Video Summarization Via Multiview Representative Selection

Jingjing Meng, *Member, IEEE*, Suchen Wang, Hongxing Wang, *Member, IEEE*, Junsong Yuan, *Senior Member, IEEE*, and Yap-Peng Tan, *Senior Member, IEEE*

***Abstract*— Video contents are inherently heterogeneous. To exploit different feature modalities in a diverse video collection for video summarization, we propose to formulate the task as a multiview representative selection problem. The goal is to select visual elements that are representative of a video consistently across different views (i.e., feature modalities). We present in this paper the multiview sparse dictionary selection with centroid co-regularization method, which optimizes the representative selection in each view, and enforces that the view-specific selections to be similar by regularizing them towards a consensus selection. We also introduce a diversity regularizer to favor a selection of diverse representatives. The problem can be efficiently solved by an alternating minimizing optimization with the fast iterative shrinkage thresholding algorithm. Experiments on synthetic data and benchmark video datasets validate the effectiveness of the proposed approach for video summarization, in comparison with other video summarization methods and representative selection methods such as K-medoids, sparse dictionary selection, and multiview clustering.**

***Index Terms*— Video summarization, multi-view, representative selection.**

I. Introduction

VIDEO summarization can be seen as a representative selection problem. Although the resulting visual summaries can take many different forms, such as key objects [1]–[3], keyframes [4]–[9], key shots [10], [11], montages [12], dynamic synopses [13], etc., the common goal is essentially to select representative visual elements that well delineate the essence of a video. However, the *representativeness* of the selected visual elements can be highly dependent on their representations, *i.e.*, the specific features used to describe them. For instance, when a video is represented by appearance features, the resulting summary could be quite different from that obtained from motion features.

To incorporate multiple features, the conventional solution is to concatenate them in to a single one before selecting representatives. However, this simple concatenation does not always produce optimal summaries, as the underlying data distributions in individual views (*i.e.*, feature modalities) can be drastically different. In addition, if the feature dimensions differ greatly, high dimensional features may become dominant thus *shadowing* low dimensional ones. Moreover, noisy features could adversely affect the selection results.

Although multi-view sparse subspace/dictionary learning approaches have been proposed [14]–[19], they require feature fusion to be conducted in advance to learn a unified data representation for representative selection. However, when there are discrepancies between different views, *e.g.*, when data points belong to different groups in different views, it is difficult for the unified representation to maintain the underlying distribution of the data across multiple feature spaces, thus affecting the performance of the subsequent representative selection.

In view of the above limitations, we propose to formulate video summarization as a multi-view representative selection problem, which aims to find a consensus selection of visual elements that is agreeable with all views (*i.e.*, feature modalities). Figure 1 illustrates the idea in comparison with direct concatenation. Specifically, we present the multi-view sparse dictionary selection with centroid co-regularization (MSDS-CC) method. It optimizes the representative selection in each view, and enforces that the view-specific selections to be similar by regularizing them towards a consensus selection (*i.e.*, centroid co-regularization). Different from our previous work [20], we introduce a diversity regularizer to encourage coverage of diverse visual elements in the resulting summaries. Our formulation provides the following benefits:

1) It can produce a consensus selection of visual elements across different views, resulting in summaries that are consistently representative across multiple feature modalities.
2) As we directly optimize for the consensus selection weights based on the view-specific selection weights optimized view-wise, which follow view-specific distributions, our formulation is better at preserving the underlying data distributions of individual views and handling unbalanced feature lengths.
3) Our formulation can better handle noisy features by incorporating view-specific selection priors (Sec. III-F) to guide the representative selection towards more relevant visual elements. This permits the use of external

Manuscript received June 10, 2017; revised October 11, 2017; accepted December 17, 2017. Date of publication January 3, 2018; date of current version February 9, 2018. This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the Infocomm Media Development Authority, Singapore. This work was supported in part by the National Natural Science Foundation of China under Grant 61602069, in part by the Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2016jcyjA0468, and in part by the Singapore Ministry of Education Academic Research Fund Tier 2 under Grant MOE2015-T2-2-114. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aydin Alatan. *(Corresponding author: Jingjing Meng.)*

J. Meng, S. Wang, J. Yuan, and Y.-P. Tan are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jingjing.meng@ntu.edu.sg; wang.sc@ntu.edu.sg; jsyuan@ntu.edu.sg; eyptan@ntu.edu.sg).

H. Wang is with the School of Software Engineering, Chongqing University, Chongqing 401331, China (e-mail: ihxwang@cqu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2789332

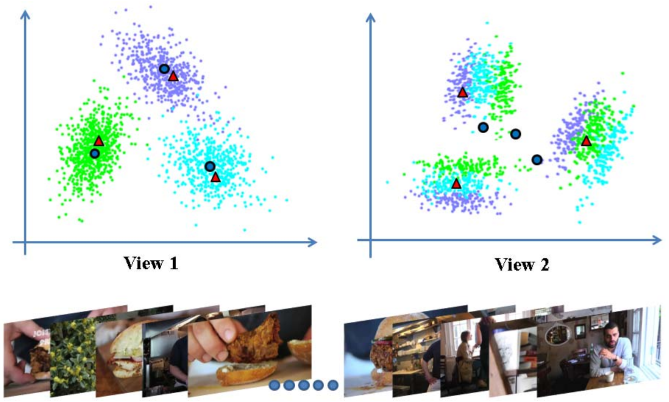


Fig. 1. An illustration of the proposed video summarization via multi-view representative selection. The top row shows two views of a video's frames side by side. There are 3 clusters (key visual concepts) in each view, and we hope to capture all the 6 clusters by selecting only 3 representatives. Note that the features in the two views have different distributions, *e.g.*, the green cluster in View 1 is scattered in View 2 (similarly for the blue and cyan clusters). Therefore, representatives selected after concatenating the two views may miss clusters in View 2 (*e.g.*, the 3 dark blue circles ○ in each view). In comparison, red triangles △ are better representatives as they cover the 3 clusters in both views.

data or/and supervision to improve summarization quality, which has been shown to be effective in [3], [4], and [21]–[23].

4) Comparing with multi-view clustering, which needs to be re-run to generate a summary of different size, the proposed multi-view sparse dictionary selection offers better scalability in that we can produce summaries of various sizes by analyzing the video only once.

Our formulation can be solved efficiently via an alternating minimizing optimization with the fast iterative shrinkage thresholding algorithm (FISTA) [24]. Comparative experiments on synthetic and challenging benchmark datasets demonstrate the efficacy of the proposed approach.

II. Related Work

A. Video Summarization

Previous work in video summarization can be roughly grouped into domain-specific [25]–[29], supervised [10], [11], [23], [30]–[33] and unsupervised methods [1], [4], [21], [22], [34]–[40]. Domain-specific methods focus on summarizing videos in specific genres such as surveillance [41], [42], sports [25], [28] and egocentric videos [26], [27]. Supervised methods generate summaries by learning from human annotations. For instance, to make a structured prediction, sub-modular functions are trained with user created summaries [31]. Gygli *et al.* [10] train a linear regression model to estimate the interestingness score of shots. More recently, Zhang *et al.* [11], Sharghi *et al.* [32], and Gong *et al.* [43] define novel models to learn from human-created summaries to select representative and diverse subsets. In addition, Zhang *et al.* [23] show that summary structures can be transferred between videos that are semantically consistent. Unsupervised methods summarize videos by seeking the visual relevance and structure. One popular method is to select representative frames by learning a dictionary from videos [36], [37], [44], [45]. Another popular trend is to leverage additional information from other sources such as video titles and web images [5], [22], [38]. Recently, video co-summarization [21], [46], [47] has also been proposed, which summarizes shots that co-occur among multiple videos of the same topic.

B. Representative Selection

There are two main categories of methods to find representatives: clustering based methods and subspace learning based methods. Existing clustering based methods include, for example, K-medoids algorithm [48], sparse selection of clustering centers [49], [50], affinity propagation [51], [52] and density peak search [53]. For these methods, representatives are determined by clustering centers. Subspace learning based methods are motivated in a different way, where representatives are required to approximate the data matrix in the sense of linear reconstruction. Such circumstances fall into dictionary learning and selection [3], [37], [41], [54]–[57]. Despite the advances in representative selection, most of the methods are not applicable to multiple features. Feature fusion such as [14]–[19] and [58] has to be conducted in advance so that unified data representations can be learned for representative selection. However, it is difficult for the unified representations to keep the underlying distribution information of the data in multiple feature spaces, thus challenging the subsequent representative selection.

III. The Proposed Method

We formulate the problem of video summarization as multi-view representative selection. Given n shots extracted from a video sequence, each of them can be represented by V views of features. Our goal is to find a subset of shots that are representatives across the multiple views. Below, we arrange the υ-th view of features as the columns of the matrix $\mathbf{X}^{(\upsilon)} \in \mathbb{R}^{d^{(\upsilon)}\times n}$, and denote by $\mathbf{w}^{(\upsilon)} = [w_1^{(\upsilon)}, w_2^{(\upsilon)}, \dots, w_n^{(\upsilon)}]^{\mathrm{T}} \in \mathbb{R}^n$ the vector of selection weights corresponding to the υ-th view. In addition, we use $\mathbf{w} = [w_1, w_2, \dots, w_n]^{\mathrm{T}} \in \mathbb{R}^n$ to denote the vector of consensus selection weights resulting from multiple views.

A. Preliminaries: Feature Concatenation

Let $\mathbf{Y} = [\mathbf{X}^{(1)}; \mathbf{X}^{(2)}; \cdots; \mathbf{X}^{(V)}] \in \mathbb{R}^{\sum_{\upsilon=1}^{V} d^{(\upsilon)}\times n}$ be the concatenated feature matrix of multiple views. Then, we have, $\forall \mathbf{C} \in \mathbb{R}^{n\times n}$,

$$\|\mathbf{Y} - \mathbf{Y}\mathbf{C}\|_{\mathrm{F}}^2 = \sum_{\upsilon=1}^{V} \|\mathbf{X}^{(\upsilon)} - \mathbf{X}^{(\upsilon)}\mathbf{C}\|_{\mathrm{F}}^2. \tag{1}$$

As a result, the following representative selection objective in (2) is tantamount to that of feature concatenation for sparse dictionary selection [41].

$$\min_{\mathbf{C}\in\mathbb{R}^{n\times n}} \sum_{\upsilon=1}^{V} \frac{1}{2}\|\mathbf{X}^{(\upsilon)} - \mathbf{X}^{(\upsilon)}\mathbf{C}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{C}\|_{1,2}, \tag{2}$$

where $\|\mathbf{C}\|_{1,2} = \sum_{i=1}^{n} \|\mathbf{C}_{i.}\|_2$, associated with the parameter λ as a regularization to the sum of reconstruction errors of multiple views, and $\|\mathbf{C}_{i,\cdot}\|_2$ is the l_2 norm of the i-th row of the selection matrix $\mathbf{C}$. In this case, $w_i = \|\mathbf{C}_{i,\cdot}\|_2$, measuring the selection confidence to the i-th sample. The solution to (2) can be obtained by the proximal gradient method [24]. Finally, exemplars can be found by ranking the consensus selection weights w_i, for $i = 1, 2, \cdots, n$.

B. Centroid Co-Regularization

It is worth noting that in (2), features in different views are treated equally to learn a consensus selection matrix. However, different views of features can be dramatically different, which heavily influences the selection result. To better handle multiple features, we propose to learn individual selection matrices $\mathbf{C}^{(v)}, v = 1, 2, \cdots, V$ for different features, and simultaneously unify them to a consensus weighting vector $\mathbf{w}$, with its i-th entry w_i measuring the selection confidence of the i-th sample. We thus formulate our objective function as multi-view sparse reconstruction with centroid co-regularization:

$$\min_{\mathbf{C}^{(v)},\mathbf{w}} \sum_{v=1}^{V} \left\{ \frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{C}^{(v)}\|_{\mathrm{F}}^2 + \lambda^{(v)}\|\mathbf{C}^{(v)}\|_{1,2} + \frac{1}{2}\eta\|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 \right\}, \quad (3)$$

where the weighting vector $\mathbf{w}^{(v)}$ consists of the l_2 norms of rows of $\mathbf{C}^{(v)}$, with the i-th entry $w_i^{(v)} = \|\mathbf{C}_{i,\cdot}^{(v)}\|_2$, and the parameters for selection learning and consensus are $\{\lambda^{(v)}\}_{v=1}^{V}$, η, and τ. By solving Problem (3), we optimize a sparse reconstruction objective for each view to make sure the selection weights fit the distribution of the features. The final centroid co-regularization term further enforces selection weights to match all feature modalities.

C. Diversity Regularizer

As pointed out by Elhamifar *et al.* [37], sparse dictionary selection tends to keep the vertices of the convex hull spanned by the data to ensure a low reconstruction cost. Therefore, nearby data points at the vertices of convex hull are likely to be selected even though they are similar. To encourage a diverse selection of dissimilar representatives, we design the following diversity regularizer

$$\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\} = \sum_{i}^{n}\sum_{j}^{n} \mathbf{S}_{i,j}^{(v)}\mathbf{w}_i\mathbf{w}_j^{(v)}, \quad (4)$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace operator, and $\mathbf{S}^{(v)}$ is the similarity matrix of data points in the v-th view. In practice, we compute $\mathbf{S}_{i,j}^{(v)}$ using the cosine distance between data points, *i.e.*, $\mathbf{S}_{i,j}^{(v)} = \frac{\mathbf{X}_i^{(v)}\mathbf{X}_j^{(v)}}{\|\mathbf{X}_j^{(v)}\|_2\|\mathbf{X}_j^{(v)}\|_2}$. When the selected data points are dissimilar, (4) is small. Otherwise (4) is large . After adding the diversity regularizer (4), the objective function becomes (3):

$$\min_{\mathbf{C}^{(v)},\mathbf{w}} \sum_{v=1}^{V} \left\{ \underbrace{\frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{C}^{(v)}\|_{\mathrm{F}}^2 + \lambda_1^{(v)}\|\mathbf{C}^{(v)}\|_{1,2}}_{J_1} + \underbrace{\frac{1}{2}\eta\|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2}_{J_2} + \underbrace{\lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\}}_{J_3} \right\}, \quad (5)$$

D. Optimization

The objective function in (5) ($\mathcal{O}$ for short) can be solved by iterating between: (1) optimizing $\mathbf{C}^{(v)}$ by fixing $\mathbf{C}^{(u)}$ $(u \neq v)$ and $\mathbf{w}$, and (2) optimizing $\mathbf{w}$ by fixing $\mathbf{C}^{(v)}$ $(v = 1, 2, \ldots, V)$.

1) Optimize $\mathbf{C}^{(v)}$ by Fixing $\mathbf{C}^{(u)}$ $(u \neq v)$ and $\mathbf{w}$: We rewrite the objective function in (5) as $\mathcal{O} = \sum_{v=1}^{V}\mathcal{O}^{(v)}$, where

$$\mathcal{O}^{(v)} = \frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{C}^{(v)}\|_{\mathrm{F}}^2 + \lambda_1^{(v)}\|\mathbf{C}^{(v)}\|_{1,2} + \frac{1}{2}\eta\|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\}. \quad (6)$$

Therefore, $\min_{\mathbf{C}^{(v)}}\mathcal{O} \Leftrightarrow \min_{\mathbf{C}^{(v)}}\mathcal{O}^{(v)}$ when $\mathbf{C}^{(u)}$ $(u \neq v)$ and $\mathbf{w}$ are fixed. Moreover, $\mathcal{O}^{v}$ can be rewritten as

$$\begin{aligned}\mathcal{O}^{(v)} &= \frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{C}^{(v)}\|_{\mathrm{F}}^2 + \lambda_1^{(v)}\|\mathbf{C}^{(v)}\|_{1,2} \\ &\quad + \frac{1}{2}\eta(\|\mathbf{C}^{(v)}\|_{\mathrm{F}}^2 + \|\mathbf{w}\|_2^2 - 2\mathbf{w}^{(v)^{\mathrm{T}}}\mathbf{w}) \\ &\quad + \lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\} \\ &= \frac{1}{2}\mathrm{tr}\{\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)} - 2\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\mathbf{C}^{(v)} \\ &\quad + \mathbf{C}^{(v)^{\mathrm{T}}}\left(\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\right)\mathbf{C}^{(v)}\} \\ &\quad + (\lambda_1^{(v)}\mathbf{1} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)} + \frac{1}{2}\eta\|\mathbf{w}\|_2^2 \\ &\quad + \lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\}. \end{aligned} \quad (7)$$

Then, we let

$$f(\mathbf{C}^{(v)}) = \frac{1}{2}\mathrm{tr}\{\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)} - 2\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\mathbf{C}^{(v)} + \mathbf{C}^{(v)^{\mathrm{T}}}\left(\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\right)\mathbf{C}^{(v)}\} + \frac{1}{2}\eta\|\mathbf{w}\|_2^2, \quad (8)$$

and

$$g(\mathbf{C}^{(v)}) = (\lambda_1^{(v)}\mathbf{1} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)} + \lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\}, \quad (9)$$

which leads to

$$\mathcal{O}^{(v)} = f(\mathbf{C}^{(v)}) + g(\mathbf{C}^{(v)}). \quad (10)$$

Since $\mathcal{O}^{(v)}$ is decomposed into two convex functions, with f smooth and g non-smooth, the problem becomes iteratively solving the following using the proximal gradient method, FISTA [24]:

$$\mathrm{prox}_{\mathcal{R}}(\mathbf{Z}) = \operatorname*{arg\,min}_{\mathbf{C}^{(v)}\in\mathbb{R}^{n\times n}} \frac{1}{2}\left\|\mathbf{C}^{(v)} - \mathbf{Z}\right\|_{\mathrm{F}}^2 + \frac{1}{L^{(v)}}g(\mathbf{C}^{(v)}), \quad (11)$$

where

$$\begin{aligned}\mathbf{Z} &= \mathbf{C}^{(v)} - \frac{1}{L^{(v)}}\frac{\partial}{\partial\mathbf{C}^{(v)}}f(\mathbf{C}^{(v)}) \\ &= \mathbf{C}^{(v)} - \frac{1}{L^{(v)}}\left\{-\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)} + \left(\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\right)\mathbf{C}^{(v)}\right\}. \end{aligned} \quad (12)$$

Here $L^{(v)}$ is the smallest Lipschitz constant of $\frac{\partial}{\partial\mathbf{C}^{(v)}}f(\mathbf{C}^{(v)})$, which is the spectral radius (r(.)) of $\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}$, *i.e.*,

$$L^{(v)} = r(\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}) = \eta + r(\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}). \quad (13)$$

Algorithm 1 Multi-View Representative Selection via Centroid Coregulerization (5)

```
Require: features $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, parameters $\{\lambda_1^{(v)}\}_{v=1}^{V}$,$\{\lambda_2^{(v)}\}_{v=1}^{V}$, $\eta$
Ensure: selection matrices for each view $\{\mathbf{C}^{(v)}\}_{i=1}^{V}$, consensus weighting vector $\mathbf{w}$
// Initialization
$\mathbf{w} = \mathbf{0}$
for $v \in [1, V]$ do
    $L^{(v)} \leftarrow \eta + r\left(\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\right)$    (Eq. (13))
end for
// Iteratively solve the objective function (Eq. (5))
repeat
    // Optimize $\mathbf{C}^{(v)}$ by fixing $\mathbf{C}^{(u)}$ $(u \neq v)$ and $\mathbf{w}$
    for $v \in [1, V]$ do
        $\mathbf{C}^{(v)} \leftarrow \mathbf{0}$, $\mathbf{V} \leftarrow \mathbf{C}^{(v)}$, $t \leftarrow 1$
        repeat
            $\mathbf{Z} \leftarrow \mathbf{V} + \frac{1}{L^{(v)}}\left\{\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)} - \left(\eta\mathbf{I} + \mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\right)\mathbf{V}\right\}$    (Eq. (12))
            $\mathbf{U} \leftarrow \mathbf{C}^{(v)}$, $\mathbf{C}_{i,\cdot}^{(v)} \leftarrow \mathbf{Z}_{i,\cdot}\max\{(1 - \frac{\hat{\lambda}_{1i}^{(v)} + \hat{\lambda}_{2i}^{(v)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}, i \in [1, n]$    (Eq. (16))
            $q = t - 1$, $t \leftarrow (1 + \sqrt{1 + 4t^2})/2$
            $\mathbf{V} \leftarrow \mathbf{C}^{(v)} + q(\mathbf{C}^{(v)} - \mathbf{U})/t$
        until convergence
    end for
    // Optimize $\mathbf{w}$ while fixing $\mathbf{C}^{(v)}$
    $\mathbf{w} \leftarrow \max\{\frac{1}{V}\sum_{v=1}^{V}(\mathbf{I} - \frac{\lambda_2^{(v)}}{\eta}\mathbf{S}^{(v)})\mathbf{w}^{(v)}, 0\}$    (Eq. (18))
until convergence
```

Following the proximal decomposition [59], we can equivalently decompose problem (11) into n pairs of proximal operators, i.e., for $i = 1, 2, \ldots, n$,

$$\mathbf{C}_{i,\cdot}^{(v)} = \arg\min_{\mathbf{c}\in\mathbb{R}^n} \frac{1}{2}\left\|\mathbf{c} - \mathbf{Z}_{i,\cdot}\right\|_2^2 + \left(\hat{\lambda}_{1i}^{(v)} + \hat{\lambda}_{2i}^{(v)}\right)\|\mathbf{c}\|_2, \quad (14)$$

where

$$\begin{aligned}\hat{\lambda}_{1i}^{(v)} &= \frac{1}{L^{(v)}}\left(\lambda_1^{(v)} - \eta w_i\right), \\ \hat{\lambda}_{2i}^{(v)} &= \frac{1}{L^{(v)}}\lambda_2^{(v)}\mathbf{S}_{\cdot,i}^{(v)^{\mathrm{T}}}\mathbf{w} \end{aligned} \quad (15)$$

For problem (14), after applying soft-thresholding [60], we have,

$$\mathbf{C}_{i,\cdot}^{(v)} = \mathbf{Z}_{i,\cdot}\max\{(1 - \frac{\hat{\lambda}_{1i}^{(v)} + \hat{\lambda}_{2i}^{(v)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}. \quad (16)$$

2) Optimize **w** *While Fixing* $\mathbf{C}^{(v)}$*:* Denote the first term in the objective function (5) as J_1, the second term as J_2 and the third term as J_3, then $\min_{\mathbf{w}} \mathcal{O} \Leftrightarrow \min_{\mathbf{w}} \sum_{v=1}^{V} J_2 + J_3$ when fixing $\mathbf{C}^{(v)}$, and

$$J_2 + J_3 = \frac{1}{2}\eta\|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \lambda_2^{(v)}\,\mathrm{tr}\{\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\}. \quad (17)$$

By applying soft-thresholding, we obtain

$$\mathbf{w} = \max\{\frac{1}{V}\sum_{v=1}^{V}(\mathbf{I} - \frac{\lambda_2^{(v)}}{\eta}\mathbf{S}^{(v)})\mathbf{w}^{(v)}, 0\} \quad (18)$$

We show the optimization procedure in Algorithm 1, where we adopt an alternating minimizing strategy and integrate decomposed soft-thresholding into the proximal gradient iteration.

E. Parameter Setting

1) Dictionary Selection Parameter $\lambda_1^{(v)}$ *in the* v*-th View:* We introduce this parameter to control the sparsity of dictionary selection in each single view. As indicated by the thresholding of **Z** in (16), when $\lambda_1^{(v)}$ is large enough, we have $\mathbf{C}^{(v)} = \mathbf{0}$, which results in an empty selection. To avoid such an empty selection in the initialization, we let $\lambda_1^{(v)} \le \lambda_{1\max}^{(v)}$ and solve $\lambda_{1\max}^{(v)}$ by substituting $\mathbf{C}^{(v)} = \mathbf{0}$ into (16) as follows:

$$\lambda_{1\max}^{(v)} = L^{(v)}\max_{1\le i\le n}\|\mathbf{Z}_{i,\cdot}\|_2. \quad (19)$$

It is worth noting that in Algorithm 1, we initialize $\mathbf{C}^{(v)}$ by a zero matrix, and **w** by a zero vector. Then according to (12), after the first iteration, we have

$$\mathbf{Z} = \frac{1}{L^{(v)}}\mathbf{X}^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}. \quad (20)$$

Therefore,

$$\lambda_{1\max}^{(v)} = \max_{1\le i\le n}\|\mathbf{x}_i^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\|_2. \quad (21)$$

In our experiments, we let $\lambda_1^{(v)} = \frac{\lambda_{1\max}^{(v)}}{\alpha_{\lambda_1}}$ and tune the hyper-parameter α_{λ_1}. Given $\lambda_1^{(v)}$, a smaller α_{λ_1} indicates a larger $\lambda_1^{(v)}$, which implies a sparser selection.

2) Centroid Co-Regularization Parameter η*:* As shown in (5), this parameter trades-off the first dictionary selection term J_1 and the second centroid co-regularization term J_2 (17). When $\eta \to 0$, we will immediately reach the consensus by feeding individual dictionary selection results into (18). When $\eta \to +\infty$, minimizing (5) will lead to a zero J_2, thus making $\mathbf{w}^{(v)}(v \in [1, V])$ and **w** to be **0**. As a result, we cannot select anything from the data. Furthermore, we can see from (15), η balances the contributions of $\lambda_1^{(v)}$ and **w** to the dictionary selection of the v-th view in (16). For ease of tuning η, we let

$$\eta = \frac{\min_{v\in[1,V]}\{\lambda_1^{(v)}\}}{\alpha_\eta}, \quad (22)$$

3) Diversity Regularizer Parameter $\lambda_2^{(v)}$*:* We introduce this parameter to control the diversity of dictionary selection in each view. As shown in (16) and (18), when $\lambda_2^{(v)} \to +\infty$, minimizing (5) will lead to $\mathbf{C}^{(v)} = \mathbf{0}$ and $\mathbf{w} = \mathbf{0}$. Note that in (18), $\lambda_2^{(v)}$ balances the contribution of η to the consensus weighting vector **w**. Thus, we let $\lambda_2^{(v)} = \frac{\eta}{\alpha_{\lambda_2}}$ and tune the hyper-parameter α_{λ_2}.

F. Extension to Incorporating Priors

As selection priors such as canonical viewpoints [4], visual co-occurrence [21] and objectness scores [3] have been shown to improve results, we also extend our method to a weighted multi-view representative selection to capture view-specific selection priors. Formally, we propose the new objective as follows:

$$\begin{aligned}\min_{\mathbf{C}^{(v)},\mathbf{w}} \sum_{v=1}^{V} \Bigg\{ &\frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{C}^{(v)}\|_F^2 + \lambda_1^{(v)}\sum_{i=1}^{n}\rho_i^{(v)}w_i^{(v)} \\ &+ \frac{1}{2}\eta\|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \lambda_2^{(v)}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_j^{(v)}\mathbf{S}_{i,j}^{(v)}\mathbf{w}_i\mathbf{w}_j^{(v)}\Bigg\} \end{aligned} \quad (23)$$

TABLE I

AVERAGE RECALL OF SYNTHETIC DATA ON 2 VIEWS. IN EACH VIEW, DATA POINTS ARE PROJECTED TO 3 CLUSTERS, THE FEATURE DIMENSION OF EACH VIEW IS INDICATED BY D: $[d_1, d_2]$. RESULTS ARE AVERAGED OVER 100 TRIALS

D	Single View Selection				Concatenated View Selection				Multi-View Selection		
	KM	SMRS	SDS	LLR-SDS	KM	SMRS	SDS	LLR-SDS	AASC	CMSC	MSDS-CC
[2,2]	0.87	0.70	0.74	0.71	0.79	0.78	0.71	0.69	0.78	0.78	**0.90**
[5,5]	0.84	0.85	0.74	0.73	0.81	0.78	0.72	0.74	0.83	0.77	**0.94**
[2,10]	0.86	0.84	0.73	0.73	0.79	0.77	0.72	0.71	0.82	0.76	**0.92**
[10,10]	0.85	0.83	0.74	0.72	0.82	0.77	0.71	0.73	0.83	0.76	**0.87**
[5,100]	0.85	0.83	0.73	0.72	0.72	0.75	0.62	0.70	0.82	0.73	**0.89**
[100,100]	0.84	0.82	0.72	0.73	0.82	0.75	0.71	0.72	0.81	0.73	**0.94**

where prior $\rho_i^{(v)}$ is the selection cost for the i-th sample according to v-th view of features, where the smaller the $\rho_i^{(v)}$, the more likely it will be selected as the representative. In this way, those shots of smaller cost in multiple views are more likely to be selected for summaries.

The optimization of (23) follows a similar procedure as shown in Subsections III-D1 and III-D2. We only need to update the non-smooth term $g(\mathbf{C}^{(v)})$ (shown in (9)) to suit the new objective function in (23) by

$$g(\mathbf{C}^{(v)}) = (\Lambda_1^{(v)} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)} + \mathrm{tr}\{\Lambda_2^{(v)}\mathbf{S}^{(v)^{\mathrm{T}}}\mathbf{w}\mathbf{w}^{(v)^{\mathrm{T}}}\} \tag{24}$$

where $\Lambda_1^{(v)}, \Lambda_2^{(v)} \in \mathbb{R}^n$, and its i-th element is $\lambda_1^{(v)}\rho_i^{(v)}$ and $\lambda_2^{(v)}\rho_i^{(v)}$ respectively. Therefore, the solution to $\mathbf{C}^{(v)}$ is still given by (16), but with a different $\hat{\lambda}_{1i}^{(v)}$, $\hat{\lambda}_{2i}^{(v)}$ compared to (15):

$$\begin{aligned}\hat{\lambda}_{1i}^{(v)} &= \frac{1}{L^{(v)}}\left(\lambda_1^{(v)}\rho_i^{(v)} - \eta w_i\right),\\ \hat{\lambda}_{2i}^{(v)} &= \frac{1}{L^{(v)}}\lambda_2^{(v)}\rho_i^{(v)}\mathbf{S}_{\cdot,i}^{(v)^{\mathrm{T}}}\mathbf{w}\end{aligned} \tag{25}$$

For equal prior selection costs with $\rho_i^{(v)} = 1$, (25) and (15) become the same. Problem (23) will perfectly degenerate into Problem (5).

To facilitate setting parameters $\{\lambda_1^{(v)}\}_{v=1}^V$, $\{\lambda_2^{(v)}\}_{v=1}^V$ and η, we also refine the calculation of $\lambda_{1\max}^{(v)}$ in Subsection III-E when optimizing (23) with the addition of priors. According to (25) and (16), we calculate $\lambda_{1\max}^{(v)}$ by

$$\begin{aligned}\lambda_{1\max}^{(v)} &= L^{(v)}\max_{1\le i\le n}\frac{1}{\rho_i^{(v)}}\|\mathbf{Z}_{i,\cdot}\|_2\\ &= L^{(v)}\max_{1\le i\le n}\frac{1}{\rho_i^{(v)}}\|\mathbf{x}_i^{(v)^{\mathrm{T}}}\mathbf{X}^{(v)}\|_2.\end{aligned} \tag{26}$$

IV. EXPERIMENTS

A. Baselines

We refer to the proposed method as Multi-view Sparse Dictionary Selection with Centroid Co-regularization (MSDS-CC), and compare with the below baselines.

1) Clustering-Based: baselines include the standard K-medoids [48] and two multi-view spectral clustering methods: Affinity aggregation spectral clustering (AASC) [16] and Co-regularized multi-view spectral clustering (CMSC) [15]. We use the centroid-based co-regularization for CMSC.

2) Subspace Learning Based: baselines include the state-of-the-art Sparse Modeling Representative Selection (SMRS) [37], Sparse Dictionary Selection (SDS) [41] and Locally Linear Reconstruction induced Sparse Dictionary Selection (LLR-SDS) [3].

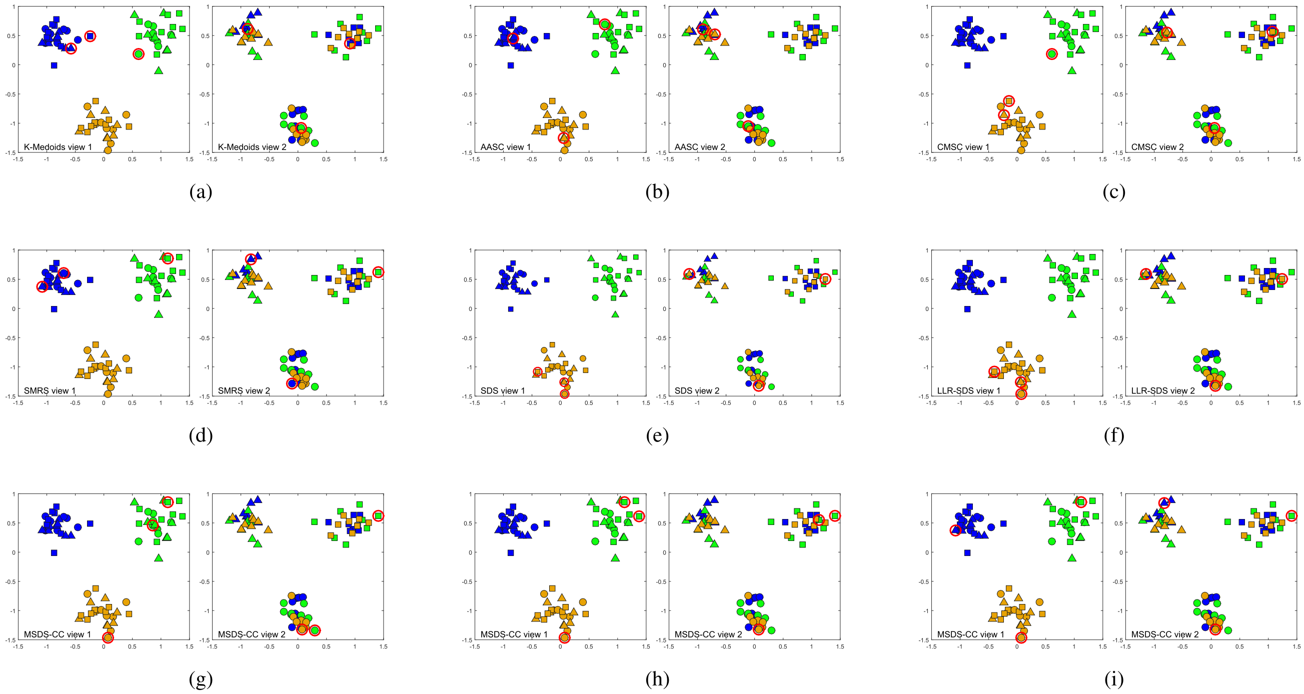
For the two multi-view clustering methods, AASC and CMSC, we adapt them for multi-view representative selection by selecting representatives from the embedding feature space, where representatives are the closest points to the cluster centers in that space. For the other baselines, feature concatenation is performed before representative selection.

In our experiments, we use the authors' implementation of each method, except for the K-medoids, for which we used the MATLAB implementation. α for SMRS and SDS, and α_1 for LLR-SDS are tested on a range of $\{5, 8, 10, 20, 30\}$. In addition, for LLR-SDS, we use the default $k = 3$ to construct the locality prior matrix and tune α_2 in a range of $\{1.5, 1, 0.5, 0\}$. The default $\lambda = 0.5$ is used for CMSC. For our proposed MSDS-CC, we tune the hyper-parameters $\alpha_{\lambda_1} \in \{3, 5, 10, 20, 30\}$, $\alpha_\eta \in \{0.01, 0.1, 1, 10\}$ and $\alpha_{\lambda_2} \in \{0.1, 0.2, 0.5, 1, 10\}$. And we report the best result for each method.

B. Experiments on Synthetic Data

We first evaluate the effectiveness of our proposed method on synthetic data in multiple views while varying the number of clusters and data dimensions (Table I). For simplicity, we consider the representative selection on two views and randomly generate $(d_1 + d_2)$-dimensional data points, where d_i is the dimension of the ambient space of the v-th view. In each view, data points are uniformly projected to n clusters whose centers are drawn uniformly from a unit-norm ball. Each data point is corrupted with independent Gaussian noise of standard deviation $\varepsilon = 0.1$. Following [3], we evaluate the performance of the top n representatives by the average recall and the results are averaged over 100 trials.

We have a few observations from Table I. First, when the feature dimensions in the two views differ greatly (*i.e.*, $D = [5, 100]$), the performance of representative selection from the concatenated view deteriorates (middle section in Table I). This can be attributed to the *shadowing* effects caused by the higher dimensional feature type. In fact, when $D = [5, 100]$, results from the concatenated views are worse than those from the single view (left section in Table I) regardless of the baseline method chosen (*i.e.*, K-medoids, SMRS, SDS, LLR-SDS). Second, multi-view clustering baselines



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 2. Simulation results of data in two views. Data points are projected to three color (blue, yellow and green) and shape (◯, △ and □) clusters respectively. Representatives (red circles ◯) are found by (a)K-medoids (b) AASC (c) CMSC (d) SMRS (e) SDS (f) LLR-SDS and (i) Ours. Results without (g) centroid co-regularizer or (h) diversity regularizer are also shown.

(AASC and CMSC in Table I) are not as sensitive to the discrepancies in feature dimensions as other baselines. This is understandable, as they first learn a unified data representation before clustering and representative selection. Consequently, neither view would dominate. However, both AASC and CMSC perform worse than the proposed MSDS-CC. It can be attributed to the difficulty for them to handle the disagreement in the feature distributions in different views, *e.g.*, when data points belong to different groups in different views. This situation may arise in our simulation as data points in each view are independently generated. Overall, the proposed MSDS-CC outperforms all baselines regardless of the feature dimensions in each view.

Fig. 2 visualizes a case of 4D data points projected to two 2D views. As can be seen, in view 1, points are clustered by color (blue, yellow and green), while in view 2, points are clustered by shape (◯, △ and □), respectively. In the ideal case, we should be able to capture all the 6 properties of data points (*i.e.*, 3 colors and 3 shapes) by only 3 representatives. It is shown that the top 3 representatives of AASC (△, ◯ and △) miss the square shape □, and CMSC (◯, △ and □) miss the blue color. This is because AASC and CMSC assume that different views have the same underlying clustering of the data. Therefore, they can not well handle the presence of view disagreement. On the other hand, the representatives of K-medoids (□, △ and ◯), SMRS (◯, △ and □) and LLR-SDS (◯, □ and △) fail to capture all properties likely because the concatenated view breaks the underlying distribution of individual views. In contrast, representatives of our proposed MSDS-CC (△, □ and ◯) capture all properties, since our method preserves the distributions of individual views and produce a consensus selection. We also visualize the results after removing the centroid co-regularization term J_2 (Fig. 2 (g)) or the diversity regularizer J_3 (Fig. 2 (h)) in (5). It shows that both terms are necessary to ensure a consistent and diverse selection in the two views. In fact, the proposed method MSDS-CC outperforms all baselines in all tested cases (Table I).

C. Proof of Concept

1) Subject and Pose Selection: We further validate the effectiveness of the proposed multi-view representative selection on the EPFL stereo face dataset [61]. The dataset consists of 100 subjects, each recorded from 8 different viewpoints by a pair of calibrated stereo cameras. We randomly select n subjects and n poses to form a dataset of $n \times n$ images, where $n \in \{2, 4, 6, 8, 10, 12, 14, 16\}$. Our goal is to capture all the subjects and poses by selecting a few representative faces. For example, when we randomly select 4 subjects and 4 poses to form a dataset of 16 images, in the ideal case, as few as 4 face images should capture all the 4 subjects and the 4 poses. Similar to [3], we evaluate the performance of representative selection by the average recall of the subjects and poses.

To capture the face appearance and pose, for each face image we extract both the 4096D CNN feature extracted from the fc7 layer of the pre-trained model VGG-Face [62] and the 136D facial landmark/fiducial points extracted from dlib (68 face landmarks with (x,y) coordinates). As can be seen in Fig. 3, in addition to the highly different feature dimensions, the distribution of the two types of features are drastically different as well.

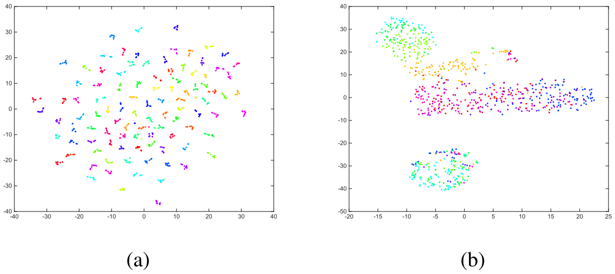


Fig. 3. Distribution of the (a) appearance features and (b) pose features on EPFL stereo face dataset (visualized by t-SNE [63]). Features of the same color indicate the same subject in (a) and the same pose in (b).

TABLE II

AVERAGE RECALL ON EPFL STEREO FACE DATASET. RANDOMLY SELECT N SUBJECTS AND N POSES TO FORM A DATASET OF $n \times n$ IMAGES. THE RESULTS ARE AVERAGED OVER 100 RUNS

n	KM	AASC	CMSC	SMRS	SDS	LLR-SDS	MSDS-CC
2	0.858	0.825	**0.918**	0.883	0.818	0.773	0.880
4	0.739	0.763	0.774	0.789	0.723	0.648	**0.820**
6	0.725	0.768	0.730	0.780	0.661	0.603	**0.801**
8	0.708	0.766	0.717	0.762	0.609	0.591	**0.781**
10	0.694	0.735	0.698	0.760	0.557	0.606	**0.782**
12	0.678	0.708	0.677	0.753	0.531	0.612	**0.762**
14	0.674	0.708	0.690	0.746	0.501	0.620	**0.765**
16	0.662	0.705	0.668	0.736	0.489	0.618	**0.743**

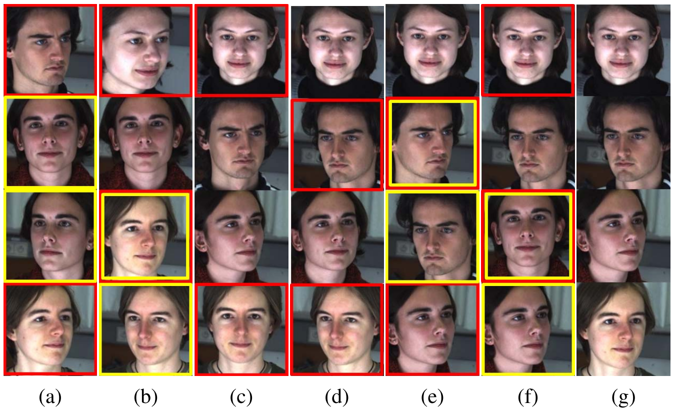


Fig. 4. EPFL stereo face: visualization of the first 4 representatives selected by each method (column-wise). Duplicate subjects or poses in each column are highlighted by bounding boxes of the same color. (a) K-medoids, (b) AASC, (e) SDS and (f) LLR-SDS capture 3 subjects and 3 poses (Average Recall@4 = 0.75). (c) CMSC and (d) SMRS capture all 4 subjects but only 3 poses (Average Recall@4 = 0.875). In comparison, (g) our proposed MSDS-CC captures all the 4 subjects and 4 poses (Average Recall@4 = 1).

Table. II compares the performance of our approach in terms of average recall on the EPFL stereo face dataset. For each n, we tune the parameters and average the results over 100 trails. As shown, except for the case of 2 subjects and 2 poses, where CMCS and SMRS outperform the proposed MSDS-CC, ours achieves the highest recall in all other cases.

Fig. 4 shows an example of qualitative results from different approaches when selecting 4 representative faces with corresponding Average Recall@4. Our approach captures all the 4 subjects and 4 poses with the 4 selected faces, outperforming the other methods with an Aver age Recall@4 = 1. In comparison, K-medoids, AASC, SDS and LLR-SDS capture 3 subjects but only 3 poses (Average Recall@4 = 0.75); both CMSC and SMRS capture 4 subjects but 3 poses (Average Recall@4 = 0.875).

TABLE III

COMPARISON OF RECALL WITH VARYING NUMBER OF REPRESENTATIVES ON THE UCF SPORTS DATASET

	Action Recall			Scene Recall			Average Recall		
No.	5	8	10	5	8	10	5	8	10
K-medoids	0.5	0.7	0.7	0.5	0.7	0.7	**0.50**	0.70	0.70
AASC	0.4	0.7	0.8	0.5	0.7	0.7	0.45	0.70	0.80
CMSC	0.5	0.8	0.7	0.5	0.7	0.8	**0.50**	0.75	0.75
SMRS	0.5	0.8	0.8	0.4	0.7	0.8	0.45	0.70	0.80
SDS	0.5	0.6	0.7	0.5	0.5	0.5	**0.50**	0.55	0.60
LLR-SDS	0.5	0.7	0.8	0.5	0.8	0.8	**0.50**	0.75	0.80
MSDS-CC	0.5	0.8	0.9	0.5	0.8	0.8	**0.50**	**0.80**	**0.85**

2) Action and Scene Selection: Next we evaluate the effectiveness of the proposed MSDS-CC on UCF Sports dataset [64]. The UCF Sports dataset consists of 150 videos of 10 actions in different scenes. Our goal is to capture the variety of actions and scenes with as few representatives as possible.

We manually label the scenes with a list of semantic categories (as shown in Fig. 5) and remove scene categories with fewer than three videos. This leads to a subset of 127 videos with 10 actions and 10 scenes. Fig 6 visualizes the correlation between the actions and scenes. It is shown that some human actions (*e.g.*, Running) may take place in different scenes (*e.g.*, Golf Course, Soccer Field, Road). On the other hand, the same scene may have different actions, *e.g.*, Golf Swing, Walking and Running can all happen at a Golf Course.

For the representation of actions, we first extract the improved dense trajectory features [65]. Then we perform PCA on the resulting HOG, HOF and MBH descriptors to reduce their dimensions to 1/4 of the original sizes, which results in 24D, 27D and 48D vectors, respectively. After concatenating them into 99D descriptors, we encode this concatenated descriptors by the Fisher Vector with a Gaussian mixture model (GMM) of 128 Gaussians, producing 25, 344 dimensional features [38]. For the scene representation, we use pre-trained VGG model on scene recognition [66] and take the 4096D features from the output of the fc7 layer. The scene representation of each video is obtained by averaging the CNN features of all frames. Fig. 7 plots the distributions of the two feature types.

Similar to [3], we evaluate the performance by the average recall of the actions and scenes. As shown in Table III, MSDS-CC outperforms all baselines on the UCF sports dataset. Specifically, MSDS-CC captures diverse actions and scenes with no duplicates with 8 representatives, and only misses one action and two scenes with 10 representatives.

D. Video Summarization

1) Datasets: To evaluate the effectiveness of our approach on video summarization, we experiment on two benchmark datasets, TVSum [22] and SumMe [10]. TVSum contains 50 videos within 10 categories representing various genres (*e.g.*, news, how-tos, documentaries and egocentric). It also provides shot-level importance scores obtained from user

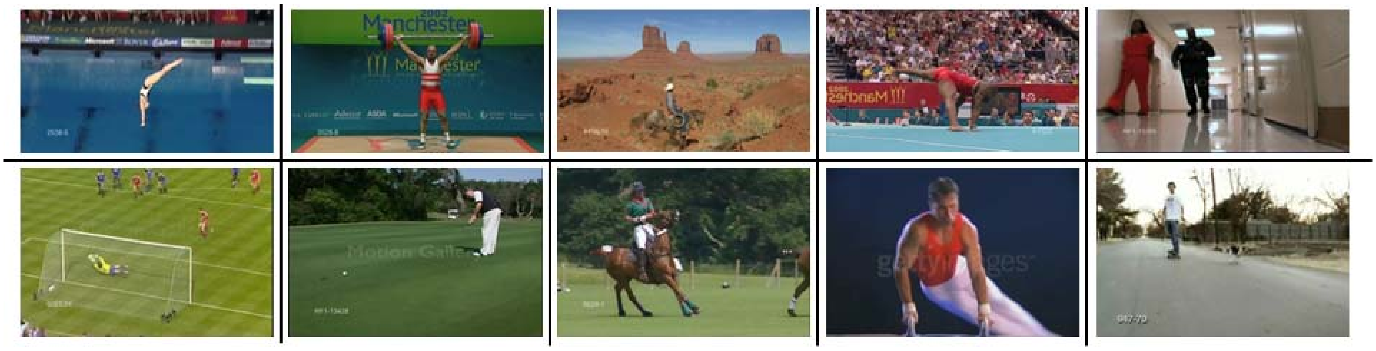


Fig. 5. Samples from the UCF Sports scene categories (From left to right, top to bottom): Diving Hall, Stadium, Field Wild, Arena, Corridor, Soccer Stadium, Golf Course, Racecourse, None and Road.

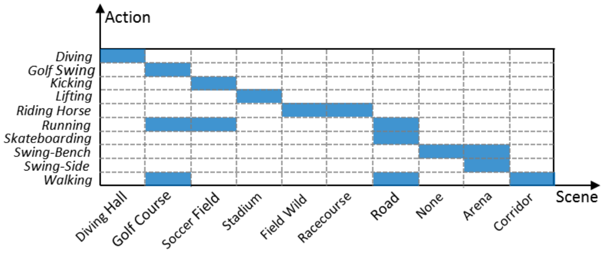
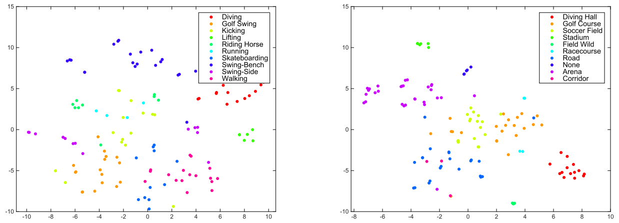


Fig. 6. Action/Scene correlation. Blue blocks indicate the co-existence of a specific action and scene in some videos. It is shown that human actions may take place in different scenes and vice versa.



(a)

(b)

Fig. 7. Distribution of the (a) action features and (b) scene features on UCF Sports dataset (visualized by t-SNE [63]). Features of the same color indicate the same action in (a) and the same scene in (b).

annotations. SumMe consists of 25 short user videos covering a variety of events. Each video has multiple user-summaries in the form of key shots. The average duration of ground-truth is 13.1% of the video length. In our experiments, we summarize videos into key shots and evaluate the performance accordingly to facilitate comparisons with prior works on these two datasets [10], [11], [22], [23], [31].

2) Settings: To capture both low-level visual information and high-level semantics, which complement each other well, we extract both frame-wise GIST features [67] and C3D features [68] with a step size of 16 frames. The former has been shown to be capable of capturing the gist of a scene by summarizing the gradient information at different scales and orientations. The latter is a deep feature learned by a 3D convolutional network, which has been shown to be effective in exploiting the semantic information in videos [68]. The GIST descriptors are computed with 32 Gabor filters at 4 scales, 8 orientations and 4 × 4 blocks, resulting in 512D features. The C3D features (4096D) are extracted from the fc6 layer of the pre-trained model.

Since neither of the datasets provides ground-truth temporal segmentations, we first temporally segment videos into disjoint intervals by Kernel Temporal Segmentation(KTS) method [30]. The average length of these segments (*i.e.*, shots) are around 5 seconds. Both GIST and C3D features are averaged within each shot to produce the shot features.

To generate a video summary of length l, we follow [10], [22] to solve the knapsack problem:

$$\max \sum_{i=1}^{s} u_i \phi_i \quad \text{s.t.} \quad \sum_{i=1}^{s} u_i n_i \le l, \;\; u_i \in \{0, 1\} \tag{27}$$

where s is the total number of shots, ϕ_i is the importance score of the i-th shot, and n_i is the length of the i-th shot. The summary is produced by concatenating shots with $u_i = 1$ chronologically. As in [10], [22], and [31], we set the length budget l to be 15% in duration of the original video for both datasets.

3) Implementation Details: Similar to the protocols in [22], for the subspace learning based baselines (i.e., SMRS [37], LLR-SDS [3]) and our proposed MSDS-CC, we predicts the importance score ϕ_i of each shot directly. The shot-level scores ϕ_i in (27) is calculated by the consensus weighting $\mathbf{w}$ in (5).

We follow [46] to evaluate clustering based baselines (*i.e.*, K-medoids, AASC [16] and CMSC [15]). As in [46], clustering is performed on the shots with the number of clusters set to 20. Then the summary is generated by selecting the shots that are closest to the centroids of top largest clusters, with a length budget l.

4) Evaluation: Following prior work [10], [11], [22], [23], we evaluate the generated summaries by the F-score (F). Pairwise precision (P) and recall (R) are computed between the resulting summary and each human-created summary according to the temporal overlap. Then F-score is computed as $F = \frac{P \times R}{0.5(P+R)}$. As in [11], we follow [10], [22] to compute the metrics when there are multiple human-created summaries of a video.

5) Results: We first validate the effectiveness of multiview representative selection for video summarization on both datasets, in comparison with single-view selection. Summarization by a single feature modality (*i.e.*, GIST or C3D) is

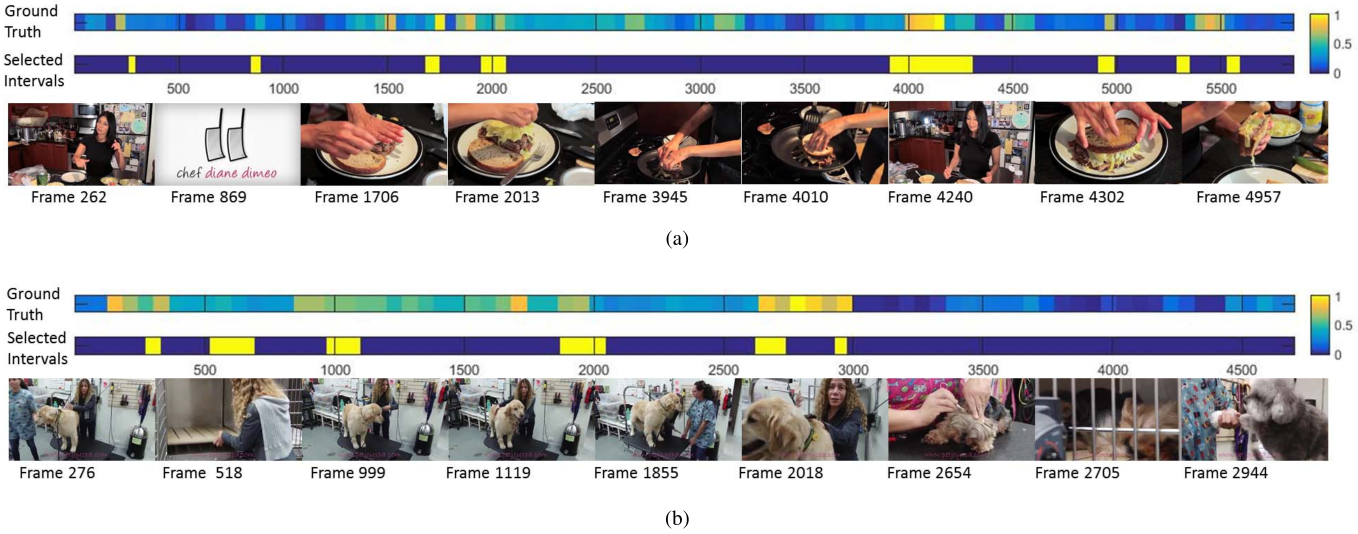


Fig. 8. Sample results on TVSum. For each video, the first colorbar shows the Ground Truth (*i.e.*, user annotated importance scores); the second colorbar shows our summarization results, where yellow intervals indicate shots selected by our proposed MSDS-CC. The bottom row shows sampled frames from selected shots. (a) Making Sandwich (MS). The averaged F-score is 53.4. (b) Grooming an Animal (GA). The averaged F-score is 60.1.

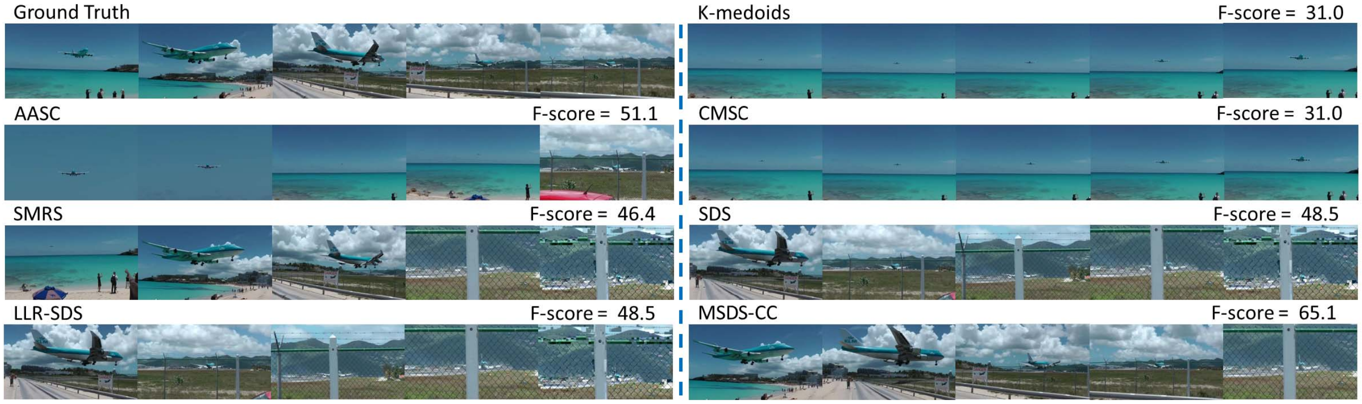


Fig. 9. Visual comparisons of the summaries produced by different methods on SumMe. The summary from the proposed MSDS-CC method better matches the ground truth and is able to capture shots that are representative both in motion and appearance with high diversity.

TABLE IV
PERFORMANCE COMPARISON OF SINGLE AND MULTIPLE FEATURES

Features	TVSum	SumMe
GIST	48.7	40.8
C3D	53.2	40.2
GIST + C3D	**53.8**	**41.5**

performed by solving $\min_{\mathbf{C}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{C}\|_{\mathrm{F}}^{2}+\lambda\|\mathbf{C}\|_{1,2}$ [41] and the multi-view selection (*i.e.*, GIST + C3D) is solved by the proposed MSDS-CC. The results are presented in Table IV and shows that our multi-view approach achieves better performance than either view alone on both datasets.

Next, we compare the proposed MSDS-CC with the baselines on TVSum and SumMe (Table V). Our method outperforms all clustering-based and subspace-based baselines on both datasets. For comparison, we also report results of other summarization methods from published prior works [10], [11], [11], [22], [31]. It is shown that the proposed MSDS-CC performs competitively without relying on external images [22] or learning from user annotated summaries [10], [11], [31].

Specifically, on TVSum, our approach performs better than the TVSum benchmark results [22], which uses additional title-based image search results to help identify canonical visual concepts shared between the video and images. Although dppLSTM (Canonical) [11] performs slightly better than ours, it uses the user annotations on 80% videos from TVSum for training and the remaining 20% for testing. Fig. 8 shows sample visual results of our method in comparison with the ground truth annotations.

On SumMe, our proposed MSDS-CC outperforms the SumMe benchmark results [10], Submodular [31] and dppLSTM (Canonical) [11], and is comparable to Summary Transfer [23], all of which use additional videos for training. Fig. 9 shows visual comparisons of the video summarization results from different methods.

Additional results from the proposed MSDS-CC on the two datasets are shown in Fig. 10.

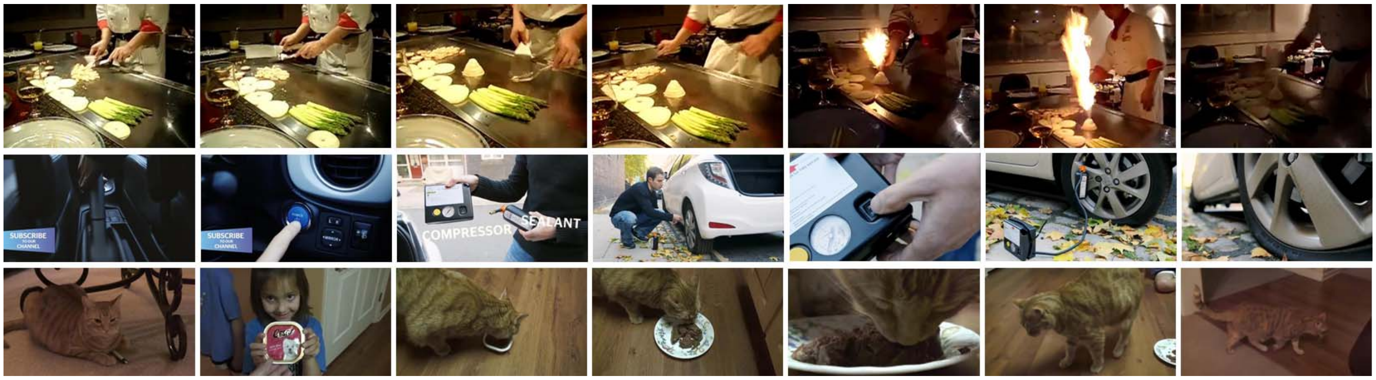


Fig. 10. Additional summarization results on SumMe (1^{st} row) and TVSum (2^{nd} and 3^{rd} rows). The proposed method is able to capture representative visual elements that capture the important appearance and motion in these videos.

TABLE V

PERFORMANCE (F-SCORE) OF VARIOUS VIDEO SUMMARIZATION METHODS ON TVSUM AND SUMME. THE TOP SECTION LISTS THE PERFORMANCE OF CLUSTERING-BASED AND SUBSPACE BASED METHODS. THE BOTTOM SECTION LISTS RESULTS FROM PUBLISHED WORKS. †DENOTES METHODS THAT USE ADDITIONAL WEB IMAGES AND ‡ DENOTES METHODS THAT USE ANNOTATED VIDEO SUMMARIES FOR TRAINING. DASHES DENOTE UNAVAILABLE DATASET-METHOD COMBINATIONS

Methods	TVSum	SumMe
K-medoids [48]	29.1	31.5
AASC [16]	28.1	35.2
CMSC [15]	27.1	33.7
SMRS [37]	36.6	35.0
SDS [41]	53.2	40.2
LLR-SDS [3]	53.2	40.0
MSDS-CC (ours)	**53.8**	**41.5**
TVSum† [22]	50.0	-
SumMe‡ [10]	-	39.4
Submodular‡ [31]	-	39.7
Summary Transfer‡ [23]	-	40.9
dppLSTM(Canonical)‡ [11]	54.7	38.6

E. Discussions

Comparing to the baselines that use concatenated features (*i.e.*, K-medoids, SMRS and LLR-SDS), the proposed MSDS-CC method is better at (1) preserving the underlying data distributions of individual views and (2) managing unbalanced feature lengths. Therefore, it performs better when different feature modalities have disparate distributions (*e.g.*, Sec. IV-B, Sec. IV-C1, Sec. IV-C2, Sec. IV-D), and/or have highly different dimensions (*e.g.*, Sec. IV-C1, Sec. IV-C2).

Comparing with multi-view spectral clustering baselines (AASC and CMSC), MSDS-CC can better handle the discrepancies across views, *e.g.*, when data points belong to different groups in different views (Sec. IV-B). This is because our method encourages a consensus selection based on the view-specific selection weights, which are optimized view-wise thus respect view-specific distributions. Contrarily, AASC and CMSC look for a consensus clustering via the embedding feature space, thus may not handle the disagreement in different views well.

Similar to SMRS and SDS, a drawback of the proposed method is that it may be sensitive to the outliers at the border of the convex hull. A possible solution is to perform outlier removal (similar to [37]) after obtaining the view-specific selection matrix $\mathbf{C}^{(v)}$ in each iteration.

It is also worth noting that using low-level features alone may produce sub-optimal results for the task of video summarization. For instance, we have examined the performance of the proposed MSDS-CC when using GIST and HOF features on TVSum and SumMe. Although the proposed multi-view selection based summarization using two low-level features (*i.e.*, GIST + HOF) performs better than each single feature alone (GIST or HOF), it performs worse than that achieved by GIST + C3D (low-level + high-level features). Empirically, we have found that a combination of low-level and high-level features perform well for the task of video summarization, as they can complement each other well. As shown in Table IV, the low-level feature GIST performs better than the high-level feature C3D on the SumMe dataset, while the opposite is observed on the TVSum dataset. Using the proposed MSDS-CC approach, we are able to produce a consensus selection of visual elements across multiple feature modalities, producing better summaries than using each single feature alone on both datasets.

V. CONCLUSION

Video summaries can be produced by selecting representative visual elements (*e.g.*, objects, frames, shots) from a video. However, as the representativeness depends on the visual representation (*i.e.*, features), the question becomes how to derive a consensus selection across multiple views (*i.e.*, feature modalities). To this end, we propose to formulate the video summarization problem as the multi-view sparse dictionary selection with centroid co-regularization (MSDS-CC), which optimizes the selection in each individual view while regularizing the view-specific selections towards a consensus selection (*i.e.*, centroid co-regularization). Experimental results on synthetic and challenging benchmark datasets demonstrate the effectiveness of the proposed approach for video summarization and its applicability to other applications.

REFERENCES

[1] D. Liu, G. Hua, and T. Chen, “A hierarchical visual model for video object summarization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010.

[2] C. Lu, R. Liao, and J. Jia, "Personal object discovery in first-person videos," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5789–5799, Dec. 2015.

[3] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: Video summarization by representative object proposal selection," in *Proc. CVPR*, Jun. 2016, pp. 1039–1048.

[4] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using Web-image priors," in *Proc. CVPR*, 2013, pp. 2698–2705.

[5] G. Kim, L. Sigal, and E. P. Xing, "Joint summarization of large-scale collections of Web images and videos for storyline reconstruction," in *Proc. CVPR*, 2014, pp. 4225–4232.

[6] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. CVPR*, Jun. 2012, pp. 1346–1353.

[7] C. Dang and H. Radha, "RPCA-KFE: Key frame extraction for video using robust principal component analysis," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3742–3753, Nov. 2015.

[8] O. Morére, H. Goh, A. Veillard, V. Chandrasekhar, and J. Lin, "Co-regularized deep representations for video summarization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3165–3169.

[9] D.-J. Jeong, H. J. Yoo, and N. I. Cho, "A static video summarization method based on the sparse coding of features and representativeness of frames," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1–14, 2016.

[10] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. ECCV*, 2014, pp. 505–520.

[11] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. ECCV*, 2016, pp. 766–782.

[12] M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Salient montages from unconstrained videos," in *Proc. ECCV*, 2014, pp. 472–488.

[13] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," in *Proc. ICCV*, Oct. 2007, pp. 1–8.

[14] X. Cai, F. Nie, H. Huang, and F. Kamangar, "Heterogeneous image feature integration via multi-modal spectral clustering," in *Proc. CVPR*, Jun. 2011, pp. 1977–1984.

[15] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. NIPS*, 2011, pp. 1413–1421.

[16] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Affinity aggregation for spectral clustering," in *Proc. IEEE CVPR*, Jun. 2012, pp. 773–780.

[17] H. Wang, C. Weng, and J. Yuan, "Multi-feature spectral clustering with minimax optimization," in *Proc. CVPR*, 2014, pp. 4106–4113.

[18] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang, "Diversity-induced multi-view subspace clustering," in *Proc. CVPR*, 2015, pp. 586–594.

[19] H. Gao, F. Nie, X. Li, and H. Huang, "Multi-view subspace clustering," in *Proc. ICCV*, 2015, pp. 4238–4246.

[20] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multi-view representative selection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1189–1198.

[21] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proc. CVPR*, 2015, pp. 3584–3592.

[22] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing Web videos using titles," in *Proc. CVPR*, Jun. 2015, pp. 5179–5187.

[23] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. CVPR*, Jun. 2016, pp. 1059–1067.

[24] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[25] F. Chen and C. De Vleeschouwer, "Formulating team-sport video summarization as a resource allocation problem," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 193–205, Feb. 2011.

[26] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.

[27] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *Int. J. Comput. Vis.*, vol. 114, no. 1, pp. 38–55, 2015.

[28] G. Zhu *et al.*, "Trajectory based event tactics analysis in broadcast sports video," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 58–67.

[29] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. CVPR*, 2013, pp. 2714–2721.

[30] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. ECCV*, 2014, pp. 540–555.

[31] M. Gygli, H. Grabne, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3090–3098.

[32] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. ECCV*, 2016, pp. 3–19.

[33] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.

[34] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid, "Unsupervised object discovery and tracking in video collections," in *Proc. ICCV*, 2015, pp. 3173–3181.

[35] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. CVPR*, 2015, pp. 1201–1210.

[36] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. CVPR*, 2014, pp. 2513–2520.

[37] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. CVPR*, Jun. 2012, pp. 1600–1607.

[38] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for large-scale image classification," in *Proc. ECCV*, 2010, pp. 143–156.

[39] C. T. Dang and H. Radha, "Heterogeneity image patch index and its application to consumer video summarization," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2704–2718, Jun. 2014.

[40] T. Mei, L.-X. Tang, J. Tang, and X.-S. Hua, "Near-lossless semantic video summarization and its applications to video analysis," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 3, 2013, Art. no. 16.

[41] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.

[42] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5469–5478, Nov. 2016.

[43] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proc. NIPS*, 2014, pp. 2069–2077.

[44] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.

[45] Y. Cong, J. Liu, G. Sun, Q. You, Y. Li, and J. Luo, "Adaptive greedy dictionary selection for Web media summarization," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 185–195, Jan. 2017.

[46] R. Panda and A. K. Roy-Chowdhury "Collaborative summarization of topic-related videos," in *Proc. CVPR*, Jun. 2017, pp. 4274–4283.

[47] R. Panda, N. C. Mithun, and A. K. Roy-Chowdhury, "Diversity-aware multi-video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4712–4724, Oct. 2017.

[48] L. Kaufman and P. J. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Y. Dodge, Ed. Amsterdam, The Netherlands: North Holland, 1987, pp. 405–416.

[49] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," in *Proc. NIPS*, 2012, pp. 19–27.

[50] E. Elhamifar, G. Sapiro, and S. S. Sastry. (2014). "Dissimilarity-based sparse subset selection." [Online]. Available: https://arxiv.org/abs/1407.6810

[51] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[52] B. J. Frey and D. Dueck, "Mixture modeling by affinity propagation," in *Proc. NIPS*, 2005, pp. 1–8.

[53] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[54] C. Yang, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," in *Proc. CVPR*, Jun. 2012, pp. 1122–1129.

[55] H. Liu, Y. Liu, Y. Yu, and F. Sun, "Diversified key-frame selection using structured $L_{2,1}$ optimization," *IEEE Trans. Ind. Inf.*, vol. 10, no. 3, pp. 1736–1745, Aug. 2014.

[56] F. Dornaika and I. K. Aldine, "Decremental sparse modeling representative selection for prototype selection," *Pattern Recognit.*, vol. 48, no. 11, pp. 3714–3727, 2015.

[57] H. Wang, Y. Kawahara, C. Weng, and J. Yuan, "Representative selection with structured sparsity," *Pattern Recognit.*, vol. 63, pp. 268–278, Mar. 2017.

[58] C. Lu, S. Yan, and Z. Lin, "Convex sparse spectral clustering: Single-view to multi-view," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2833–2843, Jun. 2016.

[59] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proc. KDD*, 2012, pp. 1095–1103.

[60] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statist. Soc., Series B, Statist. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[61] R. Fransens, C. Strecha, and L. Van Gool, "Parametric stereo for multipose face recognition and 3D-face modeling," in *Proc. Int. Workshop Anal. Modeling Faces Gestures*, 2005, pp. 109–124.

[62] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–12.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[64] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[65] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. ICCV*, 2013, pp. 3551–3558.

[66] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[67] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[68] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. ICCV*, 2015, pp. 4489–4497.

Jingjing Meng (M'09) received the B.E. degree from the Huazhong University of Science and Technology, China, in 2003, the M.S. degree from Vanderbilt University, Nashville, TN, USA, in 2006, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2017. She was a Senior Research Staff Engineer with the Motorola Applied Research Center, Schaumburg, IL, USA. She is currently a full-time Researcher with the School of Electrical and Electronic Engineering, NTU. Her current research interests include big image and video data analytics, computer vision, and human–computer interaction. She received the Best Paper Award from the IEEE Transactions on Multimedia in 2016. She is an Associate Editor of *The Visual Computer Journal* and served as the Financial Chair for the IEEE Conference on Visual Communications and Image Processing (VCIP'15).

Suchen Wang received the B.Eng. degree from the School of Information and Electronics, Beijing Institute of Technology, China, in 2016. He is currently a Researcher with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include computer vision and image and video analytics.

Hongxing Wang received the B.S. and M.S. degrees from Chongqing University, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He was a Research Fellow/an Associate with the School of Electrical and Electronic Engineering, Nanyang Technological University, and a Visiting Student with The Institute of Scientific and Industrial Research, Osaka University, Japan. He is currently a Faculty Member with the School of Software Engineering, Chongqing University. His research interests include computer vision, pattern recognition, and machine learning.

Junsong Yuan (M'08–SM'14) received the bachelor's degree from the Special Class for the Gifted Young Program of Huazhong University of Science and Technology, Wuhan, China, in 2002, the M.Eng. degree from the National University of Singapore, and the Ph.D. degree from Northwestern University.

He is currently an Associate Professor with the School of Electrical and Electronics Engineering, Nanyang Technological University (NTU). His research interests include computer vision, video analytics, gesture and action analysis, and large-scale visual search and mining. He received the Best Paper Award from the International Conference on Advanced Robotics (ICAR17), the 2016 Best Paper Award from the IEEE Transactions on Multimedia, the Doctoral Spotlight Award from the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), the Nanyang Assistant Professorship from NTU, and the Outstanding EECS Ph.D. Thesis Award from Northwestern University.

He is currently a Senior Area Editor of the *Journal of Visual Communications and Image Representation* and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He served as a Guest Editor for the *International Journal of Computer Vision*. He is the Program Co-Chair of ICME18 and the Area Chair of ACM MM, CVPR, ICIP, ICPR, and ACCV.

Yap-Peng Tan (S'95–M'97–SM'04) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 1995 and 1997, respectively. From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, USA, and the Sharp Laboratories of America, Camas, WA, USA. In 1999, he joined Nanyang Technological University, Singapore, where he is currently a Professor and the Associate Chair (academic) with the School of Electrical and Electronic Engineering. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, machine learning, and data analytics. He served as the Chair for the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society from 2012 to 2014, the Chair for the Membership and Election Subcommittee of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2017, the Chair for the Nominations and Elections Subcommittee of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society from 2012 to 2013, a Voting Member for the IEEE International Conference on Multimedia and Expo (ICME) Steering Committee from 2011 to 2012, and the Chairman for the IEEE Signal Processing Singapore Chapter from 2009 to 2010. He has also served as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, and the IEEE ACCESS, an Editorial Board Member for the *EURASIP Journal on Advances in Signal Processing* and the *EURASIP Journal on Image and Video Processing*, a Guest Editor for special issues of several journals including the IEEE TRANSACTIONS ON MULTIMEDIA, a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society, and a member of the Image, Video, and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society. He was the Finance Chair of ICIP2004, the General Co-Chair of ICME2010, the Technical Program Co-Chair of ICME2015, and the General Co-Chair of VCIP2015. He is the Technical Program Co-Chair of ICME2018 and ICIP2019 and the Chair of the ICME Steering Committee from 2018 to 2019.