From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection

Jingjing Meng[1] Hongxing Wang[1,2] 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,jsyuan,eyptan}@ntu.edu.sg, ihxwang@cqu.edu.cn

Abstract

We propose to summarize a video into a few key objects by selecting representative object proposals generated from video frames. This representative selection problem is formulated as a sparse dictionary selection problem, i.e.*, choosing a few representatives object proposals to reconstruct the whole proposal pool. Compared with existing sparse dictionary selection based representative selection methods, our new formulation can incorporate object proposal priors and locality prior in the feature space when selecting representatives. Consequently it can better locate key objects and suppress outlier proposals. We convert the optimization problem into a proximal gradient problem and solve it by the fast iterative shrinkage thresholding algorithm (FISTA). Experiments on synthetic data and real benchmark datasets show promising results of our key object summarization apporach in video content mining and search. Comparisons with existing representative selection approaches such as K-mediod, sparse dictionary selection and density based selection validate that our formulation can better capture the key video objects despite appearance variations, cluttered backgrounds and camera motions.*

1. Introduction

With videos becoming the biggest big data, there has been increasing need to summarize, index and browse the large corpus of video content. As a common practice, videos are often summarized by keyframes, i.e., a set of representative video frames [19, 20, 21]. Although such a keyframe-based summarization can capture the important scenes, it often does not pick out the key objects from less informative backgrounds in a video.

In this work, we propose to summarize videos into key objects instead of keyframes, as illustrated in Fig. 1. Comparing with keyframes, summarizing videos into a collection of key objects can be attractive to many applications. For example, the summarized key objects can serve as icons

Figure 1: Video summarization by representative object proposal selection. A pool of object proposals (middle row) are first generated from video keyframes (top). Representative object proposals (bottom) are selected from the pool to summarize the video.

to establish a quick impression of the video by telling what are there. They also provide a small footprint for indexing, browsing and search, *e.g.*, retrieving videos by matching the key objects. Besides object-level summarization and search, as these key objects are essential components of higher level semantics in videos, once identified, they can also be used to recover or help understand more complicated semantics of videos, *e.g.*, tracking candidate objects for spatio-temporal action localization [37], constructing story based video summarization by analyzing the interactions among the key objects [26] and egoncentric analysis [27, 24].

Motivated by the recent successes of category independent object proposals for detection and weekly supervised learning [14, 7], we propose to summarize videos by object proposals. It has been shown that generating multiple overlapping object proposals in general provides more accurate object regions than segmentation by sidestepping the harder problem of full segmentation [47, 1]. Start with a pool of frame-level object proposals produced by a high recall object proposal method [47], we formulate video sum-

marization as a representative selection problem: selecting a few exemplar object proposals to reconstruct the whole pool. Although representative selection methods have been applied to video keyframe selection [11], directly applying them to object-level summarization faces many challenges. First of all, the appearance of the same object may change significantly across frames due to pose and scale variations, partial occlusions, etc. Therefore the key objects do not necessary locate at the densest regions in the feature space. Consequently, classic density based representative selection method may not work well [31]. Moreover, since object proposals are just candidates of key objects, there may be many irrelevant and noisy proposals in the proposal pool. These outliers may significantly affect the representative selection methods based on sparse reconstruction [6, 11]. In such a case, even a further filtering of outliers as post-processing [11] may be less effective if most representatives are outliers. Without prior knowledge of the object, it is difficult to locate the key objects accurately.

To address the above challenges, we propose a new formulation of sparse reconstruction based representative selection, which has the following advantages. First, it can incorporate object proposal priors when selecting the representatives. Therefore, object proposals of high prior weights, *e.g.*, high objectiveness scores, are more likely to be selected as key objects, while background clutters of low weights can be suppressed. Second, our new formulation also considers the local affinity structure of the data samples by introducing a locality prior matrix to regularize the selection matrix. As the outcome, it prefers popular object proposals that appear more frequently while outlier proposals are likely to be suppressed. Third, although complex constraints are introduced, we convert our optimization into a proximal gradient problem and solve it by the fast iterative shrinkage thresholding algorithm (FISTA). As is well known, FISTA has a fast convergence rate, which is $O(1/m^2)$ in m iterations [3].

We evaluate our proposed method on both synthetic data and two benchmark video datasets in comparison with existing representative selection approaches such as K-mediod, sparse diction selection [11, 6], and density based selection [31]. The favourable results validate that our formulation fits better to the key video object summarization problem, and the selected proposals can better capture key video objects despite object appearance variations, background clutters and camera motions.

2. Related Work

Object-driven Video Summarization. Visual summaries of a video can take many forms such as keyframes [19, 20, 21], skims [15, 26, 28, 5], montages [34] and dynamic synopses [29, 30]. Recently there has been increasing interests in object-driven approaches to produce the above forms of summaries [29, 22, 21, 26, 4, 24, 40, 39, 43]. Some object-driven video summarization methods require prior knowledge. For instance, in [22, 36], frame-level labels are required to help identify the object of interest. By learning to predict important object regions in egocentric videos using egocentric and saliency cues, concise visual summaries for egocentric videos can be produced driven by those regions [21]. In [26] object-like windows (*i.e.*, object proposals) are taken from each frame as the initial pool of objects, based on which relationships between sub-events are discovered. However, these objects only act as an intermediate to help select sub-shots that construct a story. The ultimate goal is not to summarize videos into key objects and it is not fully unsupervised. Although [44] [45] can discover objects via topical models in an unsupervised way, it relies on image segments instead of object proposals. Also it targets at grouping the segments instead of selecting representative ones. Although [4] utilizes object proposals to address unsupervised discovery and localization of primary objects with multiple object classes, it targets for noisy image collections instead of a single video clip. Moreover, it only discovers a single object instance per image.

Representative Selection. Representative selection can be roughly categorized into clustering based methods and subspace learning based methods. As for clustering based methods, K-medoids algorithm is a representative one [18], which selects K clustering centroids as representatives to minimize within cluster distances. Instead of using one centroid to represent each sample, Elhamifar *et al.* improves K-medoids clustering, so that each sample is able to be represented by multiple centroids [10]. Representatives are also centroid-like in the methods of affinity propagation [12, 13] and density peak search [31]. There has also been recent interest in applying linear subspace learning to find representatives from data, *e.g.*, dictionary learning in [38, 25] and dictionary selection in [6, 11, 8, 23, 42]. These methods usually require that each sample can be linearly expressed by representatives at a low reconstruction error. To filter outliers, [11] ranks the representatives based on the norms of the rows in the coefficient matrix and only pick the top ones as final representatives. Another recent work [9] can find a subset of the source set to efficiently describe the target set, given pairwise dissimilarities between two sets. However, it does not consider the prior weight of the data samples in their applications.

3. Problem Formulation

We formulate video summarization using object proposals as the representative selection problem. Given n object proposals extracted from a video sequence, each of the object proposal can be represented by a feature vector $\in \mathbb{R}^d$. These feature vectors are arranged as the columns of the data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$. Our goal is to find a compact subset of the n data points that are representative of $\mathbf{X} \in \mathbb{R}^{d \times n}$.

3.1. Preliminaries: sparse dictionary selection [6]

Sparse dictionary selection was originally proposed for abnormal event detection [6]. Instead of learning a dictionary of arbitrary atoms, it requires that all atoms of the dictionary must come from the actual data points. In other words, the dictionary is a compact subset of the data matrix $\mathbf{X}$. Denote the selection matrix by $\mathbf{S} \in \mathbb{R}^{n\times n}$, it solves

$$\min_{\mathbf{S}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2+\lambda_1\|\mathbf{S}\|_{1,2}, \tag{1}$$

where $\|\mathbf{S}\|_{1,2} = \sum_{i=1}^{n}\|\mathbf{S}_{i,\cdot}\|_2$, associated with the regularization parameter λ_1, and $\|\mathbf{S}_{i,\cdot}\|_2$ is the l_2 norm of the i^{th} row of $\mathbf{S}$. Once (1) is solved by the proximal gradient method [3], the dictionary is constituted by selecting data points whose corresponding $\|\mathbf{S}_{i,\cdot}\|_2 \neq 0$.

Note that the selected data points can also be seen as representatives of the dataset. Consequently, their corresponding object proposals can be used to summarize the video. Similar to [11], we can adapt [6] to selecting any number of representatives by measuring and ranking the selection confidence of the i^{th} data point $\mathbf{x}_i$ according to $\|\mathbf{S}_{i,\cdot}\|_2$.

3.2. Weighted sparse dictionary selection

Note that in (1), all data points are treated equally. However, a good video summarization can certainly benefit from prior knowledge from application domain or user specifications. For instance, when summarizing egocentric videos, objects that the subject interact with are usually more important than others [21], while in surveillance videos from a fixed camera, moving foreground objects likely carry more weights than static background objects. To better leverage priors, we propose a simple extension to [6] for representative selection

$$\min_{\mathbf{S}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2+\lambda_1\sum_{i=1}^{n}\frac{1}{\rho_i+\epsilon}\|\mathbf{S}_{i,\cdot}\|_2, \tag{2}$$

where ρ_i is the prior selection weight for the i^{th} sample, and ϵ is a tiny number to avoid dividing by zero.

Similar to [6], the problem of weighted sparse dictionary selection can also be optimized by the proximal gradient iteration [3], but needs a proximal decomposition [46].

3.3. Locally linear reconstruction (LLR) induced sparse dictionary selection

As indicated in [11], sparse dictionary selection prefers keeping the vertices of convex hull spanned by input data to make sure each sample can be reconstructed at a low cost. It is thus extremely sensitive to noise. To mitigate this issue, we encourage local reconstruction for each sample to improve the robustness of representative selection.

3.3.1 Locality prior of linear reconstruction

Inspired by locally linear embedding [32], we build a locality prior matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ of representative selection based on that each sample $\mathbf{x}_i$ is only allowed to be locally linear reconstructed by its k-NNs, $\mathcal{N}(\mathbf{x}_i)\setminus\mathbf{x}_i$:

$$\begin{aligned}\min_{\mathbf{W}\in\mathbb{R}^{n\times n}} \quad & \tfrac{1}{2}\textstyle\sum_{i=1}^{n}\|\mathbf{x}_i-\sum_{j:\mathbf{x}_j\in\mathcal{N}(\mathbf{x}_i)\setminus\mathbf{x}_i} w_{ji}\mathbf{x}_j\|_2^2,\\ \text{s.t.} \quad & \sum_{j:\mathbf{x}_j\in\mathcal{N}(\mathbf{x}_i)\setminus\mathbf{x}_i} w_{ji} = 1,\\ & w_{ji} = 0, \forall \mathbf{x}_j\notin\mathcal{N}(\mathbf{x}_i)\setminus\mathbf{x}_i.\end{aligned} \tag{3}$$

Problem (3) can be solved by a constrained least squares optimization [32].

3.3.2 LLR-induced sparse dictionary selection

To introduce locality information of data for representative selection, we propose a new optimization problem combined with the locality prior martix $\mathbf{W}$ in the following:

$$\min_{\mathbf{S}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2+\lambda_1\sum_{i=1}^{n}\frac{1}{\rho_i+\epsilon}\|\mathbf{S}_{i,\cdot}\|_2+\lambda_2\|\mathbf{S}-\mathbf{W}\|_{\mathrm{F}}^2, \tag{4}$$

where the third term regularizes the selection matrix $\mathbf{S}$ by $\mathbf{W}$, and λ_2 is a locality regularization parameter. Hence, data samples are preferable to be reconstructed by nearby representatives. Moreover, dense samples are more likely to be selected as representatives than sparse noise, as the former can contribute more to the reconstruction of surrounding samples in comparison with the latter.

Problem (4) is complex due to three optimization terms. But we will show that it can be converted into a proximal gradient optimization and solved by the FISTA method [3] through a proximal decomposition [46], which converges fast with rate $O(1/m^2)$ in m iterations.

3.3.3 Optimization

To solve our optimization problem (4), we first expand the objective function ($\mathcal{O}$ for short) and rewrite it as

$$\begin{aligned}\mathcal{O} =& \frac{1}{2}\mathrm{tr}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}-2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda_2\mathbf{W}^T\right)\mathbf{S}\\ &+\mathbf{S}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda_2\mathbf{I}\right)\mathbf{S}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\}\\ &+\lambda_1\sum_{i=1}^{n}\frac{1}{\rho_i+\epsilon}\|\mathbf{S}_{i,\cdot}\|_2.\end{aligned} \tag{5}$$

We then let

$$\begin{aligned}f(\mathbf{S}) =& \frac{1}{2}\mathrm{tr}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}-2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda_2\mathbf{W}\right)\mathbf{S}\\ &+\mathbf{S}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda_2\mathbf{I}\right)\mathbf{S}+\mathbf{W}^{\mathrm{T}}\mathbf{W}\},\end{aligned} \tag{6}$$

and

$$g(\mathbf{S}) = \lambda_1\sum_{i=1}^{n}\frac{1}{\rho_i+\epsilon}\|\mathbf{S}_{i,\cdot}\|_2. \tag{7}$$

Thus, we decompose the objective function $\mathcal{O}$ into two convex functions, with f smooth and g nonsmooth, *i.e.*,

$$\mathcal{O} = f(\mathbf{S})+g(\mathbf{S}), \tag{8}$$

Algorithm 1 LLR-induced Weighted Sparse Dictionary Selection (4).

Input: $\mathbf{X}, \{\rho_i\}_{i=1}^n, k, \lambda_1, \lambda_2$
Output: $\mathbf{S}$

$L \leftarrow \lambda_2 + r\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$
▷ Lipschitz constant (Equation (11))
$\mathbf{S} \leftarrow \mathbf{0}, \mathbf{V} \leftarrow \mathbf{S}, t \leftarrow 1$
▷ Initialization
repeat
  $\mathbf{Z} \leftarrow \mathbf{V} + \frac{1}{L}\left\{-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{W}\right) + \left(\lambda_2\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\right\}$
  ▷ Equation (10)
  $\mathbf{U} \leftarrow \mathbf{S}, \mathbf{S}_{i,\cdot} \leftarrow \mathbf{Z}_{i,\cdot}\max\{(1 - \frac{\frac{\lambda_1}{L(\rho_i+\epsilon)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}, i = 1, 2, \cdots, n$
  ▷ Equation (13)
  $\tau = t - 1, t \leftarrow (1 + \sqrt{1 + 4t^2})/2$
  $\mathbf{V} \leftarrow \mathbf{S} + \tau(\mathbf{S} - \mathbf{U})/t$
until convergence

for which, we can apply the proximal gradient method, FISTA [3]. It then becomes iteratively solving

$$\mathrm{prox}_{\mathcal{R}}\left(\mathbf{Z}\right) = \underset{\mathbf{S}\in\mathbb{R}^{n\times n}}{\arg\min}\ \frac{1}{2}\left\|\mathbf{S} - \mathbf{Z}\right\|_{\mathrm{F}}^2 + \frac{1}{L}g(\mathbf{S}), \tag{9}$$

where

$$\begin{aligned}\mathbf{Z} &= \mathbf{S} - \frac{1}{L}\frac{\partial}{\partial\mathbf{S}}f(\mathbf{S}) \\ &= \mathbf{S} + \frac{1}{L}\left\{-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{W}\right) + \left(\lambda_2\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{S}\right\},\end{aligned} \tag{10}$$

and L is the smallest Lipschitz constant, which equals to the spectral radius ($r(\cdot)$) of $\lambda_2\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}$, *i.e.*,

$$L = r\left(\lambda_2\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = \lambda_2 + r\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right). \tag{11}$$

We next follow the decomposition tactic in [46], then Problem (9) is solvable, and for $i = 1, 2, \cdots, n$,

$$\mathbf{S}_{i,\cdot} = \underset{\mathbf{s}\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\left\|\mathbf{s} - \mathbf{Z}_{i,\cdot}\right\|_2^2 + \frac{\lambda_1}{L(\rho_i + \epsilon)}\|\mathbf{S}_{i,\cdot}\|_2, \tag{12}$$

After applying soft-thresholding [41] to the above n group lasso signal approximators, we have, for $i = 1, 2, \cdots, n$,

$$\mathbf{S}_{i,\cdot} = \mathbf{Z}_{i,\cdot}\max\{(1 - \frac{\frac{\lambda_1}{L(\rho_i+\epsilon)}}{\|\mathbf{Z}_{i,\cdot}\|_2}), 0\}. \tag{13}$$

We show the representative selection procedure in Algorithm 1, where we integrate a decomposed soft-thresholding strategy into an accelerated proximal gradient procedure, which is known to have a fast convergence rate $O(1/m^2)$ in m iterations.

3.3.4 Parameter setting

Sparsity regularization parameter λ_1. As in Algorithm 1, we initialize $\mathbf{S}$ by a zero matrix. Then according to (10), after the first iteration, we have

$$\mathbf{Z} = \frac{1}{L}\left\{-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{W}\right)\right\}. \tag{14}$$

As indicated by the thresholding of $\mathbf{Z}$ in (13), when λ_1 is large enough, *e.g.*, $\lambda \geq \lambda_1^{\max}$, we obtain $\mathbf{S} = \mathbf{0}$, which means we select nothing. Therefore, to avoid an empty selection, we let $\lambda_1 \leq \lambda_1^{\max}$ and solve $\lambda_1^{\max}$ by substituting $\mathbf{S} = \mathbf{0}$ into (13) as follows:

$$\begin{aligned}\lambda_1^{\max} &= L\max_{0\leq i\leq n}\left\{(\rho_i + \epsilon)\|\mathbf{Z}_{i,\cdot}\|_2\right\} \\ &= \max_{0\leq i\leq n}\left\{(\rho_i + \epsilon)\|\mathbf{x}_i^{\mathrm{T}}\mathbf{X} + \lambda_2\mathbf{W}_{i,\cdot}\|_2\right\}.\end{aligned} \tag{15}$$

In our experiments, we let $\lambda_1 = \frac{\lambda_1^{\max}}{\alpha_1}$ and tune α_1 between the interval $[2, 30]$. Given λ_2, a smaller α_1 indicates a larger λ_1, which implies a sparser selection.

Locality regularization parameter λ_2. Let us consider the LLR-induced sparse selection in (4). When $\lambda_2 = 0$, the problem becomes a weighted sparse dictionary selection as in (2). When $\lambda_2 = +\infty$, it sparsely selects representatives based on the rows of $\mathbf{W}$. Furthermore, as shown in (10), λ_2 balances the contributions of $\mathbf{W}$ and $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ to the proximal operation in (9). For ease of tuning λ_2, we let

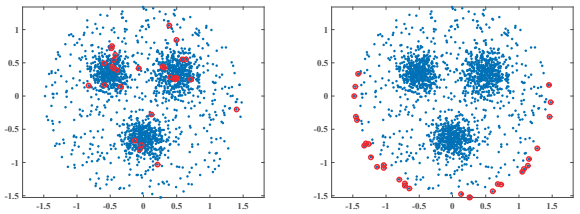
$$\lambda_2 \leftarrow \kappa \times \frac{r(\mathbf{X}^{\mathrm{T}}\mathbf{X})}{r(\mathbf{W})}, \tag{16}$$

where we set $\kappa = 0.02 \times 5^{\alpha_2}$ and α_2 between $[-3, 1]$ in our experiments.

4. Experiments on Synthetic Data

We first evaluate the effectiveness of the locality prior in handling outliers on synthetic data, in comparison with the Sparse Dictionary Selection (SDS) [6] and Sparse Modeling Representative Selection (SMRS) [11]. We refer to our proposed method as Locally Linear Reconstruction induced Sparse Dictionary Selection (LLR-SDS).

We consider the noisy data shown in Fig. 2, which consists of data points in three clusters and uniform background noise. The top 30 representatives found from the $2,018$ points by each compared method are shown in Figs. 2 (a)-(c). As can be seen, both SDS and SMRS select the outlier points at the border of the convex hull, showing these two methods are sensitive to noise. This is because those points contribute a lower linear reconstruction cost to the dataset than others, which meets the requirement of dictionary selection. As SMRS has a post-processing to filter outliers [11], we also run its outlier removal for comparison in Fig. 2 (d). Since most selection of SMRS are outliers as shown in Fig. 2 (c), it is difficult to improve the results by post-processing. In contrast, our proposed LLR-SDS method considers a locality constraint in addition to the linear reconstruction cost for dictionary selection. As a result, it can reject most noisy outliers and select the points in the clusters. For our method, we use $k = 5$ nearest neighbors to build the locality prior matrix $\mathbf{W}$, and set the sparsity regularization parameter $\alpha_1 = 5$ for λ_1, and the locality



(a) LLR-SDS (b) SDS [6]

(c) SMRS [11] (d) SMRS with outlier removal [11]

Figure 2: Data points in three clusters with background uniform noise (blue dots) and the representatives (red circles) found by (a) the proposed LLR-SDS method (b) SDS [6], (c) SMRS [11] and (d) SMRS with outlier removal [11] .

regularization parameter $\alpha_2 = 1$ for λ_2. The same sparsity regularization parameter is used for SDS and SMRS.

5. Experiments on Real Datasets

Next, we test our approach on two real datasets to evaluate its applicability to topical object discovery and object search in videos.

5.1. Settings

Although our approach is agnostic to the object proposal method, we use Edge Boxes [47] with default settings to generate category independent object proposals for a controlled comparison with existing methods on all datasets. In light of the good performance reported using features from the top layers of deep convolutional neural networks (CNNs) for many computer vision tasks such as object detection [14], image retrieval [2] and object instance search [35], we choose to represent each object proposal by the CNN feature after dimension reduction. Specifically, we take the 4096-dimensional output from the fully connected layer 6 of Caffe [17], using the pre-trained model on ILSVRC 2012 without fine-tuning. The CNN features are then reduced to 256-dimension by PCA and whitening [16]. The same set of CNN-PCA features are used for comparison with existing work unless noted otherwise.

5.2. Baseline algorithms

We compare our proposed LLR-SDS with a variety of different methods for representative selection, including Objectness [47], K-medoids [18], Density [31], Sparse Modeling Representative Selection (SMRS) [10], Sparse Dictionary Selection (SDS) [6], Locally Linear Reconstruction (LLR) and Latent Dirichelet Allocation with Word Co-occurrence prior (LDA-WCP) [44].

Objectness refers to directly ranking object proposals based on the objectness scores from Edge Boxes. For Density, representatives are ranked and selected according to the parameter γ, which is the product of local density and minimum distance between the point and any other point with higher density [31]. For SMRS, we follow the authors to tune its parameter α, which is similar to α_1 of our method. For a fair comparison, results reported are without the outlier detection post-processing. SDS refers to ranking and selecting object proposals according to the l_2 norms of the rows of the selection matrix $\mathbf{S}$ obtained from the algorithm proposed in [6] (Sec. 3.1). LLR refers to selecting object proposals according to the l_2 norms of the rows of the LLR prior matrix $\mathbf{W}$ (Sec. 3.3.1). Therefore, it favors data points with sufficient density over the outliers. For LDA-WCP, we adapt it to object proposal selection by selecting the highest score segment in the entire video for each topic to summarize the video.

5.3. Evaluation Metric

The effectiveness of all representative selection methods are evaluated by the average recall. Denote the set of selected object proposals from a video as $\mathbf{P}$ and assume a video contains t different key objects. For the i_{th} key object, denote the set of ground truth bounding boxes in all keyframe as $\mathbf{G}_i$, and the best intersection over union (IoU) score $\mathbf{S}$ with the i_{th} key object is defined as

$$\mathbf{S}(\mathbf{P}, \mathbf{G}_i) = \max_{\substack{p \in \mathbf{P} \\ g \in \mathbf{G}_i}} \mathbf{S}(p, g) = \max_{\substack{p \in \mathbf{P} \\ g \in \mathbf{G}_i}} \frac{p \cap g}{p \cup g} \tag{17}$$

The recall of a video is determined by the number of key objects that are recalled, i.e., $\frac{\sum_{i=1}^{t} \mathbb{I}(\mathbf{S}(\mathbf{P},\mathbf{G}_i) >= \theta)}{t}$, where θ is the overlap threshold and $\mathbb{I}(\cdot)$ is the indicator function. The average recall is the mean of the recall of all videos.

5.4. Topical Video Object Discovery

We first demonstrate the effectiveness of the proposed LLR-SDS in summarizing multiple topical objects in videos. We run our experiments on the "multiple" object subset of the topical video object discovery dataset [44], which consists of 10 commercial video sequences from YouTube. Each video contains multiple well-defined topical objects such as the product logos and has multiple shots. As in [44], keyframes from each video are sampled at two frames per second. For each keyframe, we take the top 100 object proposals according to the objectness score [47] and summarize from them. The overlap threshold θ is set to 0.5.

Note that LDA-WCP requires a predefined number of topics for each video, which is set to 8 for all videos in [44]. Hence, for a fair comparison, we also select 8 object proposals by each of the other methods to summarize a video.

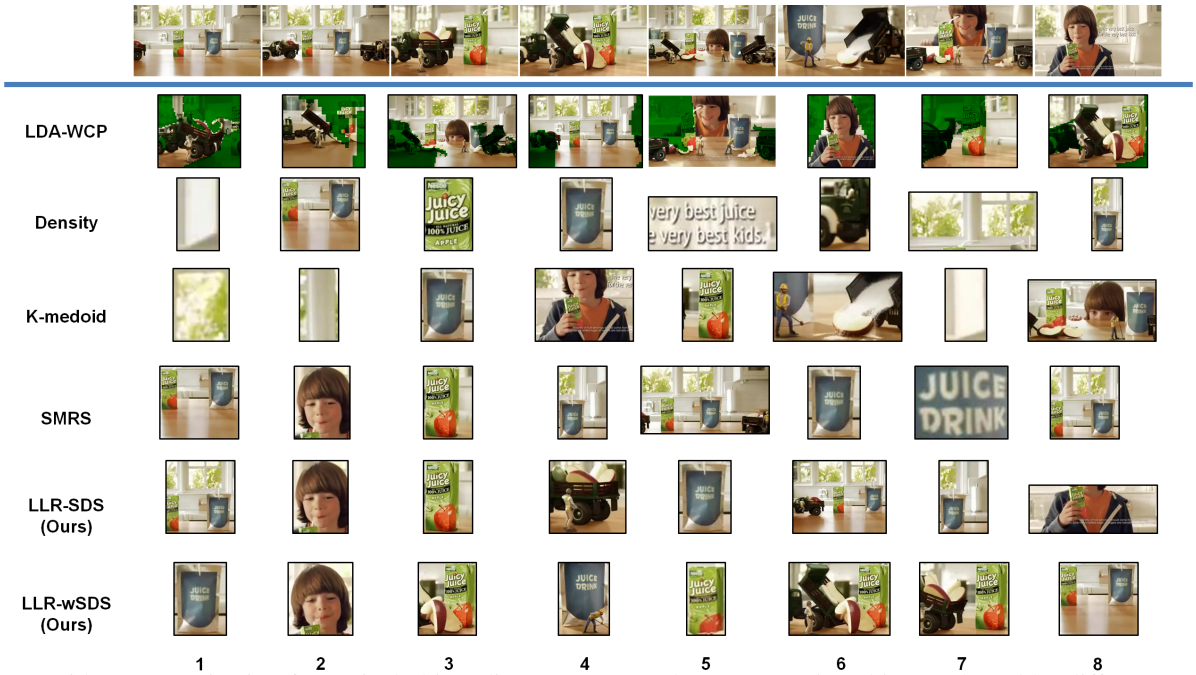


Figure 3: Video summarization for topical object discovery: example representative objects selected by different methods. Keyframes are shown in the top row. Each following row shows the 8 representative object proposals selected by different methods in rank order (except for K-medoids and LDA-WCP).

	Average Recall@8
Objectness [47]	0.173
LLR	0.243
Density [31]	0.360
LDA-WCP [44]	0.360
K-medoids	0.410
SDS [6]	0.417
SMRS [11]	0.430
LLR-SDS (Ours)	0.430
LLR-wSDS (Ours)	0.547

Table 1: Average recall when selecting 8 object proposals. Methods are sorted by increasing average recall@8. Our method with objectness as prior weights (LLR-wSDS) achieves the best average recall among all.

SMRS is tested on a range of $\alpha \in \{2, 3, 5, 10, 20, 30\}$ and we report the best average recall obtained. For our proposed LLR-SDS, we fix $k = 3$ nearest neighbors to construct the LLR prior matrix $\mathbf{W}$ (Sec. 3.3.1), $\alpha_1 = 2$ for λ_1 and $\alpha_2 = -1$ for λ_2 for all videos (Sec. 3.3.4). The same $k = 3$ is used for LLR and the same $\alpha_1 = 2$ is used for SDS. We also evaluate the effectiveness of prior weights by simply using the objectness score from Edge Boxes [47] as ρ_i (Eq. 4) for each object proposal. We refer to our method with objectness weights as LLR-wSDS.

Table 1 compares our approach with other methods in terms of the average recall of all videos. Without objectness prior weights, our proposed LLR-SDS and SMRS achieve the same highest average recall of 0.43. Although using the objectness score directly for summarization performs poorest among all, integrating it into our LLR-SDS formulation as prior weights further improves LLR-SDS and SMRS by 27.2%. A close examination of the Objectness results reveals that since object proposals are highly redundant across frames, an object proposal that is scored highest in one frame usually scores the highest in other frames as well. Therefore when selecting few (*i.e.*, top 8) object proposals purely based on the objectness scores, multiple snapshots of one or two dominant object(s) are often picked out but the other topical objects are missed. Note that LDA-WCP is based on quantized SIFT features, while the others except for Objectness are based on CNN features. This could account for the mediocre performance of LDA-WCP.

Fig. 3 shows an example video with representative object proposals selected by different methods. The 8 object proposals selected by each method are displayed in rank order in each row, except for LDA-WCP and K-medoids, which produce no ranks for the selection. There are 5 topical objects in this commercial: the blue juice drink, the green juicy juice, the boy, and 2 toy trucks. Both LLR-SDS and LLR-wSDS capture 4 out of 5 (missing one of the toy trucks) using only 8 proposals. With objectness weights, LLR-wSDS seems to improve the ranking of object proposals with more accurate coverage of the ground truth over LLR-SDS. For instance, the 1_{st} object proposal selected by LLR-wSDS provides a more accurate coverage of the blue juice drink than the 1_{st} object selected by LLR-SDS.

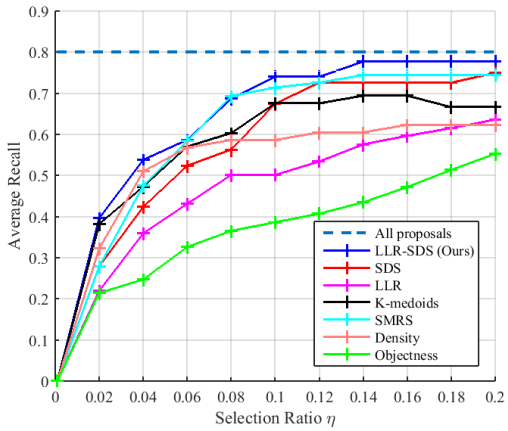


Figure 4: Average recall with selection ratio $\eta \in [0.001, 0.2]$.

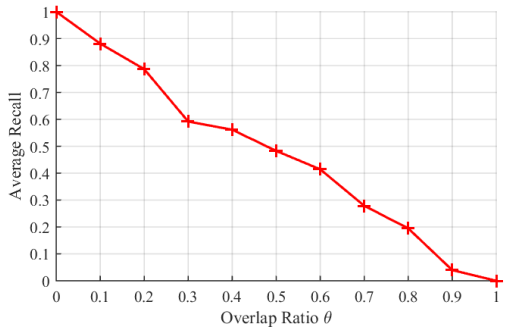


Figure 5: Evaluation of the relationship between the average recall and overlap ratio θ ($\eta = 0.01$).

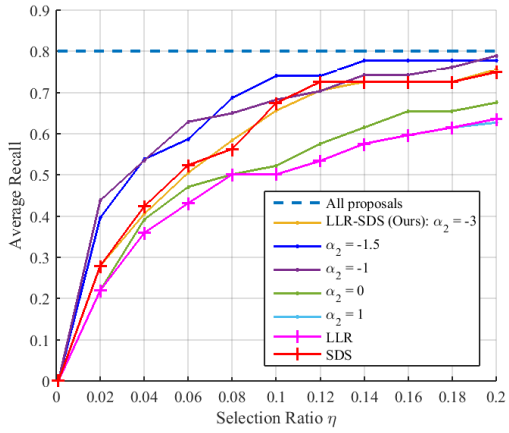


Figure 6: Average recall when the regularization parameter $\alpha_2 \in [-3, 1]$. A small α_2 (*i.e.*, -3) leads to a small λ_2, which produces a curve similar to that of SDS; while $\alpha_2 = 1$ leads to a solution similar to that of LLR (Sec. 3.3.4).

5.5. Video Summarization for Object Search

Next we evaluate the applicability of LLR-SDS to visual object search by summarizing shots of a feature movie by object proposals. We take a benchmark dataset for object retrieval, the Groundhog Day [33], which consists of $5,640$ keyframes (752 shots) from the full length feature film "Groundhog Day". The ground truth test set consists of 6 object queries. Contrary to the Topical Video Object Discovery dataset, the target objects in this dataset are usually small and the scenes are much cluttered.

For each keyframe, we extract the top 200 object proposals with an aspect ratio $\in [0.25, 4]$, and a minimal size of 30×30 pixels. Because object locations in the keyframes are not provided, we manually annotate bounding box locations in all ground truth keyframes. We run our algorithm on all ground truth shots that have ≥ 8 keyframes (Table 2). SDS is tested in a range of $\alpha \in \{2, 3, 5, 10, 15, 20, 30\}$ and $\alpha = 10$ is selected for comparison because it produces the best average recall. For our LLR-SDS, we fix $k = 5$ nearest neighbors to construct the LLR prior matrix $\mathbf{W}$ and $\alpha_1 = 15$ for calculating λ_1. We have also tested the objectness score as prior weights as in Sec. 5.4, which, however, does not boost the performance on this dataset. It is likely due to the scene clutters and generally low resolutions of the target objects in the Groundhog Day dataset. Therefore, we evaluate LLR-SDS using equal weights on this dataset. It is worth noting though that other priors could be effective on this dataset such as object size and scene context. Unless noted otherwise, we set the overlap threshold θ to 0.7 in all following experiments (Eq. 17), as a greater overlap with the ground truth generally leads to a higher matching score and increases the chance for an object to be retrieved.

5.5.1 Results

Fig. 4 illustrates the average recall on all ground truth shots of the six objects with the selection ratio $\eta \in [0.001, 0.2]$. It is shown that the proposed LLR-SDS outperforms the other methods and achieves an average recall of 0.74 with as few as 10% of the object proposals. With 14% of the object proposals, it achieves an average recall of 0.78, while the average recall obtained using all object proposals is 0.80. It also improves upon selection by SDS or LLR alone. In addition, Fig. 5 shows the relationship between the average recall and overlap ratio θ of our proposed LLR-SDS, when $\eta = 0.01$.

To evaluate the sensitivity of α_2 for λ_2 (Sec. 3.3.4), we test LLR-SDS with $\alpha_2 \in [-3, 1]$. Fig. 6 plots the average recall curves obtained with respect to the two extreme cases discussed in Sec. 3.3.4: *i.e.*, (1) $\lambda_2 = 0$ and (2) $\lambda_2 = +\infty$. The former is equivalent to the sparse dictionary selection (SDS) (Eq. 2), while the latter produces a selection matrix

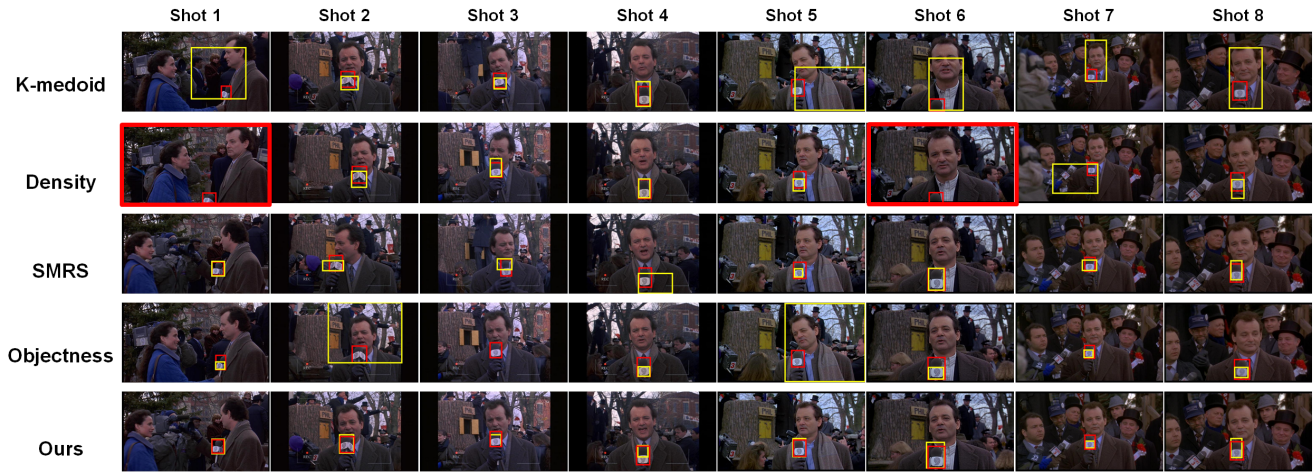


Figure 7: Video summarization for object search: visual results of the 8 shots of Microphone from Groundhog Day ($\eta = 0.5\%$ of all object proposals are selected). For each shot, the object proposal with the best IoU with the ground truth (GT) is shown in the yellow bounding box and the GT is in red. Missed shots are highlighted by red rectangles. Overall our method produces summarizations that more accurately capture the GT. Note that given a shot, the best IoU object proposal selected by different methods may come from different keyframes. Best viewed in color and magnification.

$\mathbf{S}$ similar to the LLR prior matrix $\mathbf{W}$, where we rank and select object proposals by the l_2 norms of the rows of $\mathbf{W}$ (LLR). We fix $\alpha_1 = 15$ for LLR-SDS and SDS, and $k = 5$ for LLR-SDS and LLR. It is observed that when $\alpha_2 \leq -3$, the recall curve of LLR-SDS converges to that of SDS. On the other hand, when $\alpha_2 \geqslant 0$, the recall curve of LLR-SDS almost entirely overlaps with that of LLR. Experimentally, on the Groundhog dataset, LLR-SDS achieves higher average recall than either SDS or LLR when $\alpha_2 \in [-2, -0.5]$.

We further evaluate the effectiveness of LLR-SDS in terms of the percentage of proposals required to provide accurate localization of the ground truth object, in comparison with SDS and LLR. An object proposal is considered to accurately locate the ground truth if its IoU with the ground truth $\geqslant 0.7$ or it achieves the best IoU among all object proposals in a video. Table 2 shows that on average, with as few as 6.60% of all object proposals, LLR-SDS is able to cover the object of interest when summarizing a shot. Except for the Frames Sign, LLR-SDS requires fewer object proposals than both SDS and LLR to ensure accurate localization of the ground truth, while LLR requires the most.

Fig. 7 shows visual comparisons of our results with others on all shots of Microphone, when selecting $\eta = 0.5\%$ of all object proposals. For each shot, we visualize the object proposal that has the highest IoU with the ground truth among all selected. In general, our method produces summarizations that more accurately locate the ground truth than others in all 8 shots of Microphone.

Object	#shots	SDS	LLR	LLR-SDS
Red clock	9	23.31 %	38.64 %	**12.48%**
Black clock	8	15.86%	16.68%	**12.99%**
Frames sign	6	**1.65%**	47.88%	2.25%
Digital clock	14	7.92%	9.20%	**5.49%**
Phil sign	9	6.51%	23.18%	**5.43%**
Microphone	8	7.13%	4.77%	**0.97%**
Average	9	10.40%	23.39%	**6.60%**

Table 2: Average percentage of object proposals required to cover the ground truth object.

6. Conclusions

In this work we summarize videos into key objects using object proposals. These key objects can serve as video icons and establish a brief impression of the video. By telling what objects appear in each video, we can use these key objects to help search, browse, and index large video volume. To select key objects, we propose a new formulation of sparse dictionary selection to select representative object proposals, *i.e.*, locally linear reconstruction induced sparse dictionary selection (LLR-SDS). The new formulation considers both object proposal priors and locality priors in the feature space thus can better handle outlier proposals when identifying the key objects. Our results on synthetic data and two benchmark real datasets validate the advantages of our approach in comparison with existing representative selection methods.

Acknowledgements. 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 National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012. 1

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*. 2014. 5

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIIMS*, 2(1):183–202, 2009. 2, 3, 4

[4] 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

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

[6] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 2011. 2, 3, 4, 5, 6

[7] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 1

[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. pages 19–27, 2012. 2, 5

[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, 3, 4, 5, 6

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

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

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 5

[15] 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. 2

[16] H. Jégou and O. Chum. Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening. In *ECCV*. 2012. 5

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[18] 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

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

[20] 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

[21] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 1, 2, 3

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

[23] 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

[24] C. Lu, R. Liao, and J. Jia. Personal object discovery in first-person videos. *TIP*, 2015. 1, 2

[25] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 2

[26] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 1, 2

[27] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013. 1

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

[29] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. In *ICCV*, 2007. 2

[30] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *CVPR*. 2

[31] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 2014. 2, 5, 6

[32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 3

[33] J. Sivic and A. Zisserman. Efficient visual search of videos cast as text retrieval. *TPAMI*, 2009. 7

[34] M. Sun, A. Farhadi, B. Taskar, and S. Seitz. Salient montages from unconstrained videos. In *ECCV*. 2014. 2

[35] R. Tao, E. Gavves, C. Snoek, and A. Smeulders. Locality in generic instance search from one example. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2099–2106, June 2014. 5

[36] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *ECCV*. 2014. 2

[37] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *ICCV*, 2015. 1

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

[39] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, and J. Brandt. Discovering primary objects in videos by saliency fusion and iterative appearance estimation. *T-CSVT*, 2015. 2

[40] 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

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

[42] L. Zhang, Y. Xia, K. Mao, H. Ma, and Z. Shan. An effective video summarization framework toward handheld devices. *Industrial Electronics, IEEE Transactions on*, 2015. 2

[43] G. Zhao and J. Yuan. Discovering thematic patterns in videos via cohesive sub-graph mining. In *ICDM*, pages 1260–1265. IEEE, 2011. 2

[44] G. Zhao, J. Yuan, and G. Hua. Topical video object discovery from key frames by modeling word co-occurrence prior. In *CVPR*, 2013. 2, 5, 6

[45] G. Zhao, J. Yuan, G. Hua, and J. Yang. Topical video object discovery from key frames by modeling word co-occurrence prior. *TIP*, 24(12):5739–5752, Dec 2015. 2

[46] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *KDD*, pages 1095–1103, 2012. 3, 4

[47] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 1, 5, 6