Camera network video summarization

Panda, Rameswar, Roy-Chowdhury, Amit

Rameswar Panda, Amit K. Roy-Chowdhury, "Camera network video summarization," Proc. SPIE 10223, Real-Time Image and Video Processing 2017, 102230I (1 May 2017); doi: 10.1117/12.2262408

SPIE.

Event: SPIE Commercial + Scientific Sensing and Imaging, 2017, Anaheim, California, United States

Camera Network Video Summarization

Rameswar Panda[a] and Amit K. Roy-Chowdhury[a]

[a]University of California, Riverside, USA

ABSTRACT

Networks of vision sensors are deployed in many settings, ranging from security needs to disaster response to environmental monitoring. Many of these setups have hundreds of cameras and tens of thousands of hours of video. The difficulty of analyzing such a massive volume of 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. Much progress has been made in developing a variety of ways to summarize a single video in form of a key sequence or video skim. However, generating a summary from a set of videos captured in a multi-camera network still remains as a novel and largely underaddressed problem.

In this paper, with the aim of summarizing videos in a camera network, we introduce a novel representative selection approach via joint embedding and capped $\ell_{2,1}$-norm minimization. The objective function is two-fold. The first is to capture the structural relationships of data points in a camera network via an embedding, which helps in characterizing the outliers and also in extracting a diverse set of representatives. The second is to use a capped $\ell_{2,1}$-norm to model the sparsity and to suppress the influence of data outliers in representative selection. We propose to jointly optimize both of the objectives, such that embedding can not only characterize the structure, but also indicate the requirements of sparse representative selection. Extensive experiments on standard multi-camera datasets well demonstrate the efficacy of our method over state-of-the-art methods.

Keywords: Camera Network, Video Summarization, Sparse Optimization

1. INTRODUCTION

Summarizing a video sequence is of considerable practical importance as it helps the user in several video analysis applications like content-based search, interactive browsing, retrieval and semantic storage, among others. Most summarization methods are designed to extract a brief yet informative representation of a single-view video in form of a key-frame sequence or a video skim.[1,3,4,9,13,14,16,18,19,27,30] However, another important problem and rarely addressed in this context is to find an informative summary from *multi-view* videos in a camera network.[5,20,24–26] Similar to the single-video summarization problem, the multi-view video summarization in a camera network seeks to take a set of input videos captured from different cameras focusing on roughly the same fields-of-view (Fovs) from different viewpoints and produce a reduced set of output videos or key-frame sequence that presents the most important portions of the inputs within a short duration.

Summarizing multi-view videos in a camera network is different from single-view videos in two important ways. First, since all cameras capture the scenes with overlapping portions from different viewpoints, these videos have large amount of inter-view statistical dependencies.[5] So, intra-view as well as inter-view content correlations across multiple views need to be properly modeled in order to obtain an informative summary. Second, different

Further author information: (Send correspondence to R.P.)

R.P.: E-mail: rpand002@ucr.edu, & A.R.C.: E-mail: amitrc@ece.ucr.edu.

Real-Time Image and Video Processing 2017, edited by Nasser Kehtarnavaz, Matthias F. Carlsohn,
Proc. of SPIE Vol. 10223, 102230I · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2262408

environmental factors like difference in illumination, pose and synchronization issues among the cameras pose a great challenge in multi-view settings. So, techniques that attempt to find informative summary from single-view videos usually do not produce an optimal summary while summarizing videos in a camera network.

Our current work is a novel framework in the direction of camera network video summarization, where we formulate the problem of preserving content correlations in multi-view videos as an embedding problem where the goal is to embed all the frames in an unified space where the locations of the frames preserve both intra-view and inter-view correlations. This is achieved by minimizing an objective function that has two terms; one due to intra-view correlations and another due to inter-view correlations across the multiple views. Such an embedding space acts as a new unified feature space that makes the multi-view video summarization problem in a camera network tractable in the light of sparse coding. We then utilize a capped ℓ_{21}-norm based sparse representative selection approach to extract an informative video summary. The proposed representative selection approach prefers data points that are coherent with each other in the input videos while outlier samples are likely to be suppressed. Finally, to better leverage the embedding and selection mechanism, we learn the embedding and optimal representatives jointly. Specifically, instead of simply using embedding to characterize the structural relationships 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. We demonstrate the effectiveness of our summarization approach on several standard multi-view camera network datasets including both indoor and outdoor environments.

2. METHODOLOGY

In this section, we first present details on learning the latent embedding in a camera network (Section 2.1) and then introduce our robust variant of sparse representative selection by utilizing a capped $\ell_{2,1}$-norm (Section 2.2). Finally, we present a scheme for joint embedding and representative selection (Section 2.3).

2.1 Learning the Latent Embedding in a Camera Network

Consider a set of K different videos captured from different cameras in a camera network, where $\mathbf{X}^{(k)} = \{\mathbf{X}^{(k)}_{.,i} \in \mathbb{R}^D, i = 1, \cdots, N_k\}, k = 1, \cdots, K$. Each $\mathbf{X}_{.,i}$ (i.e., i-th column of $\mathbf{X}$) represents the feature descriptor of a video frame in D-dimensional feature space. As the videos* are captured non-synchronously, the number of frames in each video might be different. We use N_k to denote the number of frames in each video and N to denote the total number of frames in all videos. Since the videos are of different sizes, there is no optimal one-to-one correspondence that can be assumed.

Given a set of videos, our goal is to find a latent embedding for all the frames while satisfying the locality and correlations. Specifically, we are seeking a set of embedded coordinates $\mathbf{Y}^{(k)} = \{\mathbf{Y}^{(k)}_{.,i} \in \mathbb{R}^d, i = 1, \cdots, N_k\}, k = 1, \cdots, K$, where, d ($<< D$) is the dimensionality of the embedding space. The embedding should satisfy the following two constraints: (1) The content correlations† between frames of a video should be preserved in the embedding space. (2) The frames 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. Motivated by this observation, we reach the following objective function on the embedded points $\mathbf{Y}$.

$$\min_{\mathbf{Y}} \sum_{i,j} ||\mathbf{Y}_{.,i} - \mathbf{Y}_{.,j}||^2 \mathbf{W}(i,j) = \min_{\mathbf{Y}, \mathbf{Y}\mathbf{Y}^T = \mathbf{I}} tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T) \tag{1}$$

*We use the word *view* and *video* interchangeably throughout this paper.

†In this paper, *correlations* represents the feature similarity between video frames.

where

$$\mathbf{W}^{(s,t)}(i,j) = \begin{cases} \mathbf{W}^{(k)}_{intra}(i,j) & \text{if } s = t = k \\ \mathbf{W}^{(m,n)}_{inter}(i,j) & \text{otherwise} \end{cases} \tag{2}$$

The matrix $\mathbf{W}$ in (1) represents a $N \times N$ similarity matrix where diagonal blocks represent the intra-view correlations and off-diagonal blocks represent inter-view correlations, as in (2). $\mathbf{W}^{(k)}_{intra}(i,j)$ represents the intra-view pair-wise correlation score between two frames i and j in k-th video whereas $\mathbf{W}^{(m,n)}_{inter}(i,j)$ represents the inter-view pair-wise correlation score between two frames i and j in m-th and n-th video respectively. $\mathbf{L} \in \mathbb{R}^{N \times N}$ is the graph Laplacian matrix of $\mathbf{W}$, i.e., $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix defined as $\mathbf{D}_{i,i} = \sum_j \mathbf{W}_{i,j}$. There are a lot of ways to construct $\mathbf{W}_{intra}$ and $\mathbf{W}_{inter}$. In this paper, we employ the Gaussian kernel to measure the correlations, since it is simple to implement and it performs well in practice. Note that optimizing the first part in (1) reduces to the problem of Laplacian embedding[2] of the data points defined by the feature affinity matrix $\mathbf{W}$. Hence, minimizing this objective function is a generalized eigenvector problem: $\mathbf{Ly} = \lambda\mathbf{Dy}$ 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 computer vision tasks.[7,21,23]

2.2 Robust Sparse Representative Selection

Once the embedding is obtained, our next goal is to find an optimal subset of all the embedded frames, such that each frame can be described as a weighted linear combination of a few of the frames from the subset. The subset is then referred as the informative summary of the multi-view videos in a camera network.

Sparse representative selection solves the following problem to estimate a selection matrix $\mathbf{Z} \in \mathbb{R}^{N \times N}$:

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{YZ}\|_F^2 \quad s.t. \quad \|\mathbf{Z}\|_{2,0} \leq \tau \tag{3}$$

where $||\mathbf{Z}||_{2,0}$ gives the number of nonzero rows of the matrix $\mathbf{Z}$. This is a NP-hard problem since it requires searching over every subset of the τ columns of $\mathbf{Y}$. A standard relaxation to the problem (4) is given by

$$\min_{\mathbf{Z}} \|\mathbf{Y} - \mathbf{YZ}\|_F^2 + \lambda\|\mathbf{Z}\|_{2,1} \tag{4}$$

where $||\mathbf{Z}||_{2,1} = \sum_{i=1}^{n} ||\mathbf{Z}_{i,.}||_2$ denotes the $\ell_{2,1}$-norm and $||\mathbf{Z}_{i,.}||_2$ is the ℓ_2-norm of the i-th row of $\mathbf{Z}$. $\lambda > 0$ is a regularization parameter that controls the level of sparsity in the reconstruction. Once (4) is solved, the representatives are selected as the frames whose corresponding $||\mathbf{Z}_{i,.}||_2 \neq 0$.

A major drawback of the existing sparse representative selection approach in (4) are with the convex $\ell_{2,1}$ relaxation on the sparse coefficient matrix $\mathbf{Z}$. Specifically, all frames are weighted equally in the $\ell_{2,1}$-regularized sparsity term while selecting representative frames for constructing a video summary. As a result, it may incorrectly keep the irrelevant rows (which should have been zero rows) or shrink the relevant rows (which should have been large rows) to zero vectors.[8] However, a robust formulation should treat the true samples and the outliers unequally by assigning different penalties in the sparsity term. Motivated by this observation, we use a non-convex[‡] capped $\ell_{2,1}$-norm[28,29] to model the sparsity, as follows.

[‡]Although we can only find a local optimal solution with the use of non-convex relaxation, it is conclusively shown that under appropriate assumptions, the (local) solution computed by the nonconvex regularization often achieves better performance than the standard convex relaxation in many practical applications.[6,8,10,12,17,28]

$$\min_{\mathbf{Z}\in\mathbb{R}^{N\times N}} \ ||\mathbf{Y}-\mathbf{YZ}||_F^2+\lambda\sum_{i=1}^{n}\min(\|\mathbf{Z}_{i,.}\|_2,\theta) \tag{5}$$

where θ is a thresholding parameter which says that we use ℓ_2 penalization when $\|\mathbf{Z}_{i,.}\|_2$ is sufficiently small, but the penalty does not increase when $\|\mathbf{Z}_{i,.}\|_2$ is larger than the threshold θ. In other words, minimization of (5) tends to shrink the rows of $\mathbf{Z}$ which have very few non-zero entries. This is logical since outliers are incoherent with respect to the true data and take part in the representation of only a few other outliers or data points. However, true data points choose points among themselves as representatives since they are more coherent with each other. Note that, when θ approaches ∞, problem (5) changes to a standard $\ell_{2,1}$-norm regularized objective with same reconstruction error.

Moreover, if we define a concave function $\mathbf{g}(z)=\min(\sqrt{z},\theta)$ with $\theta>0$, we have

$$\mathbf{g}'(z)=\begin{cases}\frac{1}{2\sqrt{z}}, & \text{if } 0<z<\theta^2\\ 0, & \text{if } z>\theta^2\end{cases} \tag{6}$$

Therefore, we can reformulate (5) as follows.

$$\min_{\mathbf{Z}}||\mathbf{Y}-\mathbf{YZ}||_F^2+\lambda\sum_{i=1}^{n}\mathbf{P}_{i,i}\|\mathbf{Z}_{i,.}\|_2^2 \tag{7}$$

$$\mathbf{P}_{i,i}=\frac{1}{2||\mathbf{Z}_{i,.}||_2}\mathbb{I}_{(||\mathbf{Z}_{i,.}||_2\leq\theta)} \tag{8}$$

$\mathbb{I}$ is an indicator function, which is equal to 1 if $||\mathbf{Z}_{.,i}||_2\leq\theta$ and 0 otherwise. Furthermore, we can expand (7) as

$$\min_{\mathbf{Z}}||\mathbf{Y}-\mathbf{YZ}||_F^2+\lambda tr(\mathbf{Z}^T\mathbf{PZ}) \tag{9}$$

where $\mathbf{P}$ is defined as in (8).

Now, we would like to explain why we can select robust representatives by minimizing (9). From the definition of $\mathbf{P}_{i,i}$ in (8), we can see that if $||\mathbf{Z}_{i,.}||_2$ is smaller than θ, then $\mathbf{P}_{i,i}$ is large and minimization of (9) tends to derive $\mathbf{Z}_{i,.}$ with much smaller ℓ_2-norm. After several iterations, norms of such $\mathbf{Z}_{i,.}$s are close to zero such that the corresponding data points are not selected as representatives.

2.3 Joint Embedding Learning and Sparse Representative Selection

We now discuss our proposed method to jointly optimize low-dimensional embedding and sparse representation to select a diverse set of representative frames. 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 (1) and (9), our joint objective function becomes:

$$\min_{\mathbf{Y},\mathbf{Z},\mathbf{YY}^T=\mathbf{I}} tr(\mathbf{YLY}^T)+\alpha\big(||\mathbf{Y}-\mathbf{YZ}||_F^2+\lambda tr(\mathbf{Z}^T\mathbf{PZ})\big) \tag{10}$$

where $\alpha>0$ is a parameter that controls the trade-off between the two objectives. The first term of the cost function projects the input frames into a latent embedding by capturing the meaningful structure of data, whereas the second term helps in selecting a robust set of representative frames 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. OPTIMIZATION

The optimization problem in (10) is non-convex. Solving it is thus more difficult than (4) due to the capped norm and the additional embedding variable $\mathbf{Y}$. An efficient iterative optimisation algorithm is formulated in this work to solve it. As will be explained later, through this kind of optimization procedure, we update the embedding $\mathbf{Y}$ and the sparse coefficient matrix $\mathbf{Z}$ alternatively to jointly optimize both of the objectives.

Note that the problem (10) is convex separately with respect to $\mathbf{Y}, \mathbf{Z}$, and $\mathbf{P}$. Hence, we can solve (10) alternatively with the following three steps w.r.t $\mathbf{P}, \mathbf{Y}$, and $\mathbf{Z}$, respectively.

(1) Solving for P: When $\mathbf{Z}$ is fixed, we can update $\mathbf{P}$ by employing the formulation in Eq. 8 directly.

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

$$\begin{aligned} &\min_{\mathbf{Y},\mathbf{Y}\mathbf{Y}^T=\mathbf{I}} tr(\mathbf{Y}\mathbf{L}\mathbf{Y}^T) + \alpha tr((\mathbf{Y}-\mathbf{Y}\mathbf{Z})(\mathbf{Y}-\mathbf{Y}\mathbf{Z})^T) \\ &= \min_{\mathbf{Y},\mathbf{Y}\mathbf{Y}^T=\mathbf{I}} tr(\mathbf{Y}(\mathbf{L}+\alpha(\mathbf{I}-2\mathbf{Z}+\mathbf{Z}\mathbf{Z}^T))\mathbf{Y}^T) \end{aligned} \tag{11}$$

Eq. 11 can be solved by eigen-decomposition of the matrix $(\mathbf{L}+\alpha(\mathbf{I}-2\mathbf{Z}+\mathbf{Z}\mathbf{Z}^T))$. We pick up the eigenvectors corresponding to the m smallest eigenvalues.

(3) Solving for Z: For a given $\mathbf{P}$ and $\mathbf{Y}$, solve the following objective to estimate $\mathbf{Z}$:

$$\min_{\mathbf{Z}} \alpha\big(tr((\mathbf{Y}-\mathbf{Y}\mathbf{Z})(\mathbf{Y}-\mathbf{Y}\mathbf{Z})^T) + \lambda tr(\mathbf{Z}^T\mathbf{P}\mathbf{Z})\big) \tag{12}$$

By setting the derivative of (12) with respect to $\mathbf{Z}$ to zero, the optimal solution can be computed by solving the following linear system.

$$(\mathbf{Y}^T\mathbf{Y}+\lambda\mathbf{P})\mathbf{Z} = \mathbf{Y}^T\mathbf{Y} \tag{13}$$

We continue to alternately solve for $\mathbf{P}, \mathbf{Y}$, and $\mathbf{Z}$ 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 embedding by solving (1) using an eigen decomposition and $\mathbf{P}$ 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.

4. EXPERIMENTS

Datasets and Performance Measures. We conduct experiments using two datasets: (i) **Office** dataset captured with 4 stably-held web cameras in an *indoor* environment, and (ii) **Campus** dataset taken with 4 handheld video cameras in an *outdoor* scene. We use three quantitative measures on all experiments, including Precision, Recall and F-measure.[5,15] We employ the ground truth of important events reported in[5] to compute the performance measures. In our approach, an event is taken to be correctly detected if we get a representative frame from the set of ground truth frames between the start and end of the event.

Features. We utilize Pycaffe with the "BVLC CaffeNet" pretrained model[11] to extract a 4096-dim CNN feature vector (i.e. the top layer hidden unit activations of the network) for each video frame. We use deep features, as they are the state-of-the-art visual features and have shown best performance on various recognition tasks.

Compared Methods. We compare our approach with total of seven existing approaches including four baseline methods (`Con-Att`,[22] `Con-SRS`,[4] `Att-Con`,[22] `SRS-Con`[4]) that use single-video summarization approach over multi-view datasets to generate summary and three methods (`RandomWalk`,[5] `RoughSets`,[20] `BipartiteOPF`[15]) which are specifically designed for multi-view video summarization. Note that the first two baselines (`Con-Att`, `Con-SRS`)

Table 1. Performance comparison with both single and multi-view methods applied on two camera network datasets. All the reported values are in percentage. We highlight the **best** and second best baseline method.

Datasets	Measures	Con-Att	Con-SRS	Att-Con	SRS-Con	RandomWalk	RoughSets	BipartiteOPF	Ours
Office	Precision	100	100	100	93	100	100	100	**100**
	Recall	38	43	46	57	61	61	69	**81**
	F-measure	55.07	59.46	63.01	71.30	76.19	76.19	81.79	**89.36**
Campus	Precision	56	62	40	58	70	69	75	**87**
	Recall	48	55	28	52	55	57	69	**76**
	F-measure	51.86	58.61	32.66	54.49	61.56	62.14	71.82	**80.77**

Figure 1. 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. As per the ground truth: A0 represents a girl with a black coat, A1 represents the same girl with a yellow sweater and B0 indicates another girl with a black coat. C and D are two boys. D wears a black topcoat and C wears a dark yellow sweater. E is a old man and F is a young guy about thirty years old. The sequence of events in our summary are: E1: A0, B, and D go out of the room, E2: A0 enters the room, E3: A0 stands in cubicle 1, E4: A0 sits in cubicle 1, E6: A0 leaves the room, E7: A1 enters the room and stands in Cubicle 1, E8: A1 sits in cubicle 1, E9: A1 and C leave the room one after another, E10: B0 enters the room, E11: C enters the room. Limited to the space, we only present 10 events arranged in temporal order, as per the ground truth in.[5] Best viewed in color.

concatenate all the views into a single video and then apply a summarization approach, whereas in the other two baselines (`Att-Con`, `SRS-Con`), an approach is first applied to each view and then the resulting summaries are combined along the time line to form a single summary. The purpose of comparing with single-video summarization methods is to show that techniques that attempt to find informative summary from a single-video usually do not produce an optimal set of representatives while summarizing multiple videos.

Results. We have the following observations from Tab. 1: (i) our approach produces summaries with same precision as `RandowWalk` and `BipartiteOPF` for Office dataset. 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. (ii) for all methods, including ours, performance on Campus dataset is not that good as compared to the other dataset. This is obvious since the Campus dataset contains many trivial events as it was captured in an outdoor environment, thus making the summarization more difficult. Nevertheless, for this challenging dataset, F-measure of our approach is about 9% better than that of the recent `BipartiteOPF`. Overall, on both datasets, our approach outperforms all the baselines in terms of F-measure. This corroborates the fact that the proposed approach is more robust and produces informative multi-view summaries in contrast to the state-of-the-art methods (See Fig. 1 for an illustrative example).

Moreover, while comparing with several single-video summarization approaches (`Con-Att`, `Con-SRS`, `Att-Con`, `SRS-Con`, `RandomWalk`, `RoughSets`, `BipartiteOPF`), Tab. 1 reveals that summaries produced using these methods contain a lot of redundancies (simultaneous presence of most of the events) since they fail to exploit the complicated inter-view correlations present in multi-view videos. However, our approach significantly outperforms these methods due to its ability to model multi-view correlations via an embedding and to suppress irrelevant frames via capped norm minimization.

5. CONCLUSIONS

We addressed the problem of summarizing multi-view videos in a camera network using a joint embedding learning and robust sparse representative selection. The embedding helps in preserving structural correlations and also in extracting diverse representatives whereas the capped $\ell_{2,1}$-norm suppresses outlier samples by iteratively assigning weights during the optimization. An alternate minimization algorithm is used to optimize the non-convex objective. Experiments on two standard camera network datasets well demonstrate the robustness of our method in extracting informative video summaries compared to the state-of-the-art methods.

REFERENCES

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

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

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

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

[5] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou. Multi View Video Summmarization. *TMM*, 2004.

[6] H. Gao, F. Nie, W. Cai, and H. Huang. Robust capped norm nonnegative matrix factorization: Capped norm nmf. In *CIKM*, 2015.

[7] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely–laplacian sparse coding for image classification. In *CVPR*, 2010.

[8] P. Gong, J. Ye, and C.-s. Zhang. Multi-stage multi-task feature learning. In *NIPS*, 2012.

[9] 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.

[10] A. Jalali, S. Sanghavi, C. Ruan, and P. K. Ravikumar. A dirty model for multi-task learning. In *NIPS*, 2010.

[11] 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*, 2014.

[12] W. Jiang, F. Nie, and H. Huang. Robust dictionary learning with capped l1-norm. In *AAAI*, 2015.

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

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

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

[16] S. K. Kuanar, R. Panda, and A. 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.

[17] G. Lan, C. Hou, and D. Yi. Robust feature selection via simultaneous capped l2-norm and l2,1-norm minimization. In *ICBDA*, 2016.

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

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

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

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

[22] Y. F. Ma, X. S. Hua, and H. J. Zhang. A Generic Framework of User Attention Model and Its Application in Video Summarization. *TMM*, 2005.
[23] F. Nie, H. Wang, H. Huang, and C. Ding. Unsupervised and semi-supervised learning via 1-norm graph. In *ICCV*, 2011.
[24] S.-H. Ou, C.-H. Lee, V. Somayazulu, Y.-K. Chen, and S.-Y. Chien. On-Line Multi-View Video Summarization for Wireless Video Sensor Network. *IEEE Journal of Selected Topics in Signal Processing*, 9(1):165–179, 2015.
[25] R. Panda, A. Das, and A. K. Roy-Chowdhury. Embedded sparse coding for summarizing multi-view videos. In *ICIP*, 2016.
[26] R. Panda, A. Das, and A. K. Roy-Chowdhury. Video summarization in a multi-view camera network. 2016.
[27] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.
[28] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *JMLR*, 2010.
[29] T. Zhang et al. Multi-stage convex relaxation for feature selection. *Bernoulli*, 2013.
[30] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014.