Discovering Thematic Objects in Image Collections and Videos

Junsong Yuan, *Member, IEEE*, Gangqiang Zhao, *Member, IEEE*, Yun Fu, *Senior Member, IEEE*, Zhu Li, *Senior Member, IEEE*, Aggelos K. Katsaggelos, *Fellow, IEEE*, and Ying Wu, *Senior Member, IEEE*

***Abstract*—Given a collection of images or a short video sequence, we define a thematic object as the key object that frequently appears and is the representative of the visual contents. Successful discovery of the thematic object is helpful for object search and tagging, video summarization and understanding, etc. However, this task is challenging because 1) there lacks *a priori* knowledge of the thematic objects, such as their shapes, scales, locations, and times of re-occurrences, and 2) the thematic object of interest can be under severe variations in appearances due to viewpoint and lighting condition changes, scale variations, etc. Instead of using a top–down generative model to discover thematic visual patterns, we propose a novel bottom–up approach to gradually prune uncommon local visual primitives and recover the thematic objects. A multilayer candidate pruning procedure is designed to accelerate the image data mining process. Our solution can efficiently locate thematic objects of various sizes and can tolerate large appearance variations of the same thematic object. Experiments on challenging image and video data sets and comparisons with existing methods validate the effectiveness of our method.**

***Index Terms*—Image data mining, thematic object discovery.**

I. Introduction

GIVEN a collection of images or a video sequence, can we discover the thematic objects that are representative of the visual contents? As two examples shown in Fig. 1, from a collection of web images sharing the same tag of "Oxford Museum," it is of great interest to locate the thematic object, i.e., the Oxford Museum; another example is the discovery of thematic object in a commercial video. Solving this emerging image data mining problem will benefit a number of applications such as video summarization [1], [2], visual object search and detection [3], [4], and image cosegmentation [5]–[7].

Fig. 1. Examples of thematic objects. (Top) Finding the "Oxford Museum" in several web images that are labeled with this tag. (Bottom) Finding the product object in a commercial video.

The thematic object must be the common object that is frequently highlighted in the visual scene. To automatically discover thematic objects, however, there are two major challenges. First of all, there lacks *a priori* knowledge of the thematic visual pattern, thus not known in advance 1) the shapes and appearances of the thematic objects; 2) the locations and scales of thematic objects; and 3) the total number of thematic objects. Moreover, the same thematic object can look quite different when presented from different viewpoints, scales, or under different lighting conditions, not to mention partial occlusions. It is not trivial to handle its variations and accurately locate its occurrences. Although invariant local features greatly improve image matching, accurate localization of thematic objects remains a challenging problem.

Motivated by the previous success in mining text data, one popular solution to image data mining is to transfer an image into a "visual document" by clustering the local visual features into "visual words." Then, traditional text mining methods can be directly applied to image data. Despite moderate successes of this approach, as images significantly differ from texts, such a solution has several limitations. First of all, unlike text data that are composed of a finite vocabulary, images are usually characterized by high-dimensional features, thus having a considerably higher uncertainty than text. The visual vocabulary inevitably introduces quantization errors in characterizing the local features, thus affecting the matching accuracy. Moreover, image patterns have spatial descriptions, whereas text data do not. Thus, unlike text analysis, the spatial configuration among visual primitives should not be ignored. Finally, although the bag-of-words scheme has been successfully applied to discover object categories from images [8]–[12], it is less suitable to discover thematic objects in a very limited number of images, for example, several or tens of images. In such a case, the visual document representation is less effective due to the small

Manuscript received August 05, 2010; revised February 06, 2011, July 18, 2011 and December 03, 2011; accepted December 03, 2011. Date of publication December 26, 2011; date of current version March 21, 2012. This work was supported in part by the Nanyang Assistant Professorship (SUG M5804001) to Dr. J. Yuan. Dr. Y. Fu was supported in part by Futurewei (Huawei) Technologies Inc. and the Intelligence Community Postdoctoral Research Fellowship under Award 2011-11071400006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xuelong Li.

J. Yuan and G. Zhao are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jsyuan@ntu.edu.sg; gqzhao@ntu.edu.sg).

Y. Fu is with the Department of Computer Science and Engineering, University at Buffalo–State University of New York, Buffalo, NY 14260-2500 USA (e-mail: yunfu@buffalo.edu).

Z. Li is with the Media Networking Laboratory, Core Networks Research, Futurewei (Huawei) Technologies, Bridgewater, NJ 08807 USA (e-mail: zhu.li@ieee.org).

A. K. Katsaggelos and Y. Wu are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: aggk@eecs.northwestern.edu; yingwu@ece.northwestern.edu).

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

Digital Object Identifier 10.1109/TIP.2011.2181952

Fig. 2. Data mining steps of discovering the thematic object. (1) First row: Each ellipse represents an extracted local feature. (2) Second row: Red " + " represents the local region with a positive detection score, whereas green "o" represents the local region with a negative detection score. (3) Third row: Find a bounding box to crop the thematic object (shown in both second and third rows) of maximum score. The images are from a public data set in [13].

number of visual primitives for training the visual vocabulary. Thus, alternative methods should be considered.

To address the above problems, we present a novel bottom–up approach for thematic object discovery. Our emphasis is on the accurate localization of the objects despite the background clutters. Fig. 2 illustrates the major steps of our method. First, each image is characterized by a collection of local features, which we referred to as *visual primitives*. We match visual primitives and gradually expand them spatially to recover the whole thematic object. In the initialization phase, "uncommon" visual primitives that are of limited matches in other images are discarded because they will not belong to any common pattern, i.e., the thematic object. For each remained visual primitive, we consider its local spatial neighborhood as a larger *visual group* and check the commonness score of this spatial pattern. Following multilayer commonness checking of different spatial scales, each local feature is finally assigned a commonness score, which indicates its likelihood of belonging to a thematic object. The commonness score of any subimage is the summation of the scores of its local features. By searching the subimage of highest commonness score in each image, we can locate and crop the thematic object. Unlike top–down generative models that rely on a visual vocabulary for topic discovery, our method only requires matching of visual primitives. It can automatically discover and locate several thematic objects without requiring knowledge of the total number of such objects in advance. The scales and locations of thematic objects are also automatically determined. Moreover, with the help of local invariant features, it can handle object variations such as scale and slight point-of-view changes, color and lighting condition variations, and it is insensitive to partial occlusion. Finally, it does not require a large number of images for data mining, and it works well to detect thematic object in a very limited number of images.

The remainder of this paper is organized as follows. We discuss related work in Section II and then explain our proposed method in Section III, followed by the description of the further acceleration of each pruning step in Section IV. Experimental results are presented in Section V, and we conclude this paper in Section VI.

II. Related Work

To discover common visual patterns in images, some previous works characterize an image as a graph composed of visual primitives, such as corners, interest points, and image segments [14]–[18], and perform graph-based matching for object discovery.

Tan and Ngo [14] use a color histogram to describe image segments. A common object is discovered by finding the maximum flows between two images. In order to consider the spatial relations among these visual primitives, Hong and Huang [15] apply an attribute relational graph model, where each visual primitive denotes a vertex of the graph, and spatial relations among visual primitives are represented as edges of the graph. Common object discovery is formulated as the common subgraph discovery problem, where each graph specifies an image. However, the proposed EM algorithm to solve the graph matching problem relies on its initialization and does not guarantee a globally optimal solution [14], [15].

Todorovic and Ahuja [16] represent an image as a tree of multiscale image segments, where each segment corresponds to a tree node. Common patterns are discovered by finding the maximally matching subtrees among the image set. By fusing the discovered subtrees into a tree union, a canonical category model can be obtained to detect and segment the object from a new image. Liu and Yan [18] employ spatially coherent constraints for graph matching. However, the proposed solution is specifically designed for a pair of images, and not for a collection of images. In general, for the graph-based representation of images, matching and mining subgraphs are computationally demanding.

To accelerate the data mining process, Yuan et al. [19] and Zhao and Yuan [20] propose a bottom–up candidate pruning method for common object discovery. Yuan and Wu [21] propose a randomized algorithm to discover common objects. Using multiple spatial random partitions, it can discover and locate multiple common objects in a collection of images. Another randomized algorithm is proposed in the work of Chum and Matas [22] to cluster spatially overlapping images. It proposes to apply the min-Hash algorithm for fast detection of pairs of images with spatial overlap, and then find the clusters of images that have partial overlaps. However, as the goal is to find image clusters, it does not require to locate the common object. Heath et al. [23] build the image web by mining common regions between any pair of images using affine cosegmentation. It also focuses on mining the connectivity of large-scale image collections. Bagon et al. [24] propose a method that can detect the common object in a very few number of images, where the discovered common object is depicted by a binary sketch. Because the common object can significantly vary, fully unsupervised object discovery under realistic circumstances remains a challenging problem.

In addition to mining common objects, the discovery of object categories is also of great interest in the literature of Tuytelaars et al. [12]. In some early work, Fergus et al. [25] and Weber et al. [26] learn a part-based model for object categories. However, these methods need to specify the structure of the model, for example, the number of parts and their spatial relations; thus, they are not fully unsupervised. By representing an image as a "bag-of-word," Sivic et al. [8] and Russell et al. [9] apply top–down text mining methods to the image data, such as probabilistic latent semantic analysis [27] and latent Dirichlet allocation [28]. In some recent works [10], [29] [30], frequent item set mining is applied to the discovery of visual patterns in visual documents. However, one common limitation of the bag-of-words approaches is that the data mining results rely on the quality of the visual vocabulary. Due to the visual variations and uncertainties of local features, it is however difficult to obtain a good universal visual vocabulary. To alleviate the quantization effects of the induced visual vocabulary, Grauman and Darrell [31] propose an unsupervised learning method that relies on visual primitive matching to discover object categories in unlabeled images. The pairwise affinities between the images are measured by their partial-match feature correspondences. Then, images sharing common objects are grouped via spectral clustering. In the work of Wu et al. [32], a generative model, called active basis model, is proposed to learn a deformable template to sketch the common object from images. The active basis consists of a small number of Gabor wavelet elements at selected locations and orientations. Overall, the aforementioned methods focus on discovering object categories instead of thematic objects. They often assume that every image contains one object, which dominates the whole image; thus, localization of the object is not required.

In addition to mining common patterns in images, there is recent work on discovering common objects in video sequences and image pairs as well [1]–[3], [7], [18], [33]. The work of Sivic and Zisserman [3] discovers interesting visual objects that frequently occur in a movie. Liu and Chen [1] propose a probabilistic framework for discovering common objects in video, where small objects in low-resolution videos can be automatically discovered and tracked. Drouin et al. [33] can track a moving deformable object through discovering its rigid parts. In the work of Liu et al. [2], common objects are discovered for video summarization. However, this method needs user labeling to initialize the search; thus, it is not fully unsupervised. Both Li and Ngan [7] and Liu and Yan [18] focus on the common object discovery from a pair of images.

III. Proposed Method

A. Overview

Given a collection of T images $\mathcal{D} = \{\mathbf{I}_i\}$, we characterize every image $\mathbf{I}_i = \{p_1, \ldots, p_m\}$ by a number of *visual primitives*, i.e., local invariant features [34], [35]. Each visual primitive $p = \{x, y, \mathbf{d}\}$ corresponds to a local image patch; (x, y) denotes its spatial location; and $\mathbf{d} \in \mathbb{R}^l$ is the descriptor of the visual primitive. A visual object $\mathcal{R}^*$ is called a λ-frequent object if it appears in more than $100 \times \lambda\%$ images, where λ is the user-specified parameter to quantify how frequent $\mathcal{R}^*$ is.

Instead of finding $\mathcal{R}^*$ directly, we propose to gradually discard uncommon primitives $p \in \mathbf{I}$ to recover $\mathcal{R}^*$. In the initialization step, we discard uncommon primitives p that find few matches among the rest of images in $\mathcal{D}$. We denote the remained set of primitives as $\mathcal{D}^1 \subseteq \mathcal{D}$. Each $p \in \mathcal{D}^1$ has the potential to act as a compositional element of a common object but needs a further check. Although each $p \in \mathcal{D}^1$ appears often, the evidence that it belongs to a common object is only local. To further validate each $p \in \mathcal{D}^1$, we expand it spatially and form a larger pattern. For each $p \in \mathcal{D}$, its spatial neighbors form a *visual group* $\mathbf{G}_p = \{p, p_1^{NN}, \ldots, p_k^{NN}\}$, where p_i^{NN} is one of the nearest neighbors of p in the image. The larger the spatial size of $\mathbf{G}_p$ and the more re-occurrences it can find, the stronger the evidence implied by a common pattern. Therefore, we can gradually increase the size of $\mathbf{G}_p$ for a multilayer checking. After that, a commonness score $C(p)$ will be assigned to each visual primitive p, which reflects the likelihood that p belongs to the common object. The formal definition of $C(p)$ will be discussed in Section III-B. Intuitively, a primitive p has a positive commonness score if its visual group $\mathbf{G}_p$ frequently appears among the data set $\mathcal{D}$, and vice versa, whereas it has a negative commonness score if it rarely repeats. Once the commonness score $C(p)$ of each p is obtained, the thematic object can be located as the subimage region $\mathcal{R} \subset \mathbf{I}$ that contains the most common primitives. Localization of the thematic object will be discussed in Section III-C.

B. Multilayer Candidate Pruning

To gradually prune primitive candidates, it is important to assign an appropriate commonness score $C(p)$ to each primitive p. Intuitively, the commonness score $C(p)$ indicates the likelihood that p belongs to a common object and should depend on the frequency of p's occurrence. Namely, the more the instances of p in the image data set $\mathcal{D}$, the higher its commonness score $C(p)$. However, due to the visual pattern variations, a direct count of p's occurrence would not be reliable. In addition, a texture region may generate a high frequency of textons p, but they may not belong to any common object. Compared with the re-occurrences of an individual p, the re-occurrence of a group of primitives that keep their spatial relation is more reliable. Therefore, instead of estimating the commonness of individual primitives, we also evaluate the commonness of p by considering its k-spatial nearest neighbors (k-SNNs). As neighborhood size k increases, we gradually enlarge the spatial support of p and perform multilayer commonness checking for p. Such a multilayer procedure has two advantages. First of all, it avoids an exhaustive search of all groups by considering the redundancy among the nodes in the tree. Thus, it accelerates the data mining process. Second, by finding the appropriate groups, it also helps to determine the spatial size of the common object.

In the following, we briefly explain the multilayer pruning procedure. For each p, its k-SNNs in the image form a visual group $\mathbf{G}_p$. Given a database of primitives $\mathcal{D} = \{\mathbf{I}_i\}$, we gradually prune visual primitives p and decrease the size of the candidate set $\mathcal{D}$. In the first step, uncommon primitives p with few matches are discarded from $\mathcal{D}$. After rejecting uncommon primitives, $\mathcal{D}^1 \subseteq \mathcal{D}$ is a smaller set. For each remaining primitive $p \in \mathcal{D}^1$, we need to check the commonness of its generated

group $\mathbf{G}_p$. It is now important to define the similarity measure (or matching) between two groups rather than two individual visual primitives. Suppose that the number of SNN is k for groups in $\mathcal{D}^1$, the similarity between two sets $\mathbf{G}_q$ and $\mathbf{G}_p$, $\mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q)$, can be defined as a matching problem [36]

$$\mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q) \triangleq \max_{f \in \mathcal{F}} \sum_{i=1}^{|\mathbf{G}_p|} s\left(p_i, f(p_i)\right) \tag{1}$$

where f denotes a matching between two point sets $\mathbf{G}_p$ and $\mathbf{G}_q$, specifying which primitive $p \in \mathbf{G}_p$ matches to $f(p_i) \in \mathbf{G}_q$, and $\mathcal{F}$ is the complete set of all possible matching. Given a group $\mathbf{G}_p$, its *supportive set* consists of the groups in the rest of images that match $\mathbf{G}_p$, i.e.,

$$\mathbf{S}_p = \{\mathbf{G}_q : \mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q) > \theta\} \tag{2}$$

where $\theta > 0$ is the matching threshold. For each $p \in \mathcal{D}^1$, it will be discarded if its group $\mathbf{G}_p$ does not have enough supportive groups, i.e., $|\mathbf{S}_p| < \lambda T$, where λ is the parameter of minimum frequency and T is the total number of images.

After discarding uncommon groups, we obtain an even smaller candidate set $\mathcal{D}^2$. In the next layer, we only check the remained candidates in $\mathcal{D}^2$, but using a larger spatial neighborhood. Suppose that there are in total L layers and denote by $\mathcal{D}^L$ the final set, we obtain a filtration $\mathcal{D}^L \subseteq \ldots \subseteq \mathcal{D}^1 \subseteq \mathcal{D}$ and the corresponding spatial neighborhood size $k^L > \ldots > k^1 > 0$. Compared with $\mathcal{D}^1$, a visual primitive $p \in \mathcal{D}^l$ $(2 \leq l \leq L)$ corresponds to a larger spatial neighborhood and is more likely to be a part of a common object.

Based on multilayer checking, each visual primitive is assigned a commonness score. For each $p \in \{\mathcal{D}^L, \ldots, \mathcal{D}^1\}$, its commonness score is a positive value. The more layers p can pass, the higher the commonness score it has. For the primitives in $\mathcal{D} \setminus \mathcal{D}^1 = \{p : p \in \mathcal{D}, p \notin \mathcal{D}^1\}$, its commonness score is a negative value because these primitives are nonrepetitive by themselves. Finally, we assign the commonness score to each p as

$$C(p) = \begin{cases} k^l & \text{if } p \in \{\mathcal{D}^l \setminus \mathcal{D}^{l+1}\}, 1 \leq l \leq L \\ \tau & \text{if } p \in \{\mathcal{D} \setminus \mathcal{D}^1\} \end{cases} \tag{3}$$

where τ is the predefined negative vote value. We will specify these parameters in the experiments. Overall, they are not sensitive to different image data sets. Our multilayer pruning algorithm is summarized in Algorithm 1.

C. Detecting Thematic Object

After obtaining the commonness score $C(p)$, we can locate the thematic object in each image using a bounding box. More formally, for each image $\mathbf{I}_i$, we search for the bounding box $\mathcal{R}^*$ with the maximum commonness score

$$\mathcal{R}^* = \arg\max_{R \subseteq I_i} \sum_{p \in R} C(p) = \arg\max_{R \in \Lambda} F(R) \tag{4}$$

where $F(R) = \sum_{p \in R} C(p)$ is the objective function and Λ denotes the candidate set of all valid subimages in $\mathbf{I}_i$.

To speed up this localization process, we apply the branch-and-bound search proposed in [37]. The target bounding box $\mathcal{R}^*$ is determined by four parameters, i.e., top, bottom, left, and right positions in the image.

Algorithm 1: Discovery of Thematic Objects.

Input: T unlabeled images with extracted local primitives $\mathcal{D} = \mathbf{I}_1 \bigcup \mathbf{I}_2 \bigcup \ldots \bigcup \mathbf{I}_T$, and threshold λ

Output: thematic objects $\mathcal{R}_i^* \in \Omega^*$ in each image $\mathbf{I}_i$

/ $*$ Multilayer Candidate Pruning

1 foreach $p \in \mathcal{D}$ **do**

2 **if** $\mathbf{M}_p \neq \emptyset$ **then**

3 add p to $\mathcal{D}^1$

4 for $2 \leq l \leq L$ **do**

5 **for** $p, q \in \mathcal{D}^{l-1}$ **do**

6 $\mathbf{S}_p = \{q : \tilde{\mathrm{Sim}}(\mathbf{N}_p, \mathbf{N}_q) > 0.5k^{l-1}\}$

7 **if** $|\mathbf{S}_p| > \lambda T$ **then**

8 add p to $\mathcal{D}^l$

9 for $2 \leq l \leq L$ **do**

10 **if** $p \in \mathcal{D}^l \setminus \mathcal{D}^{l+1}$ **then**

11 $C(p) = k^l$

12 **if** $p \in \mathcal{D} \setminus \mathcal{D}^1$ **then**

13 $C(p) = \tau$

/ $*$ Thematic Object Localization

14 foreach $\mathbf{I}_i$ **do**

obtain $\mathcal{R}_i^* = \arg\max_{R \in \Lambda} F(R)$

15

16 add $\mathcal{R}_i^*$ to a set $\mathbf{\Lambda}^*$

D. Differentiation of Multiple Thematic Objects

Although the branch-and-bound search can detect all thematic objects, we need to further differentiate them if there are multiple thematic objects in the same image set. We thus perform object clustering to differentiate thematic objects. As each detected object is characterized by a subimage region $\mathcal{R}$, we estimate the affinity relationship between two subimages $\mathcal{R}_i^*$ and $\mathcal{R}_j^*$ as follows. First, for each subimage, we count the number of primitives whose k-SNN is supported by other subimages. For layer l, this number is calculated as $N_i^l = |\{p(x,y) : p \in \mathcal{D}^l \wedge (x,y) \in R_i^*\}|$, where (x,y) denotes p's spatial location. Furthermore, for subimage R_i^*, we count the number of primitives whose k-SNN is supported by the other R_j^* and denote this number as $S_{i\_j}^l$. After counting these numbers for all layers, the pairwise similarity is defined as

$$A_{i,j} = \sum_{l=2}^{L} \frac{k^l \times S_{i\_j}^l}{N_i^l}.$$

If no visual primitive is matched between $\mathcal{R}_i^*$ and $\mathcal{R}_j^*$, their affinity value is set to be a negative value, i.e., -1. Once affinity matrix A is obtained, we start to group one common object by finding a dense subgraph $\Omega \subseteq \mathbf{\Lambda}^*$ as

$$\Omega^* = \arg\max_{\Omega \subseteq \mathbf{\Lambda}^*} W(\Omega) \tag{5}$$

where

$$W(\Omega) = \sum_{\mathcal{R}_i^*, \mathcal{R}_j^* \in \Omega} A_{i,j}.$$

We employ a fixed-point iteration procedure in [38] to solve (5) and to extract the densest subgraphs one by one. For example, after finding the densest subgraph, we can remove all of its nodes and continue to find the second densest one, until no qualified subgraph can be found, i.e., the obtained subgraph only contains a single node.

IV. Approximate Similarity Matching for Fast Pruning

Based on the description of our multilayer pruning algorithm in Section III, it is important to measure the commonness, or repetitiveness, of both *individual* visual primitives and a *group* of primitives. Given a primitive p or a group $\mathbf{G} = \{p_j\}$, we need to count the total number of its re-occurrences in the rest of images in $\mathcal{D}$. Such a commonness checking should be computationally efficient. In the following, we first discuss how to speed up the individual primitive matching in Section IV-A, then propose a method to match visual groups efficiently in Section IV-B.

A. Pruning Visual Primitives

To measure the commonness of a visual primitive $p \in \mathbf{I}_i \subset \mathcal{D}$, we define its *matching set* as its ϵ-nearest neighbors (ϵ-NN) in the feature space

$$\mathbf{M}_p = \{q : \|\mathbf{d}_q - \mathbf{d}_p\| \leq \epsilon, q \in \mathcal{D} \backslash \mathbf{I}_i\} \tag{6}$$

where $\mathbf{M}_p$ denotes all matches of p in $\mathcal{D}$, except those appearing in the same image $\mathbf{I}_i$ as p; $\epsilon \geq 0$ is a matching threshold; and $\|\cdot\|$ denotes the Euclidean distance. $\forall p, q \in \mathcal{D}$, we define their similarity measure $s(p, q)$ based on their matching sets as

$$s(q,p) = \begin{cases} \exp^{-\frac{\|\mathbf{d}_q - \mathbf{d}_p\|^2|}{\alpha}} & \text{if } p \in \mathbf{M}_q \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\alpha > 0$. $s(p, q)$ is a symmetric measure as $p \in \mathbf{M}_q \Leftrightarrow q \in \mathbf{M}_p$. After performing ϵ-NN query for each $p \in \mathcal{D}$, we can build a *matching graph* $\mathbb{G} = \{\mathcal{D}, \mathcal{E}\}$ to describe the matching relations among visual primitives. Here, $\mathcal{D}$ denotes the vertex set and $\mathcal{E}$ denotes the edge set. For each $p \in \mathcal{D}$, we denote it as a node and the edge is defined on each pair of nodes $e = \{q, p\} \in \mathcal{E}, \forall q, p \in \mathcal{D}$. The weight of an edge $s(q, p) \in [0, 1]$ reflects the similarity between two primitives.

To search for the ϵ-NN and obtain the matching set $\mathbf{M}_p$, an exhaustive search of the whole data set $\mathcal{D}$ is of linear complexity $O(|\mathcal{D}|)$. This is computationally demanding as we need to check all of the primitives $p \in \mathcal{D}$, which results in a quadratic complexity $O(|\mathcal{D}|^2)$. In order to speed up the ϵ-NN query, we apply the locality-sensitive hashing (LSH) for similarity search in a high-dimensional feature space [39]. LSH provides a randomized solution to the ϵ-NN query. The query process is accelerated by compromising the results: Instead of performing the exact ϵ-NN query, LSH performs an approximate ϵ-NN query. Performing such an approximate NN-query for each $q \in \mathcal{D}$ has two benefits. First, it can prune those uncommon visual primitives. Moreover, for the remaining visual primitives, their best matches in $\mathcal{D}$ are found, and this can be further used to help pruning uncommon visual groups $\mathcal{G}$, which will be discussed in Section IV-B.

B. Fast Pruning of Visual Groups Using Approximate Similarity Matching

To match two spatial groups, it is computationally demanding to solve (1). Suppose $|\mathbf{G}_q| = |\mathbf{G}_p| = k$, and f specifies a matching between $\mathbf{G}_q$ and $\mathbf{G}_q$

$$f = [f(p_1), f(p_2), \cdots, f(p_k)]$$

where $\forall i, j$, $f(p_i) \neq f(p_j)$. Since f is the permutation of indices, the number of all possible such matching is $|\mathcal{F}| = k!$. Given two groups with k primitives, the exhaustive search of f is of complexity $O(k!)$, and the Hungarian algorithm computes the optimal matching with complexity $O(k^3)$ [36]. In checking uncommon groups, we need to evaluate all pairs of $\mathbf{G}_p$ and $\mathbf{G}_q$, $\forall p, q \in \mathcal{D}^1$. Thus, the overall complexity for pruning all uncommon groups is of complexity $O(|\mathcal{D}^1|^2 k!)$ if using exhaustive search for matching, or $O(|\mathcal{D}^1|^2 k^3)$ if using the Hungarian algorithm for matching.

To speed up pruning of uncommon groups, we present an approximate set-to-set matching by taking advantage of the previously built matching graph $\mathbb{G}$. Instead of solving the optimal matching in (1), we use the approximate similarity score, which can be efficiently calculated. For a visual group $\mathbf{G}$, we can obtain its *matching group* as

$$\mathbf{M}_\mathbf{G} = \bigcup_{p \in \mathbf{G}} \mathbf{M}_p \tag{8}$$

which is the union of all matching sets of its primitives. Given two groups $\mathbf{G}_p$ and $\mathbf{G}_q$, the *approximate similarity score* is defined as the size of the intersection between two sets $\mathbf{G}_p$ and $\mathbf{M}_{\mathbf{G}_q}$, i.e.,

$$\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q) \triangleq \left|\mathbf{G}_p \cap \mathbf{M}_{\mathbf{G}_q}\right|. \tag{9}$$

It is worth noting that, compared with the optimal matching, the approximate similarity score is generally nonsymmetric, i.e., $\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q) \neq \tilde{\mathrm{Sim}}(\mathbf{G}_q, \mathbf{G}_p)$. In comparison, the original matching problem in (1) can be formulated as a maximum flow problem in a bipartite graph; thus, the similarity score is symmetric, i.e., $\mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q) = \mathbf{Sim}(\mathbf{G}_q, \mathbf{G}_p)$.

To justify the approximate similarity score, we show in Theorem 1 that it is an upper bound of the optimal matching score.

Theorem 1: Approximate Similarity Matching: Given two primitives p and q and their corresponding groups

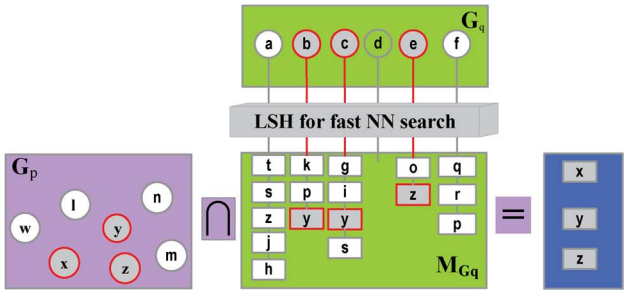


Fig. 3. Matching two groups $\mathbf{G}_q = \{a, b, c, d, e, f\}$ and $\mathbf{G}_p = \{l, m, n, w, x, y, z\}$ through approximate similarity matching. Each letter denotes an individual primitive. Each primitive in $\mathbf{G}_q$ has a matching set denoted as a string of blocks under that primitive. For example, primitive $b \in \mathbf{G}_q$ has a matching set $\mathbf{M}_b = \{k, p, y\}$, whereas d cannot find any matches, with $\mathbf{M}_d = \emptyset$. The approximate similarity score between $\mathbf{G}_q$ and $\mathbf{G}_p$ is the size of set intersection of $\mathbf{G}_p$ and $\mathbf{M}_{\mathbf{G}_q}$, $\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q) = |\mathbf{G}_p \cap \mathbf{M}_{\mathbf{G}_q}| = 3 \geq \mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q)$.

$\mathbf{G}_p$ and $\mathbf{G}_q$, let $\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q) \triangleq |\mathbf{G}_p \cap \mathbf{M}_{\mathbf{G}_q}|$, and $\mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q) \triangleq \max_{f \in \mathcal{F}} \sum_{i=1}^{|\mathbf{G}_p|} s(p_i, f(p_i))$, then we have

$$\min\left\{\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q), \tilde{\mathrm{Sim}}(\mathbf{G}_q, \mathbf{G}_p)\right\} \geq \mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q).$$

The proof of Theorem 1 is in the Appendix. Fig. 3 illustrates an example of matching two groups $\mathbf{G}_p$ and $\mathbf{G}_q$ using the approximate similarity score. The matching can be formulated as a maximum flow problem between two sets, which is bounded by the total number of edges connecting them. Based on Theorem 1, we can safely prune uncommon groups with the approximate matching score. If $\min\{\tilde{\mathrm{Sim}}(\mathbf{G}_p, \mathbf{G}_q), \tilde{\mathrm{Sim}}(\mathbf{G}_q, \mathbf{G}_p)\} < \theta$, it implies that group $\mathbf{G}_q$ is *not* a valid re-occurrence of $\mathbf{G}_p$ because $\mathbf{Sim}(\mathbf{G}_p, \mathbf{G}_q) < \theta$ as well. Therefore, if $\mathbf{G}_p$ cannot find enough valid re-occurrences even with the exaggerated approximate similarity score, it must be an uncommon group.

To calculate the intersection of two sets $\mathbf{G}_p$ and $\mathbf{M}_{\mathbf{G}_q}$, it is of linear complexity $O(|\mathbf{G}_p| + |\mathbf{M}_{\mathbf{G}_q}|)$ by using two binary vectors to index the elements in $\mathbf{G}_p$ and $\mathbf{M}_{\mathbf{G}_q}$. Suppose $|\mathbf{G}_p| = |\mathbf{G}_q| = k$, the average size of $\mathbf{M}_{\mathbf{G}_q}$, $|\mathbf{M}_{\mathbf{G}_q}|$, can be estimated as

$$|\mathbf{M}_{\mathbf{G}_q}| = \left|\bigcup_{p \in \mathbf{G}_q} \mathbf{M}_p\right| \leq \sum_{p \in \mathbf{G}_q} |\mathbf{M}_p| = kh \qquad (10)$$

where $h = (1/|\mathcal{D}^1|) \sum_{p \in \mathcal{D}^1} |\mathbf{M}_p|$ is the average size of $\mathbf{M}_p$. In general, h is a small constant controlled by ϵ in an ϵ-NN query. The worst case is when $\epsilon = \infty$, $h = |\mathcal{D}|$. By using the approximate similarity matching, we largely reduce the complexity of matching two sets from $O(k^3)$ to $O(kn)$. Because $\forall p, q \in \mathcal{D}^1$, we need to measure the approximate similarity score; the overall complexity for pruning uncommon groups is now of $O(|\mathcal{D}^1|^2 kn)$.

V. Experiments

A. Experimental Setting

To evaluate our algorithm, we collect 10 image collections and 30 commercial video clips for thematic object mining. The number of images in these 40 data sets ranges from 4 to 89. The same thematic object can be varied due to rotation, partial occlusion, scale, viewpoint, and lighting condition changes. It is possible that one image contains multiple thematic objects and some images do not contain any thematic objects. For all of our experiments, the multilayer candidate pruning has four layers, and the corresponding sizes of the spatial neighborhood are $k^1 = 1$, $k^2 = 5$, $k^3 = 10$, and $k^4 = 15$, respectively. The other parameter τ in (3) is set to be -2. From each image, we extract and compare two types of visual primitives, namely, scale-invariant feature transformation (SIFT) [34] and maximally stable extremal regions (MSER) [35]. SIFT corresponds to local square patches, whereas MSER corresponds to local ellipse patches. Both types of local features are characterized by the SIFT descriptor of 128 dimensions. Matching threshold $\epsilon = 180$ is fixed. If the data set contains more than 60 images, we set $\epsilon = 150$ to enable a more efficient matching via LSH. Without otherwise mentioned, we specify frequency parameter $\lambda = 0.1$. During the branch-and-bound search, to prevent including empty pixels into the thematic objects, we assign a small negative commonness score to each empty pixel that does not locate any visual primitive. In our experiments, we test four different negative values, i.e., -0.01, -0.02, -0.03, and -0.04, and selected the best one as our cropping result. All of the experiments are performed on a standard P4-3.19-GHz PC (2-GB memory). The algorithm is implemented in C++.

To quantify the performance, we manually label the ground truth bounding boxes of the thematic objects in each image. Let DR and GT be the discovered subimages and the bounding boxes of ground truth, respectively. The performance is measured by two criteria, i.e., precision $= |GT \cap DR|/|DR|$ and recall $= |GT \cap DR|/|GT|$. Both precision and recall measure the accuracy of thematic object discovery. By combining precision and recall, we further use a single F-measure as the metric for performance evaluation [40]. F-measure $= 2 \times$ recall $\times$ precision/recall + precision is the weighted harmonic mean of precision and recall.

B. Discovering Thematic Objects in Image Collections

We first test the mining algorithm on ten image collections. Fig. 4 presents mining results on nine image collections. The discovered thematic objects are highlighted by the red bounding boxes. From top to bottom, we describe our data set as $\text{Dataset}_1 - \text{Dataset}_9$. Dataset_{10} contains a collection of logos, and two images contain two thematic objects. $\text{Dataset}_1 - \text{Dataset}_4$ are image collections of four Oxford buildings. $\text{Dataset}_5 - \text{Dataset}_9$ are image collections of another five thematic objects. In these data sets, the thematic objects are under different variations such as rotation, scale, viewpoint, and lighting condition changes. Two images in Dataset_6 do not contain the thematic object.

Table I summarizes the performance of our method. For each data set, the number of images (*ImageNo.*) and the occurrence number of thematic objects (*ObjectNo.*) are shown in the first and second rows, respectively. The number of correctly detected thematic objects (*CorrectNo.*) and its ratio to the total thematic objects (*CorrectRa.*) are shown in the third and fourth rows, respectively. Finally, the number of falsely detected thematic objects (*FalseNo.*) and its ratio to the total thematic objects

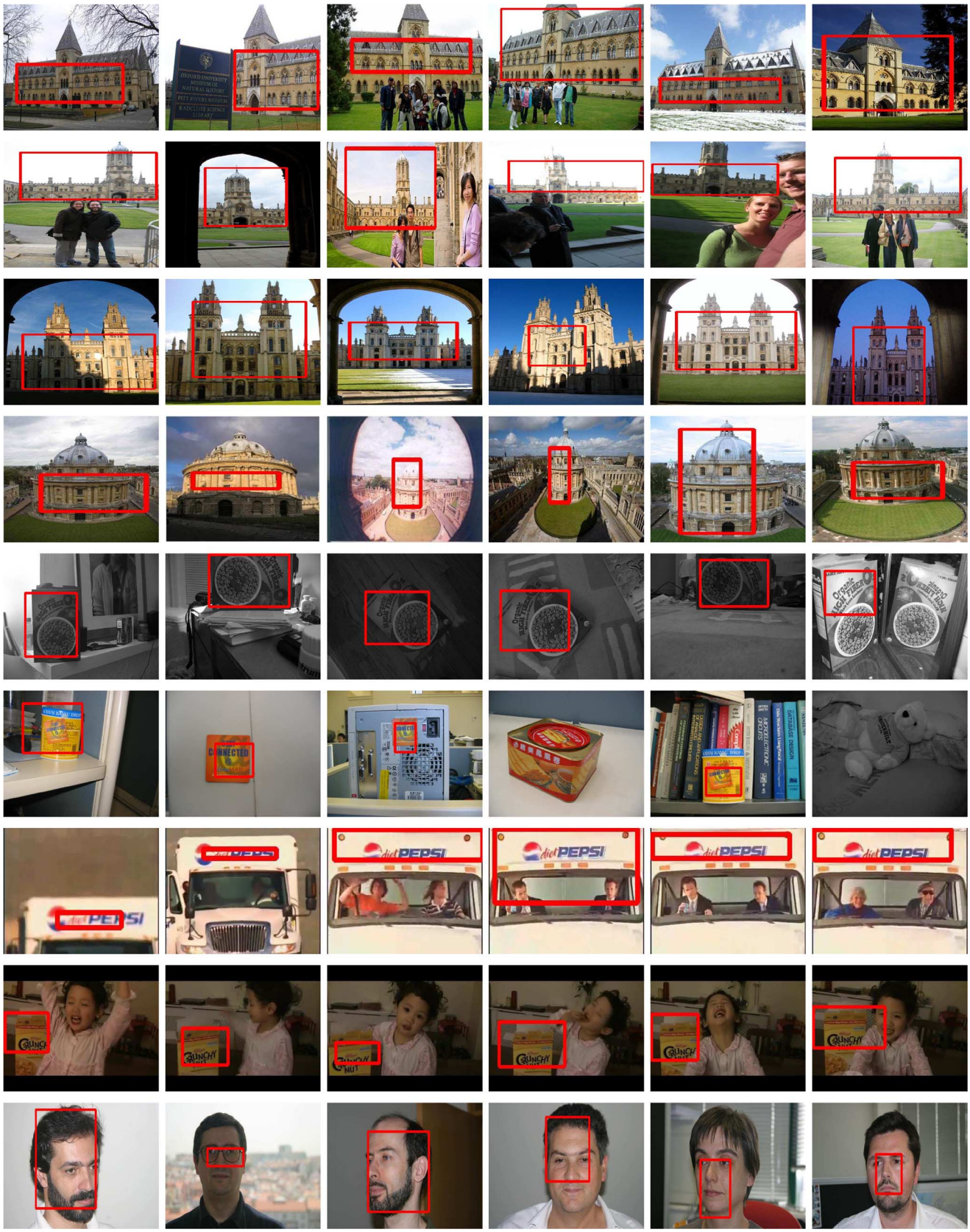


Fig. 4. Sample results of common object mining from image collections. Each row corresponds to an image collection (at least six images), and the common object is discovered by the red bounding box. The common objects are under different variations such as rotation (row 5), partial occlusion (rows 2 and 8), scale, viewpoint, and lighting condition changes (rows 3, 4, 5, and 9).

(*FalseRa.*) are given in the last two rows. These mining results show that the proposed approach performs well for mining identical thematic objects from image collections. Fig. 5 shows the F-measure of all ten testing data sets. We compare the

TABLE I
EVALUATION OF OUR METHOD WITH TEN IMAGE COLLECTIONS

Dataset	1	2	3	4	5	6	7	8	9	10	Tot.	Avg.
ImageNo.	29	12	11	9	6	6	14	6	22	4	119	11.9
ObjectNo.	29	12	11	9	6	4	14	6	22	6	119	11.9
CorrectNo.	28	11	11	9	6	4	13	6	15	6	109	10.9
CorrectRa.	0.97	0.92	1.00	1.00	1.00	1.00	0.93	1.00	0.68	1.00	0.92	0.95
FalseNo.	0	0	0	0	0	0	1	0	0	0	1	0.1
FalseRa.	0	0	0	0	0	0	0.07	0	0	0	0.009	0.007

TABLE II
EVALUATION OF OUR METHOD WITH 30 VIDEO SEQUENCES

Dataset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ImageNo.	59	40	59	46	59	27	49	60	58	19	14	28	32	89	24	30
ObjectNo.	30	35	17	22	20	15	21	13	16	18	13	19	22	13	21	15
CorrectNo.	22	30	16	22	19	15	21	11	14	18	13	19	20	12	7	10
CorrectRa.	0.73	0.85	0.94	1.00	0.95	1.00	1.00	0.84	0.87	1.00	1.00	1.00	0.90	0.92	0.80	0.66
FalseNo.	0	0	0	0	4	3	7	8	0	0	0	0	0	0	0	0
FalseRa.	0.00	0.00	0.00	0.00	0.20	0.20	0.33	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dataset	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	Tot.	Avg.
ImageNo.	49	59	59	59	82	57	31	58	59	30	40	32	59	33	1400	46.7
ObjectNo.	19	43	21	9	33	15	17	57	16	22	18	17	22	15	634	21.1
CorrectNo.	16	34	18	8	29	15	17	37	16	21	18	17	22	13	560	18.6
CorrectRa.	0.84	0.79	0.85	0.88	0.87	1.00	1.00	0.64	1.00	0.95	1.00	1.00	1.00	0.86	0.88	0.90
FalseNo.	0	11	1	0	0	0	0	0	0	0	0	0	2	0	36	1.2
FalseRa.	0.00	0.25	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.06	0.05

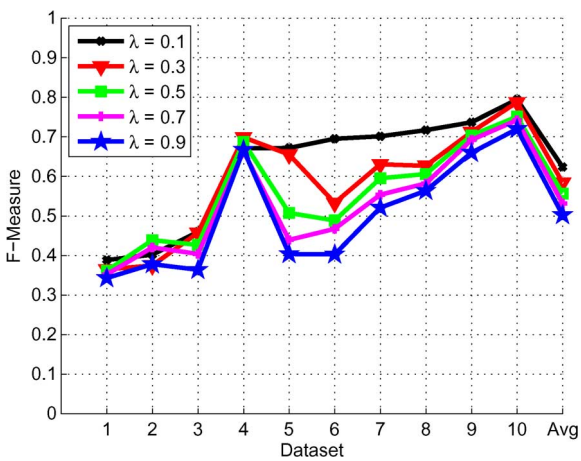


Fig. 5. Evaluation of our method using different thresholds λ on ten image collections.

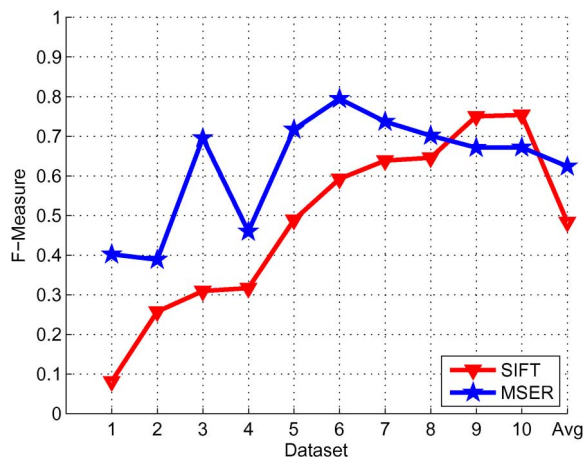


Fig. 6. Evaluation of our method using different types of visual primitives on ten image collections.

F-measure under different thresholds λ. It can be seen that the best performance is obtained when $\lambda = 0.1$, whereas the worst performance is obtained when $\lambda = 0.9$. This is due to the large variations among different occurrences of a thematic object. Therefore, a smaller λ is preferred when pruning candidates. Fig. 6 shows the evaluation of our method with different types of visual primitives. It can be seen that MSER performs better in data sets of buildings. This is because MSER is more robust to view changes. On the other hand, SIFT performs well on mining planar objects without large viewpoint changes (e.g., a graffiti wall) [41].

C. Discovering Thematic Objects in Videos

To evaluate the proposed approach on mining videos, we further test our method with 30 video clips downloaded from YouTube.com. Most of the clips are commercial videos, and each video contains at least one thematic object, e.g., the Starbucks logo in a commercial video of Starbucks Coffee. As the key object to be highlighted, such a thematic object frequently appears, and the discovery of it is essential to understand and summarize the video contents.

In each video, the occurrences of the thematic objects range from 7 to 45 s. We sample key frames at 2 frames/s, and then discover thematic objects from the extracted key frames. Fig. 7 shows some sample results of the discovered thematic objects.From top to bottom, the test video data sets are Dataset_1 – Dataset_{10}. In the video scenes, the thematic objects are also subject to different variations such as partial occlusions, scale, viewpoint, and lighting condition changes. It is possible that one video clip contains multiple thematic objects and some frames do not contain any thematic objects. Dataset_1 and Dataset_2 are also used in [2].

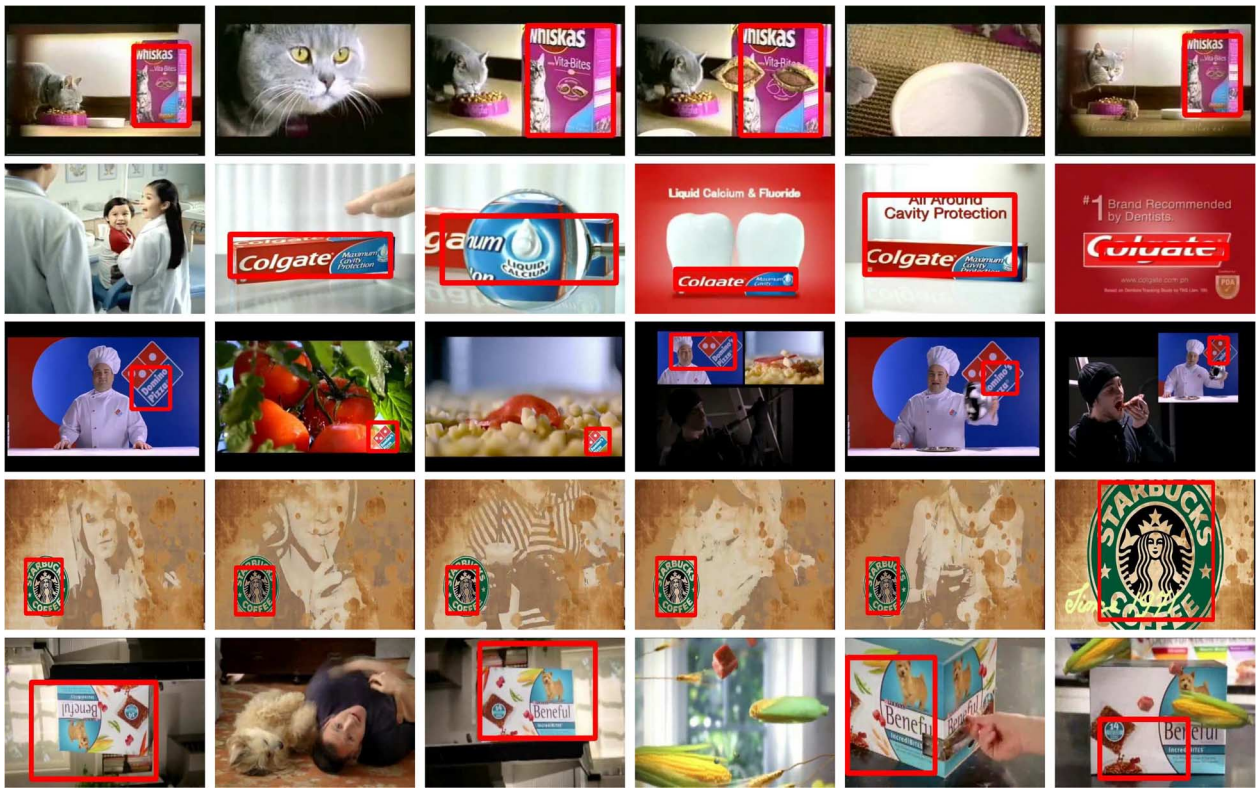


Fig. 7. Sample results of common object mining in video key frames. Each video sequence may contain multiple common objects. Each row shows a video sequence, and the discovered common object is highlighted by the red bounding box. The common objects are under different variations such as rotation (row 5), partial occlusion (rows 1, 2, and 3), scale, viewpoint, and lighting condition changes (rows 4 and 5).

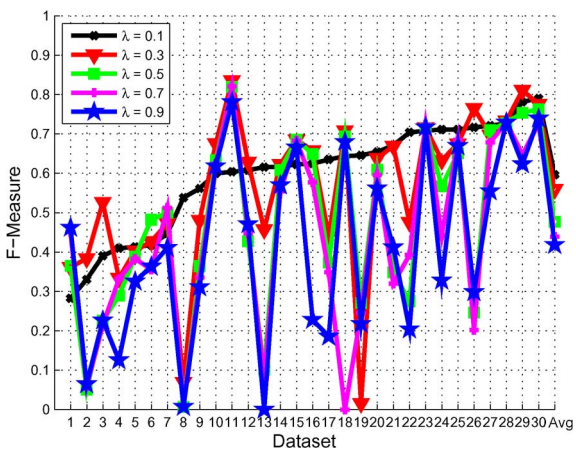


Fig. 8. Evaluation of our method using different thresholds λ on 30 video sequences.

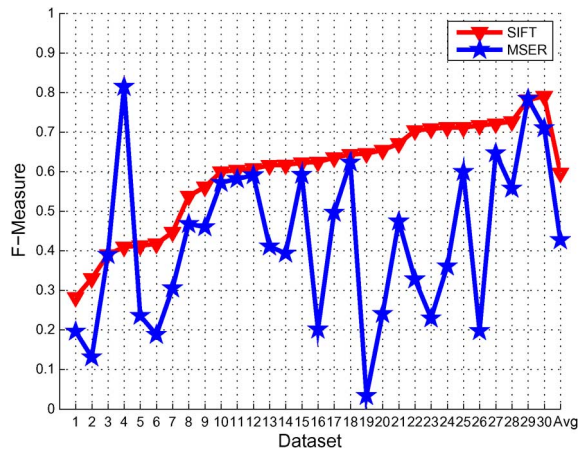


Fig. 9. Evaluation of our method using different types of visual primitives on 30 video sequences.

Table II summarizes the performance of our method on 30 video sequences. Fig. 8 shows the F-measure of all 30 testing data sets. We compare the F-measure under different thresholds λ. Fig. 9 shows the evaluation of our method with different types of visual primitives. It can be seen that SIFT performs better in most data sets, whereas MSER performs better only in several specific data sets. This is because SIFT is appropriate to match planar objects such as the commercial products or their logos.

Finally, Fig. 10 evaluates the influence of the image data set size. In the testing data set, each image contains only one thematic object. It can be seen that the data mining performance can be improved by increasing the number of images. Unsurprisingly, with more instances of the thematic object in the data set, it can help to handle the object variations.

D. Computational Cost

To quantify the efficiency of our multilayer pruning, Table III presents the results of mining the image data set, as shown in Fig. 4. It contains 14 images, and the total number of visual primitives is $|\mathcal{D}| = 5200$. By setting $\epsilon = 180$ for ϵ-NN query, the average number of matches that each primitive finds

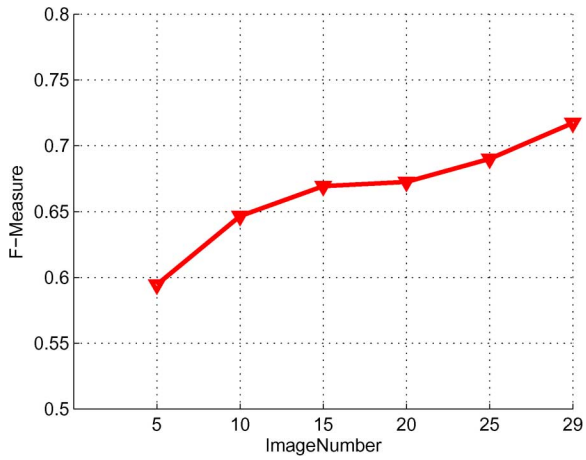


Fig. 10. Evaluation of our method with different number of images.

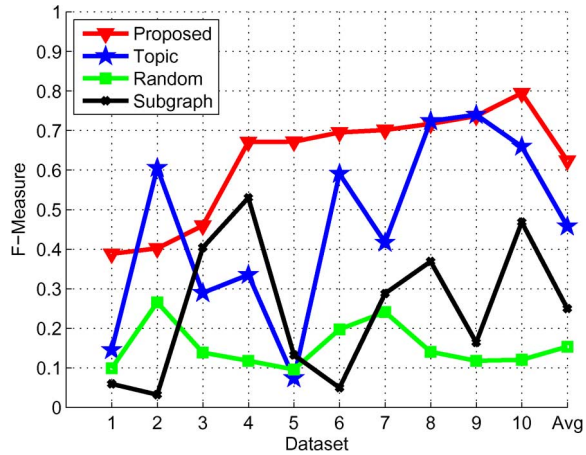


Fig. 11. Performance comparison of our approach (Proposed), topic discovery approach (Topic) [9], subgraph mining approach (Subgraph) [38], and random bounding box approach (Random) on ten image collections.

TABLE III
MULTILAYER CANDIDATE PRUNING

	# of visual primitives
Initial #	$\|\mathcal{D}\| = 5,200$
after the initial pruning	$\|\mathcal{D}^1\| = 1,270$
after the 1_{st} layer pruning	$\|\mathcal{D}^2\| = 483$
after the 2_{nd} layer pruning	$\|\mathcal{D}^3\| = 363$
after the 3_{rd} layer pruning	$\|\mathcal{D}^4\| = 228$

TABLE IV
RUNTIME COST ANALYSIS (IN SECONDS)

# of images	14	20	30
Initial pruning	16.13	18.76	27.86
Multi-layer pruning	7.10	11.49	83.13
Localization	3.19	4.58	6.83
Clustering	0.1	0.15	0.25
Total Cost	26.52	34.98	118.07

is $h = 15$. We also notice that the initial pruning of visual primitive is efficient using LSH. Once the index is built for the database, the average query time for each $p \in \mathcal{D}$ is around only 1 ms and the total query time for $|\mathcal{D}| = 5200$ is around 15 s. After initial pruning, the number of candidate primitives is reduced to $|\mathcal{D}^1| = 1270$. For the other three layers of commonness checking, the number of remained candidate visual primitives is reduced to 483, 363, and 228, respectively, with checking times of 3.21, 2.10, and 1.79 s, respectively. This shows that the approximate group matching is computationally efficient when using the data set with a limited number of images. Finally, the object localization using branch-and-bound search is also efficient. It costs only 3.19 s for all of 14 images. The total CPU cost for mining the thematic object from the 14 images is 26.52 s.

Without considering the cost of SIFT/MSER feature extraction, the computational cost of our mining method includes three parts: 1) individual visual primitive matching using LSH; 2) multilayer commonness checking; and 3) thematic object localization and clustering. Table IV presents the computational cost when mining image data sets of different sizes. It can be seen that, as the size of data set becomes larger, the multilayer pruning algorithm requires much more time.

E. Comparison With Other Approaches

To compare our proposed method, we test three other methods: 1) random bounding box; 2) topic discovery; and 3) subgraph mining. For the first method, we randomly generate a bounding box to guess the location and scale of the thematic object. The top-left and bottom-right corners of the bounding box are randomly selected with a uniform distribution. The second method is proposed in [9], which finds common objects via topic discovery. First, each image is segmented multiple times, with different number of segments K. Given multiple segmentations, it is expected that one of the segments will contain the common object. By clustering visual primitives into a visual vocabulary, each segment is represented as a bag-of-words. After obtaining a pool of segments (visual documents) from all of the images, common topics are discovered using latent Dirichlet allocation [28]. Finally, for each discovered topic, all image segments are sorted by how well they are explained by that topic. The segment at the highest rank is selected as the common object. In our implementation, we segment each image into $K = 3, 5, 7, 9, 11,$ and 13 segments, respectively. We perform normalized cut [42] in both original images and the downsampled images of half size of the original image. This generates a total of 96 segments per image. To obtain the bag-of-words presentation, SIFT descriptors are used and quantized into 1000 visual words by k-means clustering. The third method is proposed in [38], which also employs the bag-of-words presentation. By representing the pairwise relations among all words as an affinity graph, this method formulates the common pattern discovery problem as a cohesive subgraph mining problem. After discovering a subgraph of the spatially collocated word, the instances of the common pattern are located by finding the bounding boxes that contain the common words.

Fig. 11 presents the quantitative comparison of different approaches for ten image collections. Overall, our proposed approach outperforms all other methods in terms of the

F-measure, with an average score of 0.63 (Proposed) compared with 0.15 (Random), 0.46 (Topic Discovery), and 0.24 (Subgraph). It clearly shows the advantages of combining the multilayer candidate pruning strategy and the branch-and-bound search algorithm. As expected, the random guess of the bounding box has the lowest F-measure in most image collections. Its performance depends on the size of common objects. When the common object occupies a large area in the image, a randomized bounding box would well capture the common object. Otherwise, it is very likely to fail. For the topic discovery approach, it does not consider the spatial relationship among the visual primitives and is affected by the quantization error in building the visual vocabulary [13]. Moreover, the performance of topic discovery highly depends on the performance of the image segmentation. Although each image is segmented multiple times with different scales and number of segments, it is often that the common object is not well extracted (e.g., Dataset_1 and Dataset_3). Due to the background clutters, obtaining reliable image segmentation is not a trivial task. Thus, the topic discovery approach only obtains a coarse recovery of the common object, which is far from satisfactory. The subgraph mining approach does not work well for this data set either. The performance also relies on the quality of the visual vocabulary. Moreover, as it only considers pairwise relationship between two words, high-order relations among the visual primitives are not utilized.

In addition to testing on image collections, we also test our method and compare it with the above three methods on 30 video sequences. Fig. 12 presents the quantitative comparison. Overall, our proposed approach outperforms all other methods in terms of the F-measure, with an average score of 0.58 (Proposed) compared with 0.15 (Random), 0.32 (Topic Discovery), and 0.38 (Subgraph). It clearly shows that our proposed method outperforms the other three methods again. Due to the difficulty of obtaining the reliable video key frame segmentation, the topic discovery approach only obtains a coarse recovery of the common object, which is outperformed by the subgraph mining approach.

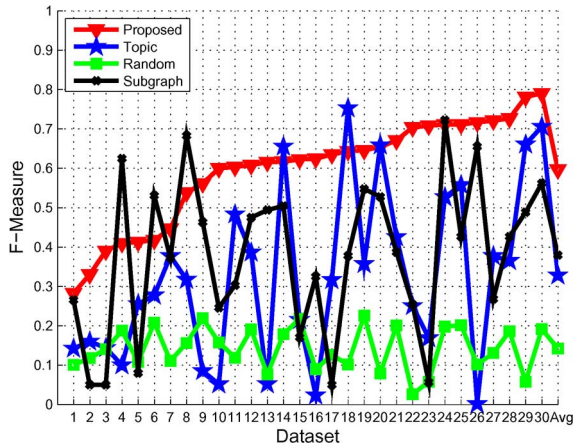


Fig. 12. Performance comparison of our approach (Proposed), topic discovery approach (Topic) [9], subgraph mining approach (Subgraph) [38], and random bounding box approach (Random) on 30 video sequences.

VI. Conclusion

We have proposed a novel bottom–up images data mining approach for thematic object discovery in both images and videos. Instead of modeling each image as a visual document and discover thematic patterns through conventional top–down generative models, we directly match visual primitives and gradually expand them to recover the thematic object of larger spatial support. Our method does not rely on visual vocabulary and considers the spatial structure of visual patterns. To overcome the computational cost of searching the huge solution space, we propose a multilayer candidate pruning method to efficiently prune unqualified candidates of thematic patterns. With each visual primitive obtaining a commonness score, these pieces of local evidence of thematic patterns are accumulated through searching the bounding box of the highest commonness score. To further accelerate the visual pattern matching and mining, we propose the approximate group matching for fast matching of two sets of visual primitives.

Experiments on both image collections and video sequences show that our method can automatically determine the scale, location, and number of thematic objects. It is able to handle object variations due to scale, viewpoint, color, and lighting condition changes, even partial occlusion. Using the branch-and-bound search, our localization is efficient and robust to the cluttered backgrounds. The proposed method is applicable to different types of local features. Our future work includes how to further improve the feature matching and candidate pruning efficiency, such that it can handle web-scale image and video data set.

Appendix A
Proof of Theorem 1

Suppose that f^* is the optimal matching, we have

$$\mathbf{Sim}(\mathbf{G}_p,\mathbf{G}_q) = \max_f \sum_{p_i\in\mathbf{G}_p} s\left(p_i, f(p_i)\right) \tag{11}$$

$$= \sum_{p_i\in\mathbf{G}_p} s\left(p_i, f^*(p_i)\right) \tag{12}$$

$$\le \left|\{p : p\in\mathbf{G}_p, s\left(p, f^*(p)\right)\neq 0\}\right| \tag{13}$$

$$\le \left|\{p : p\in\mathbf{G}_p, \exists q\in\mathbf{G}_q, s(p,q)\neq 0\}\right| \tag{14}$$

$$= \left|\mathbf{G}_p \cap \mathbf{M}_{\mathbf{G}_q}\right| \tag{15}$$

$$= \tilde{\mathrm{Sim}}(\mathbf{G}_p,\mathbf{G}_q) \tag{16}$$

where $|\{p : p\in\mathbf{G}_p, s(p,f^*(p))\neq 0\}|$ in (13) denotes the total number of primitives involved in the optimal matching; $|\{p : p\in\mathbf{G}_p, \exists q\in\mathbf{G}_q, s(p,q)\neq 0\}|$ in (14) denotes the number of valid nodes in $\mathbf{G}_p$, who has matches in $\mathbf{G}_q$. Equation (13) is obtained because $0\le s(p,f(p))\le 1$ according to (7).

Similarly, we can also prove that $\mathbf{Sim}(\mathbf{G}_q,\mathbf{G}_p) \le \tilde{\mathrm{Sim}}(\mathbf{G}_q,\mathbf{G}_p)$. Because $\mathbf{Sim}(\mathbf{G}_q,\mathbf{G}_p) = \mathbf{Sim}(\mathbf{G}_p,\mathbf{G}_q)$, we finally have

$$\mathbf{Sim}(\mathbf{G}_p,\mathbf{G}_q) \le \min\left\{\tilde{\mathrm{Sim}}(\mathbf{G}_p,\mathbf{G}_q), \tilde{\mathrm{Sim}}(\mathbf{G}_q,\mathbf{G}_p)\right\}.$$

REFERENCES

[1] D. Liu and T. Chen, "Discov: A framework for discovering objects in video," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 200–208, Feb. 2008.

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

[3] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, pp. I-488–I-495.

[4] Y. J. Lee and K. Grauman, "Object-graphs for context-aware category discovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1–8.

[5] H.-T. Chen, "Preattentive co-saliency detection," in *Proc. IEEE Conf. Image Process.*, 2010, pp. 1117–1120.

[6] M. Cho, Y. M. Shin, and K. M. Lee, "Co-recognition of image pairs by data-driven Monte Carlo image exploration," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 144–157.

[7] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[8] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. IEEE Intl. Conf. Comput. Vis.*, 2005, pp. 370–377.

[9] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using multiple segmentation to discover objects and their extent in image collections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 1605–1614.

[10] J. Yuan, Y. Wu, and M. Yang, "From frequent itemsets to semantically meaningful visual patterns," in *Proc. Int. Conf. Know. Discov. Data Mining*, 2007, pp. 864–873.

[11] L. Cao and F.-F. Li, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[12] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 284–302, Jun. 2010.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[14] H.-K. Tan and C.-W. Ngo, "Localized matching using earth mover's distance towards discovery of common patterns from small image samples," *Image Vision Comput.*, vol. 27, no. 10, pp. 1470–1483, Sep. 2009.

[15] P. Hong and T. S. Huang, "Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs," *J. Discrete Appl. Math.*, vol. 139, no. 1–3, pp. 113–135, Apr. 2004.

[16] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2158–2174, Dec. 2008.

[17] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2122–2129.

[18] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1609–1616.

[19] J. Yuan, Z. Li, Y. Fu, Y. Wu, and T. S. Huang, "Common spatial pattern discovery by efficient candidate pruning," in *Proc. IEEE Conf. Image Process.*, 2007, pp. I-165–I-168.

[20] G. Zhao and J. Yuan, "Mining and cropping common objects from images," in *Proc. ACM Multimedia*, 2010, pp. 975–978.

[21] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[22] O. Chum and J. Matas, "Large scale discovery of spatially related images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 371–377, Feb. 2010.

[23] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. J. Guibas, "Imagewebs: Computing and exploiting connectivity in image collections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3432–3439.

[24] S. Bagon, O. Brostovski, M. Galun, and M. Irani, "Detecting and sketching the common," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 33–40.

[25] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2003, pp. II-264–II-271.

[26] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 18–32.

[27] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1/2, pp. 177–196, Jan./Feb. 2001.

[28] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[29] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool, "Efficient mining of frequent and distinctive feature configurations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[30] S. Nowozin, K. Tsuda, T. Uno, T. Kudo, and G. Bakir, "Weighted substructure mining for image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[31] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Nov. 2006, pp. 19–25.

[32] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu, "Learning active basis model for object detection and recognition," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 198–235, Nov. 2010.

[33] S. Drouin, P. Hebert, and M. Parizeau, "Incremental discovery of object parts in video sequences," *Comput. Vis. Image Understand.*, vol. 110, no. 1, pp. 60–74, Apr. 2008.

[34] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[35] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. Brit. Mach. Vis. Conf.*, 2002, pp. 384–393.

[36] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA: MIT Press, 2001.

[37] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.

[38] G. Zhao, J. Yuan, J. Xu, and Y. Wu, "Discovering the thematic object in commercial videos," *IEEE MultiMedia* vol. 18, no. 3, pp. 56–65, Mar. 2011 [Online]. Available: http://dx.doi.org/10.1109/MMUL.2011.40

[39] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.

[40] S. Todorovic and N. Ahuja, "Extracting subimages of an unknown category from a set of images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 927–934.

[41] P.-E. Forssén and D. Lowe, "Shape descriptors for maximally stable extremal regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[42] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

Junsong Yuan (S'06–M'08) received the M.Eng. degree from the National University of Singapore, Singapore, and the Ph.D. degree from Northwestern University, Evanston, IL.

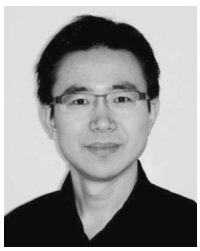
He is currently a Nanyang Assistant Professor at Nanyang Technological University (NTU), Singapore. He is also the Program Director of Video Analytics in the Infocomm Centre of Excellence, NTU. Before that, he graduated from the Special Program for the Gifted Young in Huazhong University of Science and Technology, Wuhan, China. He was a Research Intern with Microsoft Research Redmond, Redmond, WA; Kodak Research Laboratories, Rochester, NY; and Motorola Applied Research Center, Schaumburg, IL. From 2003 to 2004, he was a Research Scholar at the Institute for Infocomm Research, Singapore. He has filed three U.S. patents. His current research interests include computer vision, video analytics, multimedia search and data mining, vision-based human computer interaction, biomedical image analysis, etc.

Dr. Yuan was a recipient the Outstanding EECS Ph.D. Thesis Award from Northwestern University and the Doctoral Spotlight Award from IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09). He was also a recipient of the Elite Nanyang Assistant Professorship in 2009. In 2001, he was awarded the National Outstanding Student and Hu Chunan Scholarship by the Ministry of Education of China. He currently serves as an Editor of the KSII Transactions on Internet and Information Systems.

Gangqiang Zhao (M'08) received the B.Eng. degree in computer science from Qingdao University, Qingdao, China, in 2003, and the Ph.D. degree in computer science from Zhejiang University (ZJU), Hangzhou, China, in 2009.

From 2003 to 2009, he was a Research Assistant in the Pervasive Computing Laboratory, ZJU. Since March 2010, he has been a Research Fellow at Nanyang Technological University, Singapore. His current research interests include computer vision, multimedia data mining, and image processing.

Dr. Zhao is also a member of Association for Computing Machinery. He was the recipient of the Hewlett-Packard (HP) Distinguished Chinese Student Scholarship in 2008.



Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2004, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2007 and 2008, respectively.

He was a Research Intern with Mitsubishi Electric Research Laboratories, Cambridge, MA, in summer 2005, and with the Multimedia Research Laboratory, Motorola Laboratories, Schaumburg, IL, in summer 2006. He joined BBN Technologies, Cambridge, as a Scientist in 2008. He held a Part-Time Lecturer position at the Department of Computer Science, Tufts University, Medford, MA, in the spring of 2009. He joined the Department of Computer Science and Engineering, University at Buffalo-State University of New York, Buffalo, as an Assistant Professor in 2010. His research interests include interdisciplinary research in machine learning, social media analytics, human–computer interaction, and cyber–physical systems.

Dr. Fu is a Lifetime Member of Association for Computing Machinery, International Society for Optics and Photonics (SPIE), and the Institute of Mathematical Statistics. He was also a Beckman Graduate Fellow in 2007–2008. He was the recipient of the 2002 Rockwell Automation Master of Science Award, Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, the 2003 Hewlett-Packard (HP) Silver Medal and Science Scholarship, the 2007 Chinese Government Award for Outstanding Self-Financed Students Abroad, the IEEE International Conference on Image Processing 2007 Best Paper Award, the 2007–2008 Beckman Graduate Fellowship, the 2008 M. E. Van Valkenburg Graduate Research Award, the ITESOFT Best Paper Award of 2010 International Association for Pattern Recognition International Conferences on the Frontiers of Handwriting Recognition, the 2010 Google Faculty Research Award, IEEE International Conference on Multimedia and Expo 2011 Quality Reviewer, the LSVA Best Paper Award of the IEEE International Conference on Data Mining 2011 Workshop on Large Scale Visual Analytics, and the 2011 IC Postdoctoral Research Fellowship Award. He is the Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Zhu Li (SM'07) received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, in 2004.

He was an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University, from 2008 to 2010, and a Principal Staff Research Engineer with the Multimedia Research Laboratory, Motorola Laboratories, from 2000 to 2008. He is currently a Senior Staff Researcher with the Core Networks Research, FutureWei (Huawei) Technologies, Bridgewater, NJ, where he leads the Media Analytics Group. He has 20 issued or pending patents and more than 70 publications in book chapters, journals, conference proceedings, and standards contributions related to his areas of interest. His research interests include audio–visual analytics and machine learning with its application in large-scale video repositories annotation, mining, and recommendation, as well as video adaptation, source–channel coding, and distributed optimization issues of the wireless video networks.

Dr. Li was elected Vice-Chair of the IEEE Multimedia Communication Technical Committee 2008–2010. He was the recipient of the Best Poster Paper Award at the IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, in 2006, and the Best Paper (DoCoMo Laboratories Innovative Paper) Award at the IEEE International Conference on Image Processing, San Antonio, TX, in 2007.

Aggelos K. Katsaggelos (F'98) received the Diploma degree in electrical and mechanical engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1979 and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, where he is currently a Professor and holder of the AT&T Chair. He was previously the holder of the Ameritech Chair of Information Technology from 1997 to 2003. He is also the Director of the Motorola Center for Seamless Communications, Northwestern University; a member of the Academic Staff of NorthShore University Health System, an affiliated faculty of the Department of Linguistics, Northwestern University; and he has an appointment with Argonne National Laboratory, Lemont, IL. He has extensively published in the areas of multimedia processing and communications (180 journal papers, more than 400 conference papers, and 40 book chapters), and he is the holder of 19 international patents. He is the coauthor of *Rate-Distortion Based Video Compression* (Kluwer, 1997), *Super-Resolution for Images and Video* (Claypool, 2007), and *Joint Source–Channel Video Transmission* (Claypool, 2007).

Dr. Katsaggelos is a Fellow of International Society for Optics and Photonics (SPIE) (2009). Among his many professional activities, he was also the Editor-in-Chief of the IEEE Signal Processing Magazine (1997–2002), a BOG Member of the IEEE Signal Processing Society (1999–2001), and a member of the Publication Board of the IEEE Proceedings (2003–2007). He was the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE International Conference on Multimedia and Expo Paper Award (2006), an IEEE International Conference on Image Processing Paper Award (2007), and an International Student Paper of the Year Award (2009). He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008).

Ying Wu (SM'06) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994, the M.S. degree from Tsinghua University, Beijing, China, in 1997, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, in 2001.

From 1997 to 2001, he was a Research Assistant in the Beckman Institute for Advanced Science and Technology, UIUC. During summer 1999 and 2000, he was a Research Intern with Microsoft Research Redmond, Redmond, WA. In 2001, he joined the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, as an Assistant Professor. He is currently an Associate Professor of electrical engineering and computer science at Northwestern University. His current research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human–computer interaction.

Dr. Wu serves as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, SPIE Journal of Electronic Imaging, and IAPR Journal of Machine Vision and Applications. He was the recipient of the Robert T. Chien Award at UIUC in 2001 and the National Science Foundation CAREER Award in 2003.