UNIVERSITY OF CALIFORNIA
RIVERSIDE

Visual Learning with Weak Supervision: Applications in Video Summarization and Person Re-identification

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Rameswar Panda

December 2018

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Ertem Tuncel
Dr. Nael Abu-Ghazaleh
Dr. Salman Asif

The Dissertation of Rameswar Panda is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The satisfaction that accompanies my completion of PhD would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts with success. First and foremost, I would like to thank my advisor Dr. Amit K. Roy-Chowdhury for his constant motivation and help during the course of this dissertation. He always encouraged me to come up with new and simple solutions to complex problems. From the beginning to end, I was fully allowed to have necessary freedom to exercise thoughtful and scientific approaches to different problems and have free discussions on them. I consider myself fortunate to have him as my advisor.

I would also like to express my heartfelt gratitude to my dissertation committee members, Dr. Ertem Tuncel, Dr. Nael Abu-Ghazaleh, and Dr. Salman Asif for giving me thoughtful feedback and constructive comments in improving the quality of this dissertation. I found Dr. Tuncel as an excellent teacher and learned a lot from his courses. Working with Dr. Abu-Ghazaleh and Dr. Asif while writing research proposals have been invaluable experiences for me, and I cannot thank them enough for that. Special thanks are reserved for my masters advisor Dr. Ananda S. Chowdhury from Jadavpur University, whose constant motivation and valuable comments helped me to enhance my works. I am also thankful to Dr. Jay Farrell and Dr. Anastasios Mourikis for their kind help at different times. I also owe a lot to all my internship mentors Dr. Ziyan Wu, Dr. Jan Ernst from Siemens, Dr. Jianming Zhang from Adobe, and Dr. Xiang Yu, Dr. Samuel Schulter, Dr. Manmohan Chandraker from NEC labs, for their continued support and encouragement.

I was fortunate to get some of the bright students as my labmates in the Video

Computing Group at UC Riverside. I would first like to thank Dr. Abir Das, an Assistant Professor at IIT Kharagpur, for helping me at various critical junctures during my research. He has been an amazing researcher and a very good mentor. I also want to extend my gratitude to Dr. Mahmudul Hasan, Dr. Jawadul Hasan, Tahmida Mahmud, Sujoy Paul, Sourya Roy, and Amran Hossen who were helpful at various instants. Many thanks to Niluthpol Chowdhury Mithun for the long intellectual discussions we had in the lab.

My friends at UCR deserve special mention as I had fun throughout the ups and downs of my life as a PhD researcher. They were like my family so far away from home who never let me feel homesick. I especially thank my friends for spending memorable moments with me—Tushar Jain, Anguluri Rajasekhar (Angy), Devashree Tripathy and Abhishek Aich. I still relish the numerous crazy discussions with Angy that we had over time.

Last but not the least, I would like to express my heartfelt regards to my parents for their support and faith that always motivates me to work harder and better. It is their way of life where I find the inspiration to fight the odd and move forward. I am grateful to my brother Nrusingha Panda for his support and love. I also wish to thank my loving and supportive fiancée, Monalisha Panigrahi, who gave me perpetual inspiration.

I would like to thank the National Science Foundation (IIS-1316934, IIS-1724341, CPS-1544969) and Adobe for their grants to Dr. Amit K. Roy-Chowdhury, which partially supported my research. I also thank Victor Hill of UCR CS for setting up the computing infrastructure used in most of the works presented in this thesis.

Acknowledgment of previously published or accepted materials: The text of this dissertation, in part or in full, is a reprint of the material as appeared in three previously

published papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all publications, directed and supervised the research which forms the basis for this dissertation. The papers are as follows.

1. Rameswar Panda, Amit K. Roy-Chowdhury, Collaborative Summarization of Topic-Related Videos, CVPR 2017, Hawaii, USA.

2. Rameswar Panda, Amit K. Roy-Chowdhury, Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization, IEEE Transactions on Multimedia (TMM), Vol. 19, No. 9, 2017.

3. Rameswar Panda, Amran Bhuiyan, Vittorio Murino, Amit K. Roy-Chowdhury. Unsupervised Adaptive Re-identification in Open World Dynamic Camera Networks, CVPR 2017, Hawaii, USA. The co-first author Amran Bhuiyan Das contributed to experimentation and co-author Vittorio Murnio supervised the research.

To my parents.

ABSTRACT OF THE DISSERTATION

Visual Learning with Weak Supervision: Applications in Video Summarization and Person Re-identification

by

Rameswar Panda

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2018
Dr. Amit K. Roy-Chowdhury, Chairperson

Many of the recent successes in computer vision have been driven by the availability of large quantities of labeled training data. However, in the vast majority of real-world settings, collecting such data sets by hand is infeasible due to the cost of labeling data or the paucity of data in a given domain. One increasingly popular approach is to use weaker forms of supervision that are potentially less precise but can be substantially less costly than producing explicit annotation for the given task. Examples include domain knowledge, weakly labeled data from the web, constraints due to physics of the problem or intuition, noisy labels from distant supervision, unreliable annotations obtained from the crowd workers, and transfer learning settings. In this thesis, we explore two important and highly challenging problems in computer vision, namely video summarization and person re-identification, where learning with weak supervision could be extremely useful but remains as a largely under-addressed problem in the literature.

One common assumption of many existing video summarization methods is that videos are independent of each other, and hence the summarization tasks are conducted

separately by neglecting relationships that possibly reside across the videos. In the first approach, we investigate how topic-related videos can provide more knowledge and useful clues to extract summary from a given video. We develop a sparse optimization framework for finding a set of representative and diverse shots that simultaneously capture both important particularities arising in the given video, as well as, generalities identified from the set of topic-related videos. In the second approach, we present a novel multi-view video summarization framework by exploiting the data correlations through an embedding without assuming any prior correspondences/alignment between the multi-view videos, e.g., uncalibrated camera networks. Via extensive experimentation on different benchmark datasets, we validate both of our approaches and demonstrate that our frameworks are able to extract better quality video summaries compared to the state-of-the-art alternatives.

Most work in person re-identification has focused on a fixed network of cameras. However, in practice, new camera(s) may be added, either permanently or on a temporary basis. In the final part of the dissertation, we show that it is possible to on-board new camera(s) to an existing network using domain adaptation techniques with limited additional supervision. We develop a domain perceptive re-identification framework that can effectively discover and transfer knowledge from the best source camera (already installed) to a newly introduced target camera(s), without requiring a very expensive training phase. Our approach can greatly increase the flexibility and reduce the deployment cost of new cameras in many real-world dynamic camera networks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many of the recent successes in computer vision have been driven by the availability of large quantities of labeled training data. However, in the vast majority of real-world settings, collecting such data sets by hand is infeasible due to the cost of labeling data or the paucity of data in a given domain. Let us consider the case of video summarization [170] which aim to automatically extract a brief informative summary from a given video, as an example. Majority of the recent works leverage human-crafted training data in form of video-summary pairs or importance annotations for summarizing long videos. These approaches assume the availability of large amount of human-created video-summary pairs, which are in practice difficult to obtain for unconstrained web videos. Without supervision, summarization methods rely on different heuristically designed criteria in an unsupervised way but often fail to produce semantically meaningful video summaries. Similarly, person re-identification [279], which has become a very active research area in the last few years, has relied mostly on a supervised training phase where a transformation between the obser-

vations at two cameras is learned from labeled data. However, relying on manually labeled data for each camera pair limits scalability to large networks and adaptability to changing environmental conditions, application domains, and network configurations, due to the burden of obtaining extensive manual labels. Thus, there is an urgent need to develop methods with limited supervision which can be scaled up as more and more new data are generated.

In recent years, one increasingly popular approach is to use weaker forms of supervision that are potentially less precise but can be substantially less costly than producing explicit annotation for the given task. There exists many different forms of weak supervision that can be efficiently utilized for a specific task in hand. Examples include domain knowledge, weakly labeled data from the web, constraints due to physics of the problem or intuition, noisy labels from distant supervision, unreliable annotations obtained from the crowd workers, and transfer learning settings. The difficulties associated with fully supervised learning and the availability of weak supervision in many different forms motivate us to develop efficient algorithms and frameworks which can obtain equivalent performance of fully supervised methods by only leveraging limited human supervision.

In this thesis, we explore two important and highly challenging problems in computer vision that are video summarization and person re-identification, where learning with weak supervision could be extremely useful but remains as a largely under-addressed problem in the literature. In the first chapter, we investigate how topically close videos can provide more knowledge and useful clues to extract summary from a given video without requiring human-crafted training data in form of video-summary pairs or importance annotations. Existing works summarize videos by either exploring different heuristically

designed criteria in an unsupervised way [110, 58, 214, 39], or developing fully supervised algorithms [126, 86, 77, 194]. However, unsupervised methods are blind to the video category and often fail to produce semantically meaningful video summaries. On the other hand, acquisition of large amount of training data in supervised approaches is non-trivial and may lead to a biased model. Different from existing works, we introduce a weakly supervised approach that exploits visual context from a set of topic-related videos to extract an informative summary of a given video. Our method is motivated by the observation that *similar videos have similar summaries*. For instance, suppose we have a collection of videos of "surfing". It is quite likely good summaries for those videos would all contain segments corresponding to riding with surfboard, floating on water, and off the lip surfing, etc. Thus, we hypothesize that additional topic-related videos can provide visual context to identify the important parts of the video being summarized. We develop a sparse optimization approach for finding a set of representative and diverse shots that simultaneously capture both important particularities arising in the given video, as well as, generalities identified from the set of topic-related videos. Specifically, we formulate the task of finding summaries as an ℓ_{21}-norm optimization problem where the nonzero rows of a sparse coefficient matrix represent the relative importance of the corresponding shots. We conduct rigorous experiments on two challenging benchmark datasets to demonstrate the effectiveness of our framework. An important advantage of our method is that it learns the notion of importance from a set of videos belonging to a category (weak supervision) which are readily available on the web, and hence provides much greater scalability in extracting summaries from web videos.

Most traditional video summarization methods (including our approach in chap-

ter 2 are designed to generate effective summaries for single-view videos [194, 273, 50, 111, 120]. However, with the proliferation of surveillance cameras, a major problem is to figure out how to extract useful information from the videos captured by these cameras. Most of the prior works simply extend the single-view video summarization approaches to extract an informative summary from the multi-view videos. However, they fail to produce an optimal summary because of the large amount data correlations due to the locations and fields of view of the cameras. Moreover, these videos are captured with different view angles, and depth of fields, for the same scenery, resulting in a number of unaligned videos. Some recent approaches have focused on utilizing strong supervision in form of inter-camera frame correspondence while summarizing multi-view videos [178, 119, 182]. It becomes infeasible and unrealistic to manually align the long and unstructured videos in uncontrolled settings. To address the challenges encountered in a camera network, we propose a novel summarization framework in chapter 3 by exploiting the data correlations as one form of weak supervision 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 [37, 173], which typically aims to obtain a latent subspace shared by multiple views by assuming that these views are generated from this subspace. Specifically, to better characterize the multi-view structure, we first project the data points into a latent embedding which is able to preserve both the correlations and then propose a sparse representative selection method over the learned embedding to summarize the multi-view videos. Finally, to better leverage the multi-view embedding and the selection mechanism, we learn the embedding and optimal representatives jointly. Experiments on six challenging datasets

demonstrate that our framework achieves superior performance over some mono-view summarization approaches as well as state-of-the-art multi-view summarization methods.

Continuing on learning with weak supervision, the third work addresses multi-camera person re-identification from a different perspective. In particular, we investigate how the re-identification models can be updated as new cameras are added, with limited additional supervision. Existing approaches for person re-identification have concentrated on either designing the best feature representation or learning optimal matching metrics in a static setting where the number of cameras are fixed in a network [248, 164, 259]. Most approaches have neglected the dynamic and open world nature of the problem, where one or multiple new cameras may be temporarily on-boarded into an existing system to get additional information or added to expand an existing network. Given a newly introduced camera, traditional re-identification methods will try to relearn the inter-camera transformations/distance metrics using a costly training phase. This is impractical since labeling data in the new camera and then learning transformations with the others is time-consuming, and defeats the entire purpose of temporarily introducing the additional camera. Thus, we propose a novel approach for adapting existing multi-camera re-identification frameworks with limited supervision. First, we formulate a domain perceptive re-identification method based on geodesic flow kernel that can effectively find the best source camera (already installed) to adapt with newly introduced target camera(s), without requiring a very expensive training phase. Second, we introduce a transitive inference algorithm that can exploit the information from best source camera to improve the accuracy across other camera pairs in a network of multiple cameras. Third, we develop a target-aware sparse prototype selection

strategy for finding an informative subset of source camera data for data efficient learning in resource constrained environments. We perform extensive experiments on five benchmark datasets, which well demonstrate the efficacy of our proposed framework for on-boarding camera(s) without requiring any labeled data from the newly introduced camera(s).

Organization of the Thesis. The rest of the thesis is organized as follows. In chapter 2, we present a collaborative representative selection approach using sparse ℓ_{21} optimization and evaluate the framework for summarizing topic-related videos. We propose our multi-view video summarization approach without assuming any prior correspondences/alignment between multi-view videos in chapter 3. Finally, in chapter 4, we propose a domain perceptive re-identification method based on geodesic flow kernel to discover and transfer knowledge from existing source cameras to a newly introduced target camera, without requiring a very expensive training phase. We conclude the thesis in chapter 5 by providing some future directions related to the problem of video summarization and person re-identification.

Chapter 2

Collaborative Video Summarization

2.1 Introduction

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

Much progress has been made in developing a variety of ways to summarize videos, by exploring different design criteria (representativeness [110, 58, 273, 41, 214, 39], interestingness [64, 160, 195], importance [84, 252]) in an unsupervised manner, or developing supervised algorithms [126, 86, 77, 194, 219]. However, with the notable exception of [39], one common assumption of existing methods is that videos are independent of each other,

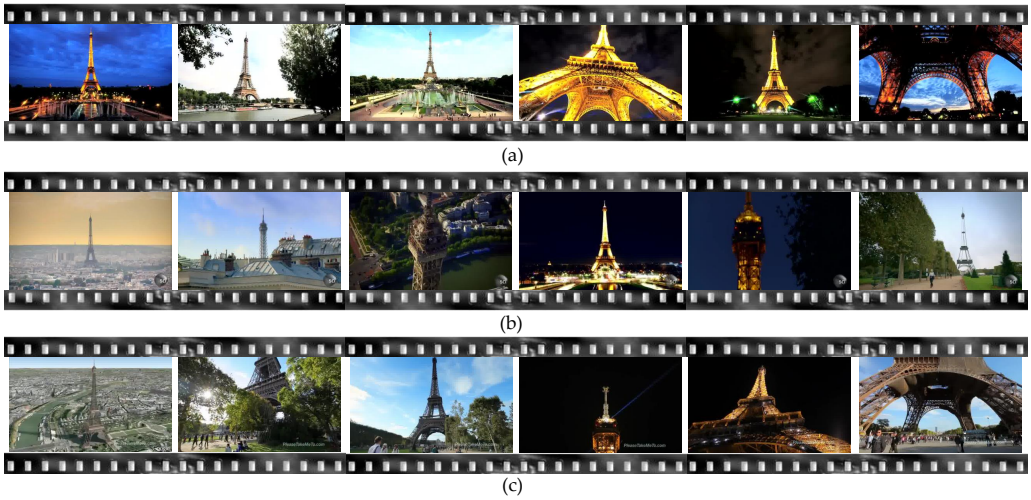


Figure 2.1: Consider three videos of the topic "*Eiffel Tower*". Each row shows six uniformly sampled shots represented by the middle frame, from the corresponding video. It is clear that all these videos have mutual influence on each other since many visual concepts tend to appear repeatedly across them. We therefore hypothesize that such topically close videos can provide more knowledge and useful clues to extract summary from a given video. We build on this intuition to propose a summarization algorithm that exploits topic-related visual context from video (b) & (c) to automatically extract an informative summary from a given video (a). Best viewed in color.

and hence the summarization tasks are conducted separately by neglecting relationships that possibly reside across the videos.

Let us consider the video in Fig. 2.1.a. The video is represented by six uniformly sampled shots. Now consider the videos in Fig. 2.1.b and 2.1.c along with the video in Fig. 2.1.a. Are these videos independent of each other or something common exists across them? The answer is clear: all of these videos belong to the same topic "*Eiffel Tower*". As a result, the summaries of these videos will have significant common information with each other. Thus, the context of additional topic-related videos can be beneficial by providing more knowledge and additional clues for extracting a more informative summary from a specified video. We build on this intuition, presenting a new perspective to summarize a

video by exploiting the neighborhood knowledge from a set of topic-related videos.

In this work, we propose a *Collaborative Video Summarization (CVS)* approach that exploits visual context from a set of topic-related videos to extract an informative summary of a given video. Our work builds upon the idea of collaborative techniques [9, 149, 253] from information retrieval (IR) and natural language processing (NLP), which typically use the attributes of other similar objects to predict the attribute of a given object. We achieve this by *finding a sparse set of representative and diverse shots that simultaneously capture both important particularities arising in the given video, as well as, generalities identified from the set of topic-related videos.* Our underlying assumption is that a few topically close videos actually have mutual influence on each other since many important visual concepts tend to appear repeatedly across them. Note that in this work, we assume that additional topic-related videos are available beforehand. One can easily use either clustering [217] or additional video meta data to obtain such topic-relevant videos.

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 2.3.1). Then, we develop a novel collaborative sparse representative selection strategy by exploiting visual context from topic-related videos (Section 2.3.2). 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 2.3.3). Note that the summary will be of the one video of interest

only, while exploiting visual context from additional topic-related videos.

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

- We propose a novel approach to extract an informative summary of a specified video by exploiting additional knowledge from topic-related videos. The additional topic-related videos provide visual context to identify what is important in a video.
- We develop a collaborative representative selection strategy by introducing a consensus regularizer that simultaneously captures both important particularities arising in the given video, as well as, generalities identified from the topic-related videos.
- We present an efficient optimization algorithm based on half-quadratic function theory to solve the non-smooth objective, where the minimization problem is simplified to two independent linear system problems.
- We demonstrate the effectiveness of our approach in two video summarization tasks– topic-oriented video summarization and multi-video concept visualization. With extensive experiments on both CoSum [39] and TVSum50 [214] video datasets, we show the superiority of our approach over competing methods for both summarization tasks.

2.2 Related Work

Video summarization has been studied from multiple perspectives. Here, we focus on some representative methods closely related to our work. Interested readers can check [170, 227] for a more comprehensive survey. While the approaches might be supervised or unsupervised, the goal of summarization is nevertheless to produce a compact

visual summary that encapsulates the most informative parts of a video.

Much work has been proposed to summarize a video using supervised learning. Representative methods use category-specific classifiers for importance scoring [194, 219] or learn how to select informative and diverse video subsets from human-created summaries [86, 77, 208, 270] or learn important facets, like faces and objects [126, 156, 20]. Although these methods have shown impressive results, their performance largely depends on huge amount of labeled examples which are difficult to collect for unconstrained web videos. In addition, it is generally feasible to have only a limited number of users to annotate training videos, which may lead to a biased summarization model. Our CVS approach, on the other hand, exploits visual context from topic-related videos without requiring any labeled examples, and thus can be easily applied to summarize large scale web videos with diverse content.

Without supervision, summarization methods rely on low-level visual indices to determine the relevance of parts of a video. Various strategies have been studied, including clustering [3, 50, 82], interest prediction [160, 84], and energy minimization [196, 65]. Leveraging crawled web images is also another recent trend for video summarization [110, 214, 111]. However, all of these methods summarize videos independently by neglecting relationships that possibly reside across them. The use of neighboring topic-related videos to improve summarization still remains as a novel and largely under-addressed problem.

The most relevant work to ours is the video co-summarization approach (CoSum) [39]. It aims to find visually co-occurring shots across videos of the same topic based on the idea of commonality analysis [38]. However, CoSum and our approach have significant differences. CoSum constructs weighted bipartite graphs for each pair of videos in

order to find the maximal bicliques, which can be computationally inefficient given a large collection of topic-related videos. Our approach, on the other hand, offers a more flexible way to find most representative and diverse video shots through a collaborative sparse optimization framework that can be efficiently solved to handle large number of web videos simultaneously. In addition, CoSum employs a computationally-intensive shot-level feature representation, namely a combination of both observation and interaction features [98], which involves extracting low-level features such as CENTRIST, Dense-SIFT and HSV color moments. By contrast, our approach utilizes deep learning features which are more computationally efficient and more accurate in characterizing both appearance and motion.

Our focus on the sparse coding as the building block of CVS is largely inspired by its appealing property in modeling sparsity and representativeness in data summarization. In contrast to prior works [41, 58, 273], we develop a novel collaborative sparse optimization that finds shots which are informative about the given video, as well as, the set of of topic-related videos. In addition, we introduce a novel regularizer in the optimization to obtain a diverse set of representatives, instead of manually filtering redundant shots from the extracted summary as some existing methods.

In recent years, collaborative techniques have been successfully applied to several IR and NLP tasks: collaborative recommendation [9, 204], collaborative filtering [253], collaborative ranking [10] and text summarization [233, 231, 232]. The common idea underlying all of these works, including ours, is to make use of the interactions among multiple objects under the assumption that similar objects will have similar behaviors and characteristics. An earlier work [8] uses a collaborative system by merging results of various

segmentation approaches to obtain a summary. By contrast, our approach builds on the idea of collaboration among the topic-related videos to efficiently summarize a given video.

2.3 Collaborative Video Summarization

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

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

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

2.3.1 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 descibe these steps in the following.

Temporal Segmentation. Our approach starts with segmenting videos using an existing algorithm [39]. We segment each video into multiple shots by measuring the amount of changes between two consecutive frames in the RGB and HSV color spaces. We added an additional constraint in 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 video summary.

Feature Representation. Deep convolutional neural networks (CNNs) have recently been successful at large-scale object recognition [200, 118]. Beyond the object recognition task itself, recent advancement in deep learning has revealed that features extracted from a CNN are generic features that have good transfer learning capabilities across different domains [210, 286, 108]. An advantage of using deep learning features is that there exist accurate, large-scale datasets such as Imagenet [200], and Sports-1M [108] from which they can be extracted. Moreover, GPU-based extraction of such features are much faster than that for the traditional hand crafted features such as CENTRIST, and Dense-SIFT.

In the case where the input is a video clip, C3D features [226] have recently shown better performance compared to the features extracted using each frame separately [225, 264]. 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 [226]. This is followed by a temporal mean pooling scheme to maintain the local ordering structure within a shot. Then the pooling result serves as the final feature vector of a shot (4096 dimensional) to be used in the sparse optimization. We will discuss the performance benefits of employing C3D features later in our experiments.

2.3.2 Collaborative Sparse Representative Selection

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

Formulation. Sparse optimization approaches [41, 58] find the representative shots from a video itself by minimizing the linear reconstruction error as

$$\min_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \ \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 + \lambda_s \|\mathbf{Z}\|_{2,1} \tag{2.1}$$

where $||\mathbf{Z}||_{2,1} = \sum_{i=1}^{n} ||\mathbf{Z}_{i,.}||_2$ and $||\mathbf{Z}_{i,.}||_2$ is the ℓ_2-norm of the i-th row of $\mathbf{Z}$. $\lambda_s > 0$ is a regularization parameter that controls sparsity in the reconstruction. Once the problem (2.1) is solved, the representatives are selected as the points whose corresponding $||\mathbf{Z}_i||_2 \neq 0$.

Clearly, the above formulation summarizes a video neglecting mutual relationships

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

$$\min_{\mathbf{Z},\,\tilde{\mathbf{Z}}} \ \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big)+\lambda_s\big(\|\mathbf{Z}\|_{2,1}+\|\tilde{\mathbf{Z}}\|_{2,1}\big) \tag{2.2}$$

where parameter $\alpha > 0$ balances the penalty between errors in the reconstruction of video v and errors in the reconstruction of the remaining videos in the collection $\tilde{v}$[1]. The objective function is intuitive: minimization of (4.8) favors selecting a sparse set of representative shots that simultaneously reconstructs the target video $\mathbf{X}$ via $\mathbf{Z}$, as well as the set of topic related videos $\tilde{\mathbf{X}}$ via $\tilde{\mathbf{Z}}$, with high accuracy.

Diversity Regularization. The data reconstruction and sparse optimization formulations in (4.8) tend to select shots that can cover a specified video, as well as the set of topic-related videos. However, there is no explicit tendency to select diverse shots capturing different but also important information described in the set of videos. Prior works [41, 58] handle this issue by manually filtering redundant shots from the extracted summary which can be unreliable while summarizing large scale web videos. Recent works on sparse representative selection [254, 239, 147] also addresses this diversity problem by explicitly adding non-convex regularizers in the objective which makes it difficult to optimize.

Inspired by the recent work on convex formulation for active learning [59] and

[1]Note that we use a common α to weight the reconstruction term related to the topic-related videos in (4.8) for simplicity of exposition. However, if we have some prior information on which video is more informative about the topic or close to the given video, we can assign different αs for different videos.

document compression [263], we introduce two diversity regularization functions, $f_d(\mathbf{Z})$ and $f_d(\tilde{\mathbf{Z}})$ to select a sparse set of representative and diverse shots from the video. Our motivation is that, rows in sparse coefficient matrices corresponding to two similar shots are not nonzero at the same time. This is logical since the representative shots should be non-redundant capturing diverse aspects of the input video.

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

$$\begin{aligned} f_d(\mathbf{Z}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}Z_{ij} = tr(\mathbf{D}^T\mathbf{Z}), \\ f_d(\tilde{\mathbf{Z}}) &= \sum_{i=1}^{n}\sum_{j=1}^{\tilde{n}} \hat{d}_{ij}\tilde{Z}_{ij} = tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}) \end{aligned} \tag{2.3}$$

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

After adding the diversity regularization functions into problem (4.8), we have the objective function as follows:

$$\min_{\mathbf{Z},\,\tilde{\mathbf{Z}}} \ \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2 + \alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big) + \lambda_s\big(\|\mathbf{Z}\|_{2,1} + \|\tilde{\mathbf{Z}}\|_{2,1}\big) + \lambda_d\big(tr(\mathbf{D}^T\mathbf{Z}) + tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big) \tag{2.4}$$

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

Consensus Regularization. The objective function (2.4) favors selecting a sparse set of representative and diverse shots from a target video $\mathbf{X}$ by exploiting visual context from

additional topic-related videos $\tilde{\mathbf{X}}$. Specifically, rows in $\mathbf{Z}$ provide information on relative importance of each shot in describing the video $\mathbf{X}$, while rows in $\tilde{\mathbf{Z}}$ give information on relative importance of each shot in $\mathbf{X}$ in describing $\tilde{\mathbf{X}}$. Given the two sparse coefficient matrices, our next goal is to select a unified set of shots that simultaneously cover the important particularities arising in the target video, as well as the generalities arising in the video collection. To achieve the above goal, we propose to minimize the following function:

$$\begin{aligned} \min_{\mathbf{Z},\,\tilde{\mathbf{Z}}} \ & \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big)+\lambda_s\big(\|\mathbf{Z}\|_{2,1}+\|\tilde{\mathbf{Z}}\|_{2,1}\big) \\ & +\lambda_d\big(tr(\mathbf{D}^T\mathbf{Z})+tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big)+\beta||\mathbf{Z}_c||_{2,1} \quad s.t.\ \ \mathbf{Z}_c=[\mathbf{Z}|\tilde{\mathbf{Z}}],\ \mathbf{Z}_c\in\mathbb{R}^{n\times(n+\tilde{n})} \end{aligned} \tag{2.5}$$

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

Optimization. Since problem (2.5) is non-smooth involving multiple $\ell_{2,1}$-norms, it is difficult to optimize directly. Half-quadratic optimization techniques [91, 92] have shown to be effective in solving these sparse optimizations in several computer vision applications [236, 190, 242, 154, 16]. Motivated by such methods, we devise an iterative algorithm to efficiently solve (2.5) by minimizing its augmented function alternatively. Specifically, if we define $\phi(x)=\sqrt{x^2+\epsilon}$ with ϵ being a constant, we can transform $\|\mathbf{Z}\|_{2,1}$ to $\sum_{i=1}^{n}\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}$, according to the analysis of $\ell_{2,1}$-norm in [91, 154]. With this transfor-

mation, we can optimize (2.5) efficiently in an alternative way as follows.

According to the half-quadratic theory [91, 92, 75], the augmented cost-function of (4.9) can be written as follows.

$$\begin{aligned}\min_{\mathbf{Z},\,\tilde{\mathbf{Z}}}\ \frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2\big)+\lambda_s\big(tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z})+tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}})\big)\\ +\lambda_d\big(tr(\mathbf{D}^T\mathbf{Z})+tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})\big)+\beta\big(tr(\mathbf{Z}_c^T\mathbf{R}\mathbf{Z}_c)\big)\end{aligned} \tag{2.6}$$

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

$$\mathbf{P}_{i,i}=\frac{1}{2\sqrt{||\mathbf{Z}_i||_2^2+\epsilon}},\qquad \mathbf{Q}_{i,i}=\frac{1}{2\sqrt{||\tilde{\mathbf{Z}}_i||_2^2+\epsilon}},\mathbf{R}_{i,i}=\frac{1}{2\sqrt{||\mathbf{Z}_{ci}||_2^2+\epsilon}} \tag{2.7}$$

where ϵ is a smoothing term, which is usually set to be a small constant value. Optimizing (2.6) over $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ is equivalent to optimizing the following two problems.

$$\min_{\mathbf{Z}}\ \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{Z}\|_F^2+\lambda_d tr(\mathbf{D}^T\mathbf{Z})+\lambda_s tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z})+\beta tr(\mathbf{Z}^T\mathbf{R}\mathbf{Z}) \tag{2.8}$$

$$\min_{\tilde{\mathbf{Z}}}\ \frac{\alpha}{2}\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}\|_F^2+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}})+\lambda_s tr(\tilde{\mathbf{Z}}^T\mathbf{Q}\tilde{\mathbf{Z}})+\beta tr(\tilde{\mathbf{Z}}^T\mathbf{R}\tilde{\mathbf{Z}}) \tag{2.9}$$

Now with fixed $\mathbf{P},\ \mathbf{Q},\ \mathbf{R}$, the optimal solution of (2.8) and (2.9) can be computed by solving the following two linear systems:

$$\begin{aligned}(\mathbf{X}^T\mathbf{X}+2\lambda_s\mathbf{P}+2\beta\mathbf{R})\mathbf{Z}=(\mathbf{X}^T\mathbf{X}-\lambda_d\mathbf{D})\\ (\alpha\mathbf{X}^T\mathbf{X}+2\lambda_s\mathbf{Q}+2\beta\mathbf{R})\tilde{\mathbf{Z}}=(\alpha\mathbf{X}^T\tilde{\mathbf{X}}-\lambda_d\tilde{\mathbf{D}})\end{aligned} \tag{2.10}$$

Algorithm 1 summarizes the alternative minimization procedure to optimize (2.5). In step 1, we compute the auxiliary matrices $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ which play an important role in representative selection, according to the half-quadratic analysis for $\ell_{2,1}$-norm [91]. In step 2, we find the optimal sparse coefficient matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ by solving two linear systems as defined in (2.10). Step 3 corresponds to the consensus matrix, which is expected to uncover the shared knowledge of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ by enforcing same sparse pattern using a joint $\ell_{2,1}$-norm.

Algorithm 1 Algorithm for Solving Problem (2.5)

Input: Video feature matrices $\mathbf{X}$ and $\tilde{\mathbf{X}}$;

Parameters $\alpha, \lambda_s, \lambda_d, \beta$, set $t = 0$;

Construct $\mathbf{D}$ and $\hat{\mathbf{D}}$ using inner product similarity;

Initialize $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ randomly, set $\mathbf{Z}_c$= [$\mathbf{Z}$ | $\tilde{\mathbf{Z}}$] ;

Output: Optimal sparse coefficient matrix **Zc**.

while *not converged* **do**

1. Compute $\mathbf{P}^t$, $\mathbf{Q}^t$ and $\mathbf{R}^t$ using (3.15);
2. Compute $\mathbf{Z}^{t+1}$ and $\tilde{\mathbf{Z}}^{t+1}$ using (2.10);
3. Compute $\mathbf{Z}_c^{t+1}$ as: $\mathbf{Z}_c^{t+1}$= [$\mathbf{Z}^{t+1}$ | $\tilde{\mathbf{Z}}^{t+1}$];
4. $t = t + 1$;

end while

2.3.3 Summary Generation

Above, we described how we compute the optimal sparse coefficient matrix $\mathbf{Z}_c$ by exploiting visual context from the topic-related videos. The consensus matrix of coefficients, $\mathbf{Zc} = [\mathbf{Z}|\hat{\mathbf{Z}}]$ provides information about the contribution of each shot in $\mathbf{X}$ to summarize each video in the collection. To generate a summary, we first sort the shots by decreasing importance according to the ℓ_2 norms of the rows in $\mathbf{Z}_c$ (resolving ties by favoring shorter video shots), and then construct the optimal summary from the top-ranked shots that fit in the length constraint. Note that this also provides scalability to our approach as the ranked list of shots can be used as a scalable representation to provide summary of different lengths as per user request *(analyze once, generate many)*.

2.4 Convergence Analysis

In this section, we will first prove the convergence of our proposed algorithm and then discuss the computational complexity of our method. Since, we have solved (2.5) using an alternating minimization, we would like to show its convergence behavior.

Theorem 1. *Algorithm 1 will monotonically decrease the objective value of (2.5) until it achieves an optimal solution.*

Proof. As seen from (2.6), when we fix $\{\mathbf{P}, \mathbf{Q}, \mathbf{R}\}$ as $\{\mathbf{P}^t, \mathbf{Q}^t, \mathbf{R}^t\}$ in t-th iteration and compute $\mathbf{Z}^{t+1}$, $\tilde{\mathbf{Z}}^{t+1}$, $\mathbf{Z}_c^{t+1}$, the following inequality holds,

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})\\
&+\lambda_s tr((\mathbf{Z}^{t+1})^T\mathbf{P}^t\mathbf{Z}^{t+1})+\lambda_s tr((\tilde{\mathbf{Z}}^{t+1})^T\mathbf{Q}^t\tilde{\mathbf{Z}}^{t+1}))+\beta\big(tr((\mathbf{Z}_c^{t+1})^T\mathbf{R}^t\mathbf{Z}_c^{t+1})\big)\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})\\
&+\lambda_s tr((\mathbf{Z}^{t})^T\mathbf{P}^t\mathbf{Z}^{t})+\lambda_s tr((\tilde{\mathbf{Z}}^{t})^T\mathbf{Q}^t\tilde{\mathbf{Z}}^{t}))+\beta\big(tr((\mathbf{Z}_c^{t})^T\mathbf{R}^t\mathbf{Z}_c^{t})\big)
\end{aligned}\tag{2.11}$$

Adding $\sum_{i=1}^n \frac{\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}$ to both sides of (2.11), we have

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})\\
&+\lambda_s\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}+\lambda_s\sum_{i=1}^n\frac{||\tilde{\mathbf{Z}}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\tilde{\mathbf{Z}}_i^t||_2^2+\epsilon}}+\beta\sum_{i=1}^n\frac{||\mathbf{Z}_{ci}^{t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_{ci}^t||_2^2+\epsilon}}\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})\\
&+\lambda_s\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}+\lambda_s\sum_{i=1}^n\frac{||\tilde{\mathbf{Z}}_i^{t}||_2^2+\epsilon}{2\sqrt{||\tilde{\mathbf{Z}}_i^t||_2^2+\epsilon}}+\beta\sum_{i=1}^n\frac{||\mathbf{Z}_{ci}^{t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_{ci}^t||_2^2+\epsilon}}
\end{aligned}\tag{2.12}$$

According to the *Lemma* in [174]:

$$\sum_{i=1}^n\sqrt{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}-\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t+1}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}\leq\sum_{i=1}^n\sqrt{||\mathbf{Z}_i^{t}||_2^2+\epsilon}-\sum_{i=1}^n\frac{||\mathbf{Z}_i^{t}||_2^2+\epsilon}{2\sqrt{||\mathbf{Z}_i^t||_2^2+\epsilon}}\tag{2.13}$$

Subtracting Eq. (2.13) from Eq. (2.12), we have

$$\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t+1}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t+1}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t+1})\\
+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t+1})&+\lambda_s\big(||\mathbf{Z}^{t+1}||_{2,1}+||\tilde{\mathbf{Z}}^{t+1}||_{2,1}\big)+\beta||{\mathbf{Z}_c}^{t+1}||_{2,1}\\
&\leq\frac{1}{2}\big(\|\mathbf{X}-\mathbf{X}\mathbf{Z}^{t}\|_F^2+\alpha\|\tilde{\mathbf{X}}-\mathbf{X}\tilde{\mathbf{Z}}^{t}\|_F^2\big)+\lambda_d tr(\mathbf{D}^T\mathbf{Z}^{t})\\
+\lambda_d tr(\tilde{\mathbf{D}}^T\tilde{\mathbf{Z}}^{t})&+\lambda_s\big(||\mathbf{Z}^{t}||_{2,1}+||\tilde{\mathbf{Z}}^{t}||_{2,1}\big)+\beta||{\mathbf{Z}_c}^{t}||_{2,1}
\end{aligned}\tag{2.14}$$

which establishes that the objective function (2.5) monotonically decreases in each iteration. Note that the objective function has lower bounds, so it will converge. Empirical results show that the convergence is fast and only a few iterations are needed to converge. Therefore, the proposed method can be applied to large scale problems in practice.

Computational Complexity. For the computational cost of our method, the major bottleneck lies on solving linear systems, where the time complexity is $\mathcal{O}(n^3)$. It can be reduced to $\mathcal{O}(n^{2.376})$ using the Coppersmith-Winograd algorithm [42]. Thus, the total time complexity of our method is $\mathcal{O}(kn^{2.376})$ approximately, where k is the total number of iterations needed to converge. Though developing scalable algorithm is not the main concern of this work, it will be an interesting future work.

2.5 Experiments

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

Datasets. We evaluate the performance of our approach using two benchmark datasets: (i) the CoSum dataset [39] and (ii) the TVSum50 dataset [214]. To the best of our knowledge, these are the only two publicly available summarization datasets of multiple videos orga-

nized into groups with a topic keyword. Both of the datasets are extremely diverse: while CoSum dataset consists of 51 videos covering 10 topics from the SumMe benchmark [84], the TVSum50 dataset contains 50 videos organized into 10 topics from the TRECVid task [212].

Implementation details. For all the videos, we first segment them into multiple shots using the method described in Sec. 3.3.1. Raw features are extracted from the FC6 layer of the C3D network [226]. We then apply temporal mean pooling within each shot to compute a 4096 dimensional feature vector. Our results can be reproduced through the following parameters. The regularization parameters λ_s and β are taken as λ_0/γ where $\gamma > 1$ and λ_0 is analytically computed from the data [58]. The other parameters α and λ_d are empirically set to 0.5 and 0.01 respectively and kept fixed for all results.

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

- **Clustering (`CK` and `CS`):** We first clustered the shots using k-means (`CK`) and spectral clustering (`CS`), with k set to 20 [39]. We then generate a summary by selecting shots that are closest to the centroid of top largest clusters.

- **Sparse Coding (`SMRS` and `LL`):** We tested two approaches: Sparse Modeling Representative Selection (`SMRS`) [58] and LiveLight (`LL`) [273]. `SMRS` finds the representative shots using the entire video as the dictionary and selecting key shots based on the zero patterns of the coding vector. Note that [41] also uses the same objective function as in [58] for summarizing consumer videos. The only difference lies in the algorithm used to solve the objective function (Proximal vs ADMM). Hence, we compared only with [58]. LL generates a summary over time by measuring the redundancy using a

dictionary of shots updated online. We implemented it using SPAMS library [163] with dictionary of size 200 and the threshold $\epsilon_0 = 0.15$, as in [273].

- **Co-occurrence Statistics (`CoC` and `CoSum`):** We compared with two baselines that leverage visual co-occurrence across the topic-related videos to generate a summary. Co-clustering (`CoC`) [54] generates a summary by partitioning the graph into co-clusters such that each cluster contains a subset of shot-pairs with high visual similarity. On the other hand, `CoSum` finds maximal bicliques from the complete bipartite graph using a block coordinate descent algorithm. We generate a summary by selecting top-ranked shots based on the visual co-occurrence score and set the threshold to select maximal bicliques to 0.3, following [39].

All methods (including the proposed one) use the same C3D feature as described in Sec. 2.3.1. Such an experimental setting can give a fair comparison for various methods.

2.5.1 Topic-oriented Video Summarization

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

Solution. The objective function (2.5) extracts summary of a specified video by exploiting the visual context of topic-related videos. Given a set of videos, our approach can find summaries of each video by exploiting the additional knowledge from the remaining videos. Moreover, one can easily parallelize the computation for more computational efficiency given our alternating minimization in Algorithm 1. This provides scalability to our approach in

Table 2.1: Experimental results on CoSum dataset. Numbers show top-5 AP scores averaged over all the videos of the same topic. We highlight the **best** and second best baseline method. Overall, our approach, CVS, performs the best.

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

processing large number of web videos simultaneously.

Evaluation. Motivated by [39, 110], we assess the quality of an automatically generated summary by comparing it to human judgment. In particular, given a proposed summary and a set of human selected summaries, we compute the pairwise average precision (AP) and then report the mean value motivated by the fact that there exists not a single ground truth summary, but multiple summaries are possible. Average precision is a function of both precision and change in recall, where precision indicates how well all the representative shots match with the reference summaries and recall indicates how many and how accurately are the representative shots returned in the retrieval result.

For CoSum dataset, we follow [39] and compare each video summary with five human created summaries, whereas for TVSum50 dataset, we compare each summary with twenty ground truth summaries that are created via crowdsourcing. Since the ground truth annotations in TVSum50 dataset contain frame-wise importance scores, we first compute

Table 2.2: Experimental results on TVSum50 dataset. Numbers show top-5 AP scores averaged over all the videos of the same topic. We highlight the **best** and second best baseline method. Overall, our approach outperforms all the baseline methods.

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

the shot-level importance scores by taking average of the frame importance scores within each shot and then select top 50% shots for each video, as in [39].

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

Comparison with baseline methods. Tab. 2.1 shows the AP on top 5 shots included in the summaries for CoSum dataset. We can see that our method significantly outperforms all baseline methods to achieve an average performance of 85%, while the closest published competitor, `CoSum`, reaches 74%. Moreover, if we compare to the human performance, we

can see that our method even outperforms the `worst human` score of each topic in most cases. This indicates that our method produces summaries comparable to human created summaries. Similarly, for the top-15 results, our approach achieved the highest average score of 83% compared to 69% by the `CoSum` baseline.

Our approach performed particularly well on videos that have their visual concepts described well by the topic-related videos, e.g., a video of the topic *Eiffel Tower* contains shots that shows the night view of the tower and the remaining videos in the collection also depicts this well (see Fig. 2.1). While our method overall produces better summaries, it has a low performance for certain videos, e.g., videos of the topic *Surfing*. These videos contain fast motion and subtle semantics that define representative shots of the video, such as surfing on the wave or sea swimming. We believe these are difficult to capture without an additional semantic analysis [168]. Tab. 2.2 shows top-5 AP results for the TVSum50 datset. Summarization in this dataset is more challenging because of the unconstrained topic keywords. Our approach still outperforms all the alternative methods significantly to achieve an average performance of 74%. Similarly for top-15 results, our approach achieved highest score of 75% compared to 66% by the CoSum baseline.

Test of Statistical Significance. We have done t-test of our results and observe that our approach, `CVS`, statistically significantly outperforms all six compared methods ($p < .01$), except for `worst human`. To further interpret the not-statistically significant result with respect to `worst human`, we performed a statistical power analysis ($\alpha = 0.01$) and see that the power computed for top-5 mAP results on CoSum dataset is 0.279, while on combining with top-15 results, it reaches to 0.877. Similarly, the power reaches 1 for a test that

Table 2.3: Performance comparison between 2D CNN(VGG) and 3D CNN(C3D) features. Numbers show top-5 AP scores averaged over all the videos of the same topic. * abbreviates topic name for display convenience. See Tab. 2.1 for full names.

Methods	Base*	Bike*	Eiffel*	Excv*	Kids*	MLB	NFL	Notre*	Statue*	Surf*	mean
CVS(Features[39])	0.580	0.632	0.677	0.614	0.598	0.607	0.575	0.612	0.655	0.623	0.618
CVS(VGG)	0.591	0.626	0.724	0.638	0.617	0.642	0.615	0.604	0.721	0.649	0.643
CVS(C3D)	0.658	0.675	0.722	0.693	0.707	0.679	0.674	0.702	0.715	0.647	**0.687**

Table 2.4: Ablation analysis of the proposed approach with different constraints on (4.9). Numbers show top-5 AP scores averaged over all the videos of the same topic.

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

combines both top-5 and top-15 results of both of the datasets. Since, power of a high quality test should usually be > 0.80, we can conclude that our approach statistically outperforms the `worst human` for a large sample size.

Effectiveness of C3D features. We investigate the importance and reliability of C3D features by comparing with 2D shot-level deep features, and found that the later produces inferior results, with a top-5 mAP score of 0.643 on the CoSum dataset (Tab. 2.3). We utilize Pycaffe [100] with the VGG net pretrained model [211] 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. 2.3.1. We also compare with the shallow feature representation presented in [39] and observe that C3D features performs significantly better over shallow features in summarizing videos (0.618 vs 0.687). We believe this is because C3D features exploit the temporal aspects of activities typically shown in videos.

Performance of the individual components. To better understand the contribution

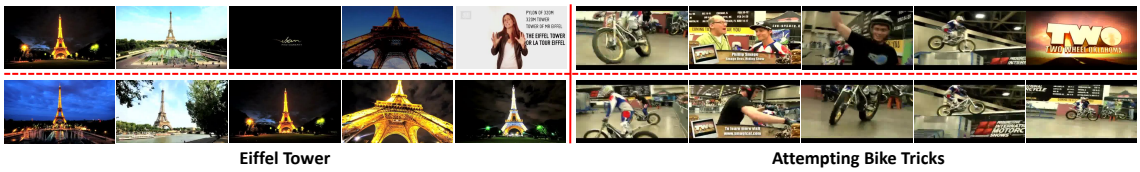


Figure 2.2: Role of topic-related visual context in summarizing a video. Top row: `CVS` w/o topic-related visual context, and Bottom row: `CVS` w/ topic-related visual context. We show two exemplar summaries of the topic *Eiffel Tower* and *Attempting Bike Tricks* from the CoSum and TVSum50 dataset respectively. As can be seen, `CVS` w/o visual context often selects some shots that are irrelevant and not truly related to the topic. `CVS` w/ visual context, on the other hand, automatically select the maximally informative shots by exploiting the information from additional neighborhood videos. Best viewed in color.

of various components in (4.9), we analyzed the performance of the proposed approach, by ablating each constraint while setting corresponding regularizer to zero (Tab. 2.4). With all the components working, the mAP for the CoSum dataset is 0.687. By turning off the neighborhood information from topic-related videos, the mAP decreases to 0.538 (`CVS-Neighborhood`). This corroborates the fact that additional knowledge of topic-related videos help in extracting better summaries, closer to the human selection (see Fig. 2.2 for qualtitative examples). Similarly, by turning off the diversity constraint, the mAP becomes 0.654 (`CVS-Diversity`). We can see that additional knowledge of topic-related videos contributes more than the diversity constraint in summarizing web videos.

2.5.2 Multi-video Concept Visualization

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

Solution. A simple option would be to combine the individual summaries generated from

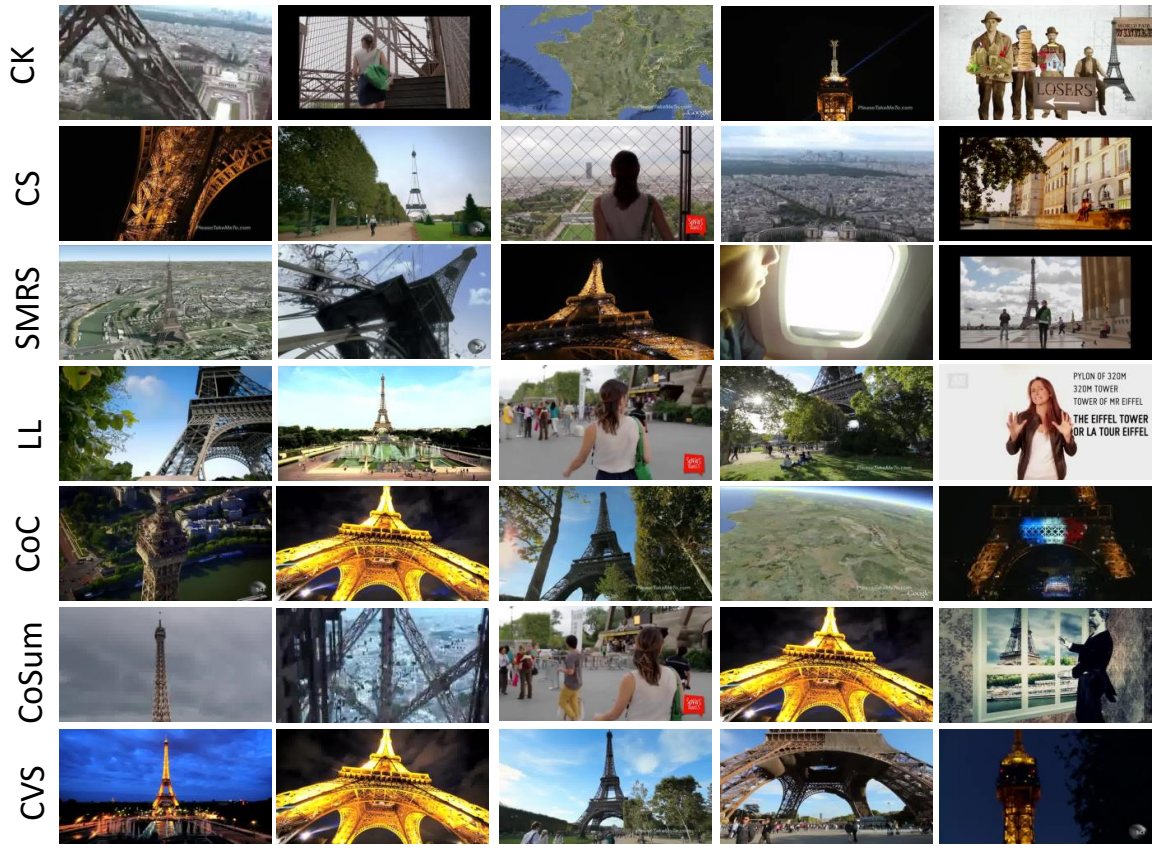


Figure 2.3: Summaries constructed by different methods for the topic *Eiffel Tower*. We show the top-5 results represented by the central frame of each shot. CoSum often select shots that are non-informative about the concept. Our approach selects a diverse set of informative shots that better visualizes the concepts of *Eiffel Tower* (bottom row).

Section. 2.5.1 and select top ranked shots, regardless of the video, as in the existing existing method [39]. However, such choice will produce a lot of redundant events which eventually reduces the quality of the final summary. We believe this is because, although the individual summaries are informative and diverse, there exists redundancy across the extracted summaries that are relevant to the topic. Our approach can handle this by combining the summaries into a single video, say $\mathbf{X}$ and then extracting a single diverse summary using the final objective function (4.9) with setting $(\alpha, \beta, \tilde{\mathbf{D}})$ equal to zero.

Table 2.5: User Study—Average expert ratings in concept visualization experiments. Our approach significantly outperforms other baseline methods in both of the datasets.

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

Evaluation. To evaluate multi-video concept visualization, we need a single ground truth summary of all the topic-related videos that describes the collection altogether. However, since there exists no such ground truth summaries for both of the datasets, we performed human evaluations using 10 experts. Given a video, the study experts were first shown the topic key word (*e.g., Eiffel Tower*) and then shown the summaries constructed using different methods. They were asked to rate the overall quality of each summary by assigning a rating from 1 (worst) to 10 (best). We did not maintain the same order of the summaries across different topics of the dataset. This is to ensure that the users will not be biased from the previous order and ratings while providing ratings for the current topic.

Results. Tab. 2.5 shows average expert ratings for both CoSum and TVSum50 datasets. Similar to the results of topic-oriented summarization, our approach significantly outperforms all the baseline methods which indicates that our method generates a more informative summary that describes the video collection altogether. Furthermore, we note that the relative rank of the different approaches are largely preserved as compared to the topic-oriented summarization results. We show a visual comparison between the summaries produced by different methods in Fig. 2.3. As can be seen, our approach, CVS, generates a summary that better estimates human's visual concepts related to the topic.

2.6 Conclusion

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

Chapter 3

Multi-View Surveillance Video Summarization

3.1 Introduction

Network 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 [99, 199]. 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

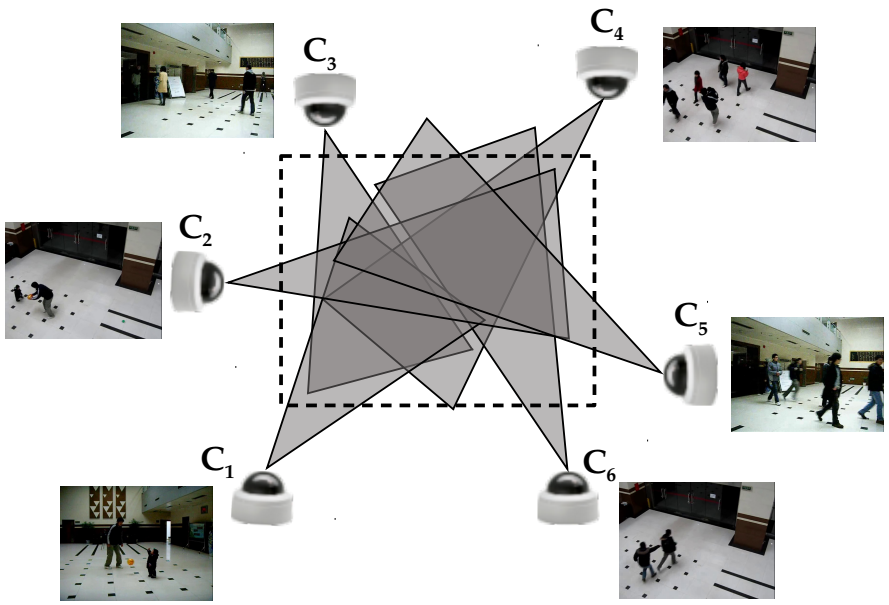


Figure 3.1: An illustration of a multi-view camera network where six cameras C_1, C_2, ..., C_6 are observing an area (black rectangle) from different viewpoints. Since the views are roughly overlapping, information correlations across multiple views along with correlations in each view should be taken into account for generating a concise multi-view summary.

single-view videos in form of a key-frame sequence or a video skim [194, 58, 273, 50, 111, 110, 120]. However, another important problem and rarely addressed in this context is to find an informative summary from *multi-view* videos [67, 132, 178, 119, 182]. *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 important portions of the inputs within a short duration (see Fig. 3.1).* In this work, 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 [37, 173], 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 sum-

maries 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.

3.2 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 [170, 227] 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 [3, 50, 82, 202], attention modeling [160, 84], saliency based linear regression model [126], kernel temporal segmentation [194], crowd-sourcing [110], energy minimization [196, 65], storyline graphs [111], submodular maximization [86], determinantal point process [77, 269], archetypal analysis [214], long short-term memory [270] and maximal biclique finding [39].

Recently, there has been a growing interest in using sparse coding (SC) to solve the problem of video summarization [58, 273, 41, 169, 56, 167] 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 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 [67] and rough sets [132] to summarize multi-view videos. A recent work in [119] 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 [178]. However, this method relies on inter-camera frame correspondence, which can be a very difficult problem in uncontrolled settings. The work in [128] and [129] also addresses a similar problem of summarization in non-overlapping camera networks. Learning from multiple information sources such as video tags [237], topic-related web videos [185, 186] and non-visual data [289, 238] is also a recent trend in multiple web video summarization.

3.3 Proposed Methodology

In this section, we start by giving main notations and definition of the 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 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 $\ell_{2,1}$-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 3.3.1). 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 3.3.2, 3.3.3, 3.3.4). 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 3.5).

3.3.1 Video Representation

Video representation 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 [39]. 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% [39]. 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 [210, 108]. An advantage of using deep learning features is that there exist accurate, large-scale datasets such as Imagenet [200] and Sports-1M [108] from which they can be extracted. For videos, C3D features [226] have recently shown better performance compared to the features extracted using each frame separately [226, 264]. 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 [226]. 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.

3.3.2 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 [62], 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* [62] 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

$$x_i^{(k)} = X^{(k)} c_i^{(k)},\ c_{ii}^{(k)} = 0, \tag{3.1}$$

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 (3.1), 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 ℓ_0 norm, i.e.,

$$\min\ ||c_i^{(k)}||_1\ \text{ s.t. }\ x_i^{(k)} = X^{(k)} c_i^{(k)},\ c_{ii}^{(k)} = 0, \tag{3.2}$$

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

$$\min\ ||C^{(k)}||_1\ \text{ s.t. } X^{(k)} = X^{(k)} C^{(k)},\ \text{diag}(C^{(k)}) = 0, \tag{3.3}$$

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 ℓ_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 ℓ_1 optimization as in intra-view similarities. Specifically, we calculate the pairwise similarity between m-th and n-th view by solving the following optimization problem:

$$\min \ ||C^{(m,n)}||_1 \ \text{ s.t. } \ X^{(m)} = X^{(n)}C^{(m,n)}, \tag{3.4}$$

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 (3.4), 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.

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

$$= \sum_{k}\sum_{i,j} ||y_i^{(k)} - y_j^{(k)}||^2 C_{intra}^{(k)}(i,j) + \sum_{\substack{m,n \\ m\neq n}}\sum_{i,j} ||y_i^{(m)} - y_j^{(n)}||^2 C_{inter}^{(m,n)}(i,j) \tag{3.5}$$

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 (3.5) can be rewritten using one similarity matrix defined over the whole set of video shots as

$$\mathcal{J}(Y) = \sum_{m,n}\sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 C_{total}^{(m,n)}(i,j) \tag{3.6}$$

where the total similarity matrix is defined as

$$C_{total}^{(m,n)}(i,j) = \begin{cases} C_{intra}^{(k)}(i,j) & \text{if } m = n = k \\ C_{inter}^{(m,n)}(i,j) & \text{otherwise} \end{cases} \tag{3.7}$$

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 (3.7) is that since each ℓ_1 optimization is solved individually, a fast parallel computing strategy can be easily adopted for efficiency. However, the matrix in (3.7) is not symmetric since in ℓ_1

optimization (3.2, 3.4), 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 (3.6) with a symmetric similarity matrix $W = C_{total} + C_{total}^T$ as

$$\mathcal{F}(Y) = \sum_{m,n} \sum_{i,j} ||y_i^{(m)} - y_j^{(m)}||^2 W^{(m,n)}(i,j) \tag{3.8}$$

With the above formulation, we make sure that x_i and x_j get connected to each other if either x_i and x_j is in the sparse representation of the other. 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 (3.8) reduces to the Laplacian embedding [13] of shots defined by the similarity matrix W. So, the optimization problem can be written as

$$Y^* = \underset{Y, YY^T=I}{\operatorname{argmin}} \; tr\big(YLY^T\big) \tag{3.9}$$

where L is the graph Laplacian matrix of W and I is an identity matrix. Minimizing (3.9) 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 [70, 155, 175].

3.3.3 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 ℓ_1 regularization on rows of the sparse coefficient matrix [41, 58]. By introducing row sparsity regularizer, the summarization problem can now be succinctly formulated as

$$\min_{Z \in \mathbb{R}^{N \times N}} \quad ||Z||_{2,1} \;\; \text{s.t.} \;\; Y = YZ \tag{3.10}$$

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, (3.10) can be written as

$$\min_{Z} \|Y - YZ\|_F^2 + \lambda \|Z\|_{2,1} \tag{3.11}$$

where λ balances the weight of the two terms. Once (3.11) 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 (3.3) and (3.10) look similar; however, the nature of sparse regularizer in both formulations are completely different. In (3.3), the objective of ℓ_1 regularizer is to induce element wise sparsity in a column whereas in (3.10), the objective of $\ell_{2,1}$ regularizer is to induce row level sparsity in a matrix.

Remark 2. Given non-uniform length of shots, (3.11) can be modified to a weighted $\ell_{2,1}$-norm based objective to consider length of shots while selecting representatives as

$$\min_{Z} \|Y - YZ\|_F^2 + \lambda \|QZ\|_{2,1} \tag{3.12}$$

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 (3.12) favors selection of shorter video shots by assigning a lower score via Q. In other words, problem (3.12) 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 video summary.

3.3.4 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 (3.9) and (3.11), the joint objective function becomes:

$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha\big(||Y - YZ||_F^2 + \lambda||Z||_{2,1}\big) \tag{3.13}$$

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.

3.4 Optimization

The optimization problem in (3.13) 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 [91, 92] have shown to be effective in solving these sparse optimizations in several vision and multimedia applications [236, 242, 190, 154]. Motivated by such methods, we devise an iterative algorithm to efficiently solve (3.13) by minimizing its augmented function alternatively[1]. Specifically, if we define $\phi(x) = \sqrt{x^2 + \epsilon}$ with ϵ 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 [91, 154]. With this transformation, we can optimize (3.13) efficiently in an alternative way as follows.

According to the half-quadratic theory [91, 92, 75], the augmented cost-function of (3.13) can be written as

$$\min_{Y,Z,YY^T=I} tr(YLY^T) + \alpha\big(||Y - YZ||_F^2 + \lambda tr(Z^T P Z)\big) \tag{3.14}$$

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

$$P_{i,i} = \frac{1}{2\sqrt{||z^i||_2^2 + \epsilon}} \tag{3.15}$$

where ϵ is a smoothing term with small constant value. With this transformation, note that the problem (3.14) is convex separately with respect to Y, Z, and P. Hence, we can solve

[1]We solve all the sparse optimization problems using Half-quadratic optimization techniques [91, 92]. Due to space limitation, we only present the optimization procedure to solve (3.13). However, the same procedure can be easily extended to solve other sparse optimizations (3.3, 3.4).

(3.14) 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:

$$\min_{Z} \alpha\big(tr((Y - YZ)(Y - YZ)^T) + \lambda tr(Z^T P Z)\big) \tag{3.16}$$

By setting derivative of (3.16) with respect to Z to zero, the optimal solution can be computed by solving the following linear system.

$$(Y^T Y + \lambda P)Z = Y^T Y \tag{3.17}$$

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

$$\begin{aligned} &\min_{Y, YY^T = I} tr(YLY^T) + \alpha tr((Y - YZ)(Y - YZ)^T) \\ &= \min_{Y, YY^T = I} tr(Y(L + \alpha(I - 2Z + ZZ^T))Y^T) \end{aligned} \tag{3.18}$$

Eq. 3.18 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. 3.15 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 (3.9) 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.

3.5 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:

- We first generate a weight curve using ℓ_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.

- 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

Table 3.1: 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

more flexibility for practical applications. More details on the summary length and scalability are included in experiments.

3.6 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.

3.6.1 Datasets and Settings

We conduct rigorous experiments using 6 multi-view datasets with 36 videos in total, which are from [67, 178] (see Tab. 3.1). 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 [67, 119, 132]. It is important to note that experiments in our prior work [183] 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.

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 [24]. (iv) The sparsity regularization parameter λ is computed as λ_0/ρ and λ_0 is analytically computed from the embedded points [58]. (v) We empirically set α to 0.05 and kept fixed for all results.

3.6.2 Performance Measures

To provide an objective comparison, we compare all the approaches using three quantitative measures, including Precision, Recall and F-measure ($\frac{2\times Precision\times Recall}{Precision+Recall}$) [67, 119]. For all these metrics, the higher value indicates better summarization quality. We set the same summary length as in [67] to generate our summaries and employ the ground truth of important events reported in [67] 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 [67, 119, 178] 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 [67, 132, 119, 182, 183].

3.6.3 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` [67]. 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` [132]. The method first adopt a SVM classifier as the key frame abstraction process and then applies rough set to remove similar frames.
- `BipartiteOPF` [119]. 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` [178]. 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` ourselves using the same video representation as the proposed one and tune their parameters to have the best performance.

Table 3.2: 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 [160]
`Sparse-Concate`	100	50	66.67	56	55	55.70	91	70	78.95	TMM2012 [41]
`Concate-Attention`	100	38	55.07	56	48	51.86	95	72	81.98	TMM2005 [160]
`Concate-Sparse`	93	58	71.30	56	62	58.63	86	70	77.18	TMM2012 [41]
`Graph`	100	26	41.26	50	48	49.13	100	58	73.41	TCSVT2006 [191]
`RandomWalk`	100	61	75.77	70	55	61.56	100	77	86.81	TMM2010 [67]
`RoughSets`	100	61	75.77	69	57	62.14	97	74	84.17	ICIP2011 [132]
`BipartiteOPF`	100	69	81.79	75	69	71.82	100	79	88.26	TMM2015 [119]
`Ours`	**100**	**81**	**89.36**	**84**	**72**	**77.78**	**100**	**86**	**92.52**	Proposed

Results. Table 3.2 shows the results on three multi-view datasets, namely Office, Campus and Lobby datasets. We have the following key observations from Table 3.2: (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. 3.2 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

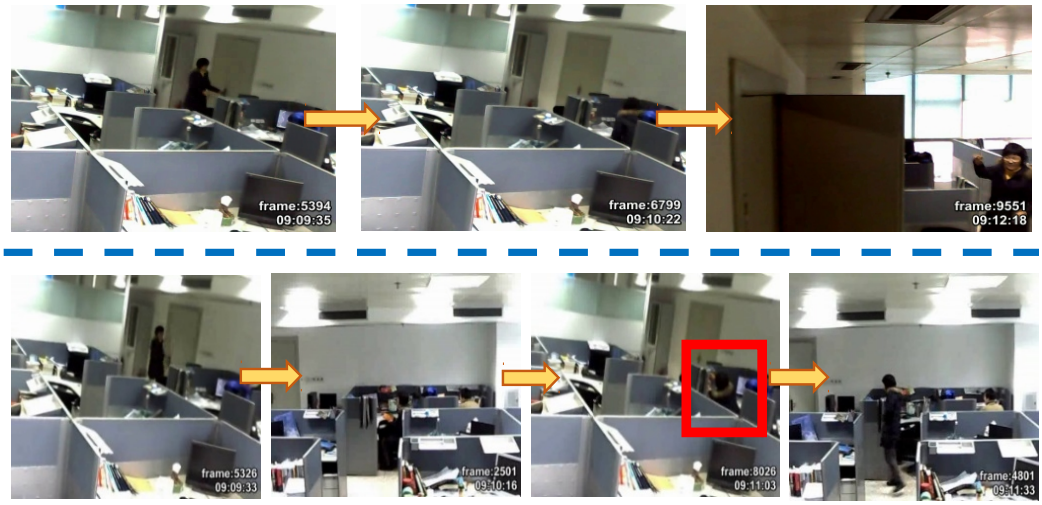


Figure 3.2: Sequence of events detected related to activities of a member (A_0) inside the Office dataset. Top row: Summary produced by method [67], and Bottom row: Summary produced by our approach. Sequence of events detected in top row: 1st: A_0 enters the room, 2nd: A_0 sits in cubicle 1, 3rd: A_0 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 [67]), and 4th: A_0 leaves the room. The event of looking for a thick book to read (as per the ground truth in [67]) is missing in the summary produced by method [67] 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 [67]. Best viewed in color.

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 3.2 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 [67]. As a result, number of events 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. 3.3 for an illustrative example).

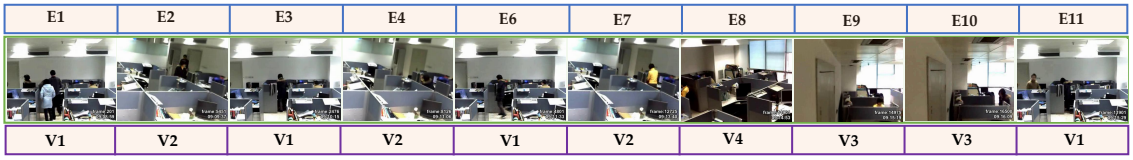


Figure 3.3: Summarized events for the Office dataset. Each event is represented by a key frame and is associated with two numbers, one above and below of the key frame. Numbers above the frame (`E1`, $\cdots$, `E26`) represent the event number whereas the numbers below (`V1`, $\cdots$, `V4`) indicate 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 [67].

Table 3.3: Performance Comparison with `GMM` baseline on BL-7F Dataset

Methods	Precision(%)	Recall(%)	F-measure(%)	Reference
GMM	58	61	60.00	JSTSP2015 [178]
Ours	73	70	71.29	Proposed

Table 3.3 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 [178]. From Table 3.3, it is clearly evident that our approach significantly outperforms `GMM` in generating more informative multi-view summaries. The F-measure of our method is about 11% better than that of `GMM` [178]. 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 [178] and compute the performance measures in the unit of frames instead of events as in Table 3.2 to make a fair comparison with the `GMM` baseline.

3.6.4 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, namely, `Attention-Concate` [160], `Sparse-Concate` [41], `Concate-Attention` [160], `Concate-Sparse` [41], and `Graph` [191] 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 (3.11) to summarize multi-view videos with out any embedding. 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 [67, 119].

Results. We have the following key findings from Table 3.2 and Fig. 3.4: (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

Figure 3.4: Some summarized events for the Lobby dataset. Top row: summary produced by `Sparse-Concate` [41], Middle row: summary produced by `Concate-Sparse` [41], 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 [67] 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.

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. 3.4 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

Figure 3.5: 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.

multi-view correlations and sparse representative selection.

3.6.5 Scalability in Generating Summaries

Scalability in generating summaries of different length has shown to be effective while summarizing single videos [93, 184]. 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. 3.5 shows the generated summaries of length 3, 5 and 7 most important shots (as determined by the weight curve described in Sec. 3.5) for Office dataset.

3.6.6 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 [211] to extract a 4096-dim feature vector of a frame and then use temporal mean pooling to compute a single shot-level feature vector, similar to C3D features described in Sec. 3.3.1. The spatio-temporal C3D features perform best, as they exploit the temporal aspects of activities typically shown in videos.

3.6.7 Performance Analysis with Video Segmentation

We examined the performance of our approach by replacing the temporal segmentation algorithm [39] 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 [39], 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., [193] in generating video summaries—we expect such advanced and complex video segmentation algorithms will only benefit our proposed approach.

3.6.8 Performance Comparison with [183]

We now compare the proposed approach with [183] to explicitly verify the effectiveness of video representation and joint optimization for summarizing multi-view videos. Table 3.4 shows the comparison with [183] on Office, Campus and Lobby datasets. Following are the analysis of the results: (i) The proposed framework consistently outperforms [183] 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 the more challenging Campus dataset which demonstrates that the current framework is more effective in summarizing videos with outdoor scenes.

Table 3.4: F-measure Comparison with [183]

Methods	Office	Campus	Lobby	Reference
[183]	84.48	75.42	88.26	ICPR2016 [183]
Ours	89.36	77.78	92.52	Proposed

Table 3.5: 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

(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 [183], 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 [183] is limited to key frame extraction only and hence may not be suitable for many surveillance applications where video skims with motion information seems better suited for obtaining significant information in short time.

3.6.9 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 3.5. 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 3.2.

3.6.10 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 (3.13) 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 [67]). 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.

3.7 Conclusion

In this work, we addressed the problem of summarizing multi-view videos via joint embedding learning and $\ell_{2,1}$ sparse optimization. The embedding helps in 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.

Chapter 4

On-boarding New Camera(s) in Person Re-identification

4.1 Introduction

Person re-identification (re-id), which addresses the problem of matching people across non-overlapping views in a multi-camera system, has drawn a great deal of attention in the last few years [105, 229, 279]. Much progress has been made in developing methods that seek either the best feature representations (e.g., [248, 143, 11, 146]) or propose to learn optimal matching metrics (e.g., [179, 140, 251, 258, 34, 267, 7]). While they have obtained reasonable performance on commonly used benchmark datasets (e.g., [73, 45, 278]), we believe that these approaches have not yet considered a fundamental related problem: *Given a camera network where the inter-camera transformations/distance metrics have been learned in an intensive training phase, how can we incorporate new camera(s) into the*

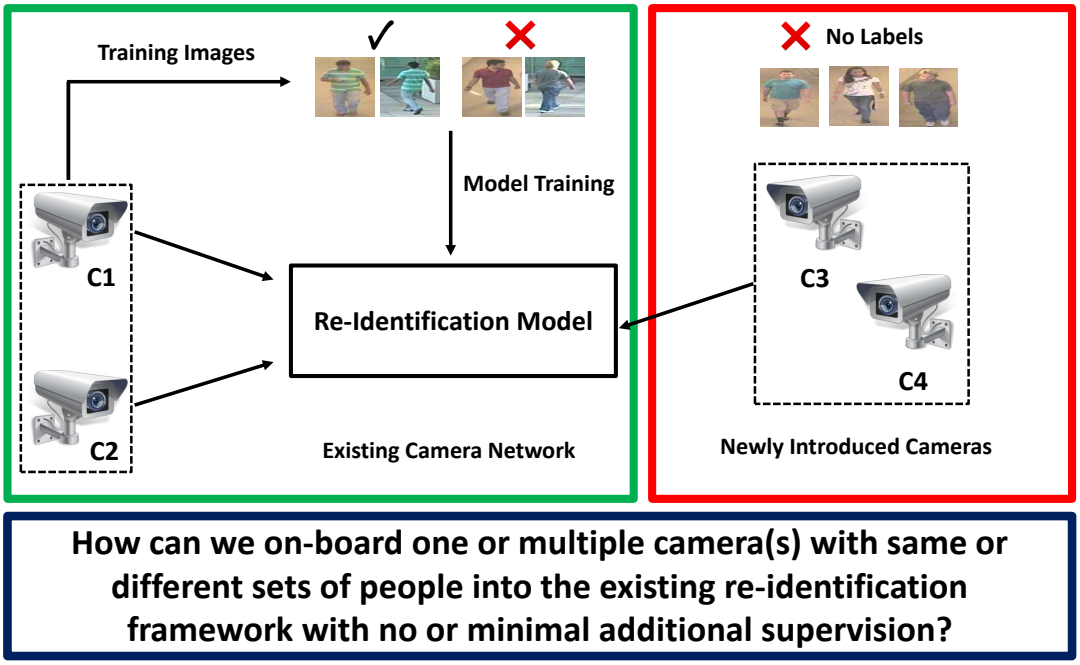


Figure 4.1: Consider an existing network with two cameras $\mathtt{C}_1$ and $\mathtt{C}_2$ where we have learned a re-id model using pair-wise training data from both of the cameras. During the operational phase, two new cameras $\mathtt{C}_3$ and $\mathtt{C}_4$ are introduced to cover a certain area that is not well covered by the existing 2 cameras. Most of the existing methods do not consider such dynamic nature of a re-id model. In contrast, we propose an unsupervised approach for on-boarding new camera(s) into the existing re-identification framework by exploring: *what is the best source camera(s) to pair with the new cameras and how can we exploit the best source camera(s) to improve the matching accuracy across the other existing cameras?*

installed system with minimal additional effort? This is an important problem to address in realistic open-world re-identification scenarios, where one or multiple new cameras may be temporarily inserted into an existing system to get additional information.

To illustrate such a problem, let us consider a scenario with $\mathcal{N}$ cameras for which we have learned the optimal pair-wise distance metrics, so providing high re-identification accuracy for all camera pairs. However, during a particular event, a new camera may be temporarily introduced to cover a certain related area that is not well-covered by the

existing network of $\mathcal{N}$ cameras (see Fig. 4.1 for an example). Despite the dynamic and open nature of the world, almost all work in re-identification assume a *static* and *closed* world model of the re-id problem where the number of cameras are fixed in a network. Given a newly introduced camera, traditional re-id methods will try to relearn the inter-camera transformations/distance metrics using a costly training phase. This is impractical since labeling data in the new camera and then learning transformations with the others is time-consuming, and defeats the entire purpose of temporarily introducing the additional camera. Thus, there is a pressing need to develop *unsupervised* approaches for integrating new camera(s) into an existing re-identification framework with limited supervision.

Domain adaptation [121, 188] has recently been successful in many vision problems such as object recognition [201, 80] and activity classification [161, 261] with multiple classes or domains. The main objective is to scale learned systems from a source domain to a target domain without requiring prohibitive amount of training data in the target domain. Considering newly introduced camera(s) as target domain, we pose an important question in this work: *Can unsupervised domain adaptation be leveraged upon for on-boarding new camera(s) into person re-identification frameworks with limited supervision?*

Unlike object recognition [201], domain adaptation for person re-identification has additional challenges. A central issue in domain adaptation is *which source to transfer from.* When there is only one source of information available which is highly relevant to the task of interest, then domain adaptation is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying relevance. Re-identification in a dynamic network falls into the latter, more difficult case. Specifically,

given multiple source cameras (already installed) and a target camera (newly introduced), *how can we select the best source camera to pair with the target camera?* The problem can be easily extended to multiple additional cameras being introduced.

Moreover, once the best source camera is identified, *how can we exploit this information to improve the re-identification accuracy of other camera pairs?* For instance, let us consider $\mathtt{C}_1$ being the best source camera for the newly introduced camera $\mathtt{C}_3$ in Fig. 4.1. Once the pair-wise distance metric between $\mathtt{C}_1$ and $\mathtt{C}_3$ is obtained, can we exploit this information to improve the re-identification accuracy across ($\mathtt{C}_2$–$\mathtt{C}_3$)? This is an especially important problem because it will allow us to now match data in the newly inserted target camera $\mathtt{C}_3$ with all the previously installed cameras.

Given a network with thousands of cameras involving large number of images, finding the best source camera for a newly introduced camera can involve intensive computation of the pair-wise kernels over the whole set of images. Thus, it is important to automatically select an informative subset of the source data to pair with the target domain data. Specifically, *can we select an informative subset of source camera data that share similar characteristics as target camera data and use those for model adaptation in resource constrained environments?* This is crucial to increase the flexibility and decrease the deployment cost of newly introduced cameras in large-scale dynamic camera networks.

Overview of Solution Strategy. We first propose an approach based on geodesic flow kernel [78, 80] that can effectively find the best source camera to adapt with a target camera. Given camera pairs, each consisting of 1 (out of $\mathcal{N}$) source camera and a target camera, we first compute a kernel over the subspaces representing the data of both cameras and then

use it to find the kernel distance across the source and target camera. Then, we rank the source cameras based on the average distance and choose the one with lowest distance as the best source camera to pair with the target camera. This is intuitive since a camera which is closest to the newly introduced camera will give the best re-identification performance on the target camera and hence, is more likely to adapt better than others. In other words, a source camera with lowest distance with respect to a target camera indicates that both of the sensors could be similar to each other and their features may be similarly distributed. Note that we learn the kernel with the labeled data from the source camera only.

We then introduce a transitive inference algorithm for person re-identification that can exploit information from best source camera to improve accuracy across other camera pairs. Reminding the previous example in Fig. 4.1 in which source camera $\mathtt{C}_1$ best matches with target camera $\mathtt{C}_3$, our proposed transitive algorithm establishes a path between camera pair ($\mathtt{C}_2$ – $\mathtt{C}_3$) by marginalization over the domain of possible appearances in best source camera $\mathtt{C}_1$. Specifically, $\mathtt{C}_1$ plays the role of a "connector" between $\mathtt{C}_2$ and $\mathtt{C}_3$. Experiments show that this approach consistently increases the overall re-identification accuracy in multiple networks by improving matching performance across camera pairs, while exploiting side information from best source camera.

Moreover, we also propose a source-target selective adaptation strategy that uses a subset of source camera data instead of all existing data to compute the kernels for finding the best source camera to pair with a target camera. Our key insight is that not all images in a source camera are equally effective in terms of adaptability and hence using an informative subset of images from the existing source cameras whose characteristics are

similar to those of the target camera can well adapt the models in resource constrained environments. We develop a target-aware sparse prototype selection strategy using $\ell_{2,1}$-norm optimization to select a subset of source data that can efficiently describe the target set. Experiments demonstrate that our source-target selective learning strategy achieves the same performance as the full set while only using about 30% of images from the source cameras. Interestingly, our approach with prototype selection outperforms the compared methods that use all existing source data by a margin of about 8%-10% in rank-1 accuracy while only requiring about 10% of source camera data while introducing new cameras.

Contributions. We address a novel and very practical problem—how to on-board new camera(s) to an existing re-identification framework without adding a very expensive training phase. Towards solving this problem, we make the following contributions.

- We develop an unsupervised approach based on geodesic flow kernel that can find the best source camera to adapt with the newly introduced target camera(s).
- We propose a transitive inference algorithm to exploit side information from the best source camera to improve the matching accuracy across other camera pairs.
- We also develop a target-aware sparse prototype selection strategy using $\ell_{2,1}$-norm optimization to select an informative subset of source camera data for data-efficient learning in resource constrained environments.

4.2 Related Work

Person re-identification has been studied from different perspectives (see [279] for a survey). Here, we focus on some representative methods closely related to our work.

Supervised Re-identification. Most existing person re-identification techniques are based on supervised learning. These methods either seek the best feature representation [248, 143, 11, 164, 277] or learn discriminant metrics/dictionaries [115, 189, 140, 283, 189, 139, 141, 222, 221, 97, 106, 276, 259, 34, 267, 7] that yield an optimal matching score between two cameras or between a gallery and a probe image. By learning listwise [31] and pairwise [283] similarities as well as mixture of polynomial kernel-based models [30], different solutions yielding similarity measures have also been investigated.

Recently, deep learning methods have shown significant performance improvement on person re-id [266, 256, 249, 55, 35, 246, 148, 287, 40, 274, 131, 197, 288]. Combining feature representation and metric learning with an end-to-end deep neural networks is also a recent trend in person re-identification [1, 135, 250]. Considering that a modest-sized camera network can easily have hundreds of cameras, these supervised re-id models will require huge amount of labeled data which are difficult to collect in real-world settings. In an effort to bypass tedious labeling of training data in supervised re-id models, there has been recent interest in using active learning for labeling examples in an interactive manner [144, 243, 46, 165, 235]. However, all these approaches consider a static camera network unlike the problem domain we consider.

Unsupervised Re-identification. Unsupervised learning models have received little attention in re-identification because of their weak performance on benchmarking datasets

compared to supervised methods. Representative methods along this direction use either hand-crafted appearance features [159, 145, 158, 36, 151] or saliency statistics [275] for matching persons without requiring huge amount of labeled data. Dictionary learning based methods have also been utilized in an unsupervised setting [113, 150, 114, 4]. Recently, Generative Adversarial Networks (GAN) has also been used in semi-supervised settings [285, 244]. Although being scalable in real-world settings, these approaches have not yet considered the dynamic nature of the re-identification problem, where new cameras can be introduced at any time to an existing network.

Open World Re-Identification. Open world recognition has been introduced in [15] as an attempt to move beyond the static setting to a dynamic and open setting where the number of training images/classes are not fixed in recognition. Recently there have been few works in re-identification [284, 26, 290] which try to address the open world scenario by assuming that gallery and probe sets contain different identities of persons. Unlike such approaches, we consider another yet important aspect of open world re-identification, i.e. the intrinsic dynamic network of cameras where a new camera has to be incorporated in the system with minimal additional effort. Unlike such approaches, we consider another yet important aspect of open world re-identification where the camera network is dynamic and the system has to incorporate a new camera with minimal additional effort.

Domain Adaptation. Domain adaptation [188], which aims to adapt a source domain to a target domain, has been successfully used in many areas of computer vision and image processing, e.g., object classification, and action recognition. Despite its applicability in classical vision tasks, domain adaptation for re-identification still remains as a challenging

and under addressed problem. Recently, domain adaptation for re-id has begun to be considered [134, 282, 124, 241, 157]. However, these studies consider only improving the re-id performance in a static camera network with fixed number of cameras. Furthermore, most of these approaches learn supervised models using labeled data from the target domain. In contrast, we propose an unsupervised approach that permit re-identification in a newly introduced camera without any labeled data.

4.3 Proposed Methodology

To on-board new camera(s) into an existing re-identification framework, we first formulate an unsupervised approach based on geodesic flow kernel which effectively finds the best source camera (out of multiple installed ones) to pair with newly introduced target camera(s) with minimal additional effort (Sec. 4.3.2). Then, to exploit information from the best source camera, we propose a transitive inference algorithm that improves the matching performance across other source-target camera pairs in a network (Sec. 4.3.3). We describe the details on target-aware sparse prototype selection to select an informative subset of source camera data in Sec. 4.3.4. Finally, we present extensions of our proposed approach to more realistic scenarios where multiple cameras are introduced to an existing network at the same time (Sec. 4.3.5) and labeled data from the newly introduced camera is available for semi-supervised adaptation (Sec. 4.3.6).

4.3.1 Initial Setup

Our proposed framework starts with an installed camera network where the discriminative distance metrics between each camera pairs is learned using a off-line intensive training phase. Let there be $\mathcal{N}$ cameras in a network and the number of possible camera pairs is $\binom{\mathcal{N}}{2}$. Let $\{(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_i^{\mathcal{B}})\}_{i=1}^{m}$ be a set of training samples, where $\mathbf{x}_i^{\mathcal{A}} \in \mathbb{R}^D$ represents feature representation of a training sample from camera view $\mathcal{A}$ and $\mathbf{x}_i^{\mathcal{B}} \in \mathbb{R}^D$ represents feature representation of the same person in a different camera view $\mathcal{B}$.

Given the training data, we follow KISS metric learning (KISSME) [116] and compute the pairwise distance matrices such that distance between images of the same individual is less than distance between images of different individuals. The basic idea of KISSME is to learn the Mahalanobis distance by considering a log likelihood ratio test of two Gaussian distributions. The likelihood ratio test between dissimilar pairs and similar pairs can be written as

$$\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \log \frac{\frac{1}{\mathcal{C}_{\mathcal{D}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{D}}^{-1} \mathbf{x}_{ij})}{\frac{1}{\mathcal{C}_{\mathcal{S}}} \exp(-\frac{1}{2}\mathbf{x}_{ij}^T \Sigma_{\mathcal{S}}^{-1} \mathbf{x}_{ij})} \tag{4.1}$$

where $\mathbf{x}_{ij} = \mathbf{x}_i^{\mathcal{A}} - \mathbf{x}_j^{\mathcal{B}}$, $\mathcal{C}_{\mathcal{D}} = \sqrt{2\pi|\Sigma_{\mathcal{D}}|}$, $\mathcal{C}_{\mathcal{S}} = \sqrt{2\pi|\Sigma_{\mathcal{S}}|}$, $\Sigma_{\mathcal{D}}$ and $\Sigma_{\mathcal{S}}$ are covariance matrices of dissimilar and similar pairs respectively. With simple manipulations, (4.1) can be written as $\mathcal{R}(\mathbf{x}_i^{\mathcal{A}}, \mathbf{x}_j^{\mathcal{B}}) = \mathbf{x}_{ij}^T \mathbf{M} \mathbf{x}_{ij}$, where $\mathbf{M} = \Sigma_{\mathcal{S}}^{-1} - \Sigma_{\mathcal{D}}^{-1}$ is the Mahalanobis distance between covariances associated to a pair of cameras. We perform an Eigen-analysis to ensure $\mathbf{M}$ is positive semi-definite, as in [116]. Note that our approach is agnostic to the choice of metric learning algorithm used to learn the optimal metrics across camera pairs in an existing network. We adopt KISSME in this work since it is simple to compute and has shown to perform satisfactorily on the person re-identification problem.

4.3.2 Discovering the Best Source Camera

Objective. Given an existing camera network where matching metrics across all camera pairs are computed using the above training phase, our first objective is to select the best source camera which has the lowest kernel distance with respect to the newly inserted camera. Towards this, we adopt an unsupervised strategy based on geodesic flow kernel [78, 80] to compute the distances without requiring any labeled data from the target camera.

Approach Details. Our approach for discovering the best source camera consists of the following steps: (i) compute geodesic flow kernels between the new (target) camera and other existing cameras (source); (ii) use the kernels to determine the distance between them; (iii) rank the source cameras based on distance with respect to the target camera and choose the one with the lowest as best source camera.

Let $\{\mathcal{X}^s\}_{s=1}^{\mathcal{N}}$ be the $\mathcal{N}$ source cameras and $\mathcal{X}^{\mathcal{T}}$ be the newly introduced target camera. To compute the kernels in an unsupervised way, we extend a previous method [78] that adapts classifiers in the context of object recognition to the re-identification in a dynamic camera network. The main idea of our approach is to compute the low-dimensional subspaces representing data of two cameras (one source and one target) and then map them to two points on a Grassmanian[1]. Intuitively, if these two points are close by on the Grassmanian, then the computed kernel would provide high matching performance on the target camera. In other words, both of the cameras could be similar to each other and their features may be similarly distributed over the corresponding subspaces. For simplicity, let us assume we are interested in computing the kernel matrix $\mathbf{K}^{\mathcal{ST}} \in \mathbb{R}^{D \times D}$ between the source camera

[1]Let d being the subspace dimension, the collection of all d-dimensional subspaces form the Grasssmanian.

$\mathcal{X}^{\mathcal{S}}$ and a newly introduced target camera $\mathcal{X}^{\mathcal{T}}$. Let $\tilde{\mathcal{X}}^{\mathcal{S}} \in \mathbb{R}^{D\times d}$ and $\tilde{\mathcal{X}}^{\mathcal{T}} \in \mathbb{R}^{D\times d}$ denote the d-dimensional subspaces, computed using Partial Least Squares (PLS) and Principal Component Analysis (PCA) on the source and target camera, respectively. Note that we can not use PLS on the target camera since it is a supervised dimension reduction technique and requires label information for computing the subspaces.

Given both of the subspaces, the closed loop solution to the geodesic flow kernel across two cameras is defined as

$$\mathbf{x}_i^{\mathcal{S}^T}\mathbf{K}^{\mathcal{S}\mathcal{T}}\mathbf{x}_j^{\mathcal{T}} = \int_0^1 (\psi(\mathbf{y})^T\mathbf{x}_i^{\mathcal{S}})^T(\psi(\mathbf{y})\mathbf{x}_j^{\mathcal{T}})\ d\mathbf{y} \tag{4.2}$$

where $\mathbf{x}_i^{\mathcal{S}}$ and $\mathbf{x}_j^{\mathcal{T}}$ represent feature descriptor of i-th and j-th sample in source and target camera respectively. $\psi(\mathbf{y})$ is the geodesic flow parameterized by a continuous variable $\mathbf{y} \in [0, 1]$ and represents how to smoothly project a sample from the original D-dimensional feature space onto the corresponding low dimensional subspace. The geodesic flow $\psi(\mathbf{y})$ over two cameras can be defined as [78],

$$\psi(\mathbf{y}) = \begin{cases} \tilde{\mathcal{X}}^{\mathcal{S}} & \text{if } \mathbf{y} = 0 \\ \tilde{\mathcal{X}}^{\mathcal{T}} & \text{if } \mathbf{y} = 1 \\ \tilde{\mathcal{X}}^{\mathcal{S}}\mathcal{U}_1\mathcal{V}_1(\mathbf{y}) - \tilde{\mathcal{X}}_o^{\mathcal{S}}\mathcal{U}_2\mathcal{V}_2(\mathbf{y}) & \text{otherwise} \end{cases} \tag{4.3}$$

where $\tilde{\mathcal{X}}_o^{\mathcal{S}} \in \mathbb{R}^{D\times(D-d)}$ is the orthogonal matrix to $\tilde{\mathcal{X}}^{\mathcal{S}}$ and $\mathcal{U}_1, \mathcal{V}_1, \mathcal{U}_2, \mathcal{V}_2$ are given by the following pairs of Singular Value Decompositions (SVDs),

$$\mathcal{X}^{\mathcal{S}^T}\mathcal{X}^{\mathcal{T}} = \mathcal{U}_1\mathcal{V}_1\mathcal{P}^T,\ \mathcal{X}_o^{\mathcal{S}^T}\mathcal{X}^{\mathcal{T}} = -\mathcal{U}_2\mathcal{V}_2\mathcal{P}^T \tag{4.4}$$

With the above defined matrices, $\mathbf{K}^{\mathcal{ST}}$ can be computed as

$$\mathbf{K}^{\mathcal{ST}} = \begin{bmatrix} \tilde{\mathcal{X}}^{\mathcal{S}}\mathcal{U}_1 & \tilde{\mathcal{X}}_o^{\mathcal{S}}\mathcal{U}_2 \end{bmatrix} \mathcal{G} \begin{bmatrix} \mathcal{U}_1^T \mathcal{X}^{\mathcal{S}^T} \\ \mathcal{U}_2^T \mathcal{X}_o^{\mathcal{S}^T} \end{bmatrix} \tag{4.5}$$

where $\mathcal{G} = \begin{bmatrix} \text{diag}[1 + \frac{\sin(2\theta_i)}{2\theta_i}] & \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] \\ \text{diag}[\frac{(\cos(2\theta_i)-1)}{2\theta_i}] & \text{diag}[1 - \frac{\sin(2\theta_i)}{2\theta_i}] \end{bmatrix}$ and $[\theta_i]_{i=1}^d$ represents the principal angles between source and target camera. Once we compute all pairwise geodesic flow kernels between a target camera and source cameras using (4.5), our next objective is to find the distance across all those pairs. A source camera which is closest to the new camera is more likely to adapt better than others. We follow [192] to compute distance between a target camera and a source camera pair. Specifically, given a kernel matrix $\mathbf{K}^{\mathcal{ST}}$, the distance between data points of a source and target camera is defined as

$$\mathbf{D}^{\mathcal{ST}}(\mathbf{x}_i^{\mathcal{S}}, \mathbf{x}_j^{\mathcal{T}}) = \mathbf{x}_i^{\mathcal{S}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_i^{\mathcal{S}} + \mathbf{x}_j^{\mathcal{T}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_j^{\mathcal{T}} - 2\mathbf{x}_i^{\mathcal{S}^T} \mathbf{K}^{\mathcal{ST}} \mathbf{x}_j^{\mathcal{T}} \tag{4.6}$$

where $\mathbf{D}^{\mathcal{ST}}$ represents the kernel distance matrix defined over a source and target camera. We compute the average of a distance matrix $\mathbf{D}^{\mathcal{ST}}$ and consider it as the distance between two cameras. Finally, we chose the one that has the lowest distance a best source camera to pair with the newly introduced camera. Algorithm 2 summarizes the procedure to discover best source camera for a newly introduced target camera.

Remark 1. Note that we do not use any labeled data from the newly introduced target camera to either compute the geodesic flow kernels in (4.5) or the kernel distance matrices in (4.6). Hence, our approach can be applied to on-board new cameras in a large-scale camera network with minimal additional effort.

Algorithm 2 Discovering the Best Source Camera

Input: Set of $\mathcal{N}$ source cameras $\{\mathcal{X}^s\}_{s=1}^{\mathcal{N}}$,;

A newly introduced target camera $\mathcal{X}^{\mathcal{T}}$;

Output: Best source camera $\mathcal{X}^{\mathcal{S}\star}$.

for $s = 1, \ldots, \mathcal{N}$ **do**

1. Compute kernel matrix $\mathbf{K}^{\mathcal{ST}}$ using (4.5);
2. Compute distance matrix $\mathbf{D}^{\mathcal{ST}}$ using (4.6);
3. Compute average distance using $\texttt{mean}(\mathbf{D}^{\mathcal{ST}})$;

end for

4. Rank cameras based on average distance and chose the one with lowest distance as the best source camera $\mathcal{X}^{\mathcal{S}\star}$;

Remark 2. We assume that the newly introduced camera will be close to at least one of the installed ones since we consider them to be operating in the same time window with same set of people appear in all camera views, as in most prior works except the work in [284]. However, our adaptation approach is not limited to this constrained setting as we compute the view similarity in a completely unsupervised manner and hence can be easily applied in real-world settings where different sets of people appear in different camera views. To the best of our knowledge, this is first work which can be employed in fully open world re-identification systems considering both dynamic network and different identity of persons across cameras (see illustrative experiments in Sec. 4.4.10).

4.3.3 Transitive Inference for Re-identification

Objective. In the previous section we have presented an unsupervised approach that can effectively find a best source camera to pair with the target camera. Once the best source camera is identified, another question that remains is: *can we exploit the best source camera information to improve the re-identification accuracy across other camera pairs?* More specifically, our objective is to exploit $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ and pair-wise optimal metrics learned in Sec. 4.3.1 to improve the overall matching accuracy of the target camera in a network.

Approach Details. Let $\{\mathbf{M}^{ij}\}_{i,j=1,i<j}^{\mathcal{N}}$ be the optimal pair-wise metrics learned in a network of $\mathcal{N}$ cameras following Section 4.3.1 and $\mathcal{S}^\star$ be the best source camera for a newly introduced target camera $\mathcal{T}$ following Sec. 4.3.2.

Motivated by the effectiveness of Schur product for improving the matrix consistency and reliability in multi-criteria decision making [117], we develop a simple yet effective transitive algorithm for exploiting information from the best source camera. Our problem naturally fits to such decision making systems since our goal is to establish a path between two cameras via the best source camera. Given the best source camera $\mathcal{S}^\star$, we compute the kernel matrix between remaining source and target camera as follows,

$$\tilde{\mathbf{K}}^{\mathcal{S}\mathcal{T}} = \mathbf{M}^{\mathcal{S}\mathcal{S}^\star} \odot \mathbf{K}^{\mathcal{S}^\star\mathcal{T}},\ \forall[\mathcal{S}]_{i=1}^{\mathcal{N}},\quad \mathcal{S} \neq \mathcal{S}^\star \tag{4.7}$$

where $\tilde{\mathbf{K}}^{\mathcal{S}\mathcal{T}}$ represents the updated matrix between source $\mathcal{S}$ and target camera $\mathcal{T}$ by exploiting information from best source camera $\mathcal{S}^\star$. The operator $\odot$ denotes Schur product of two matrices. Eq. 4.7 establishes an indirect path between camera pair ($\mathcal{S}$,$\mathcal{T}$) by marginalization over the domain of possible appearances in best source camera $\mathcal{S}^\star$. In other words, camera $\mathcal{S}^\star$ plays a role of connector between target camera $\mathcal{T}$ and all other source cameras.

Summarizing, to incorporate new camera(s) in an existing network, we use the kernel matrix $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ in (4.5) to obtain the matching accuracy across the new camera and best source camera, whereas we use the updated kernel matrices, computed using (4.7) to find the matching accuracy across the target camera and remaining source cameras.

Remark 3. While there are more sophisticated strategies to incorporate the side information, the reason to adopt a simple weighting approach as in problem (4.7) is that by doing so we can scale the re-identification models easily to a large scale network involving hundreds to thousands of cameras in real-time applications. Furthermore, the proposed transitive algorithm performs satisfactorily in several camera networks as illustrated in Sec. 4.4.

4.3.4 Learning Kernels with Prototype Selection

Objective. For many applications with limited computation and communication resources, there is an imperative need of methods that could extract an informative subset from the source camera data for computing the kernels instead of all existing data. Thus, our main objective in this section is to develop a target-aware sparse prototype selection strategy for finding a subset of source camera data that share similar characteristics as the target camera and then use those for discovering the best source camera in Sec. 4.3.2.

Approach Details. Motivated by sparse subset selection [58, 41], we develop an efficient optimization framework to extract a sparse set of images from each source camera that balances two main objectives: (a) they are informative about the given source camera, and (b) they are also informative about the target camera. Given the above stated goals, we

formulate the following objective function,

$$\min_{\mathcal{Z}^s,\mathcal{Z}^\mathcal{T}} \ \frac{1}{2}\big(\|\mathcal{X}^s - \mathcal{X}^s\mathcal{Z}^s\|_F^2 + \alpha\|\mathcal{X}^\mathcal{T} - \mathcal{X}^s\mathcal{Z}^\mathcal{T}\|_F^2\big) + \lambda\big(\|\mathcal{Z}^s\|_{2,1} + \|\mathcal{Z}^\mathcal{T}\|_{2,1}\big) \tag{4.8}$$

where α balances the penalty between errors in the reconstruction of source camera data $\mathcal{X}^s$ and errors in the reconstruction of target camera data $\mathcal{X}^\mathcal{T}$. $||\mathcal{Z}^s||_{2,1} = \sum_{i=1}^{m} ||\mathcal{Z}_i^s||_2$ and $||\mathcal{Z}_i^s||_2$ is the ℓ_2-norm of the i-th row of $\mathcal{Z}^s$. $\lambda > 0$ is a sparsity regularization parameter.

The objective function is intuitive: minimization of (4.8) favors selecting a sparse set of prototypes that simultaneously reconstructs the source camera data $\mathcal{X}^s$ via $\mathcal{Z}^s$, as well as the target camera data $\mathcal{X}^\mathcal{T}$ via $\mathcal{Z}^\mathcal{T}$, with high accuracy. Specifically, rows in $\mathcal{Z}^s$ provide information on relative importance of each image in describing the source camera $\mathcal{X}^s$, while rows in $\mathcal{X}^\mathcal{T}$ give information on relative importance of each image in $\mathcal{X}^s$ in describing target camera $\mathcal{X}^\mathcal{T}$. Given the two sparse coefficient matrices, our next goal is to select a unified set of images from source camera that share similar characteristics with target camera. To achieve the above goal, we propose to minimize the following objective function:

$$\begin{aligned}\min_{\mathcal{Z}^s,\mathcal{Z}^\mathcal{T}} \ & \frac{1}{2}\big(\|\mathcal{X}^s - \mathcal{X}^s\mathcal{Z}^s\|_F^2 + \alpha\|\mathcal{X}^\mathcal{T} - \mathcal{X}^s\mathcal{Z}^\mathcal{T}\|_F^2\big) \\ & +\lambda\big(\|\mathcal{Z}^s\|_{2,1} + \|\mathcal{Z}^\mathcal{T}\|_{2,1}\big) + \beta||\mathcal{Z}_c||_{2,1} \ s.t. \ \mathcal{Z}_c = [\mathcal{Z}^s|\mathcal{Z}^\mathcal{T}]\end{aligned} \tag{4.9}$$

where $\ell_{2,1}$-norm on the consensus matrix $\mathcal{Z}_c$ enables $\mathcal{Z}^s$ and $\mathcal{Z}^\mathcal{T}$ to have the similar sparse patterns and share the common components. In each round of the optimization, the updated sparse coefficient matrices in the former rounds can be used to regularize the current optimization criterion. Thus, it can uncover the shared knowledge of $\mathcal{Z}^s$ and $\mathcal{Z}^\mathcal{T}$ by suppressing irrelevant images, which results in an optimal $\mathcal{Z}_c$ for selecting representative source images to pair with target camera.

Optimization. Since problem (4.9) is non-smooth involving multiple $\ell_{2,1}$-norms, it is difficult to optimize directly. Motivated by the effectiveness of Half-quadratic optimization techniques [91], we devise an iterative algorithm to efficiently solve (4.9) by minimizing its augmented function alternatively. Specifically, if we define $\phi(x) = \sqrt{x^2 + \epsilon}$ with ϵ being a constant, we can transform $\|\mathcal{Z}^s\|_{2,1}$ to $\sum_{i=1}^{n} \sqrt{||\mathcal{Z}_i^s||_2^2 + \epsilon}$, according to the analysis of $\ell_{2,1}$-norm in [91, 154]. With this transformation, we can optimize (4.9) efficiently in an alternative way as shown in Algorithm 3.

Once the problem (4.9) is solved, we first sort the source camera images by decreasing importance according to the ℓ_2 norms of the rows of $\mathcal{Z}_c$ and then select the top-ranked images that fit in the budget constraint. To summarize, we learn the pair-wise kernels across all the unlabeled target camera data and selected prototypes from the source camera to discover the best camera as in Sec. 4.3.2. Second, we adopt the same transitive inference algorithm mentioned in Sec. 4.3.3 to exploit the information from best source camera to improve re-identification accuracy across other source-target camera pairs.

4.3.5 Extension to Multiple Newly Introduced Cameras

Our approach is not limited to a single camera and can be easily extended to even more realistic scenarios where multiple cameras are introduced to an existing network at the same time. Given multiple newly introduced cameras, one can follow two different strategies to adapt re-id models in dynamic camera networks. Specifically, one can easily find a common best source camera based on lowest average distance to pair with all the new cameras or multiple best source cameras, one for each target camera, in an unsupervised

Algorithm 3 Algorithm for Solving Problem (4.9)

Input: Feature matrices $\mathcal{X}^s$ and $\mathcal{X}^{\mathcal{T}}$

Parameters α, λ, β, set $t = 0$

Initialize $\mathcal{Z}^s$ and $\mathcal{Z}^{\mathcal{T}}$ randomly, set $\mathcal{Z}_c = [\mathcal{Z}^s | \mathcal{Z}^{\mathcal{T}}]$

Output: Optimal sparse coefficient matrix $\mathcal{Z}_c$.

while *not converged* **do**

1. Compute P^t, Q^t and R^t as:

$$P_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_i^s||_2^2 + \epsilon}}, \quad Q_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_i^{\mathcal{T}}||_2^2 + \epsilon}}, \; R_{ii} = \frac{1}{2\sqrt{||\mathcal{Z}_{ci}||_2^2 + \epsilon}}$$

2. Compute $\mathcal{Z}^{st+1}$ and $\mathcal{Z}^{\mathcal{T}^{t+1}}$ as:

$$\mathcal{Z}^s = (\mathcal{X}^{sT}\mathcal{X}^s + 2\lambda P + 2\beta R)^{-1}\mathcal{X}^{sT}\mathcal{X}^s$$

$$\mathcal{Z}^{\mathcal{T}} = (\alpha\mathcal{X}^{sT}\mathcal{X}^s + 2\lambda Q + 2\beta R)^{-1}\alpha\mathcal{X}^{sT}\mathcal{X}^{\mathcal{T}}$$

3. Compute $\mathcal{Z}_c^{t+1}$ as: $\mathcal{Z}_c^{t+1} = [\mathcal{Z}^{st+1} \mid \mathcal{Z}^{\mathcal{T}^{t+1}}]$;

4. $t = t + 1$;

end while

way similar to the above approach. Experiments on a large-scale network of 16 cameras show that our approach works better with multiple source cameras, one for each target camera compared to the case where a common best source camera is used for all target cameras (see illustrative experiments in Sec. 4.4.4).

4.3.6 Extension to Semi-supervised Adaptation

Although our framework is designed for unsupervised adaptation of re-id models, it can be easily extended if labeled data from the newly introduced camera become available. Specifically, the label information from target camera can be encoded while computing sub-

spaces. That is, instead of using PCA for estimating the subspaces, we can use Partial Least Squares (PLS) to compute the discriminative subspaces on the target data by exploiting the labeled information. PLS has shown to be effective in finding discriminative subspaces by projecting labeled data into a common subspace [74, 205]. This essentially leads to semi-supervised adaptation in a dynamic camera network (see experiments in Sec 4.4.5).

4.4 Experiments

In this section, we evaluate the performance of our approach by performing several illustrative experiments on multiple benchmark datasets.

4.4.1 Datasets and Settings

Datasets. We conduct experiments on five different publicly available benchmark datasets to verify the effectiveness of our framework, namely WARD [166], RAiD [45], SAIVT-SoftBio [19], Shinpuhkan2014 [109], and and Market-1501 [278]. Although there are number of other datasets (*e.g.* ViPeR [81], CAVIAR4REID [36], PRID450S [96], and CUHK [133] etc.) for evaluating the performance in re-id, these datasets do not fit our purposes since they have only two cameras or specifically designed for video-based person re-identification [240].

- **WARD [166]** has 4786 images of 70 different people captured in a real surveillance scenario from 3 non-overlapping cameras. This dataset has a huge illumination variation apart from resolution and pose changes.
- **RAiD [45]** was collected with a view to have large illumination variation that is not present in most of the publicly available benchmark datasets. In the original dataset

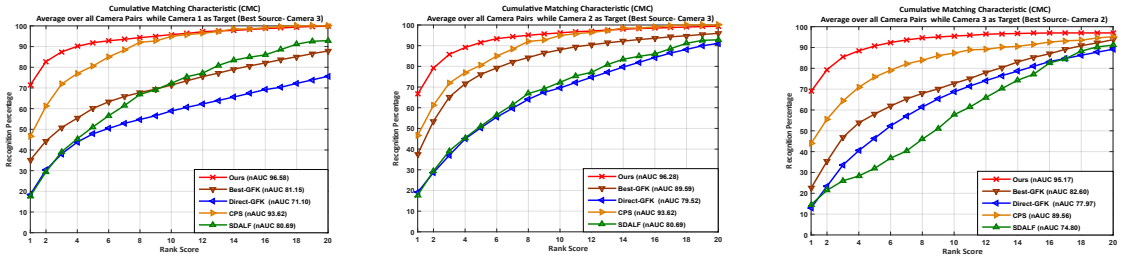
43 subjects were asked to walk through 4 cameras of which two are outdoor and two are indoor to make sure there is enough variation of appearance between cameras.

- **SAIVT-SoftBio [19]** includes annotated sequences (704 × 576 pixels, 25 frames per second) of 150 people, each of which is captured by a subset of 8 different cameras, providing various viewing angles and varying illumination conditions.

- **Shinpuhkan2014 [109]** dataset consists of more than 22,000 images of 24 people which are captured by 16 cameras installed in a shopping mall. All images are manually cropped and resized to 48 ×128 pixels, grouped into tracklets with annotation. The number of tracklets of each person is 86. To the best of our knowledge, this is the largest publicly available dataset for re-id with 16 cameras.

- **Market-1501 [278]** is one of the biggest dataset containing 32,668 images of 1501 persons that are collected by 6 cameras in front of a supermarket in Tsinghua University. Each annotated identity is present in at least two cameras, so that cross-camera search can be performed. Apart from large variations in pose and illuminations, the size of the dataset itself introduces a new level of computational challenge.

Feature Extraction and Matching. The feature extraction stage consists of extracting Local Maximal Occurrence (LOMO) feature [143] for person representation. The descriptor has 26,960 dimensions. We follow [116, 179] and apply principle component analysis to reduce the dimensionality to 100 in all our experiments. Without low-dimensional feature, it is computationally infeasible to inverse covariance matrices of both similar and dissimilar



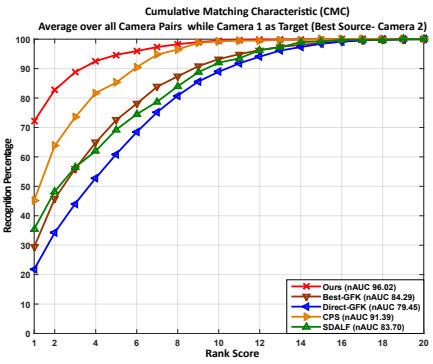
(a) Camera 1 as Target (b) Camera 2 as Target (c) Camera 3 as Target

Figure 4.2: CMC curves for WARD dataset with 3 cameras. Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a dynamic network. Please see the text in Sec. 4.4.2 for the analysis of the results.

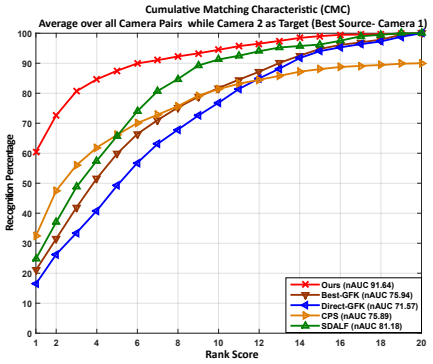
pairs as discussed in [116, 179]. To compute distance between cameras, as well as, matching score, we use kernel distance [192] (Eq. 4.6) for a given projection metric.

Performance Measures. We show results using Cumulative Matching Characteristic (CMC) curves and normalized Area Under Curve (nAUC) values [106, 45, 165, 275, 113]. CMC curve is a plot of recognition performance versus re-id ranking score and represents the expectation of finding correct match in the top k matches. nAUC gives an overall score of how well a method performs irrespective of the dataset size.

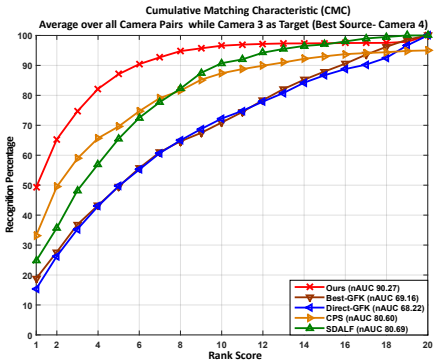
Experimental Settings. We maintain following conventions during all our experiments: All the images for each dataset are normalized to 128×64 for being consistent with the evaluations carried out by state-of-the-art methods [11, 45, 36]. Following the literature [45, 116, 143], the train and test set are kept disjoint by picking half of the available data for training set and rest of the half for testing. We repeated each task 10 times by randomly picking 5 images from each identity both for train and test time. The subspace dimension for all the possible combinations are kept 50.



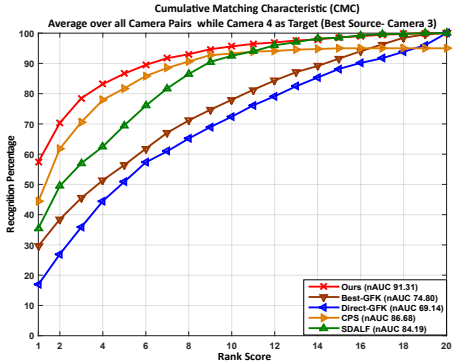
(a) Camera 1 as Target



(b) Camera 2 as Target



(c) Camera 3 as Target



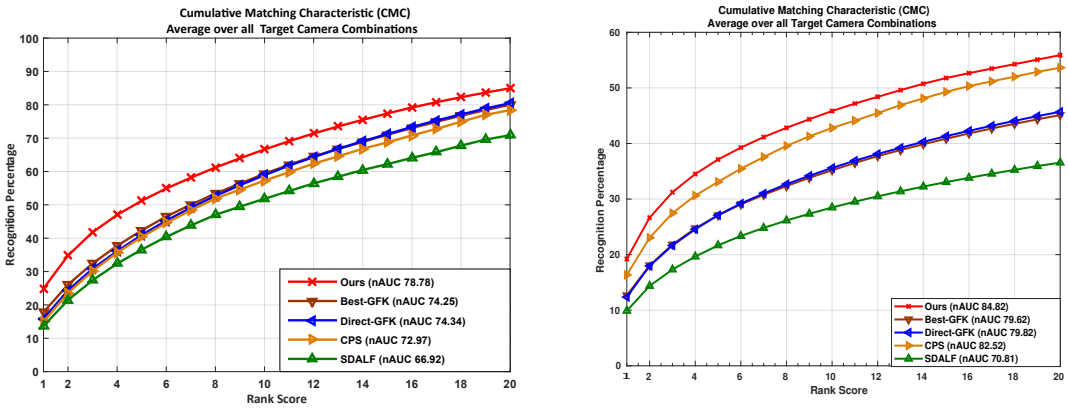
(d) Camera 4 as Target

Figure 4.3: CMC curves for RAiD dataset with 4 cameras. Plots (a, b, c, d) show the performance of different methods while introducing camera 1, 2, 3 and 4 respectively to a dynamic network. Our method significantly outperforms all the compared baselines.

4.4.2 Re-identification by Introducing a New Camera

Goal. The goal of this experiment is to analyze the performance of our unsupervised framework while introducing a single camera to an existing network where optimal distance metrics are learned using an intensive training phase.

Compared Methods. We compare our approach with several unsupervised alternatives which fall into two categories: (i) hand-crafted feature-based methods including `CPS` [36] and `SDALF` [11], (ii) two domain adaptation based methods (`Best-GFK` and `Direct-GFK`) based



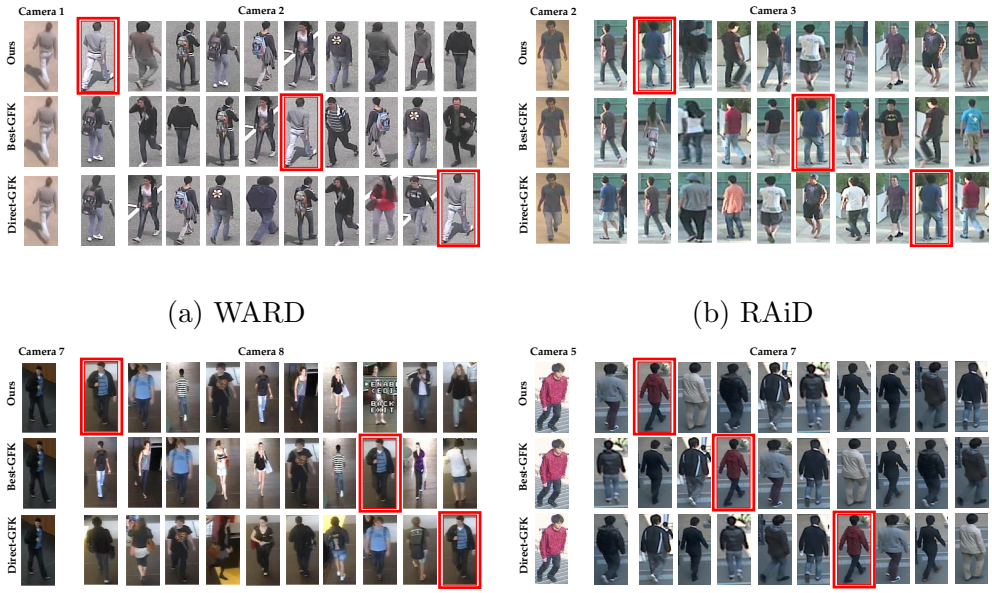
(a) SAIVT-SoftBio

(b) Market-1501

Figure 4.4: CMC curves averaged over all target camera combinations, introduced one at a time. (a) Results on SAVIT-SoftBio dataset, and (b) Results on Market-1501 dataset.

on geodesic flow kernel [78]. For `Best-GFK` baseline, we compute the re-id performance of a camera pair by applying the kernel matrix, $\mathbf{K}^{\mathcal{S}^\star\mathcal{T}}$ computed between best source and target camera [78], whereas in `Direct-GFK` baseline, we use the kernel matrix computed directly across source and target camera using (4.5). The purpose of comparing with `Best-GFK` is to show that the kernel matrix computed across the best source and target camera does not produce optimal re-id performance in computing matching performance across other source cameras and the target camera. On the other hand, the purpose of comparing with `Direct-GFK` baseline is to explicitly show the effectiveness of our transitive algorithm in improving re-id performance in a dynamic camera network.

Implementation Details. We use publicly available codes for `CPS` and `SDALF` and tested on our experimented datasets. We use the same features as the proposed one and kept the parameters same as mentioned in the published works. We also implement both `Best-GFK` and `Direct-GFK` baselines under the same experimental settings as mentioned earlier to



(a) WARD (b) RAiD

(c) SAIVT-SoftBio (e) Shinpuhkan2014

Figure 4.5: Effectiveness of transitive algorithm in re-identification on different datasets. Top row: Our matching result using the transitive algorithm. Middle row: matching the same person using `Best-GFK`. Bottom row: matching the same person using `Direct-GFK`. Visual comparison of top 10 matches shows that `Ours` perform best in matching persons across camera pairs by exploiting information from the best source camera.

have a fair comparison with our proposed method. For all the datasets, we considered one camera as newly introduced target camera and all the other as source cameras. We considered all the possible combinations for conducting experiments. We first pick which source camera matches best with the target one, and then, use the proposed transitive algorithm to compute the re-id performance across remaining camera pairs.

Results. Fig. 4.2 and Fig. 4.3 show the results for all possible combinations on the 3 camera WARD dataset and 4 camera RAiD dataset respectively, whereas Fig. 4.4 shows the average performance over all possible combinations by inserting one camera on SAIVT-SoftBio dataset and Market-1501 dataset respectively. From all three figures, the following obser-

vations can be made: (i) the proposed framework consistently outperforms all compared unsupervised methods on all three datasets by a significant margin, including the Market-1501 dataset with significantly large number of images and person identities. (ii) among the alternatives, `CPS` baseline is the most competitive. However, the gap is still significant due to the two introduced components working in concert: discovering the best source camera and exploiting its information for re-identification. The rank-1 performance improvements over `CPS` are 23.44%, 24.50%, 9.98%, and 2.85% on WARD, RAiD, SAIVT-SoftBio and Market-1501 datasets respectively. (iii) `Best-GFK` works better than `Direct-GFK` in most cases, which suggests that kernel computed across the best source camera and target camera can be applied to find the matching accuracy across other camera pairs in re-identification. (iv) Finally, the performance gap between our method and `Best-GFK` (maximum improvement of 17% in nAUC on RAiD) shows that the proposed transitive algorithm is effective in exploiting information from the best source camera while computing re-id accuracy across camera pairs (see Fig. 4.5 for some illustrative examples on different datasets).

4.4.3 Model Adaptation with Prototype Selection

Goal. The main objective of this experiment is to analyze the performance of our approach by using the selected prototypes from source camera while learning the geodesic flow kernels in resource constrained environments.

Compared Methods. We compare our approach (denoted as `Ours-Proto`) with the same methods (`CPS`, `SDALF`, `Best-GFK` and `Direct-GFK`) as we did in Sec. 4.4.2.

Implementation Details. The regularization parameters λ and β in (4.9) are taken as λ_0/γ where $\gamma = 50$ and λ_0 is analytically computed from the data [58]. The other parameter

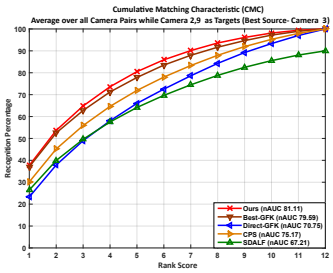
Table 4.1: Model adaptation with prototype selection. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras.

Methods	SDALF	CPS	Direct-GFK	Best-GFK	Ours-Proto	Ours
WARD	16.66	45.70	16.87	32.72	60.72	68.99
RAiD	26.80	35.35	17.63	24.74	53.67	59.84

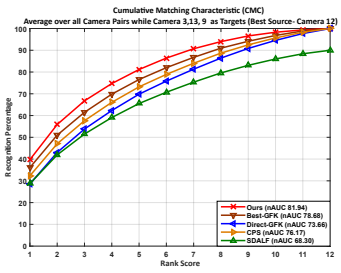
α is empirically set to 0.5 and kept fixed for all results. For each datasets, we show average rank-1 performance over all possible combinations by introducing one camera at a time.

Results. Tab. 4.1 shows the results on both WARD and RAiD datasets. We have the following observations from Tab. 4.1: (i) Our approach with prototypes (`Ours-Proto`) significantly outperforms all compared methods that use all existing source data on both datasets. The rank-1 performance improvements over `CPS` are 15.02% and 18.32% on WARD and RAiD datasets respectively. (ii) As expected, our approach works best with the use of all existing source camera data (ideal case). However, performance using prototypes is still close to the ideal case (a margin of 6%-8%) with only requiring 15%-20% of source camera data while computing the kernels. This can greatly reduce the deployment cost of new cameras in many resource constrained environments.

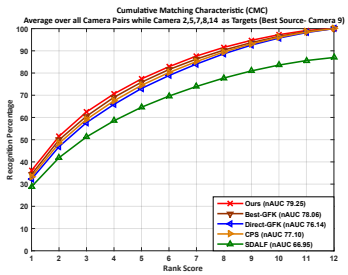
We also investigate the effectiveness of our target-aware sparse prototype selection strategy by comparing with randomly selecting same number of prototypes, and found that the later produces inferior results with rank-1 accuracy of 27.54% and 19.82% on WARD and RAiD datasets respectively. We believe this is because our prototype selection strategy efficiently exploits the information of target camera (see Eq. (4.9)) to select an informative subset of source camera data which share similar characteristics as target camera.



(a) Camera 2, 9 as Targets

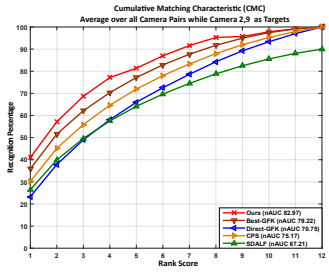


(b) Camera 3, 9, 13 as Targets

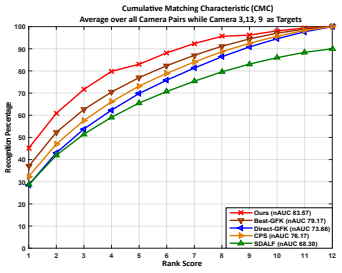


(c) Camera 2, 5, 7, 8, 14 as Targets

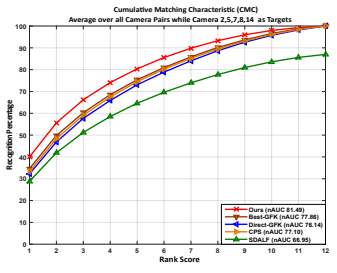
Figure 4.6: CMC curves for Shinpuhkan2014 dataset with 16 cameras. Plots (a, b, c) show the performance of different methods while introducing 2, 3 and 5 cameras respectively at the same time. We use one common best source camera for all the target cameras while computing re-id performance across a network. Please see the text in Sec. 4.4.4 for the analysis of the results. Best viewed in color.



(a) Camera 2, 9 as Targets



(b) Camera 3, 9, 13 as Targets



(c) Camera 2, 5, 7, 8, 14 as Targets

Figure 4.7: CMC curves for Shinpuhkan2014 dataset with 16 cameras. Plots (a, b, c) show the performance of different methods while introducing 2, 3 and 5 cameras respectively at the same time. We use multiple best source cameras, one for each target camera while computing re-id performance across a network. Please see the text in Sec. 4.4.4 for the analysis of the results. Best viewed in color.

4.4.4 Introducing Multiple Cameras

Goal. The aim of this experiment is to validate the effectiveness of our proposed approach while introducing multiple cameras at the same time in a dynamic camera network. As described in Sec. 4.3.5, we investigate our performance in two different scenarios such as (a) one common best source camera for all target cameras and (b) multiple best source cameras, one for each target camera while computing re-id performance across a network.

Compared Methods. We compare our approach with the same methods (`CPS`, `SDALF`, `Best-GFK` and `Direct-GFK`) as we did for single camera in Sec. 4.4.2.

Implementation Details. We conduct this experiment on Shinpuhkan2014 dataset [109] with of 16 cameras. We randomly chose 2, 3 and 5 cameras as the target cameras while remaining cameras as possible source cameras. For scenario (a), we pick the common best source camera based on the average distance and follow the same strategy as in Sec. 4.4.2 while for scenario (b), instead of using the common best source camera, we use multiple best source cameras, one for each target camera in the transitive inference.

Results. Fig. 4.6 and Fig. 4.7 show results of different methods in two different scenarios while randomly introducing 2, 3 and 5 cameras respectively on Shinpuhkan2014 dataset. The following observations can be made from the figs: (i) Similar to the results in Sec. 4.4.2, our approach outperforms all compared methods in all three scenarios. This indicates that the proposed method is very effective and can be applied to large-scale dynamic camera networks where multiple cameras can be introduced at the same time. (ii) The gap between ours and `Best-GFK` is moderate but still we improve by 4% in nAUC values, which corroborates the effectiveness of transitive inference for re-identification in a large-scale camera

network (see Fig. 4.6). (iii) The proposed adaptation approach works better with multiple best source cameras compared to a common best source camera used for transitive inference (about 5% improvement–see Fig. 4.7). This is expected since multiple best source cameras can better exploit information from best source camera to improve the re-identification accuracy. Our approach is quite generic which can handle either multiple best source cameras or a common best source camera for transitive inference in a dynamic camera network.

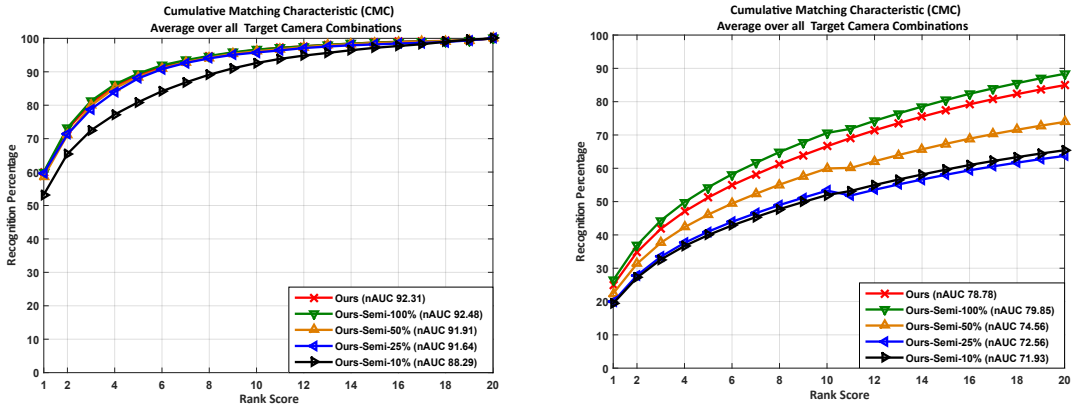
4.4.5 Extension to Semi-supervised Adaptation

Goal. As discussed in Sec. 4.3.6, the proposed method can be easily extended to semi-supervised settings when labeled data from the target camera become available. The objective of this experiment is to analyze the performance of our approach in such settings by incorporating labeled data from the target camera.

Compared Methods. We compare the proposed unsupervised approach with four variants of our method where 10%, 25%, 50% and 100% of the labeled data from target camera are used for estimating kernel matrix respectively.

Implementation Details. We follow same strategy in finding average accuracies over a camera network. However, we use PLS instead of PCA, to compute the discriminative subspaces in target camera by considering 10%, 25%, 50% and 100% labeled data respectively.

Results. We have the following key findings from Fig. 4.8: (i) As expected, the semi-supervised baseline `Ours-Semi-100%`, works best since it uses all the labeled data from target domain to compute the kernel matrix for finding the best source camera. (ii) Our method remains competitive to `Ours-Semi-100%` on both datasets (Rank-1 accuracy: 60.04% vs 59.84% on RAiD and 26.41% vs 24.92% on SAIVT-SoftBio dataset). However, it is impor-



(a) RAiD

(b) SAIVT-SoftBio

Figure 4.8: Semi-supervised adaptation with labeled data. Plots (a,b) show CMC curves averaged over all target camera combinations, introduced one at a time, on RAiD and SAIVT-SoftBio respectively. Please see the text in Sec. 4.4.5 for analysis of the results.

tant to note that collecting labeled samples from the target camera is very difficult in practice. (iii) Interestingly, the performance gap between our unsupervised method and other three semi-supervised baselines (`Ours-Semi-50%`, `Ours-Semi-25%`, and `Ours-Semi-10%`) are moderate on RAiD (see Fig. 4.8-a), but on SAIVT-SoftBio, the gap is significant (see Fig. 4.8-b). We believe this is probably due to the lack of enough labeled data in the target camera to give a reliable estimate of PLS subspaces.

4.4.6 Re-identification with LDML Metric Learning

Goal. The objective of this experiment is to verify the effectiveness of our approach by changing the initial setup presented in Sec. 4.3.1. Specifically, our goal is to show the performance of the proposed method by replacing KISSME [116] with LDML metric learning [83]. Ideally, we would expect similar performance improvement by our method, irrespective of

the metric learning used to learn the distance metrics in an existing network of cameras.

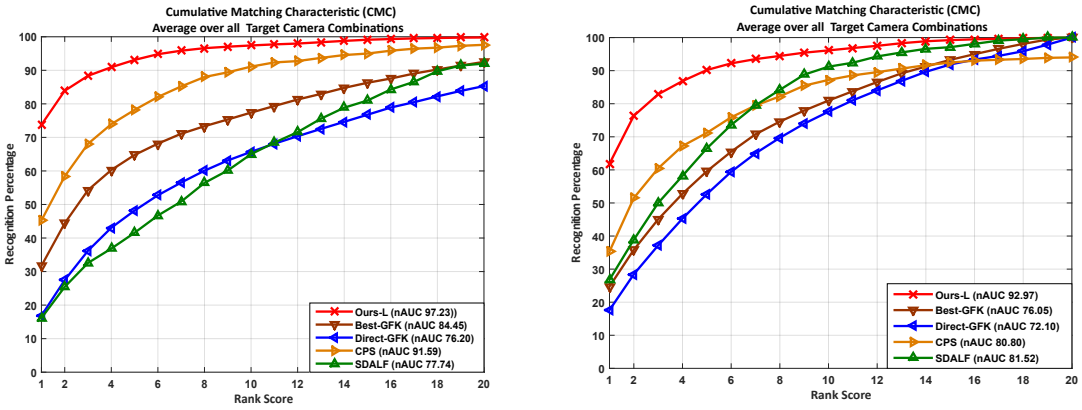
Implementation Details. We use the publicly available code of LDML to test the performances and set the parameters same as recommended in the published work.

Results. Fig. 4.9 shows results of our method on WARD and RAiD respectively. Following are the analysis of the figures: (i) Our approach outperforms all compared methods in both datasets which suggests that the proposed adaptation technique works significantly well irrespective of the metric learning method used in the existing camera network. (ii) The proposed adaptation approach works slightly better with LDML compared to KISSME on the 3 camera WARD dataset (73.77% vs 68.99% in rank-1 accuracy). However, the margin becomes smaller on RAiD (61.87 vs 59.84) which is relatively a complex re-id dataset with 2 outdoor and 2 indoor cameras. (iii) Although performance of LDML is slightly better than KISSME, it is important to note that KISSME is about 40% faster than that of LDML in learning the metrics in WARD dataset. KISSME is computationally efficient and hence more suitable for learning pairwise distance metrics in a large-scale camera network.

4.4.7 Effect of Feature Representation

Goal. The goal of this experiment is to verify the effectiveness of our approach by changing the feature representation. Specifically, our goal is to show the performance of the proposed method by replacing LOMO feature with Weighted Histograms of Overlapping Stripes (WHOS) feature representation [143].

Implementation Details. We use the publicly available code of WHOS to test the per-



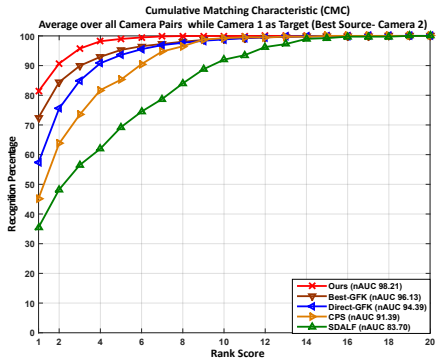
(a) WARD

(b) RAiD

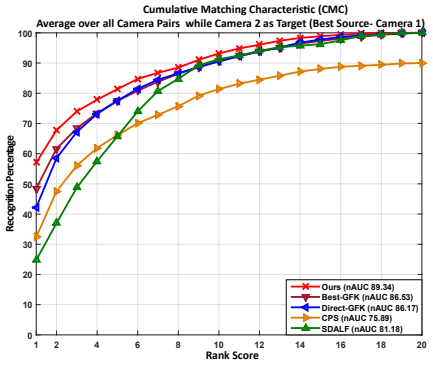
Figure 4.9: Re-id performance with LDML as initial setup. Plots (a,b) show CMC curves averaged over all target camera combinations, introduced one at a time, on WARD and RAiD respectively. Please see the text in Sec. 4.4.6 for analysis of the results.

formances and set the parameters same as recommended in the published work. Except the change in feature, we followed the same settings while comparing with other methods.

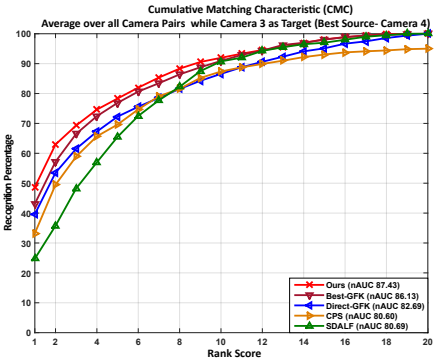
Results. Fig. 4.10 shows results for all possible 4 combinations (three source and one target) on RAiD dataset. From Fig. 4.10, the following observations can be made: (i) our approach outperforms all compared methods which suggests that the proposed adaptation technique works significantly well irrespective of the feature used to represent persons. (ii) Among the alternatives, `Best-GFK` is the most competitive. However, the gap is still significant compared to `Ours` with an average margin of about 10%. (iii) The improvement over `Best-GFK` shows that the proposed transitive algorithm is very effective in exploiting information from the best source camera irrespective of the feature representation.



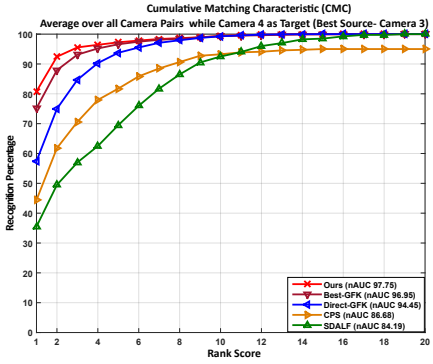
(a) Camera 1 as Target



(b) Camera 2 as Target



(c) Camera 3 as Target



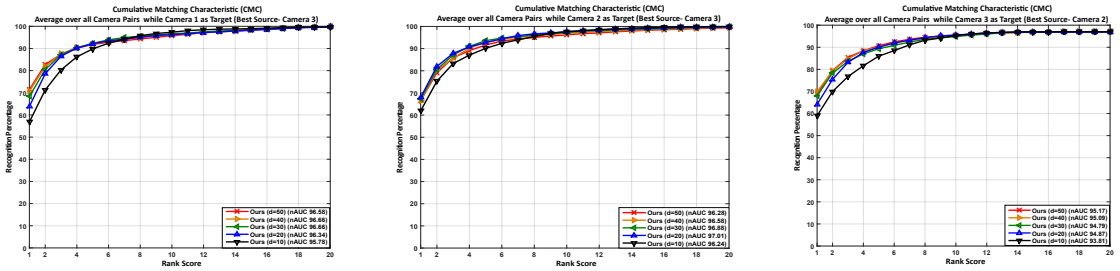
(d) Camera 4 as Target

Figure 4.10: Re-identification performance on RAiD dataset with WHOS feature representation. Plots (a, b, c, d) show CMC curves averaged over all camera pairs while introducing camera 1, 2, 3 and 4 respectively to a dynamic network.

4.4.8 Effect of Subspace Dimension

Goal. The main objective of this experiment is to analyze the performance of our method by changing the dimension of subspace used to compute the geodesic flow kernels across target and source cameras. In ideal case, we would like to see a minor change in performance with changing the dimension of subspace.

Implementation Details. We tested our approach with 5 cases of d, set to 10, 20, 30, 40 and 50. Except the change in dimension, we kept everything fixed while computing re-id



(a) Camera 1 as Target (b) Camera 2 as Target (c) Camera 3 as Target

Figure 4.11: Re-identification performance on WARD dataset with change in subspace dimension. Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a dynamic network.

performance in a dynamic camera network.

Results. We have the following observations from Fig. 4.11: (i) Dimensionality of the subspace has a little effect on the re-id performance of our method suggesting that our method is robust to the change in dimensionality of the subspace used to compute the geodesic kernels across target and source cameras. (ii) Performance of our method is comparatively less when the dimension is set to 10. We believe this is because the principal angles computed at a dimension of 10 for this dataset are very small in magnitude which suggests that variances captured in the subspace corresponding to the source cameras would not be able to transfer to the target subspace. (iii) Although we empirically set the dimension to 50 in all our experiments, we believe finding the optimal dimension specific to a dataset can provide best re-id performance in a network of cameras.

4.4.9 Comparison with Supervised Re-identification

Goal. The objective of this experiment is to compare the performance of our approach with supervised alternatives in a dynamic camera network.

Compared Methods. We compare with several supervised alternatives which fall into two categories: (i) feature transformation based methods including `FT` [164], `ICT` [6], `WACN` [166], that learn the way features get transformed between two cameras and then use it for matching, (ii) metric learning based methods including `KISSME` [116], `LDML` [83], `XQDA` [143] and `MLAPG` [141]. As mentioned in Sec. 4.4.6, our model can operate with any initial network setup and hence we show our results with both KISSME and LDML, denoted as `Ours-K` and `Ours-L`, respectively. Note that we could not compare with recent deep learning based methods as they are mostly specific to a static setting and also their pairwise camera results are not available on the experimented datasets. We did not re-implement such methods in our dynamic setting as it is very difficult to exactly emulate all the implementation details.

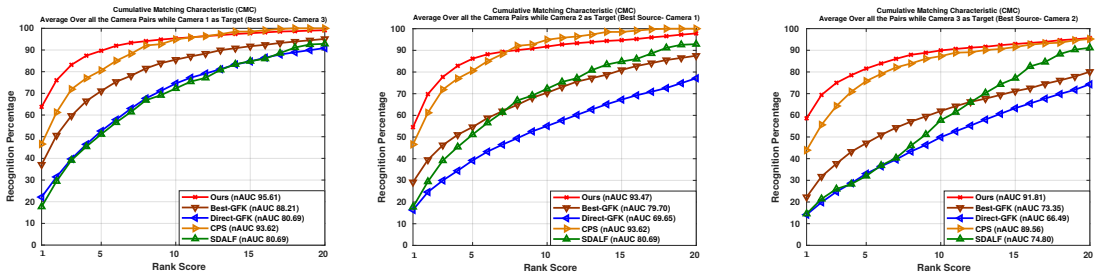
Implementation Details. To report existing feature transformation based methods results, we use prior published performances from [45]. For metric learning based methods, we use publicly available codes to test on our experimented datasets. Given a newly introduced camera, we use the metric learning based methods to relearn the pair-wise distance metrics using the same train/test split, as mentioned earlier in Sec. 4.4.1. We show the average performance over all possible combinations by introducing one camera at a time.

Results. Table 4.2 shows the rank-1 accuracy averaged over all possible target cameras introduced one at a time in a dynamic network. We have the following key findings from Table 4.2: (i) Both variants of our unsupervised approach (`Ours-K` and `Ours-L`) outperforms all the feature transformation based approaches on both datasets by a big margin. (ii) On WARD dataset with 3 cameras, our approach is very competitive on both settings: `Ours-K` outperforms `KISSME` and `LDML` whereas `Ours-L` overcomes `MLAPG`. This result suggests that

Table 4.2: Comparison with supervised methods. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Methods	**WARD**	**RAiD**	**Reference**
FT	49.33	39.81	TPAMI2015 [164]
ICT	42.51	25.31	ECCV2012 [6]
WACN	37.53	17.71	CVPRW2012 [166]
KISSME	66.95	55.68	CVPR2012 [116]
LDML	58.66	61.52	ICCV2009 [83]
XQDA	**77.20**	**77.81**	TPAMI2015 [143]
MLAPG	72.26	77.68	ICCV2015 [141]
Ours-K	*68.99*	*59.84*	Proposed
Ours-L	*73.77*	*61.87*	Proposed

our approach is more effective in matching persons across a newly introduced camera and existing source cameras by exploiting information from best source camera via a transitive inference. (iii) On the RAiD dataset with 4 cameras, the performance gap between our method and metric-learning based methods begins to appear. This is expected as with a large network involving a higher number of camera pairs, an unsupervised approach can not compete with a supervised one, especially, when the latter one is using an intensive training phase. However, we would like to point out once more that in practice collecting labeled samples from a newly inserted camera is very difficult and unrealistic in actual scenarios.



(a) Camera 1 as Target (b) Camera 2 as Target (c) Camera 3 as Target

Figure 4.12: Re-identification performance on WARD dataset with different sets of people in the target camera (0% Overlap). Plots (a, b, c) show the performance of different methods while introducing camera 1, 2 and 3 respectively to a network.

4.4.10 Re-identification with Different Sets of People

Goal. The goal of this experiment is to analyze the performance of our approach with different identities of person appearing in the target camera as in a real world setting.

Implementation Details. We consider first 15 persons in source camera and next 20 persons in target camera (0% Overlap) for training on WARD dataset while we use first 13 persons in source camera and next 10 persons in target camera for training on RAiD dataset. We also consider a scenario where partial overlap of persons exists across source and target cameras, i.e., all the persons appearing in the source camera are present in the target camera but there exists some persons that only appear in target camera and not in source cameras. We consider first 13 persons in source camera and all 23 persons in target camera for training in this partial overlap setting (50% Overlap). Note that the train and test set are still kept disjoint as in standard person re-identification setting.

Results. Fig. 4.12 shows the re-id performance of different methods on WARD dataset with completely disjoint sets of people in the target camera. Following are the key observations

Table 4.3: Performance comparison with different percentage of overlap in person identities across source and target camera. Numbers show rank-1 recognition scores in % averaged over all possible combinations of target cameras, introduced one at a time.

Datasets	**0% Overlap**	**50% Overlap**	**100% Overlap**
RAiD	50.83	56.81	59.84

from Fig. 4.12: (i) The proposed framework for re-identification consistently outperforms all compared methods by a significant margin even though completely new persons appear in the target camera. (ii) Similar to previous results with 100% overlap of persons across source and target cameras (see Fig. 4.2), `CPS` is still the most competitive. However, our approach outperforms `CPS` by a margin about 20% in rank-1 accuracy on WARD dataset. (iii) Finally, the large performance gap between our method, `Direct-GFK` and `Best-GFK` (improvement of more than 30% in rank-1 accuracy) shows that the proposed transitive algorithm is also effective in real-world scenarios where completely new person identities appear in the newly introduced camera.

Tab. 4.3 shows the performance of our approach with different percentage of overlap in person identities across source and target camera on RAiD dataset. As expected, the performance increases with increase in the percentage of overlap and achieves the maximum rank-1 accuracy of 59.84% when the same set of people appear in all camera views. This is because kernel matrices are the best measure of similarity when there is complete overlap across two data distributions. Our approach outperforms all compared methods at 0% overlap on both WARD and RAiD datasets showing it's effectiveness in real-world systems with both dynamic network and different identity of persons across cameras.

4.5 Conclusion

In this work, we presented an effective framework to adapt re-identification models in a dynamic network, where one or multiple new cameras may be temporarily inserted into an existing system to get additional information. We developed a domain perceptive re-identification method based on geodesic flow kernel to find the best source camera to pair with newly introduced camera(s), without requiring a very expensive training phase. We then introduced a simple yet effective transitive inference algorithm that can exploit information from best source camera to improve the accuracy across other camera pairs. Moreover, we develop a source-target selective adaptation strategy that uses a subset of source data instead of all existing data to compute the kernels in resource constrained environments. Extensive experiments on several benchmark datasets well demonstrate the efficacy of our method over state-of-the-art methods.

Chapter 5

Conclusions

5.1 Thesis Summary

In this thesis, we focused on one fundamental challenge in computer vision–how to learn efficient models with limited supervision for two specific applications namely video summarization and person re-identification. In the first two works, we focused on developing weakly supervised frameworks for video summarization while on the last work, we developed an effective approach for on-boarding new camera(s) into an existing person re-identification framework with limited supervision. Our proposed frameworks show the way to scale video summarization and re-identification to the sheer size of tomorrows available data or cameras.

We proposed a collaborative approach for summarizing topic-related videos in chapter 2. Our framework exploits visual context from a set of topic-related videos to extract an informative summary of a given video that simultaneously capture both important particularities arising in the given video, as well as, generalities identified from the set of topic-related videos. We show that our proposed framework while exploiting weak

supervision in form of freely available topic-related videos from the web can generated high quality video summaries by performing rigorous experiments on two standard summarization datasets. In chapter 3, we presented an unsupervised approach by exploiting data correlations for summarizing multi-view videos in a camera network. The proposed multi-view embedding helps in capturing correlations without assuming any prior correspondence between the individual ones. 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 alignment between them, e.g., uncalibrated camera networks. 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.

In chapter 4, we presented a novel approach for adapting existing multi-camera person re-identification frameworks with limited supervision through transfer learning. Specifically, we focused on the problem of on-boarding new camera(s) by discovering and transferring knowledge from installed cameras without also adding a very expensive training phase. We also developed a source-target selective adaptation strategy that uses a subset of source data instead of all the existing data to compute the kernels in resource constrained environments. This is crucial to increase the flexibility and decrease the deployment cost of newly introduced cameras in large-scale dynamic camera networks. We demonstrated that the proposed model significantly outperforms the state-of-the-art unsupervised learning based alternatives on five benchmark datasets involving large number of images and cameras whilst being extremely efficient to compute.

5.2 Future Research Directions

5.2.1 Joint Video Segmentation and Summarization

Our proposed approaches for video summarization in chapter 2 and chapter 3 use video temporal 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 [187] 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 [25]. Developing an efficient framework for joint video segmentation and summarization is an interesting practical problem—we leave this as future work, with no existing work, to the best of our knowledge.

5.2.2 Personalized Video Summarization

Most summarization approaches follow the principle of "one summary fits all" where video summaries are automatically generated without considering any user interest. However, the best summary of a long video differs among different people due to its highly subjective nature. Even for the same person, the best summary may change with time or mood. Recently, the problem of personalized video summarization has gained attention in the research community where the goal is to generate customized video summaries specific

to user interests. Many approaches has been developed with the use of attention [142], user interest modeling [85] or by including a human in the loop [88]. However, most of these approaches including our proposed works in chapter 2 and chapter 3 can only handle explicit user interest which are unreliable to specify in many applications since an user may not be interested to provide his/her interests all the time while summarizing long videos. An important question we want to ask here is whether we can implicitly infer the user interests for generating high quality personalized video summaries. With the rapid proliferation of social media, we are now very active in many social platforms such as Facebook, Twitter and many more. Thus, implicitly inferring user interest via social media analysis is an interesting direction of future research in the context of video summarization. In future, we plan to achieve this with three main steps, namely social context identification, user interest discovery and personalized summary generation. Social context identification will focus on different data mining techniques to extract related videos and like minded users from the social media platforms. Once, the related videos along with like minded users have been identified, we can perform clustering to discover latent concepts related to different users and their associated activities in the social media (eg., tagging, re-tweeting in Twitter). Finally, a factor representing correlation with the latent concepts can be integrated along with representativeness and sparsity in any summarization framework to generate personalized video summaries to enable a more efficient and engaging viewing experience.

5.2.3 Online and Distributed Video Summarization

In many applications, video summarization algorithms may be running on sensors which are equipped with limited computational resources. In recent work [122], we introduce

a reinforcement learning agent to automatically fast-forward a single-view video and present a subset of relevant frames to users on the fly. It does not require processing the entire video, but just the portion that is selected by the agent, which makes the process very computationally efficient. Fast-forwarding multi-view videos captured with different sensors in a overlapping or non-overlapping camera network still remains as a novel and largely under-addressed problem. Building upon these results for a single camera, we propose to develop multi-agent reinforcement learning approaches for fast-forwarding through multiple data streams captured using different sensors. Moreover, we expect the complexity of the problem to be much higher for mobile networks where a wider variety of conditions can be encountered. Develop scalable algorithms using Q-learning to solve these problems in an efficient manner can be an interesting future research direction. Moreover, in many applications, all the data may not be available in a central repository. Analysis would need to happen by combining local information at different camera nodes or processing nodes that assimilate information from groups of cameras. Such nodes would be able to communicate locally and summaries would have to be generated via local communication. Developing scalable algorithms using the theory of submodular maximization and MapReduce style computations in a distributed fashion will also be an interesting direction for future research.

5.2.4 Knowledge Transfer across Networks

In chapter 4, we have shown that it is possible to add a new target camera to an existing network of source cameras using transfer learning with no additional supervision for the new camera. However, transfer learning across networks is still a largely under-addressed problem with many challenges. Given multiple existing source networks

and a newly installed target network with limited labeled data, we first need to find the relevance/similarity of each source network, or parts thereof, in terms of amount of knowledge that it can transfer to a target network. Developing efficient statistical measures for finding relevance in a multi-camera network with significant changes in viewing angle, lighting, background clutter, and occlusion can be a very interesting future work. Furthermore, labeled data from source networks are often a subject of legal, technical and contractual constraints between data owners and customers. Thus, existing transfer learning approaches may not be directly applicable in such scenarios where the source data is absent. The question we want to ask here is whether learned source models instead od source data can be used as a proxy for knowledge transfer across networks. Compared to the source data, the well-trained source model(s) are usually freely accessible in many applications and contain equivalent source knowledge as well. In future, we plan to use *distillation* [95] for transferring knowledge across networks where data from the source network(s) are not either readily available or subject of several data regulations. Attention transfer techniques [268] along with distillation can also be adopted to transfer knowledge from a number of existing labeled networks to an unlabeled target network containing targets which never appeared in the source network.

5.2.5 Learning in Mobile Camera Networks

Existing re-identification works including ours in chapter 4 are conventionally formulated as a one-to-one set-matching problem between two or more fixed cameras, for which an effective model can be learned. Despite the success of these works in static platforms, considering mobile cameras (e.g., network of robots) opens up exciting new research prob-

lems in terms of thinking about learning such data association models. It is not possible to learn transformation models between every possible pair of views in two mobile cameras due to the constantly changing nature of the videos being captured. Thus, in order to efficiently learn data association models, we need the data to represent the variety of scenarios that will be encountered by the mobile cameras. Towards this, we plan to develop a semi-supervised pipeline that uses limited manual training data along with newly generated data through a generative adversarial network (GAN) [79]. One initial approach could be to use the unlabeled samples produced by a Multi-view Generative Adversarial Network (Mv-GAN) [32] in conjunction with the labeled training data to learn view-invariant features in a mobile network. Moreover, apart from generating samples, we may need to evolve the learned models over time based on the observed features.

Bibliography

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] Azadeh Alavi, Yan Yang, Mehrtash Harandi, and Conrad Sanderson. Multi-shot person re-identification via relational stein divergence. In *ICIP*, 2013.

[3] Jurandy Almeida, Neucimar J. Leite, and Ricardo da S. Torres. VISON: VIdeo Summarization for ONline applications. *Pattern Recognition Letters*, 33(4):397–409, 2012.

[4] Le An, Xiaojing Chen, Songfan Yang, and Bir Bhanu. Sparse representation matching for person re-identification. *Information Sciences*, 2016.

[5] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM symposium on Discrete algorithms*, 2007.

[6] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In *ECCV*, 2012.

[7] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[8] W. Bailer, E. Dumont, S. Essid, and B. Mérialdo. A collaborative approach to automatic rushes video summarization. In *ICIP*, 2008.

[9] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[10] S. Balakrishnan and S. Chopra. Collaborative ranking. In *ACM international conference on Web search and data mining*, 2012.

[11] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *CVIU*, 2013.

[12] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[13] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

[14] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[15] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015.

[16] Ronny Bergmann, Raymond H Chan, Ralf Hielscher, Johannes Persch, and Gabriele Steidl. Restoration of manifold-valued images by half-quadratic minimization. *arXiv preprint arXiv:1505.07029*, 2015.

[17] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. Person re-identification by discriminatively selecting parts and features. In *ECCV*, 2014.

[18] Amran Bhuiyan, Alessandro Perina, and Vittorio Murino. Exploiting multiple detections to learn robust brightness transfer functions in re-identification systems. In *ICIP*, 2015.

[19] Alina Bialkowski, Simon Denman, Sridha Sridharan, Clinton Fookes, and Patrick Lucey. A database for person re-identification in multi-camera surveillance networks. In *DICTA*, 2012.

[20] Gunhee Kim Bo Xiong and Leonid Sigal. Storyline representation of egocentric videos with an application to story-based search. In *ICCV*, 2015.

[21] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[23] Tibério S Caetano, Julian J McAuley, Li Cheng, Quoc V Le, and Alex J Smola. Learning graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):1048–1058, 2009.

[24] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression for efficient regularized subspace learning. In *ICCV*, 2007.

[25] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *ICCV*, 2015.

[26] Octavia Camps, Mengran Gou, Tom Hebble, Srikrishna Karanam, Oliver Lehmann, Yang Li, Richard Radke, Ziyan Wu, and Fei Xiong. From the lab to the real world: Re-identification in an airport camera network. *TCSVT*, 2016.

[27] J Carbonell and J Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[28] Anirban Chakraborty, Abir Das, and Amit K Roy-Chowdhury. Network consistent data association. *TPAMI*, 2016.

[29] S Chakraborty, O Tickoo, and R Iyer. Towards distributed video summarization. In *MM*, 2015.

[30] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015.

[31] Jiaxin Chen, Zhaoxiang Zhang, and Yunhong Wang. Relevance metric learning for person re-identification by exploiting listwise similarities. *TIP*, 2015.

[32] Mickaël Chen and Ludovic Denoyer. Multi-view generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 175–188. Springer, 2017.

[33] Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S Huang. Learning with l1 graph for image analysis. *Image Processing, IEEE Transactions on*, 19(4):858–866, 2010.

[34] De Cheng, Xiaojun Chang, Li Liu, Alexander G Hauptmann, Yihong Gong, and Nanning Zheng. Discriminative dictionary learning with ranking metric embedded for person re-identification. In *IJCAI*, 2017.

[35] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[36] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.

[37] Chakra Chennubhotla and Allan Jepson. Sparse coding in practice. In *Proc. of the Second Int. Workshop on Statistical and Computational Theories of Vision*, 2001.

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

[39] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.

[40] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *CVPR*, 2017.

[41] Yang Cong, Junsong Yuan, and Jiebo Luo. Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection. *IEEE Transactions on Multimedia*, 14(1):66–75, 2012.

[42] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *ACM symposium on Theory of computing*, 1987.

[43] Timothee Cour, Praveen Srinivasan, and Jianbo Shi. Balanced graph matching. *Advances in Neural Information Processing Systems*, 19:313, 2007.

[44] K Dale, E Shechtman, S Avidan, and H Pfister. Multi-video browsing and summarization. In *CVPRW*, 2012.

[45] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *ECCV*, 2014.

[46] Abir Das, Rameswar Panda, and Amit Roy-Chowdhury. Active image pair selection for continuous person re-identification. In *ICIP*, 2015.

[47] Abir Das, Rameswar Panda, and Amit K Roy-Chowdhury. Continuous adaptation of multi-camera person identification models through sparse non-redundant representative selection. *CVIU*, 2017.

[48] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[49] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[50] Sandra Eliza Fontes de Avila, Ana Paula Brando Lopes, Antonio da Luz Jr., and Arnaldo de Albuquerque Arajo. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[51] Vikas Singh Deepti Pachauri, Risi Kondor. Solving the multi-way matching problem by permutation synchronization. In *NIPS*, 2013.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[53] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011.

[54] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, pages 269–274, 2001.

[55] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[56] F Dornaika and I Kamal Aldine. Decremental sparse modeling representative selection for prototype selection. *PR*, 2015.

[57] Naveed Ejaz, Irfan Mehmood, and Sung Wook Baik. Efficient visual attention based framework for extracting key frames from videos. *SPIC*, 2013.

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

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

[60] Ehsan Elhamifar, Guillermo Sapiro, and Shankar Sastry. Dissimilarity-based sparse subset selection. *TPAMI*, 2016.

[61] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *CVPR*, 2009.

[62] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2765–2781, 2013.

[63] Erhan Baki Ermis, Pierre Clarot, Pierre-Marc Jodoin, and Venkatesh Saligrama. Activity based matching in distributed camera networks. *Image Processing, IEEE Transactions on*, 19(10):2595–2613, 2010.

[64] S. Feng, Z. Lei, D. Yi, and S.Z. Li. Online content-aware video condensation. In *CVPR*, 2012.

[65] Shikun Feng, Zhen Lei, and Stan.Z. Li. Online content-aware video condensation. In *CVPR*, 2012.

[66] Yanwei Fu. Multi-View Metric Learning for Multi-View Video Summarization. *arXiv.org*, 2014.

[67] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi View Video Summmarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2004.

[68] Marco Furini, Filippo Geraci, Manuela Montangero, and Marco Pellegrini. Stimo: Still and moving video storyboard for the web scenario. *Multimedia Tools and Applications*, 46(1):47–69, 2010.

[69] Chuang Gan, Ting Yao, Gerard de Melo, Yi Yang, and Tao Mei. Improving action recognition using web images. In *IJCAI*, 2016.

[70] Shenghua Gao, Ivor Wai-Hung Tsang, Liang-Tien Chia, and Peilin Zhao. Local features are not lonely–laplacian sparse coding for image classification. In *CVPR*, 2010.

[71] Shenghua Gao, I.W.-H. Tsang, and Liang-Tien Chia. Laplacian Sparse Coding, Hypergraph Laplacian Sparse Coding, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):92–104, 2013.

[72] Jorge Garcıa, Niki Martinel, Gian Luca Foresti, Alfredo Gardel, and Christian Micheloni. Person orientation and feature distances boost re-identification. 2014.

[73] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel. Person re-identification ranking optimisation by discriminant context information analysis. In *ICCV*, 2015.

[74] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 1986.

[75] Donald Geman and George Reynolds. Constrained restoration and the recovery of discontinuities. *TPAMI*, 1992.

[76] R Glowinski and P Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. SIAM, 1989.

[77] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014.

[78] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[79] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[80] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[81] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[82] Genliang Guan, Zhiyong Wang, Shaohui Mei, Max Ott, Mingyi He, and David Dagan Feng. A Top-Down Approach for Video Summarization. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(4):56–68, 2014.

[83] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[84] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[85] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[86] Michael Gygli and Helmut Grabner1 Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015.

[87] Michael Haenlein and Andreas M Kaplan. A beginner's guide to partial least squares analysis. *Understanding statistics*, 2004.

[88] Bohyung Han, Jihun Hamm, and Jack Sim. Personalized video summarization with human in the loop. In *WACV*, 2011.

[89] Dongyoon Han and Junmo Kim. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, 2015.

[90] A. Hanjalic and H. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster validity analysis. *IEEE Transactions on Circuit and Systems for Video Technology*, 9:1280–1289, 1999.

[91] Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. l21 regularized correntropy for robust feature selection. In *CVPR*, 2012.

[92] Ran He, Wei-Shi Zheng, Tieniu Tan, and Zhenan Sun. Half-quadratic-based iterative minimization for robust sparse representation. *TPAMI*, 2014.

[93] Luis Herranz and Jos M Martinez. A framework for scalable summarization of video. *TCSVT*, 2010.

[94] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *NIPS*, 2002.

[95] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[96] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, 2011.

[97] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

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

[99] Tiejun Huang. Surveillance video: the biggest big data. *Computing Now*, 7(2), 2014.

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

[101] Luo Jie, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, 2011.

[102] Feng Jing, Changhu Wang, Yuhuan Yao, Kefeng Deng, Lei Zhang, and Wei-Ying Ma. Igroup: web image search results clustering. In *MM*, 2006.

[103] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[104] F Sha K Zhang, W-L Chao and K Grauman. Summary transfer: exemplar-based subset selection for video summarization. In *CVPR*, 2016.

[105] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016.

[106] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[107] Srikrishna Karanam, Yang Li, and Richard J Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPRW*, pages 33–40, 2015.

[108] 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.

[109] Yasutomo Kawanishi, Yang Wu, Masayuki Mukunoki, and Michihiko Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, 2014.

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

[111] Gunhee Kim, Leonid Sigal, and Eric P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014.

[112] P Kline. The handbook of psychological testing. *Psychology Press*, 2000.

[113] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person re-identification by unsupervised ℓ_1 graph learning. In *ECCV*, 2016.

[114] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *BMVC*, 2015.

[115] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[116] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[117] Gang Kou, Daji Ergu, and Jennifer Shang. Enhancing data consistency in decision matrix: Adapting hadamard model to mitigate judgment contradiction. *European Journal of Operational Research*, 2014.

[118] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[119] S. Kuanar, K. Ranga, and A. Chowdhury. Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. *IEEE Transactions on Multimedia*, PP(99):1–1, 2015.

[120] S. K. Kuanar, R. Panda, and A.S. Chowdhury. Video Key frame Extraction through Dynamic Delaunay Clustering with a Structural Constraint. *Journal of Visual Communication and Image Representation*, 24(7):1212–1227, 2013.

[121] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

[122] Shuyue Lan, Rameswar Panda, Qi Zhu, and Amit K Roy-Chowdhury. Ffnet: Video fast-forwarding via reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6771–6780, 2018.

[123] K. Lange and T. Wu. An MM algorithm for multicategory vertex discriminant analysis. *Journal of Computational and Graphical Statistics*, 17(3):527–544, 2008.

[124] Ryan Layne, Timothy M Hospedales, and Shaogang Gong. Domain transfer for person re-identification. In *Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, 2013.

[125] Lily Lee, Raquel Romano, and Gideon Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):758–767, 2000.

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

[127] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

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

[129] C. De Leo and B. S. Manjunath. Multicamera Video Summarization and Anomaly Detection from Activity Motifs. *ACM Transaction on Sensor Networks*, 10(2):1–30, 2014.

[130] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1482–1489, 2005.

[131] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[132] Ping Li, Yanwen Guo, and Hanqiu Sun. Multi key-frame abstraction from videos. In *ICIP*, 2011.

[133] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, 2013.

[134] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[135] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[136] Y Li and B Merialdo. Multi-video summarization based on av-mmr. In *CBMI*, 2010.

[137] Y Li and B Merialdo. Multi-video summarization based on video-mmr. In *WIAMIS*, 2010.

[138] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015.

[139] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[140] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[141] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015.

[142] Wen-Nung Lie and Kuo-Chiang Hsu. Video summarization based on semantic feature analysis and user preference. In *SUTC*, 2008.

[143] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *TPAMI*, 2015.

[144] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *ICCV*, 2013.

[145] Chunxiao Liu, Shaogang Gong, and Chen Change Loy. On-the-fly feature importance mining for person re-identification. *PR*, 2014.

[146] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *ECCV*, 2012.

[147] Huaping Liu, Yunhui Liu, Yuanlong Yu, and Fuchun Sun. Diversified key-frame selection using structured optimization. *IEEE Transactions on Industrial Informatics*, 2014.

[148] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. Multi-scale triplet cnn for person re-identification. In *MM*, 2016.

[149] Wei Liu, Steven CH Hoi, and Jianzhuang Liu. Output regularized metric learning with side information. In *ECCV*, 2008.

[150] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*, 2014.

[151] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017.

[152] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *CVPR*, 2014.

[153] C Lu, H Li, Z Lin, and S Yan. Fast proximal linearized alternating direction method of multiplier with parallel splitting. In *AAAI*, 2016.

[154] Canyi Lu, Jinhui Tang, Min Lin, Liang Lin, Shuicheng Yan, and Zhouchen Lin. Correntropy induced l2 graph for robust subspace clustering. In *ICCV*, 2013.

[155] Xiaoqiang Lu, Yuan Yuan, and Pingkun Yan. Image super-resolution via double sparsity regularized manifold learning. *TCSVT*, 2013.

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

[157] Andy J Ma, Jiawei Li, Pong C Yuen, and Ping Li. Cross-domain person reidentification using domain adaptation ranking svms. *TIP*, 2015.

[158] Bingpeng Ma, Yu Su, and Frédéric Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012.

[159] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012.

[160] Y. F. Ma, X. S. Hua, and H. J. Zhang. A Generic Framework of User Attention Model and Its Application in Video Summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.

[161] Zhigang Ma, Yi Yang, Feiping Nie, Nicu Sebe, Shuicheng Yan, and Alexander G Hauptmann. Harnessing lab knowledge for real-world action recognition. *IJCV*, 2014.

[162] J. Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *NIPS*, 2002.

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

[164] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Re-identification in the function space of feature warps. *TPAMI*, 2015.

[165] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification. In *ECCV*, 2016.

[166] Niki Martinel and Christian Micheloni. Re-identify people in wide area camera network. In *CVPRW*, 2012.

[167] Shaohui Mei, Genliang Guan, Zhiyong Wang, Shuai Wan, Mingyi He, and David Dagan Feng. Video summarization via minimum sparse reconstruction. *PR*, 2015.

[168] T Mei, L. X Tang, J Tang, and X. S Hua. Near-lossless semantic video summarization and its applications to video analysis. *TOMCAP*, 9(3):16, 2013.

[169] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *CVPR*, 2016.

[170] Arthur G Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, 2008.

[171] Padmavathi Mundur, Yong Rao, and Yelena Yesha. Keyframe-based video summarization using delaunay clustering. *International Journal on Digital Libraries*, 6(2):219–232, 2006.

[172] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2005.

[173] Hien Nguyen, Vishal Patel, Nasser Nasrabadi, and Rama Chellappa. Sparse embedding: A framework for sparsity promoting dimensionality reduction. *ECCV*, 2012.

[174] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *NIPS*, 2010.

[175] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Unsupervised and semi-supervised learning via ℓ_1-norm graph. In *ICCV*, 2011.

[176] Mila Nikolova and Michael K Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 2005.

[177] J. Ortega and W. Rheinboldt. Iterative Solutions of Nonlinear Equations in Several Variables. *New York: Academic*, pages 253–255, 1970.

[178] Shun-Hsing Ou, Chia-Han Lee, V.S. Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien. On-Line Multi-View Video Summarization for Wireless Video Sensor Network. *IEEE Journal of Selected Topics in Signal Processing*, 9(1):165–179, 2015.

[179] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015.

[180] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 2010.

[181] Rameswar Panda, Amran Bhuiyan, Vittorio Murino, and Amit K Roy-Chowdhury. Unsupervised adaptive re-identification in open world dynamic camera networks. In *CVPR*, 2017.

[182] Rameswar Panda, Abir Das, and Amit K Roy-Chowdhury. Embedded sparse coding for summarizing multi-view videos. In *ICIP*, 2016.

[183] Rameswar Panda, Abir Das, and Amit K Roy-Chowdhury. Video summarization in a multi-view camera network. In *ICPR*, 2016.

[184] Rameswar Panda, Sanjay K Kuanar, and Ananda S Chowdhury. Scalable video summarization using skeleton graph and random walk. In *ICPR*, 2014.

[185] Rameswar Panda and Amit K Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017.

[186] Rameswar Panda and Amit K Roy-Chowdhury. Sparse modeling for topic-oriented video summarization. In *ICASSP*, 2017.

[187] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011.

[188] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *SPM*, 2015.

[189] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[190] Yong Peng and Bao-Liang Lu. Robust structured sparse representation via half-quadratic optimization for face recognition. *Multimedia Tools and Applications*, 2016.

[191] Yuxin Peng and Chong-Wah Ngo. Clip-based similarity measure for query-dependent clip retrieval and video summarization. *TCSVT*, 2006.

[192] Jeff M Phillips and Suresh Venkatasubramanian. A gentle introduction to the kernel distance. *arXiv preprint arXiv:1103.1625*, 2011.

[193] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014.

[194] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *ECCV*, 2014.

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

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

[197] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. *arXiv preprint arXiv:1709.05165*, 2017.

[198] Abir Das Rameswar Panda and Amit K. Roy-Chowdhury. Embedded sparse coding for summarizing multi-view videos. In *ICIP*, 2016.

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

[200] 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*, 115(3):211–252, 2015.

[201] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

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

[203] Peter Sand and Seth Teller. Video matching. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 592–599, 2004.

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

[205] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S Davis. Human detection using partial least squares analysis. In *ICCV*, 2009.

[206] G. Scott and H. Longuett-Higgins. An algorithm for associating the features of two images. *The Royal Society of London*, 1991.

[207] 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.

[208] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *ECCV*, 2016.

[209] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):885–905, 2000.

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

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

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

[213] Michael A Smith and Takeo Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *CVPR*, 1997.

[214] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015.

[215] Wei-Tsung Su, Yung-Hsiang Lu, and Ahmed S Kaseb. Harvest the information from multimedia big data in global camera networks.

[216] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

[217] Waqas Sultani and Mubarak Shah. What if we do not have multiple videos of the same action?video action localization using web images. In *CVPR*, 2016.

[218] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015.

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

[220] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, 2014.

[221] Dapeng Tao, Lianwen Jin, Yongfei Wang, and Xuelong Li. Person reidentification by minimum classification error-based kiss metric learning. *TCYB*, 2015.

[222] Dapeng Tao, Lianwen Jin, Yongfei Wang, Yuan Yuan, and Xuelong Li. Person reidentification by regularized smoothing kiss metric learning. *TCSVT*, 2013.

[223] M Torki and A Elgammal. One-shot multi-set non-rigid feature-spatial matching. In *CVPR*, 2010.

[224] M. Torki and A. Elgammal. Putting local features on a manifold. In *CVPR*, 2010.

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

[226] 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.

[227] B. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transaction on Multimedia Comput., Commun., Appl (TOMCCAP)*, 3(1), 2007.

[228] Laurens JP van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[229] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 2013.

[230] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

[231] X. Wan and J. Xiao. Exploiting neighborhood knowledge for single document summarization and keyphrase extraction. *ACM Transactions on Information Systems (TOIS)*, 28(2), 2010.

[232] X. Wan and J. Yang. Collabsum: exploiting multiple document clustering for collaborative single document summarizations. In *International ACM SIGIR conference on Research and development in information retrieval*, 2007.

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

[234] F Wang and B Merialdo. Multi-document video summarization. In *ICME*, 2009.

[235] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *ECCV*, 2016.

[236] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2013.

[237] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. Event driven web video summarization by tag localization and key-shot identification. *TMM*, 2012.

[238] Meng Wang, Guangda Li, Zheng Lu, Yue Gao, and Tat-Seng Chua. When amazon meets google: Product visualization by exploring multiple web sources. *TOIT*, 2013.

[239] Shuai Wang, Yang Cong, Jun Cao, Yunsheng Yang, Yandong Tang, Huaici Zhao, and Haibin Yu. Scalable gastroscopic video summarization via similar-inhibition dictionary selection. *Artificial Intelligence in Medicine*, 2016.

[240] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *TPAMI*, 2016.

[241] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person re-identification. *TCSVT*, 2015.

[242] Ying Wang, Chunhong Pan, Shiming Xiang, and Feiyen Zhu. Robust hyperspectral unmixing with correntropy-based metric. *TIP*, 2015.

[243] Zheng Wang, Ruimin Hu, Chao Liang, Qingming Leng, and Kaimin Sun. Region-based interactive ranking optimization for person re-identification. In *Pacific Rim Conference on Multimedia*, 2014.

[244] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. *arXiv preprint arXiv:1711.08565*, 2017.

[245] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[246] Lin Wu, Chunhua Shen, and Anton van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016.

[247] Yang Wu, Wei Li, Michihiko Minoh, and Masayuki Mukunoki. Can feature-based inductive transfer learning help person re-identification? In *ICIP*, 2013.

[248] Ziyan Wu, Y Li, and Richard J RRadke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *TPAMI*, 2016.

[249] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016.

[250] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016.

[251] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[252] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, 2015.

[253] G. R Xue, C. Lin, Q. Yang, W. Xi, H. J. Zeng, Y. Yu, and Z. Chen. Scalable collaborative filtering using cluster-based smoothing. In *International ACM SIGIR conference on Research and development in information retrieval*, 2005.

[254] Chunlei Y, Jinye P, and Jianping F. Image collection summarization via dictionary learning for sparse representation. In *CVPR*, 2012.

[255] I Yahiaoui, B Merialdo, and B Huet. Generating summaries of multi-episode video. In *ICME*, 2001.

[256] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016.

[257] Allen Y Yang, S Shankar Sastry, Arvind Ganesh, and Yi Ma. Fast 1-minimization algorithms and an application in robust face recognition: A review. In *ICIP*, 2010.

[258] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2:78, 2006.

[259] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *TIP*, 2018.

[260] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *ECCV*, 2014.

[261] Yi Yang, Zhigang Ma, Zhongwen Xu, Shuicheng Yan, and Alexander G Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013.

[262] Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Majorization-minimization for manifold embedding. In *AISTATS*, 2015, to appear.

[263] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Compressive document summarization via sparse optimization. In *AAAI*, 2015.

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

[265] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016.

[266] Dong Yi, Zhen Lei, Shengcai Liao, Stan Z Li, et al. Deep metric learning for person re-identification. In *ICPR*, 2014.

[267] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[268] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[269] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarizatio. *CVPR*, 2016.

[270] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016.

[271] Luming Zhang, Yingjie Xia, Kuang Mao, He Ma, and Zhenyu Shan. An effective video summarization framework toward handheld devices. *IEEE Transactions on Industrial Electronics*, 2015.

[272] 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.

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

[274] Liming Zhao, Xi Li, Jingdong Wang, and Yueting Zhuang. Deeply-learned part-aligned representations for person re-identification. *arXiv preprint arXiv:1707.07256*, 2017.

[275] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[276] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014.

[277] Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, Hongkai Yu, Dazhou Guo, and Song Wang. Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *ICCV*, 2017.

[278] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[279] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[280] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, and Qi Tian. Person re-identification in the wild. *arXiv preprint*, 2017.

[281] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph Regularized Sparse Coding for Image Representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

[282] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Transfer re-identification: From person to set-based verification. In *CVPR*, 2012.

[283] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *TPAMI*, 2013.

[284] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *TPAMI*, 2016.

[285] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 3, 2017.

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

[287] Sanping Zhou, Jinjun Wang, Deyu Meng, Xiaomeng Xin, Yubing Li, Yihong Gong, and Nanning Zheng. Deep self-paced learning for person re-identification. *PR*, 2018.

[288] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017.

[289] Xiatian Zhu, Chen Change Loy, and Shaogang Gong. Learning from multiple sources for video summarisation. *IJCV*, 2016.

[290] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast open-world person re-identification. *TIP*, 2018.