SPARSE MODELING FOR TOPIC-ORIENTED VIDEO SUMMARIZATION

Rameswar Panda *Amit K. Roy-Chowdhury*

Electrical and Computer Engineering Department, University of California, Riverside

ABSTRACT

While most existing video summarization approaches aim to extract an informative summary of a single video, we propose an unsupervised framework for summarizing topic-related videos by exploring complementarity within videos. We develop a novel sparse optimization method to extract a diverse summary that is both interesting and representative in describing the video collection. To efficiently solve our optimization problem, we develop an alternating minimization algorithm that minimizes the overall objective function with respect to one video at a time while fixing the other videos. Experimental results demonstrate that our approach clearly outperforms the state-of-the-art methods.

***Index Terms*—** Video Summarization, Sparse Coding

1. INTRODUCTION

Video summarization is a challenging problem with great application potential. Imagine a scenario that being unfamiliar with a place. e.g., Machu Picchu, a user performed a video search on YouTube to find out whether he/she would like the place and to discover what to expect while visiting the place. The search result consists of a set of relevant videos presenting different aspects of the place. Given that browsing through all the videos is a very time consuming task, we want to explore whether we can automatically create a preview video summary considering both the overlap and complementarity within the videos.

Much progress has been made in developing a variety of ways to summarize a single video in an unsupervised manner [1, 2, 3], or developing supervised algorithms [4, 5]. However, generating a summary from a set of topic-related videos still remains as a novel and largely under-addressed problem. Some of early works on topic-oriented video summarization focused on videos of specific genres, such as tv news [6, 7] and generated an automatic summary by frame clustering [8] or leveraging genre specific information, e.g., speech transcripts in news [9, 10]. These methods generally fail to summarize large scale open world web videos since they are unstructured and range over a wide variety of content. A recent approach on summarizing multiple sensor-rich topic-related videos can be seen in [11]. However, the system relies on meta-data sensor information related to a geographical area that are mostly unavailable while summarizing unconstrained topic-related web videos generated from a search.

In this paper, we focus on the task of summarizing a set of topic-related web videos resulting from a search[1]. We observe that each video in the set may contain some information that other videos do not have, and thus exploring the underlying complementarity is of great importance for the success of topic-oriented video summarization. We achieve this by developing a novel sparse optimization method that jointly summarizes a set of videos to find a single diverse summary to optimally describe the video collection. Our approach consider two aspects. One, it considers prior knoweldge in form shot interestingness to extract summary that is both interesting and representative of the input video. Second, we introduce a novel diversity regularizer in the optimization framework to explore the complementarity within multiple videos in extracting a high quality multi-video summary. We finally develop an efficient alternating minimization algorithm to solve our optimization problem.

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

- We propose a novel approach for topic-oriented video summarization by exploring complementarity within the videos.
- We develop a novel diversity-aware sparse optimization method based on weighted $\ell_{2,1}$-norm that can be efficiently solved by an alternating minimization algorithm.
- We obtain excellent experimental results, showing that our approach generates high quality summaries compared to the state-of-the-art methods.

2. TOPIC-ORIENTED VIDEO SUMMARIZATION

Problem Statement: Consider a set of m relevant web videos given a video search, where $X^v = \{X^v_{.,i} \in \mathbb{R}^d, i = 1, \cdots, n_v\}, v = 1, \cdots, m$. Each $X^v_{.,i}$ represents the feature descriptor of a video shot in d-dimensional feature space.

Given a set of topic-related videos, our goal is to find a summary that conveys the most *important* details of the original video collection. Specifically, it is composed of several shots that represent most important portions of the input video collection within a short duration.

Preliminaries: Sparse optimization approaches [12, 13] select representative shots from a single video by modeling sparsity and representativeness as follows:

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

where $||Z^v||_{2,1} = \sum_{i=1}^{n_v} ||Z^v_{i,.}||_2$ and $||Z^v_{i,.}||_2$ is the ℓ_2-norm of the i-th row of Z. $\lambda_s^v > 0$ is a regularization parameter

[1] We assume that videos given by a search are relevant to the topic. However, in most cases, some videos may not be relevant to the topic. One can use either clustering or additional video meta data to refine the results.

that controls the level of sparsity in the reconstruction. Once the problem (1) is solved, the representatives are selected as the shots whose corresponding $||Z^v_{i,.}||_2 \neq 0$. The affine constraint ${Z^v}^T 1 = 1$ makes the selection of representatives invariant with respect to the global translation of the data.

Introducing Prior Knowledge via Weighted $\ell_{2,1}$-Norm: Note that in problem (1), all shots are treated equally in selecting representatives. However, a good summarization method can certainly benefit from incorporating prior knowledge from the application domain or user specifications. For instance, optimizing only for representativeness risks leaving out some crucial shot(s) which can be captured in the summary by combining interestingness and representatives in summarizing videos. To better leverage prior knowledge in video summarization, we propose a weighted $\ell_{2,1}$-norm based objective function as follows:

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

where $Q^v = [diag(q^v)]^{-1}$ and $q^v \in \mathbb{R}^{n_v}$ represent the interstingness score of each video shot. It is easy to see that problem (2) favors selection of interesting shots by assigning a lower score via Q^v. We follow [22] to compute the interestingness score of each shot by considering attention, asthetic quality and presence of landmarks/persons. However, problem (2) is quite generic to employ any type of prior knowledge-we expect more sophisticated ones will only benefit our proposed approach.

Introducing Diversity of Multiple Videos: The sparse optimization (2) extracts a good summary from a single video. However, summarizing multiple topic-related videos is ubiquitous in web search, hence, extending (1) into multi-video setting is of vital importance for many multimdia applications. Unlike prior works that simply combine the multiple videos into a single one, we explore the complementary structural information within the videos to select a diverse set of representative shots. Mathematically, we have the final objective function as follows:

$$\begin{aligned} \min_{Z^1, Z^2, \cdots, Z^m} & \sum_{v=1}^{m} \|X^v - X^v Z^v\|_F^2 \\ + \lambda_s & \sum_{v=1}^{m} \|Q^v Z^v\|_{2,1} + \lambda_d \sum_{\substack{1 \leq v, w \leq m \\ v \neq w}} f_d(Z^v, Z^w) \quad (3) \\ s.t. \; & {Z^v}^T 1 = 1, \; Z^v \in \mathbb{R}^{n_v \times n_v}, \; \forall \, 1 \leq v \leq m \end{aligned}$$

where λ_s and λ_d are two tradeoffs associated with the sparsity and diversity regularization functions respectively. $f_d(.,.)$ is the regularization function for enforcing the sparse coefficient matrices of different videos to be of maximum diversity. Specifically, the objective of $f_d(.,.)$ is to penalize the condition that two correlated shots from two distinct videos are present in the summary at the same time. For example, if the i-th shot from v-th video is highly correlated to the j-th shot in w-th video, then we do not need to select both of them simultaneously. Mathematically, we define $f_d(.,.)$ as follows:

Definition 1. *Given the sparse coefficient matrices Z^v and Z^w, the diversity regularization function is defined as:*

$$f_d(Z^v, Z^w) = \sum_{i=1}^{n_v} \sum_{j=1}^{n_w} ||Z^v_{i,.}||_2 C_{i,j} ||Z^w_{j,.}||_2 = ||W^{vw} Z^v||_{2,1} \quad (4)$$

where $C_{i,j}$ measure the correlation between i-th sample from v-th view and the j-th sample in w-th view. The second equality follows from the simple manipulation as $W^{vw}_{i,i} = \sum_{j=1}^{n_w} C_{i,j} \|Z^w_{j,.}\|_{2,1}$. Minimization of (4) tries to explore the complementary information by penalizing the condition that rows of two similar shots from two distinct videos are nonzero at the same time.

There are a lot of ways to measure $C_{i,j}$. In this paper, we employ Scott and Longuet-Higgins (SLH) algorithm [14] with Gaussian kernel to measure the correlation, since it is simple to implement and it performs well in several vision tasks [15, 16]. Specifically, SLH algorithm finds an orthonormal permutation matrix by solving a trace maximization problem over the similarity matrix computed using the Gaussian kernel. The orthonormal matrix is then used as the correlation scores after setting the negative values to 0 [16].

3. OPTIMIZATION

To solve (3), we devise an alternative algorithm by minimizing the function with respect to one video at a time while fixing the other videos. Specifically, we minimize the following function with respect to Z^v while keeping others fixed:

$$\begin{aligned} \min_{Z^v} & \|X^v - X^v Z^v\|_F^2 + \lambda_s \|Q^v Z^v\|_{2,1} \\ + \lambda_d & \sum_{w=1, v \neq w}^{m} \|W^{vw} Z^v\|_{2,1} \; s.t. \; {Z^v}^T 1 = 1 \end{aligned} \quad (5)$$

Using the properties of a standard norm, it is easy to reformulate problem (5) as following:

$$\begin{aligned} \min_{Z^v} & \|X^v - X^v Z^v\|_F^2 + \lambda_s \|Q^v Z^v\|_{2,1} \\ & + \lambda_d \|W^v Z^v\|_{2,1} \; s.t. \; {Z^v}^T 1 = 1 \end{aligned} \quad (6)$$

where $W^v = \sum_{w=1, v \neq w}^{m} W^{vw}$. The reformulation follows directly from the fact that $\ell_{2,1}$-norm is a valid norm and the equality in triangle inequality holds if two matrices are positive semidefinite. Furthermore, notice that both second and third term in (6) are functions of the same variable Z^v with two tradeoffs λ_s and λ_d respectively. Hence, with the same logic and ignoring the superscripts for convenience, we can approximate (6) with one tradeoff parameter λ as following:

$$\min_{Z} \|X - XZ\|_F^2 + \lambda \|Z\|_{K,2,1} \; s.t. \; Z^T 1 = 1 \quad (7)$$

where $K = Q + W$, $\|Z\|_{K,2,1}$ denotes the weighted $\ell_{2,1}$-norm of Z and is defined as $\|Z\|_{K,2,1} = \|KZ\|_{2,1}$. When we replace X with $[X^T, \alpha * 1]^T$ where α approaches to infinity, (7) is equivalent to the following problem:

$$\min_{Z} \|X - XZ\|_F^2 + \lambda \|Z\|_{K,2,1} \quad (8)$$

We can prove equation (7) is equivalent to (8) by expanding (8) as follows:

$$\|X - XZ\|_F^2 = \|X^* - X^* Z\|_F^2 + \alpha \|1^T - 1^T Z\|_F^2 \quad (9)$$

where X^* is the original X presented in (7). When α approaches to infinity, $Z^T 1$ approaches to 1. Thus, problem (7) is equivalent to (8).

The objective function (8) is a convex weighted $\ell_{2,1}$-norm minimization problem which can be efficiently solved using ADMM [17]. Finally, the above alternating procedure over multiple videos is carried out until convergence, as in Algo. 1. In all our experiments, we monitor the convergence is reached within less than 10 iterations. Therefore, the proposed method can be applied to large scale problems in practice.

Algorithm 1 Algorithm for solving (3)

```
Input: Video feature matrices $X^1, X^2, \cdots, X^m$
for each v do
  Initialize $Z^v$ by solving (3) with $\lambda_d = 0$;
end for
while not converged do
    for each v do
     repeat
      $U \leftarrow$ Solve the linear system:
             $(X^T X + \mu I)U = X^T X + \mu Z - \Lambda$;
      $Z \leftarrow \max\left\{\|U + \Lambda/\mu\|_2 - \frac{\lambda K}{\mu}, 0\right\} \frac{U+\Lambda/\mu}{\|U+\Lambda/\mu\|_2}$;
      $\Lambda \leftarrow \Lambda + \mu(U - Z)$;
     until converges
    end for
end while
Output: Coefficient matrices $Z^1, Z^2, \cdots, Z^m$.
```

4. SUMMARY GENERATION

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

Remark 1. Since the alternating minimization can make the Algo. 1 stuck in a local minimum, it is important to have a sensible initialization. We initialize the sparse coefficient matrices of $m - 1$ videos by first solving (3) with $\lambda_d = 0$. After the initialization, the follwoing question remain: from which view we should start the alternating minimization? One possible way is to randomly start with any video and repeat the minimization over all videos until convergence. However, since we have some prior knoweldge on which video is more informative in the collection, we can start with initializing and fixing more informative videos, and optimize with respect to the least informative video. More specifically, we start with the specific Z^v which has more number of nonzero rows after the intialization since the number of nonzero rows indicate the relative importance of each video in the collection.

5. EXPERIMENTS

Dataset. To evaluate topic-oriented video summarization, we need a single ground truth summary of all the videos that describes the collection altogether. However, since there exists no such publicly available dataset that fits our need, we introduce a new dataset, we self compiled a dataset from the web. We selected 20 tourist attractions from the Tripadvisor travelers choice landmarks 2015 list[2] and collected 140 videos from YouTube under the Creative Commons license.

Performance Measures. Motivated by [18, 5, 3], we assess the quality of an automatically generated summary by comparing it to human judgment. Specifically, given a proposed summary and a set of human selected summaries, we compute the pairwise F-measure and then report the mean value motivated by the fact that there exists not a single ground truth summary, but multiple summaries are possible. We follow [5] and utilize VSUMM evaluation package [19] for finding matching pair of shots.

Video Segmentation. We first segment videos into multiple shots using an existing algorithm [2] with an constraint to ensure that the number of frames within each shot lies in the range of [32,96]. The segmented shots serve as the basic units for feature extraction and creating ground truth summaries.

Ground truth Summaries. Given the videos that were preprocessed into shots, we asked three study experts to select at least 5%, but no more than 15% shots for each video as well as a single set of diverse shots that can describe the video collection altogether. We set the summary length to be in the range [5%, 15%] of total number of shots to ensure that the input video is indeed summarized rather than being slightly shortened. While audio or embedded text can be used during generating ground truth summaries, we muted the audio to ensure that representative shots are selected based solely on visual stimuli. Moreover, we specify that if something is only mentioned in onscreen text, then it should not be labeled as important. The total user time of the study amounts to over 30 hours. To assert the consistency of human created summaries, we compute both pairwise F-measure and the Cronbach's alpha between them, as in [18, 3]. The dataset has a mean F-measure of 0.643 and mean Cronobach's alpha of 0.944. Ideally alpha is around 0.9 for a good test [20].

Features. In deep feature learning, C3D features [21] have recently shown better performance compared to the features extracted using each frame separately [22, 23]. We therefore extract C3D features, by taking sets of 16 input frames, applying 3D convolutional filters, and extracting the responses at layer FC6, as in [21]. 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 optimization. All methods (including the proposed one) use the same C3D features in representing videos. Such a setting can give a fair comparision for various methods.

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

Table 1. Quantitative results. Numbers show mean F-measures at 10% summary length, *i.e.*, summary containing 10% of total shots of a video collectionn. We highlight the **best** and second best baseline method. Overall, our approach statistically significantly outperforms all baselines $(p < .01)$. Name of the tourist places are presented in the format "name (# videos)".

Topic Names	ConcateKmeans	ConcateSpectral	ConcateSparse	KmeansConcate	SpectralConcate	SparseConcate	MultiVideoContent	MultiVideoMMR	Ours
Angkor Wat (7)	0.426	0.405	0.407	0.418	0.418	0.391	0.431	0.452	**0.567**
Machu Picchu (7)	0.336	0.367	0.379	0.373	0.394	0.427	0.438	0.507	**0.582**
Taj Mahal (7)	0.428	0.484	0.465	0.518	0.522	0.588	0.593	0.533	**0.679**
Basilica of Sagrada Familia (6)	0.423	0.415	0.461	0.382	0.427	0.478	0.488	0.492	**0.597**
St. Peter's Basilica (5)	0.437	0.458	0.497	0.533	0.526	0.575	0.586	0.602	**0.699**
Milan Cathedral (10)	0.475	0.430	0.451	0.449	0.442	0.489	0.481	0.473	**0.571**
Alcatraz (6)	0.601	0.550	0.638	0.631	0.651	0.729	0.652	0.668	**0.755**
Golden Gate Bridge (6)	0.447	0.443	0.508	0.504	0.475	0.509	0.527	0.515	**0.618**
Eiffel Tower (8)	0.408	0.390	0.460	0.401	0.427	0.448	0.436	0.446	**0.562**
Notre Dame Cathedral (8)	0.315	0.350	0.235	0.413	0.451	0.461	0.463	0.473	**0.550**
The Alhambra (6)	0.485	0.570	0.543	0.551	0.551	0.567	0.553	0.582	**0.662**
Hagia Sophia Museum (6)	0.305	0.346	0.315	0.433	0.384	0.523	0.473	0.536	**0.585**
Charles Bridge (6)	0.400	0.379	0.414	0.409	0.444	0.451	0.453	**0.534**	0.525
Great Wall at Mutiantu (5)	0.390	0.410	0.484	0.500	0.474	0.488	0.493	0.507	**0.673**
Burj Khalifa (9)	0.284	0.362	0.350	0.301	0.355	0.352	**0.450**	0.392	0.441
Wat Pho (5)	0.342	0.414	0.564	0.501	0.575	0.633	0.625	0.603	**0.722**
Chichen Itza (8)	0.337	0.361	0.430	0.413	0.426	0.507	0.514	0.492	**0.582**
Sydney Opera House (10)	0.400	0.391	0.497	0.409	0.458	0.474	0.503	0.512	**0.614**
Petronas Twin Towers (9)	0.302	0.326	0.421	0.418	0.376	0.445	0.453	0.486	**0.643**
Panama Canal (6)	0.377	0.410	0.492	0.539	0.523	0.528	0.512	0.544	**0.639**
mean	**0.396**	**0.413**	**0.450**	**0.455**	**0.465**	**0.503**	**0.506**	**0.517**	**0.613**

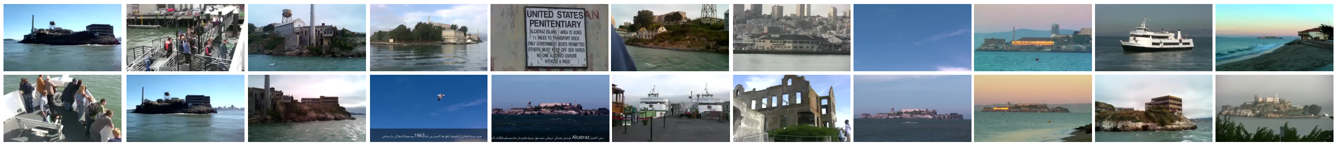
UNITED STATES
PENITENTIARY
ALCATRAZ ISLAND AREA 12 ACRES
1½ MILES TO TRANSPORT DOCK
ONLY GOVERNMENT BOATS PERMITTED
OTHERS MUST KEEP OFF 200 YARDS
NO ONE ALLOWED ASHORE
WITHOUT A PASS

Fig. 1. Video summary generated by our approach for the topic **Alcatraz**. We show the summaries at 10% length (*i.e.,* 22 shots out of total 223 shots) and represent each shot using the central frame. As can be seen from the figure, our approach produces informative shots that can describe the whole video collection in short duration. The F-measure achieved by our approach for this topic is the highest (0.755) in our experimented dataset. (Best viewed in color)

Other Details. The regularization parameter λ is taken as λ_0/γ where $\gamma > 1$ and λ_0 is computed from the data [13]. In Algo. 1, we set the stop criteria for alternating minimization over multiple videos as $\frac{|f^{(t+1)} - f^{(t)}|}{f^{(t)}} < 10^{-2}$, where $f^{(t)}$ is the objective value in the t-th iteration.

Compared Methods. We compare our approach with six baselines (ConcateKmeans, ConcateSpectral, ConcateSparse [13], KmeansConcate, SpectralConcate, SparseConcate [13]) that use single-video summarization approach over multiple videos to generate a summary and two state-of-the-art methods (MultiVideo Content [7], MultiVideoMMR [6]) which are specifically designed for topic-oriented video summarization. The first three baselines (ConcateKmeans, ConcateSpectral, ConcateSparse) concatenate all the videos into a single video and then apply k-means, spectral clustering and sparse coding [13] (i.e., applying (1) to the concatenated video) respectively, whereas in the other three baselines (KmeansConcate, SpectralConcate, SparseConcate), the corresponding approach is first applied to each video and then the resulting summaries are combined to form a single summary. MultiVideoContent [7] uses a greedy approach with a content inclusion measure to summarize multiple videos whereas MultiVideoMMR [6] extends the concept of maximal marginal relevance [24] to the video domain for the same purpose.

Results. Tab. 1 shows that our approach statistically significantly outperforms all other compared methods $(p < .01)$. Our method achieves the highest overall score of 0.613, while the strongest baseline reaches 0.517. Our method is able to find the important shots from a video collection which are comparable to manual human created summaries (see Fig. 1 for an illustrative example). Moreover, while comparing with several single-video summarization approaches (ConcateKmeans, ConcateSpectral, ConcateSparse, KmeansConcate, SpectralConcate, SparseConcate), our method significantly outperforms all the baselines $(p < .01)$ to generate high quality summaries. We observe that directly applying single-video summarization approaches to summarize multiple videos produce a lot of redundant shots in the final summary since they fail to explore the complicated inter-video content correlations. However, our approach efficiently explores these correlations to generate a more informative summary from multiple videos.

Note that our approach outperforms the naïve approach, SparseConcate, that summarizes multiple videos without any diversity constraint with a clear margin (0.613 vs 0.503). This corroborates the importance of exploring the underlyning complementarity in creating a diverse informative summary from multiple topic-related videos.

6. CONCLUSIONS

We present a novel unsupervised framework for topic-oriented video summarization by exploring the complementarity within the videos. We achieve this by developing a diversity-aware sparse optimization method that jointly summarizes a set of videos to find a single summary that is both interesting and representativeness of the input video collection. We show the effectiveness of our approach through rigorous experiments and comparison with state-of-the-art methods.

Acknowledgments: This work was partially supported by NSF grant IIS-1316934.

7. REFERENCES

[1] S. Chakraborty, O. Tickoo, and R. Iyer, "Towards distributed video summarization," in *MM*, 2015.

[2] W. S. Chu, Y. Song, and A. Jaimes, "Video cosummarization: Video summarization by visual cooccurrence," in *CVPR*, 2015.

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

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

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

[6] Y. Li and B. Merialdo, "Multi-video summarization based on video-mmr," in *WIAMIS*, 2010.

[7] F. Wang and B. Merialdo, "Multi-document video summarization," in *ICME*, 2009.

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

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

[10] 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.

[11] 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.

[12] Y. Cong, J. Yuan, and J. Luo, "Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection," *TMM*, 2012.

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

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

[15] D. Pachauri, R. Kondor, and V. Singh, "Solving the multi-way matching problem by permutation synchronization," in *NIPS*, 2013.

[16] M. Torki and A. Elgammal, "One-shot multi-set nonrigid feature-spatial matching," in *CVPR*, 2010.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, 2011.

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

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

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

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

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

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

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