Video Summarization via Multi-View Representative Selection

Jingjing Meng[1]* Suchen Wang[1]* Hongxing Wang[2] Junsong Yuan[1] Yap-Peng Tan[1]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]School of Software Engineering, Chongqing University, China
{jingjing.meng, wang.sc, jsyuan, eyptan}@ntu.edu.sg, ihxwang@cqu.edu.cn

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 multi-view 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 multi-view sparse dictionary selection with centroid co-regularization (MSDS-CC) 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. The problem can be efficiently solved by an alternating minimizing optimization with the fast iterative shrinkage thresholding algorithm (FISTA). We also show how the formulation can be applied to category-specific video summarization by incorporating visual co-occurrence priors. Experiments on benchmark video datasets validate the effectiveness of the proposed approach in comparison with other video summarization methods and representative selection methods.*

1. 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 [31, 29, 27], keyframes [22, 21, 25], key shots [18, 48], montages [40], dynamic synopses [36], 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 (Sec. 3.1). 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.

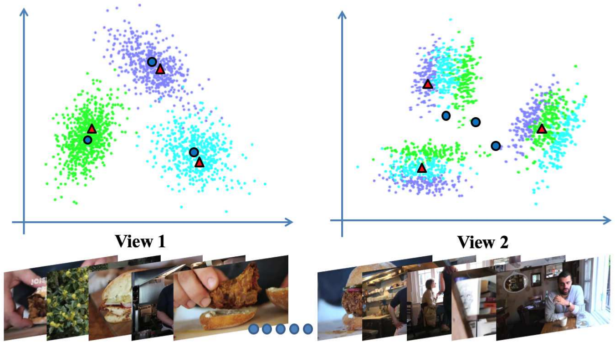


Figure 1: An illustration of the proposed video summarization via multi-view representative selection. The top row shows two views of a video's frames side by side. There are 3 clusters (key visual concepts) in each view, and our aim is to delineate all the 6 clusters by selecting only 3 representatives. Note that features in the two views have different distributions, *e.g.*, the green cluster in View 1 is scattered in View 2 (similarly for the purple 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.

Although multi-view sparse subspace/dictionary learning approaches have been proposed [2, 23, 19, 43, 3, 15], they require feature fusion to be conducted in advance to learn unified data representations for 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 representations to well characterize the underlying distribution of the data across multiple feature spaces, thus affecting the performance of the subsequent representative selection.

*Equal contributions

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). Fig. 1 illustrate the idea in comparison with 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, see Sec. 3.2). Our formulation provides the following benefits:

- It can produce a consensus selection of visual elements across different views, resulting in summaries that are consistently representative across multiple feature modalities.
- As we directly optimize for 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.
- Our formulation can better handle noisy features by incorporating view-specific selection priors (Sec. 3.4) to guide the representative selection towards more relevant visual elements. This permits the use of external data or/and supervision to improve summarization quality, a scheme that has been shown to be effective in previous work [21, 39, 31, 6, 47].
- Compared 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 summaries of various sizes can be produced 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) [1]. Comparative experiments on challenging benchmark datasets demonstrate the efficacy of the proposed MSDS-CC. We also show how it can be applied to category-specific video summarization by using visual co-occurrence as priors. The resulting category-specific video summaries reflect both the *local* representativeness within individual videos and the *global* visual commonness among multiple videos of the same topic (*i.e.*, visual concepts that appear repeatedly, Fig. 3).

2. Related work

Video summarization Previous work in video summarization can be grouped into three broad categories: domain-specific [4, 45, 26, 52, 30], supervised [17, 35, 18, 47, 48, 38] and unsupervised [24, 5, 6, 49, 21, 39, 11, 27, 34, 50] methods. Domain-specific methods focus on summarize videos in specific genres, such as surveillance [7], sports [4, 52] and egocentric videos [26, 45, 30]. Supervised methods usually generate summarizations by learning from human annotations. For instance, to make a structured prediction, submodular functions are trained with user created summaries [18]. Gygli *et al.* [17] train a linear regression model to estimate the interestingness score of shots. More recently, Gong *et al.* [16] and others [48, 38] define novel models to learn from human-created summaries for selecting representative and diverse subsets. In addition, Zhang *et al.* [47] shows summary structures can be transferred between videos that are semantically consistent. Unsupervised methods usually summarize videos by seeking the visual relevance and structure. A popular method is to select representative frames/objects by learning a dictionary from videos [11, 49, 31]. Alternatively one can leverage information from other sources such as video titles and web images [34, 39, 22]. Recently, video co-summarization [6] has also been proposed, which summarized shots that co-occur among multiple videos of the same topic.

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 [20], sparse selection of clustering centers [10, 9], affinity propagation [13, 14], and density peak search [37]. 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 [7, 11, 44, 28, 8, 31, 42]. Despite the advances in representative selection, most of the methods are not applicable to multiple features. Feature fusion such as [2, 23, 19, 43, 3, 15] 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.

3. Problem formulation

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

3.1. Preliminaries: feature concatenation

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

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

As a result, the following representative selection objective in (2) is tantamount to that of feature concatenation for sparse dictionary selection [7].

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

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

3.2. 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 may differ significantly, which can heavily influence the selection result. To better handle multiple features, we propose to learn individual selection matrices $\mathbf{S}^{(v)}, v = 1, 2, \cdots, V$ 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:

$$\begin{aligned}\min_{\mathbf{S}^{(v)}, \mathbf{w}} & \underbrace{\sum_{v=1}^{V} \left\{ \frac{1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{S}^{(v)}\|_{\mathrm{F}}^2 + \lambda^{(v)} \|\mathbf{S}^{(v)}\|_{1,2} \right\}}_{J_1} \\ & + \eta \underbrace{\left\{ \frac{1}{2} \sum_{v=1}^{V} \|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1 \right\}}_{J_2},\end{aligned} \tag{3}$$

where the weighting vector $\mathbf{w}^{(v)}$ consists of the l_2 norms of rows of $\mathbf{S}^{(v)}$, with the i^{th} entry $w_i^{(v)} = \|\mathbf{S}_{i,\cdot}^{(v)}\|_2$, 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.

The objective function in (3) ($\mathcal{O}$ for short) can be solved by iterating between: (1) optimizing $\mathbf{S}^{(v)}$ by fixing $\mathbf{S}^{(u)}$ ($u \neq v$) and $\mathbf{w}$, and (2) optimizing $\mathbf{w}$ by fixing $\mathbf{S}^{(v)}$ ($v = 1, 2, ..., V$).

3.2.1 Optimize $\mathbf{S}^{(v)}$ by fixing $\mathbf{S}^{(u)}$ ($u \neq v$) and $\mathbf{w}$

Regroup the objective function in (3) as

$$\mathcal{O} = \sum_{v=1}^{V} \mathcal{O}^{(v)} + \eta\tau \|\mathbf{w}\|_1, \tag{4}$$

where

$$\begin{aligned}\mathcal{O}^{(v)} = & \frac{1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{S}^{(v)}\|_{\mathrm{F}}^2 + \lambda^{(v)} \|\mathbf{S}^{(v)}\|_{1,2} \\ & + \frac{1}{2}\eta \|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2.\end{aligned} \tag{5}$$

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

$$\begin{aligned}\mathcal{O}^{(v)} = & \frac{1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{S}^{(v)}\|_{\mathrm{F}}^2 + \lambda^{(v)} \|\mathbf{S}^{(v)}\|_{1,2} \\ & + \frac{1}{2}\eta (\|\mathbf{S}^{(v)}\|_{\mathrm{F}}^2 + \|\mathbf{w}\|_2^2 - 2{\mathbf{w}^{(v)}}^{\mathrm{T}}\mathbf{w}) \\ = & \frac{1}{2} \mathrm{tr}\{{\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} - 2{\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)}\mathbf{S}^{(v)} \\ & + {\mathbf{S}^{(v)}}^{\mathrm{T}} \left(\eta\mathbf{I} + {\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} \right) \mathbf{S}^{(v)}\} \\ & + (\lambda^{(v)}\mathbf{1} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)} + \frac{1}{2}\eta\|\mathbf{w}\|_2^2.\end{aligned} \tag{6}$$

Then, we let

$$\begin{aligned}f(\mathbf{S}^{(v)}) = & \frac{1}{2} \mathrm{tr}\{{\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} - 2{\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)}\mathbf{S}^{(v)} \\ & + {\mathbf{S}^{(v)}}^{\mathrm{T}} \left(\eta\mathbf{I} + {\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} \right) \mathbf{S}^{(v)}\} + \frac{1}{2}\eta\|\mathbf{w}\|_2^2,\end{aligned} \tag{7}$$

and

$$g(\mathbf{S}^{(v)}) = (\lambda^{(v)}\mathbf{1} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)}, \tag{8}$$

which leads to

$$\mathcal{O}^{(v)} = f(\mathbf{S}^{(v)}) + g(\mathbf{S}^{(v)}). \tag{9}$$

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 FISTA [1]:

$$\mathrm{prox}_{\mathcal{R}}(\mathbf{Z}) = \operatorname*{arg\,min}_{\mathbf{S}^{(v)} \in \mathbb{R}^{n \times n}} \frac{1}{2} \left\| \mathbf{S}^{(v)} - \mathbf{Z} \right\|_{\mathrm{F}}^2 + \frac{1}{L^{(v)}} g(\mathbf{S}^{(v)}), \tag{10}$$

where

$$\begin{aligned}\mathbf{Z} = & \mathbf{S}^{(v)} - \frac{1}{L^{(v)}} \frac{\partial}{\partial \mathbf{S}^{(v)}} f(\mathbf{S}^{(v)}) \\ = & \mathbf{S}^{(v)} - \frac{1}{L^{(v)}} \left\{ -{\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} + \left(\eta\mathbf{I} + {\mathbf{X}^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)} \right) \mathbf{S}^{(v)} \right\}.\end{aligned} \tag{11}$$

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

$$L^{(v)} = r(\eta \mathbf{I} + {\mathbf{X}^{(v)}}^{\mathrm{T}} \mathbf{X}^{(v)}) = \eta + r({\mathbf{X}^{(v)}}^{\mathrm{T}} \mathbf{X}^{(v)}). \tag{12}$$

Follow the proximal decomposition [51], Problem (10) is solvable. For $i \in [1, n]$,

$$\mathbf{S}_{i,\cdot}^{(v)} = \arg\min_{\mathbf{s} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{s} - \mathbf{Z}_{i,\cdot}\|_2^2 + \hat{\lambda}_i^{(v)} \|\mathbf{s}\|_2, \tag{13}$$

where

$$\hat{\lambda}_i^{(v)} = \frac{1}{L^{(v)}} \left(\lambda^{(v)} - \eta w_i \right). \tag{14}$$

After applying soft-thresholding [46], we have, for $i = 1, 2, ..., n$,

$$\mathbf{S}_{i,\cdot}^{(v)} = \mathbf{Z}_{i,\cdot} \max\{(1 - \frac{\hat{\lambda}_i^{(v)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}. \tag{15}$$

3.2.2 Optimize w while fixing $\mathbf{S}^{(v)}$

Denote the first term in the objective function (3) as J_1 and the second term as J_2, then $\min_{\mathbf{w}} \mathcal{O} \Leftrightarrow \min_{\mathbf{w}} J_2$ when fixing $\mathbf{S}^{(v)}$, and

$$J_2 = \frac{1}{2} \sum_{v=1}^{V} \|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1. \tag{16}$$

By applying soft-thresholding, we obtain

$$\begin{aligned} \mathbf{w} =& \mathrm{sign}(\frac{1}{V} \sum_{v=1}^{V} \mathbf{w}^{(v)}) \odot \max\{(\frac{1}{V} |\sum_{v=1}^{V} \mathbf{w}^{(v)}|) - \frac{1}{V}\tau, 0\} \\ =& \max\{\frac{1}{V}(\sum_{v=1}^{V} \mathbf{w}^{(v)} - \tau), 0\}. \end{aligned} \tag{17}$$

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.

3.3. Parameter setting

Dictionary selection parameter $\lambda^{(v)}$ 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 $\mathbf{Z}$ in (15), when $\lambda^{(v)}$ is large enough, we have $\mathbf{S}^{(v)} = \mathbf{0}$, which results in an empty selection. To avoid such an empty selection in the initialization, we let $\lambda^{(v)} \leq \lambda_{\max}^{(v)}$ and solve $\lambda_{\max}^{(v)}$ by substituting $\mathbf{S}^{(v)} = \mathbf{0}$ into (15) as follows:

$$\lambda_{\max}^{(v)} = L^{(v)} \max_{0 \leq i \leq n} \|\mathbf{Z}_{i,\cdot}\|_2. \tag{18}$$

Algorithm 1 Multi-view Representative Selection via Centroid Coregulerization (3).

```
Input:  features ${\mathbf{X}^{(v)}\}_{v=1}^V$, parameters $\{\lambda^{(v)}\}_{v=1}^V$, $\eta$, $\tau$
Output: selection matrices for each view $\{\mathbf{S}^{(v)}\}_{i=1}^V$, 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. (12))
end for
// Iteratively solve the objective function (Eq. (3))
repeat
    // Optimize $\mathbf{S}^{(v)}$ by fixing $\mathbf{S}^{(u)}$ $(u \neq v)$ and $\mathbf{w}$
    for $v \in [1, V]$ do
    $\mathbf{S}^{(v)} \leftarrow \mathbf{0}$, $\mathbf{V} \leftarrow \mathbf{S}^{(v)}$, $t \leftarrow 1$
    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. (11))
        $\mathbf{U} \leftarrow \mathbf{S}^{(v)}$, $\mathbf{S}_{i,\cdot}^{(v)} \leftarrow \mathbf{Z}_{i,\cdot} \max\{(1 - \frac{\hat{\lambda}_i^{(v)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}, i \in [1, n]$    ( Eq. (15))
        $q = t - 1, t \leftarrow (1 + \sqrt{1 + 4t^2})/2$
        $\mathbf{V} \leftarrow \mathbf{S}^{(v)} + q(\mathbf{S}^{(v)} - \mathbf{U})/t$
    until convergence
    end for
    // Optimize $\mathbf{w}$ while fixing $\mathbf{S}^{(v)}$
    $\mathbf{w} \leftarrow \max\{\frac{1}{V}(\sum_{v=1}^V \mathbf{w}^{(v)} - \tau), 0\}$    ( Eq. (17))
until convergence
```

It is worth noting that in Algorithm 1, we initialize $\mathbf{S}^{(v)}$ by a zero matrix, and $\mathbf{w}$ by a zero vector. Then according to (11), after the first iteration, we have

$$\mathbf{Z} = \frac{1}{L^{(v)}} {\mathbf{X}^{(v)}}^{\mathrm{T}} \mathbf{X}^{(v)}. \tag{19}$$

Therefore,

$$\lambda_{\max}^{(v)} = \max_{0 \leq i \leq n} \|{\mathbf{x}_i^{(v)}}^{\mathrm{T}} \mathbf{X}^{(v)}\|_2. \tag{20}$$

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

Centroid co-regularization parameter η. As shown in (3), this parameter trades-off the first dictionary selection term J_1 and the second centroid co-regularization term J_2 (16). When $\eta \to 0$, we will immediately reach the consensus by feeding individual dictionary selection results into (17). When $\eta \to +\infty$, minimizing (3) will lead to a zero J_2, thus making $\mathbf{w}^{(v)} (v \in [1, V])$ and $\mathbf{w}$ to be $\mathbf{0}$. As a result, we cannot select anything from the data. Furthermore, we can see from (14), η balances the contributions of $\lambda^{(v)}$ and $\mathbf{w}$ to the dictionary selection of the v^{th} view in (15). For ease of tuning η, we let

$$\eta = \frac{\min_{v \in [1, V]} \{\lambda^{(v)}\}}{\alpha_\eta}, \tag{21}$$

and tune the hyper-parameter α_η.

Sparse consensus parameter τ. This parameter controls the sparsity of selection consensus by minimizing (16). According to the solution to (16) in (17), a larger τ implies a sparser selection result. To facilitate tuning τ, we introduce an auxiliary parameter α_τ and let

$$\tau = \frac{\max_{i\in[1,n]}\left\{\sum_{v\in[1,V]} w_i^{(v)}(1)\right\}}{\alpha_\tau}, \tag{22}$$

where $w_i^{(v)}(1)$ denotes the result of $w_i^{(v)}$ after the first round of optimization.

3.4. Extension to incorporate priors

As selection priors such as canonical viewpoints [21], visual co-occurrence [6] and objectness scores [31] have been shown to improve video summarizaiton 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:

$$\begin{aligned}\min_{\mathbf{S}^{(v)},\mathbf{w}} \sum_{v=1}^{V} \left\{ \frac{1}{2}\|\mathbf{X}^{(v)} - \mathbf{X}^{(v)}\mathbf{S}^{(v)}\|_F^2 + \lambda^{(v)} \sum_{i=1}^{n} \rho_i^{(v)} w_i^{(v)} \right\} \\ + \eta \left\{ \frac{1}{2} \sum_{v=1}^{V} \|\mathbf{w}^{(v)} - \mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1 \right\},\end{aligned} \tag{23}$$

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.

The optimization of (23) follows a similar procedure as shown in Subsections 3.2.1 and 3.2.2. We only need to update the non-smooth term $g(\mathbf{S}^{(v)})$ (shown in (9)) to suit the new objective function in (23) by $g(\mathbf{S}^{(v)}) = (\mathbf{\Lambda}^{(v)} - \eta\mathbf{w})^{\mathrm{T}}\mathbf{w}^{(v)}$, where $\mathbf{\Lambda}^{(v)} \in \mathbb{R}^n$, and its i^{th} element is $\lambda^{(v)}\rho_i^{(v)}$. Therefore, the solution to $\mathbf{S}^{(v)}$ is still given by (15), but with a different $\hat{\lambda}_i^{(v)}$ compared to (14), and $\hat{\lambda}_i^{(v)}$ becomes

$$\hat{\lambda}_i^{(v)} = \frac{1}{L^{(v)}}\left(\lambda^{(v)}\rho_i^{(v)} - \eta w_i\right). \tag{24}$$

For equal prior selection costs with $\rho_i^{(v)} = 1$, (24) and (14) become the same. Problem (23) will perfectly degenerate into Problem (3).

To facilitate setting parameters $\{\lambda^{(v)}\}_{v=1}^{V}$, η, and τ, we also refine the calculation of $\lambda_{\max}^{(v)}$ in Subsection 3.3 when optimizing (23) with the addition of priors. According to (24) and (15), we calculate $\lambda_{\max}^{(v)}$ by

$$\begin{aligned}\lambda_{\max}^{(v)} &= L^{(v)} \max_{0\le i\le n} \frac{1}{\rho_i^{(v)}}\|\mathbf{Z}_{i,\cdot}\|_2 \\ &= L^{(v)} \max_{0\le i\le n} \frac{1}{\rho_i^{(v)}}\|{\mathbf{x}_i^{(v)}}^{\mathrm{T}}\mathbf{X}^{(v)}\|_2.\end{aligned} \tag{25}$$

4. Experiments

4.1. 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.

Clustering-based baselines include the standard K-medoids [20] and two multi-view spectral clustering methods: Affinity aggregation spectral clustering (AASC) [19] and Co-regularized multi-view spectral clustering (CMSC) [23]. We use the centroid-based co-regularization for CMSC.

Subspace learning based baselines include the state-of-the-art Sparse Modeling Representative Selection (SMRS) [11] and Locally Linear Reconstruction induced Sparse Dictionary Selection (LLR-SDS) [31].

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 K-medoids, for which we used the MATLAB implementation. α for SMRS and α_1 for LLR-SDS are tested on a range of $\{5, 8, 10, 20, 30\}$. 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 $\alpha_\lambda \in \{3, 5, 10, 20, 30\}$, $\alpha_\eta \in \{0.1, 1, 2, 5, 10\}$ and fix $\alpha_\tau = 10$. And we report the best result for each method.

4.2. 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 1). For simplicity, we consider the representative selection on two views and randomly generated $2D$-dimensional data points, where D is the dimension of the ambient space for each 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 [31], we evaluate the performance of the top n representatives by the average recall. Results are averaged over 25 trials. As can be seen, the proposed approach outperforms all baselines in the test cases. It is worth noting that direct concatenation of multiple features does not necessarily perform better than the single-view selection. In addition, multi-view clustering methods (*i.e.*, AASC and CMSC) perform worse than ours, which can be attributed to the difficulty faced by these methods in handling the disagreement

Table 1: Average recall of synthetic data on 2 views. In each view, data points are projected to N clusters, and the feature dimension of each view is indicated by D. Results are averaged over 25 trials.

D	N	Single View Selection			Concatenated View Selection			Multi-View Selection		
		KM	SMRS	LLR-SDS	KM	SMRS	LLR-SDS	AASC	CMSC	MSDS-CC
2	3	0.87	0.70	0.71	0.79	0.78	0.69	0.78	0.78	**0.90**
3	3	0.86	0.77	0.71	0.81	0.78	0.68	0.83	0.77	**0.91**
3	5	0.84	0.85	0.68	0.81	0.77	0.62	0.81	0.73	**0.87**
5	5	0.83	0.73	0.64	0.72	0.75	0.60	0.71	0.76	**0.84**
5	7	0.82	0.69	0.64	0.69	0.78	0.60	0.69	0.81	**0.84**

in the feature distributions in different views, *e.g.*, when data points belong to different groups in different views.

4.3. Proof of concept

We further validate the effectiveness of the proposed multi-view representative selection on the EPFL stereo face dataset [12]. The dataset consists of 100 subjects, each recorded from 8 different viewpoints by a pair of calibrated stereo cameras. We randomly select 4 subjects and 4 poses to form a dataset of 16 images. Our goal is to capture all the 4 subjects and 4 poses by selecting a few representative faces. In the ideal case, as few as 4 face images should capture all the 4 subjects and the 4 poses. Similar to [31], 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 [33] and the 136D facial landmark/fiducial points extracted from dlib (68 face landmarks with (x,y) coordinates).

Fig. 2 shows qualitative comparisons of 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 Average Recall $@4 = 1$.

4.4. Video summarization

4.4.1 Datasets

To demonstrate the effectiveness of our approach on video summarization, we experiment on two benchmark datasets, TVSum [39] and SumMe [17]. TVSum consists of 50 videos within 10 categories representing various genres. It also provides shot-level importance scores obtained from user 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 the ground-truth is 13.1% of that of the original video. In our experiments, we summarize videos into key shots to facilitate comparisons with prior work [39, 17, 18, 47, 48] and evaluate the performance accordingly.

4.4.2 Settings

Features We extract GIST [32] and CNN features from each frame. GIST descriptors are computed with 32 Gabor filters at 4 scales, 8 orientations and 4×4 blocks, resulting in 512D features. CNN features(1024D) are extracted from pool 5 layer of the pre-trained GoogLeNet model [41].

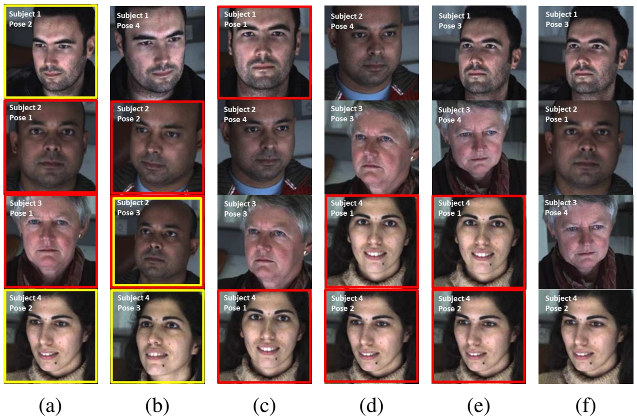


Figure 2: EPFL stereo face: visualization on the first 4 representatives selected by each method (column-wise). Duplicate subjects or poses in each column are highlighted by bounding boxes of the same color. (a) K-medoids captures all 4 subjects but only 2 poses (Average Recall@4 = 0.75). (b) AASC captures 3 subjects and 3 poses (Average Recall@4 = 0.75). (c) CMSC captures all 4 subjects but 3 poses (Average Recall@4 = 0.875). (d) SMRS and (e) LLR-SDS both select 3 subjects and 4 poses (Average Recall@4 = 0.875). In comparison, (f) our MSDS-CC captures all the 4 subjects and 4 poses (Average Recall@4 = 1).

Shot segmentation Since neither of the datasets provides ground-truth temporal segmentation, we first temporally segment videos into disjoint intervals by Kernel Temporal Segmentation(KTS) method [35]. The average length of intervals/shots are around 5 seconds. We sample 5 frames per shot to reduce the computational cost.

Summary generation To generate a video summary of length l, we follow [39, 17] to solve the knapsack problem:

$$\max \sum_{i=1}^{s} u_i \phi_i \quad \text{s.t.} \quad \sum_{i=1}^{s} u_i n_i \leq l, u_i \in \{0, 1\} \qquad (26)$$

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 prior work [39, 17, 18, 48], we set the length budget l to be 15% in duration of the original

Table 2: 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 work. † denotes methods that use additional web images and ‡ denote methods that use annotated video summaries for training. Dashes denote unavailable dataset-method combinations.

Methods	TVsum	SumMe
K-medoids [20]	31.4	29.7
AASC [19]	31.8	35.8
CMSC [23]	32.7	33.1
SMRS [11]	41.0	37.3
LLR-SDS [31]	49.7	40.4
MSDS-CC (ours)	**52.3**	**40.6**
TVsum† [39]	50.0	-
SumMe‡ [17]	-	39.4
Submodular‡ [18]	-	39.7
Summary Transfer‡ [47]	-	40.9
dppLSTM(Canonical)‡ [48]	54.7	38.6

video for both datasets.

Implementation details Similar to [39], for the subspace learning based baselines (*i.e.*, SMRS [11], LLR-SDS [31]) and our proposed MSDS-CC, we predicts the importance score of each shot ϕ_i by the importance score of its frames. Specifically, the importance score of each frame is predicted by the resulting selection weights from each method (*e.g.*, the consensus weight $\mathbf{w}$ in (3) for MSDS-CC), and the shot-level scores ϕ_i in (26) is calculated by selecting the maximum score of frames within each shot.

We follow [39] to evaluate clustering based baselines (*i.e.*, K-medoids, AASC [19] and CMSC [23]). As in [39], clustering is performed on the video frames with the number of clusters set to 100. We first compute the distance of each frame to its closest centroid. Then the shot-level distances is calculated as the average distance of the frames belonging to the most frequently assigned cluster within each shot. Finally, the summary is generated by selecting the shots closest to the centroid of the largest clusters, with a length budget l.

4.4.3 Evaluation

Following prior work [39, 17, 18, 47, 48], 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 $F = \frac{P \cdot R}{0.5(P+R)}$. As in [48], we follow [39, 18] to compute the metrics when there are multiple human-created summaries of a video.

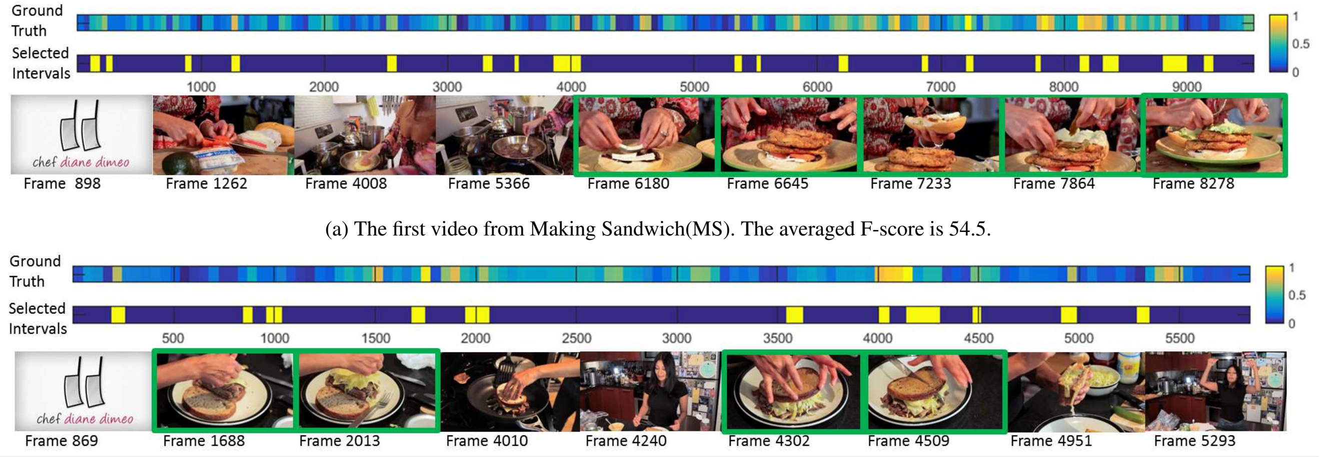
4.4.4 Results

Table 2 shows the performance of our approach (MSDS-CC) on TVSum and SumMe. Our approach outperforms all clustering-based and subspace-based baselines on both datasets. For comparison, we also report results of other summarization methods from published prior work [39, 17, 18, 48, 47]. It is shown that the proposed MSDS-CC perform competitively without relying on external images [39] or learning from user annotated summaries [17, 18, 47, 48]. Specifically, on TVSum, our approach performs better than the TVSum benchmark results [39], which uses additional title-based image search results to help identify canonical visual concepts shared between the video and images. Although dppLSTM (Canonical) [48] performs slightly better than ours, it uses the user annotations on 80% videos from TVSum for training and the remaining 20% for testing. On SumMe, our MSDS-CC outperforms the SumMe benchmark results [17], Submodular [18] and dppLSTM (Canonical) [48] and is comparable to Summary Transfer [47], which uses additional videos for training.

4.5. Category-specific video summarization with visual co-occurrence priors

Next, we show how our formulation can be applied to category-specific video summarization by using visual co-occurrence as priors in (23). This is motivated by video co-summarization [6], which aims to summarize shots that co-occur most frequently in videos of the same topic, while discarding the infrequent ones. Similarly, for videos from the same category, we would like to explore visual co-occurrence to guide the summary towards common concepts among videos of the same category. However, different from co-summarization, we would like to also keep shots that are representative of individual videos even if they do not occur in other videos.

We evaluate our approach with priors, *i.e.*, MSDS-CC (prior) in Table 3, on TVSum dataset that provides video categories. We first use the following criteria to filter videos that are visually disparate from the rest in each category. Given a feature modality (*i.e.*, view v), we first calculate the frame-level pair-wise similarity, excluding pairs from the same video. The similarity between two frame features $\mathbf{x}_i^{(v)}$ and $\mathbf{x}_j^{(v)}$ is defined as $S_{i,j}(k,l) = \exp\{-\|\mathbf{x}_i^{(v)} - \mathbf{x}_j^{(v)}\|^2/2\sigma^2\}$, where i index frames in video k and j index frames in video l, and video k, l belong to same category. σ is calculated as the average euclidean distance of the top 20% closest neighbours in a category. Then the video-level pairwise similarity is computed by averaging corresponding pairwise frame distances. We use a threshold of 0.65 to remove videos that are not similar to others in its category. After filtering, 15 videos in 6 categories remain, and we report results on this subset. For MSDS-CC (prior), the frame weight $\rho_i^{(v)}$ of video k in (23) is calculated as



(a) The first video from Making Sandwich(MS). The averaged F-score is 54.5.

(b) The second video from Making Sandwich(MS). The average F-score is 60.0.

Figure 3: Sample results for category-specific summarization: a pair of videos from the same category, Making Sandwich (MS), in TVSum. For each video, the first colorbar shows Ground Truth (*i.e.*, user annotated importance scores); the second colorbar shows our summarization results, where yellow intervals indicate shots selected by our MSDS-CC (prior). The bottom row shows sampled frames from selected shots. Co-occurring concepts are highlighted by green rectangles.

Table 3: Category specific summarization results on TVSum. Specific Categories are VU (getting Vehicle Unstuck), GA (Grooming an Animal), MS (Making Sandwich), PR (PaRade), FM (Flash Mob gathering) and BT (attempting Bike Tricks).

Cat	K-medoids	AASC	CMSC	SMRS	LLR-SDS	MSDS-CC	MSDS-CC (prior)
VU	45.1	43.1	48.5	38.3	53.0	55.4	**56.2**
GA	24.6	35.0	38.0	32.8	39.4	45.7	**48.5**
MS	43.3	38.3	36.7	37.8	51.9	56.2	**57.3**
PR	41.0	44.0	31.4	41.6	45.5	54.0	**57.1**
FM	27.7	34.9	36.1	42.0	52.4	52.3	**52.8**
BT	32.9	22.6	28.2	48.6	52.4	55.5	**57.7**
Avg	35.8	36.3	36.5	40.2	49.1	53.2	**54.9**

$\rho_i^{(v)} = \sum_l \max_j S_{i,j}(k,l)$.

Summarization results in comparison with the baselines are shown in Table 3. As seen in table 3, when using visual co-occurrence as view-specific selection priors (*i.e.*, MSDS-CC (prior)), the performance of the proposed MSDS-CC can be further improved across all categories. Both MSDS-CC (prior) and MSDS-CC outperform the baseline representative selection methods overall and in each category.

Fig. 3 shows visual examples of the category-specific video summarization by MSDS-CC (prior) on two videos in the Making Sandwich (MS) category. It shows that the produced summaries can capture both repeated visual contents that reflect the *global* commonness in a given category and *local* contents that are representative of individual videos. The weakness of the co-occurrence priors, however, is that unimportant shots may also be selected if they are similar to shots from other videos of the same category (*e.g.*, the leftmost frames of the two videos in Fig. 3).

5. Conclusions

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 challenging benchmark datasets demonstrate the effectiveness of the proposed approach for generic and category-specific video summarization.

Acknowledgements: This research was carried out at the ROSE Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. Hongxing Wang is supported in part by National Natural Science Foundation of China under Grant 61602069 and Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2016jcyjA0468).

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIIMS*, 2(1):183–202, 2009. 2, 3

[2] X. Cai, F. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, pages 1977–1984, 2011. 1, 2

[3] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015. 1, 2

[4] F. Chen and C. D. Vleeschouwer. Formulating team-sport video summarization as a resource allocation problem. *T-CSVT*, 2011. 2

[5] 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 *CVPR*, 2015. 2

[6] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, pages 3584–3592, 2015. 2, 5, 7

[7] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 2011. 2, 3

[8] F. Dornaika and I. K. Aldine. Decremental sparse modeling representative selection for prototype selection. *Pattern Recognition*, 48:3714–3727, 2015. 2

[9] E. Elhamifar, G. Sapiro, and S. S. Sastry. Dissimilarity-based sparse subset selection. *arXiv preprint arXiv:1407.6810*, 2014. 2

[10] E. Elhamifar, G. Sapiro, and R. Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *NIPS*, pages 19–27, 2012. 2

[11] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 2012. 2, 5, 7

[12] R. Fransens, C. Strecha, and L. Van Gool. Parametric stereo for multi-pose face recognition and 3d-face modeling. In *International Workshop on Analysis and Modeling of Faces and Gestures*. Springer, 2005. 6

[13] B. J. Frey and D. Dueck. Mixture modeling by affinity propagation. In *NIPS*, pages 379–386, 2005. 2

[14] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007. 2

[15] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *ICCV*, pages 4238–4246, 2015. 1, 2

[16] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2014. 2

[17] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 2, 6, 7

[18] M. Gygli and H. G. L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3090–3098, 2015. 1, 2, 6, 7

[19] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Affinity aggregation for spectral clustering. In *CVPR*, pages 773–780. IEEE, 2012. 1, 2, 5, 7

[20] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. In Y. Dodge, editor, *Statistical Data Analysis Based on the L1 Norm and Related Methods*, pages 405–416. North-Holland, 1987. 2, 5, 7

[21] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 1, 2, 5

[22] G. Kim, L. Sigal, and E. P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014. 1, 2

[23] A. Kumar, P. Rai, and H. D. III. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011. 1, 2, 5, 7

[24] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid. Unsupervised object discovery and tracking in video collections. In *ICCV*, pages 3173–3181, 2015. 2

[25] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1

[26] Y. J. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 2015. 2

[27] D. Liu, G. Hua, and T. Chen. A hierarchical visual model for video object summarization. *TPAMI*, 2010. 1, 2

[28] H. Liu, Y. Liu, Y. Yu, and F. Sun. Diversified key-frame selection using structured optimization. *IEEE Transactions on Industrial Informatics*, 10(3):1736–1745, 2014. 2

[29] C. Lu, R. Liao, and J. Jia. Personal object discovery in first-person videos. *TIP*, 2015. 1

[30] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[31] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *CVPR*, June 2016. 1, 2, 5, 6, 7

[32] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42, 2001. 6

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 6

[34] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 2

[35] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*. 2014. 2, 6

[36] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007. 1

[37] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 2014. 2

[38] A. Sharghi, B. Gong, and M. Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 2

[39] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 2, 6, 7

[40] M. Sun, A. Farhadi, B. Taskar, and S. Seitz. Salient montages from unconstrained videos. In *ECCV*. 2014. 1

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, June 2015. 6

[42] H. Wang, Y. Kawahara, C. Weng, and J. Yuan. Representative selection with structured sparsity. *Pattern Recognition*, 63:268 – 278, 2017. 2

[43] H. Wang, C. Weng, and J. Yuan. Multi-feature spectral clustering with minimax optimization. In *CVPR*, pages 4106–4113, 2014. 1, 2

[44] C. Yang, J. Peng, and J. Fan. Image collection summarization via dictionary learning for sparse representation. In *CVPR*, pages 1122–1129, 2012. 2

[45] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu. Discovering thematic objects in image collections and videos. *TIP*, 21(4):2207–2219, 2012. 2

[46] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *RSSSB*, 68(1):49–67, 2006. 4

[47] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, June 2016. 2, 6, 7

[48] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1, 2, 6, 7

[49] B. Zhao and E. P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, pages 2513–2520, 2014. 2

[50] G. Zhao, J. Yuan, and G. Hua. Topical video object discovery from key frames by modeling word co-occurrence prior. In *CVPR*, 2013. 2

[51] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *KDD*, pages 1095–1103, 2012. 4

[52] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao. Trajectory based event tactics analysis in broadcast sports video. In *Proceedings of the 15th International Conference on Multimedia*, 2007. 2