overall score of 0.613, while the strongest baseline reaches 0.517 on the Tour20 dataset. Our approach is able to find the important segments from a video collection which are comparable to manual human created summaries (See Fig. 1). (2) Surprisingly, the performance of SubMod baseline is superior compared to MultiVideoContent. It is probably because SubMod considers both interestingness and representativeness in summarizing videos whereas the later one only optimizes for representativeness which may leave out some interesting segments in the summary. (3) Our method overall produces better summaries by optimizing all the important criteria of a video summary as explained earlier. However, it has a lower performance for certain videos, e.g., videos of the topic "Burj Khalifa". These videos contain fast motion and subtle semantics that define important segments of the video, such as opening the parachute or a nice panning shot from the top of the building. We believe these are difficult to capture without an additional semantic analysis [47]; we leave this as an interesting future work.

Performance Analysis with C3D Features: We investigate the importance and reliability of C3D features by comparing with 2D segment-level deep features, and found that the later produces inferior results, with a mean F-measure of 0.572 compared to 0.613 by the C3D features. We utilize Pycaffe with the VGG net pretrained model [66] to extract a 4096-dim

(a) DiMS w/o Interestingness Prior

(b) Our approach (DiMS)

Fig. 3. Role of interestingness prior in summarizing videos of Sydney Opera House. (a) DiMS w/o interestingness prior (by setting $Q^{(v)} = I$ in problem (7)), and (b) Our approach (DiMS). We show the top segments generated using 10% summary length. As can be seen, optimizing only for representativeness misses some crucial segments (e.g., the girl taking a photo by pointing to the opera house or segments showing several persons roaming around the house), which are indeed captured in our summary by jointly considering both representativeness and interestingness in the sparse optimization.

feature vector of a frame and then use temporal mean pooling to compute a single segment-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.

Performance Analysis with Interestingness Prior: To better understand the contribution of interestingness prior in summarizing videos, we analyzed the performance of the proposed approach by setting $Q^{(v)} = I$ in problem (7), where I denote the identity matrix of appropriate dimension. By turning off the interestingness prior, the mean F-measure decreases to 0.556. This is due to the fact that sparse representative selection in (3) only consider reconstruction capability and sparsity in summarizing videos. Optimizing only for representativeness risks leaving out some crucial segment(s) which are indeed captured in the summary by combining both interestingness and representativeness in summarizing videos (See Fig. 3 for an example). So, we conjecture that interestingness is also an important factor in summarization to generate a more condensed, descriptive and aesthetically pleasing summary.

Performance Analysis with Diversity Constraint: Fig. 2 shows the advantage of our proposed diversity regularization in summarizing videos of Taj Mahal. By turning off the diversity constraint (i.e., SparseConcate baseline), the mean F-measure decreases from 0.613 to 0.503 on Tour20 dataset. Furthermore, we also compare our approach with importance weighted clustering methods [14], [49], i.e., ConcateWKmeans, ConcateWSpectral, KmeansWConcate and SpectralWConcate to explicitly show the advantage of our diversity constraint in generating informative summaries. We use the interestingness score of each video segment for weighting the segment-level C3D features and then perform clustering on the feature weighted space to generate video summaries [14], [49], [11]. We use the same method [26] to compute the interestingness score of each video segment–such a setting gives a fair comparison in our experiments. Table III shows the comparison with importance weighted clustering methods on Tour20 dataset. We have the following key findings from Table III: The performance of importance weighted clustering methods are superior

TABLE III
COMPARISON WITH IMPORTANCE WEIGHTED CLUSTERING METHODS. NUMBERS SHOW MEAN F-MEASURES AT 10% SUMMARY LENGTH.

F-measure	ConcateWKmeans	ConcateWSpectral	KmeansWConcate	SpectralWConcate	DiMS
mean	0.426	0.448	0.472	0.483	**0.613**

compared to the classical K-means and spectral clustering (a maximum improvement of about 3%). This is expected since the weighted version of K-means and spectral clustering use the interestingness prior while summarizing videos. However, the proposed method, DiMS still outperforms these methods by a significant margin which again shows the advantage of our proposed diversity regularization in selecting informative and diverse segments from a video collection.

Performance Analyis with SLH Algorithm: We examined the performance of our approach using cosine similarity instead of SLH algorithm in computing segment-level correlations and found that the later produces inferior results, with a mean F-measure of 0.471 compared to 0.613 with the SLH algorithm. We kept all the parameters fixed in both of the case. This is probably because SLH algorithm tries to maintain the consistency in computing inter-video correlations via the exclusion principle [63], [69] which preserves the spatial arrangement of each video in computing such correlations. On the other hand, cosine similarity does not obey the exclusion principle which results in removing some crucial segments in the summary. However, we also believe that learning these correlations (as a future work) via a Siamese network or multiple kernel learning will further enhance our performances.

B. Multi-View Video Summarization in a Camera Network

Goal: This experiment aims at evaluating our proposed framework in summarizing multi-view videos captured using a network of cameras with considerable overlapping field of views. Such a summary can be very beneficial in surveillance systems equipped in offices, banks, factories, and crossroads of cities, for obtaining significant information in short time.

Datasets: We conduct experiments using three publicly available datasets[6]: (i) Office dataset captured with 4 stably-held web cameras in an indoor environment, (ii) Campus dataset taken with 4 hand-held ordinary video cameras in an outdoor scene, (iii) Lobby dataset captured with 3 cameras in a large lobby area.

Performance Measures. We use three quantitative measures on all experiments, including Precision, Recall and F-measure [20], [36]. For all these metrics, the higher value indicates better summarization quality.

Compared Methods. We contrast our approach with total of ten existing approaches including seven baseline methods (ConcateKmeans [1], ConcateSpectral [72], ConcateSparse [18], KmeansConcate [1], SpectralConcate [72], SparseConcate [18], Graph [59]) that use single-view summarization approach over multi-view videos to generate summary and four state-of-the-art methods (RandomWalk [20], RoughSets [40], BipartiteOPF [36]) which are specifically designed for multi-view video summarization. Similar to the experiments in topic-oriented video summarization, the first

[6][Online] Available: http://cs.nju.edu.cn/ywguo/summarization.html

TABLE IV
PERFORMANCE COMPARISON WITH SEVERAL BASELINES INCLUDING BOTH SINGLE AND MULTI-VIEW METHODS APPLIED ON THE THREE MULTI-VIEW DATASETS. ALL THE REPORTED VALUES ARE IN PERCENTAGE. OURS PERFORM THE BEST.

Methods	Office			Campus			Lobby		
	Precision	Recall	F – measure	Precision	Recall	F – measure	Precsion	Recall	F – measure
ConcateKmeans	100	38	55.07	55	41	47.06	85	67	75.05
ConcateSpectral	100	54	66.99	59	45	50.93	93	65	76.69
ConcateSparse	100	46	63.01	62	55	58.61	86	70	77.18
KmeansConcate	100	53	68.17	56	55	55.70	91	70	78.75
SpectralConcate	100	50	66.67	54	52	52.63	88	70	77.93
SparseConcate	93	58	71.30	56	62	58.63	97	67	79.45
Graph	100	50	66.67	56	48	51.86	91	67	77.33
RandomWalk	100	61	75.77	70	55	61.56	100	77	86.81
RoughSets	100	61	75.77	69	57	62.14	97	74	84.17
BipartiteOPF	100	69	81.79	75	69	71.82	100	79	88.26
DiMS(ours)	**100**	**77**	**86.91**	**83**	**69**	**75.47**	**100**	**86**	**92.52**

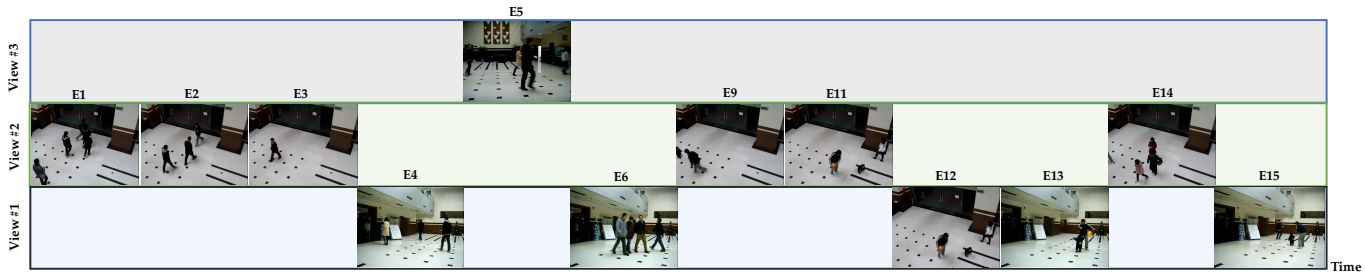

Fig. 4. Some summarized events for the Lobby dataset. X-axis denotes the time line as per the ground truth and the Y-axis represent the view (camera) from which the event is detected. Each event is represented by a key frame and an event number. The sequence of events in our summary are: E1: Five persons walk across the lobby towards the gate; a man runs to the gate, E2: Two men walks across the lobby towards the gate, and a man walks into the lobby, E3: A man run into the lobby from the gate, E4: Four persons walk into the lobby from the gate, E5: A man walks across the lobby towards the gate, E6: Three men are walking across the lobby towards the gate, E9: A man plays a ball with a baby, E11: A woman wearing a white coat walks across the lobby towards the gate, E12: A woman with a white coat passes away while a man is playing with a baby, E13: A man throws the ball towards the baby, E14: Two women and a man walk across the lobby from the gate, E15: A man plays a ball with a baby, a man with a black coat passes away.

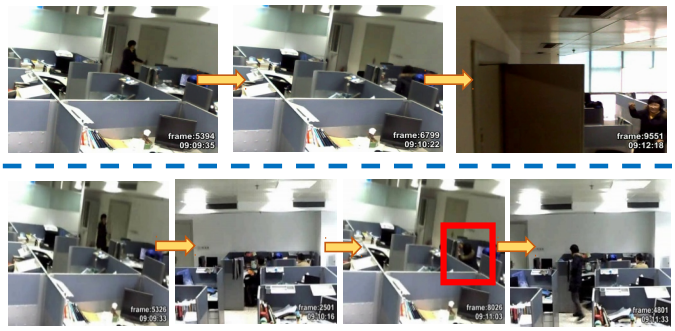

Fig. 5. Sequence of events detected related to activities of a member (A_0) inside the Office dataset. Top row: Summary produced by method [20], and Bottom row: Summary produced by our approach. Sequence of events detected in top row: 1st: A_0 enters the room, 2nd: A_0 sits in cubicle 1, 3rd: A_0 leaves 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 [20]), and 4th: A_0 leaves the room. The event of looking for a thick book to read (as per the ground truth in [20]) is missing in the summary produced by method [20] where as it is correctly detected by our approach (3rd frame: bottom row). This indicates our method captures video semantics in more informative way compared to [20].

seven single-view baselines generate a multi-view summary by either applying the method to each video separately or concatenating all the videos into a single video.

Experimental Settings. We set the same summary length as in [20] to generate our summaries and then employ the ground truth of events reported in [20] to compute the performance measures. We implement all the single-video summarization methods with the same video segmentation and feature representation as ours, whereas for the multi-view methods, we use prior published numbers when possible. In particular, for the multi-view summarization methods (RandomWalk, BipartiteOPF), we report the available results from the corresponding papers and implement RoughSets ourselves using the same video representation as the proposed one and tune their parameters to have the best performance.

Comparision with Single-View Baseline Methods: Table IV shows the results on three multi-view datasets, namely Office, Campus and Lobby datasets. We have the following observations from Table IV: (1) As expected, summaries produced using the single video-summarization methods, including the graph clustering based method (Graph) contain a lot of redundancies (simultaneous presence of most of the events) since they fail to exploit the complicated interview content correlations present in multi-view videos. (2) By using our diversity-aware sparse optimization method, such redundancy is largely reduced in contrast. Our proposed framework significantly outperforms all the single-view baseline methods in terms of precision, recall and F-measure due to its ability to model multi-view correlations.

Comparision with State-of-the-art Methods: While comparing with state-of-the-art multi-view summarization methods, we have the following observations from Table IV: (1) Our approach produces summaries with same precision as RandowWalk 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 (See Fig. 5 for one such example). The performance improvements over the recently published baseline BipartiteOPF, on three datasets are 5.12%, 3.65%, 4.26% in terms F-measure, respectively. (2) Notice that for all methods, including ours, performance on Campus dataset is not that good as compared to other two 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 4% better than that of the recent BipartiteOPF and 14% better than that of RandomWalk. Overall, on all datasets, our approach outperforms all the baselines in terms of F-measure. This corroborates the fact that our approach produces more informative multi-view summaries in contrast to the state-of-the-art methods. We present a part of the summarized events for the Lobby dataset in Fig. 4.

Scalability in Generating Summaries: Scalability in generating summaries of different length has shown to be effective while summarizing single videos [29], [57]. However, most of the previous multi-video summarization methods [42], [36] require the number of representative segments to be specified before generating the summaries which is highly undesirable in practical applications. Concretely speaking, the algorithm need to be rerun for each change in the number of representative segments that the user want to see in the summary. By contrast, our approach provides scalability in generating summaries of different length based on the user constraints without any further analysis of the input videos, similar to [29]. This is due to the fact that a ranked list of video segments can be generated after the alternating minimization which can produce 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. 6 shows the generated summaries of length 3, 4 and 7 most important events for the Office dataset.

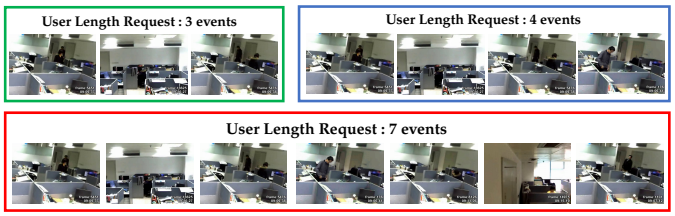

Fig. 6. The figure shows an illustrative example of scalability in generating summaries of different length based on the user constraints for the Office dataset. Each video segment is represented by a key frame and are arranged according to the summary generation rules mentioned in Sec. III-C.

VI. Conclusions and Future Works

We present an unsupervised framework for multi-video summarization by exploring the complementarity within the videos. We achieve this by developing a diversity-aware sparse optimization method that jointly summarizes a set of videos to find a single summary that is both interesting and representativeness of the input video collection. We also introduced a new dataset, Tour20, along with clear ground truth summaries to evaluate summarization algorithms in a fast and repeatable manner. We obtain excellent experimental results in two video summarization tasks such as topic-oriented video summarization and multi-view video summarization in a camera network, showing that our approach generates high quality summaries compared to the state-of-the-art methods.

In our current work, we assume that videos given by a web search are relevant to the topic. 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 [30] or additional video meta data to refine the results. Using active learning or deep CNNs [21] to get a set of topic-relevant videos is also another possibility in this regard. Moving forward, we would like to improve our method by using clustering [30] to handle such real-world scenarios while summarizing topic-related web videos. Moreover, we would like to improve our method by utilizing other types of metadata (e.g., social media images, comments, audio) while summarizing web videos.

Acknowledgment

This work was partially supported by NSF grant IIS-1316934. We would like to thank Andrew Kwon and Daniel Handojo, two current UCR undergraduate students, for helping in the annotation of the Tour20 dataset.

References

[1] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM symposium on Discrete algorithms*, 2007.
[2] G. K. Bo Xiong and L. Sigal. Storyline representation of egocentric videos with an application to story-based search. In *ICCV*, 2015.
[3] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 1996.
[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 2011.
[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.
[6] S. Chakraborty, O. Tickoo, and R. Iyer. Towards distributed video summarization. In *MM*, 2015.
[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] K. Dale, E. Shechtman, S. Avidan, and H. Pfister. Multi-video browsing and summarization. In *CVPRW*, 2012.
[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.
[11] R. C. De Amorim and B. Mirkin. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *PR*, 2012.
[12] 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.
[13] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.
[14] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *KDD*, 2004.
[15] F. Dornaika and I. K. Aldine. Decremental sparse modeling representative selection for prototype selection. *PR*, 2015.
[16] N. Ejaz, I. Mehmood, and S. W. Baik. Efficient visual attention based framework for extracting key frames from videos. *SPIC*, 2013.
[17] E. Elhamifar, G. Sapiro, and S. Sastry. Dissimilarity-based sparse subset selection. *TPAMI*, 2016.
[18] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012.
[19] S. Feng, Z. Lei, and S. Li. Online content-aware video condensation. In *CVPR*, 2012.
[20] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou. Multi View Video Summmarization. *TMM*, 2010.
[21] C. Gan, T. Yao, G. de Melo, Y. Yang, and T. Mei. Improving action recognition using web images. In *IJCAI*, 2016.
[22] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. SIAM, 1989.
[23] B. Gong, W. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014.
[24] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, and D. D. Feng. A Top-Down Approach for Video Summarization. *TOMCCAP*, 2014.
[25] M. Gygli, H. Grabner, and L. V. Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015.
[26] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *ECCV*, 2014.

[27] B. Han, J. Hamm, and J. Sim. Personalized video summarization with human in the loop. In *WACV*, 2011.
[28] D. Harwath and J. Glass. Deep multimodal semantic embeddings for speech and images. In *ASRU*, 2015.
[29] L. Herranz and J. M. Martinez. A framework for scalable summarization of video. *TCSVT*, 2010.
[30] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: web image search results clustering. In *MM*, 2006.
[31] 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.
[32] A. Khosla, R. Hamid, C. J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013.
[33] 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.
[34] P. Kline. The handbook of psychological testing. *Psychology Press*, 2000.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
[36] S. Kuanar, K. Ranga, and A. Chowdhury. Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. *TMM*, 2015.
[37] Y. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.
[38] C. D. Leo and B. S. Manjunath. Multicamera video summarization from optimal reconstruction. In *ACCV Workshop*, 2011.
[39] C. D. Leo and B. S. Manjunath. Multicamera Video Summarization and Anomaly Detection from Activity Motifs. *TOSN*, 2014.
[40] P. Li, Y. Guo, and H. Sun. Multi key-frame abstraction from videos. In *ICIP*, 2011.
[41] Y. Li and B. Merialdo. Multi-video summarization based on av-mmr. In *CBMI*, 2010.
[42] Y. Li and B. Merialdo. Multi-video summarization based on video-mmr. In *WIAMIS*, 2010.
[43] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *HLT*, 2011.
[44] S. Lu, I. King, and M. R. Lyu. Video summarization by video structure analysis and graph optimization. In *ICME*, 2004.
[45] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013.
[46] 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.
[47] T. Mei, L. X. Tang, J. Tang, and X. S. Hua. Near-lossless semantic video summarization and its applications to video analysis. *TOMCCAP*, 2013.
[48] 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.
[49] D. S. Modha and W. S. Spangler. Feature weighting in k-means clustering. *Machine learning*, 2003.
[50] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *JVCIR*, 2008.
[51] P. Mundur, Y. Rao, and Y. Yesha. Keyframe-based video summarization using delaunay clustering. *IJDL*, 2006.
[52] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 1978.
[53] 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.
[54] D. Pachauri, R. Kondor, and V. Singh. Solving the multi-way matching problem by permutation synchronization. In *NIPS*, 2013.
[55] R. Panda, A. Das, and A. K. Roy-Chowdhury. Embedded sparse coding for summarizing multi-view videos. In *ICIP*, 2016.
[56] R. Panda, A. Das, and A. K. Roy-Chowdhury. Video summarization in a multi-view camera network. In *ICPR*, 2016.
[57] R. Panda, S. K. Kuanar, and A. S. Chowdhury. Scalable video summarization using skeleton graph and random walk. In *ICPR*, 2014.
[58] R. Panda and A. K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017.
[59] Y. Peng and C.-W. Ngo. Clip-based similarity measure for query-dependent clip retrieval and video summarization. *TCSVT*, 2006.
[60] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.
[61] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007.
[62] 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.
[63] G. Scott and H. Longuett-Higgins. An algorithm for associating the features of two images. *The Royal Society of London*, 1991.
[64] J. Shao, D. Jiang, M. Wang, H. Chen, and L. Yao. Multi-video summarization using complex graph clustering and mining. *Computer Science and Information Systems*, 2010.
[65] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.
[66] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[67] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015.
[68] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014.
[69] M. Torki and A. Elgammal. One-shot multi-set non-rigid feature-spatial matching. In *CVPR*, 2010.
[70] 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.
[71] B. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *TOMCCAP*, 2007.
[72] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.
[73] F. Wang and B. Merialdo. Multi-document video summarization. In *ICME*, 2009.
[74] I. Yahiaoui, B. Merialdo, and B. Huet. Generating summaries of multi-episode video. In *ICME*, 2001.
[75] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016.
[76] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarizatio. *CVPR*, 2016.
[77] L. Zhang, Y. Xia, K. Mao, H. Ma, and Z. Shan. An effective video summarization framework toward handheld devices. *TIE*, 2015.
[78] Y. Zhang, G. Wang, B. Seo, and R. Zimmermann. Multi-video summary and skim generation of sensor-rich videos in geo-space. In *MMSys*, 2012.
[79] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014.

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.

Niluthpol Chowdhury Mithun received his Bachelors and Masters degree in 2011 and 2014 respectively from Bangladesh University of Engineering and Technology, Dhaka. 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, intelligent transportation systems and web image collection summarization.

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 Engineering at U. of California, Riverside. His research interests include image processing and analysis, computer vision, and video communications and statistical methods for signal analysis. His current research projects include intelligent camera networks, wide-area scene analysis, motion analysis in video, activity recognition and search, video-based biometrics (face and gait), biological video analysis, and distributed video compression. He is coauthor of "The Acquisition and Analysis of Videos over Wide Areas" He is the editor of the book "Distributed Video Sensor Networks". He has been on the organizing and program committees of multiple conferences and serves on the editorial boards of a number of journal. He is a Fellow of IAPR.