Discovering the Thematic Object in Commercial Videos

Gangqiang Zhao and Junsong Yuan
Nanyang Technological University

Jiang Xu and Ying Wu
Northwestern University

The rapid development of digital video capture techniques has led to an explosive growth of video data, creating the need for effective video content analysis. Despite previous success in video content understanding and search at the clip and shot level, object-level video content analysis remains a challenging problem. In this article, we address the problem of thematic video object discovery to improve object-level video understanding and search. We define the thematic object as the key object that appears frequently in a video, such as the logo of a product in a commercial. Successful discovery of the thematic object is important for understanding and summarizing the video content.

There are two major challenges for thematic object discovery. First, as the thematic object is video-dependent, we don’t have a priori knowledge about it, that is, where and when the object will appear in the video. Because no query object is provided, this data-mining problem is challenging. Second, the same thematic object might vary in its appearance due to changes in viewpoint, illumination, scale, and so on. Thus given a cluttered and dynamic video scene, it’s difficult to locate a small thematic object accurately.

To address these two challenges in thematic object discovery, some recent approaches have been proposed. In spite of a moderate success, because these methods mainly focus on mining image data sets, they might not be applicable to video data directly. For example, although a video sequence can be represented by a collection of keyframes, as mentioned earlier, we cannot assume that every keyframe contains the thematic object. It thus introduces great challenges for video object discovery. Moreover, as a video only contains a few images, large-scale image data-mining methods might not be suitable here. It remains a challenging problem to discover the thematic object in a limited number of images. Finally, different from object category discovery where the object usually dominates the whole image, the thematic object can be small and hidden in the cluttered background. It becomes difficult and time consuming, even for humans, to find all of the thematic objects in video.

We propose a novel data-mining method to discover the thematic object in videos. A video sequence is characterized by a number of keyframes and each frame is composed of a collection of local visual features. Given the bag-of-words representation of each keyframe, our goal is to find a subgroup of spatially collocated visual words, whose co-occurrence is frequent, thus corresponding to the thematic object. Our data-mining algorithm is designed to discover this subgroup of visual words by maximizing their overall mutual information scores. Finally, we perform an efficient bounding-box search to locate all instances of the thematic object.

Our testing results on challenging commercial videos show that our approach can discover the thematic object despite its appearance variations due to scale, viewpoint, color, and lighting-condition changes. In addition, our approach is insensitive to partial occlusion as well as cluttered and dynamic backgrounds. Although only tested on commercial videos, our method can be easily extended to other types of videos that contain a thematic object.

Thematic video object discovery

Figures 1a–e illustrate the main steps of our method. First, each video is presented as a sequence of keyframes, and visual features extracted from keyframes are clustered to different visual words. Then, the relation of all visual words is represented as an affinity graph.
Next, the affinity relationship between each pair of words is estimated and the affinities of all pairs of words are concatenated as the affinity matrix. After that, the word group corresponding to the thematic object is discovered by the proposed mining method. Finally, the thematic object instance in each keyframe is located by using the branch-and-bound search method.

Preliminaries

Our method first extracts a set of local visual features from the keyframes, for example, the scale-invariant feature transform (SIFT) features. Each visual feature in keyframe $I_m$ is described by a feature vector $\phi_m(x) = [x, d]$, where the 2D vector $x$ is its spatial location, and the high-dimensional vector $d$ encodes the visual appearance of this feature. Then, a keyframe $I_m$ is represented by a set of visual features $I_m = \{\phi_m(x_1), \ldots, \phi_m(x_p)\}$. Features with similar appearance might occur multiple times in the keyframe set $\{I_m\}_{m=1}^M$, where $M$ is the number of keyframes in this keyframe set. Clustering algorithms, such as $k$-means, group the features in $\{I_m\}_{m=1}^M$ together according to the similarity of their appearance vectors, yielding $N$ visual words $\{W_1, W_2, \ldots, W_N\}$. The features clustered to word $W_i$ are named as instances of word $W_i$. We denote this representation using the feature-word set $\{\psi_{mi}\}$, where $\psi_{mi} = \{\phi_{mi}\}$ represents all instances of word $W_i$ in keyframe $I_m$, and $\phi_{mi}$ is one of these features.

The relation of all visual words can be represented by an affinity graph where each edge represents the relation of two words, as shown in Figure 1b. The affinity value $A_{ij}$ is related to the chance of two words $W_i$ and $W_j$ belonging to the same thematic video object. Intuitively, $A_{ij} \geq 0$ implies that two words have high probability in the same thematic video object, and $A_{ij} \leq 0$ otherwise. The affinity values of all word pairs are concatenated as a $N \times N$ symmetric matrix $A$, as shown in Figure 1c.

Pairwise affinity estimation

Given the visual words representation, we can estimate the affinity $A_{ij}$ between each word pair $P = [W_i, W_j]$. Existing works demonstrate that word co-occurrence is an important criterion for thematic object discovery. However, due to the inherent complexity of a visual pattern, visual words that co-occur frequently don’t always suggest an accurate and meaningful affinity relationship. Even if a word pair appears frequently, it’s not clear whether such co-occurrences among the words are statistically significant or just by chance.

Therefore, inspired by other work, we employ the pointwise mutual information criterion to estimate the affinity relationship between a pair of words:

$$PMI(P) = \log \left( \frac{Pr(P)}{Pr(W_i) \times Pr(W_j)} \right)$$

where $Pr(P)$ is the joint probability of word pair $P$ and $Pr(W_i)$ is the individual probability of word $W_i$. This quantity is positive if the co-occurrences of $W_i$ and $W_j$ are positively related and negative if they are negatively related.
correlated, while negative if they are not correlated.

To estimate the probability $Pr(P)$ and $Pr(W_i)$, simply checking their occurrence frequency is far from sufficient. Even in the same keyframe, one instance of a visual word might belong to the thematic object while the other instances of the same visual word might belong to the background. This situation is serious especially for some highly frequent visual words. Therefore, we first employ the stop-list technology to suppress the most frequent visual words that occur in almost all keyframes, as these visual words have weak discriminative power.

Ideally, a pair of visual words corresponding to the same thematic object not only need to co-occur in different keyframes, but also need to maintain consistent spatial relationship. But in practice it’s difficult to catch this kind of consistency due to the appearance variations of the same thematic object, and the enormous computational cost involved in exploring the huge solution space, including the location, scale, and the number of thematic video objects. So we employ the local property of objects, that is, the features belonging to the same thematic object should be close to each other in the keyframe. As a result, the frequency of $P$ is estimated as

$$\text{frq}(P) = |\{m : \text{Dist} (\phi_{mi}, \phi_{mj}) < \Lambda_0\}| \quad (1)$$

where $\text{Dist}(\phi_{mi}, \phi_{mj})$ represents the minimal distance between instances of $W_i$ and $W_j$ in keyframe $I_m$, and $\Lambda_0$ is a threshold. If $P$ satisfies $\text{Dist}(\phi_{mi}, \phi_{mj}) < \Lambda_0$, then keyframe $I_m$ is an effective keyframe to support word pair $P$, word $W_i$, and word $W_j$. Therefore, the frequency of each word pair $\text{frq}(P)$ equals to the number of their effective keyframes. Similarly, the frequency of each word $\text{frq}(W_i)$ is also the number of its effective keyframes. Figure 2 describes the effective keyframe concept. Moreover, if instances of $W_i$ and $W_j$ satisfy $\text{Dist}(\phi_{mi}, \phi_{mj}) < \Lambda_B$, they will be added to $\Psi_{mi}^B$ and $\Psi_{mj}^B$, respectively. The set $\Psi_{mi}^B \subseteq \Psi_{mi}$ represents the visual features of keyframe $I_m$, which are not only clustered to word $W_i$ but also satisfied the local constrain in Equation 1.

On the basis of the pointwise mutual-information measure, the word pair affinity value finally becomes

$$A_{ij} = \begin{cases} \log \left( \frac{\text{frq}(P)}{M} \right) & \text{if frq}(P) > 0 \\ \tau & \text{else} \end{cases}$$

where $\tau$ is a negative value. If $W_i$ and $W_j$ have strong affinity, the value of $A_{ij}$ is positive. Otherwise, if they have weak affinity, the value of $A_{ij}$ is negative. If $W_i$ and $W_j$ don’t have any instances of co-occurrence, $A_{ij}$ is set to be a negative constant $\tau$.

**Spatially collocated word group mining**

According to our assumption, a thematic object is stable and distinctive over the whole keyframe set. In other words, on the basis of the visual word representation, the thematic object attracts its own unique words while it repels words in the background or other objects. Therefore, following this intuition, we represent the thematic object as the spatially collocated word group $\Omega \subseteq \Pi$, where all words $W_i \in \Omega$ belong to the same thematic video object. This formulation is intuitive, and our objective is to extract spatially collocated word groups one by one. We define the affinity potential function of the group $\Omega$ as

$$S(\Omega) = \sum_{W_i, W_j \in \Omega} A_{ij}$$

and the solution to the following optimization problem gives the maximum spatially collocated word group:

$$\Omega^* = \arg \max_{\Omega \subseteq \Pi} S(\Omega) \quad (2)$$
that is, the group that has the largest affinity potential is the maximum spatially collocated word group.

After obtaining the affinity matrix \( A \) for all word pairs, the subset optimization problem in Equation 2 is converted to a binary optimization problem. Given a group \( \Omega \), let \( f = \{f_i\}_{i=1}^N \) with \( f_i \in \{-1, 1\} \). When \( f_i = 1 \), word \( W_i \) belongs to group \( \Omega \), and vice versa. As the value of \( f \) and the group \( \Omega \) correspond to each other, Equation 2 can be equivalently rewritten as

\[
f^* = \arg \max \left\{ \sum_{i=1}^N \sum_{j=1}^N A_{ij}f_i f_j \right\}
\]

subject to \( f_i = 1, f_j = 1 \) \( \forall i \neq j \).

It’s easy to see that Equation 3 belongs to the unconstrained binary quadratic programming problem, which is nondeterministic polynomial-hard. As the affinity matrix \( A \) has negative values, Equation 3 can’t be solved by the state-of-art simplex-based algorithm. Therefore, to make the optimization feasible, we relax each dimension of the indicator vector \( f_i \) from \([-1, +1]\) to be real-value, but keep the \( L_2 \) norm of \( f \), that is, \( \|f\|_2 = \sqrt{N} \).

We rewrite Equation 3 as

\[
f^* = \arg \max \left\{ \left(1 + \frac{1}{2} f \right)^T A \left(1 + \frac{1}{2} f \right) \right\}
\]

subject to \( f^T f = N \) \( \forall i \neq j \).

where 1 is the all-one vector with proper dimension. To solve the optimization problem of Equation 4, we look at its Lagrangian:

\[
\mathcal{L}(f, \lambda) = \frac{1}{4}(1 + f)^T A (1 + f) - \lambda (f^T f - N)
\]

By taking the derivative and setting

\[
\frac{\partial \mathcal{L}(f, \lambda)}{\partial f} = 0
\]

we obtain

\[
\frac{1}{2} A (1 + f^*) = \lambda f^*
\]

\[
f^* f^T f^* = N
\]

Solving Equation 5 explicitly is difficult. Therefore, as suggested in other work, we employ a fixed-point iteration procedure to solve all the \( N + 1 \) unknowns. When the iteration terminates, we add all words \( W_i \) that satisfies \( f^*_i > \eta \) to \( \Omega^* \). We also select the optimal \( \eta \in [-1, 1] \) that can maximize Equation 2.

**Thematic video object localization**

After obtaining the spatially collocated word group \( \Omega^* \), we can localize the thematic object instance in keyframes by checking the occurrence of their corresponding visual features. For a pixel \( x \) whose corresponding feature is assigned to one word in the spatially collocated word group \( \Omega^* \), we set it to a positive score, indicating the likelihood of the thematic object. Otherwise, we set \( x \) to a negative score, indicating that it doesn’t belong to any thematic object. This is reasonable as these pixels have a low chance of being part of a thematic video object. We assign the commonness score to each pixel \( x \) in \( I_m \) as

\[
C(m, x) = \begin{cases} 
1 & \text{if } \phi_m(x) \in \Psi_{m} \land W_i \in \Omega^* \\
\nu & \text{else}
\end{cases}
\]

where \( \nu \) is a predefined small negative value.

After obtaining the commonness score \( C(m, x) \), we locate the candidate thematic object instance in each keyframe using a bounding box. More formally, for each keyframe \( I_m \), we search for the bounding box \( R_m^* \) with the maximum commonness score:

\[
R_m^* = \arg \max_{R \in A} \sum_{x \in R} C(m, x) = \arg \max_{R \in A} F(R)
\]

where

\[
F(R) = \sum_{x \in R} C(m, x)
\]

is the objective function, and \( A \) denotes the candidate set of all valid subimages in \( I_m \).

To speed up the localization process, we apply the branch-and-bound search proposed in other work. Due to the quantization error of the feature-clustering algorithm, some non-thematic subimages still have a chance of being selected as the candidate thematic object instance. However, the different instances of the same thematic object should keep the geometric consistency. Relying on this, we estimate the affinity relationship between different subimages by checking the geometric consistency using the \( k \)-spatial nearest neighbors (\( k \)-SNN) criterion. Finally, the correct thematic object instances are obtained by reapplying the proposed mining algorithm discussed previously.
Evaluation

To evaluate our algorithm, we test 36 commercial video sequences for thematic object discovery. We sample keyframes from each video at two frames per second, and discover thematic objects from these extracted keyframes. For each video sequence, we only discover the first thematic object, that is, the thematic object corresponding to the maximum spatially collocated word group. In our visual words representation, SIFT local features are extracted from each keyframe and characterized by the 128 dimensional descriptors. For each sequence, the local features are quantized into 1,000 visual words by the k-means clustering. The top 10 percent of commonly appearing visual words that occur in almost all keyframes are suppressed in the following experiments. To employ the local property of thematic objects, the distance threshold $\Delta_D$ is set according to the size of each keyframe, that is we set $\Delta_D = 0.33W$ where $W$ denotes the width of the keyframe. If two visual words do not have any co-occurrence in the keyframe sets, their affinity value is set to be $-3$, that is, $\tau = -3$. To localize the thematic object in each keyframe, the commonness score of noncommon pixels are set to be $-0.01$, that is, $\nu = -0.01$. All the experiments are performed on a standard Xeon 2.67-GHz PC.

To quantify the performance of the proposed approach, we select the product presented in each commercial video as the ground-truth thematic object. As one commercial video clip usually presents one specific product, this kind of ground truth has less bias. We manually labeled the ground-truth bounding boxes of the thematic objects in each keyframe. Let $DR$ and $GT$ be the discovered subimages and the bounding boxes of ground truth, respectively. The performance is measured by using an intersection-over-union metric. The

$$\text{intersection-over-union} = \frac{|GT \cap DR|}{|GT \cup DR|}$$

criterion is used to measure the overlap between the discovered subimages and the bounding boxes of the ground truth. To count the intersection-over-union value for one video, the intersection-over-union value is calculated for each keyframe and then the average value of all keyframes is used to evaluate the whole video.

Thematic object discovery from video sequences

To evaluate the proposed approach on thematic objects discovery, we test our method on 36 video sequences from YouTube. All of the sequences are commercial advertisements, where the length of videos range from 7 to 41 seconds. Figure 3 shows some sample results of thematic object discovery. In the video sequences, the thematic objects are subject to different variations, such as rotation (row two, image three); partial occlusion (row four, image five); and scale, viewpoint, and lighting condition changes (row one, image six).

Table 1 summarizes the performance of our algorithm on 36 video sequences. For each sequence, the table shows the number of keyframes (# frame), the number of total thematic object instances (# instances), the number of correctly detected thematic object instances (# correct) and its ratio to total thematic object instances (CorrectRa), and the number of falsely detected thematic object instances (CorrectRa).
and its ratio to total thematic object instances (FalseRa). These discovery results show that the proposed approach performs well for discovering identical thematic object from video sequences.

Figure 4 (next page) illustrates the intersection-over-union values of all 36 sequences. We compare the performance when the threshold $\lambda_D$ is set to different values. Although the best performance is obtained when $\lambda_D = 0.33W$, it is interesting to note that the intersection-over-union values are very close for different $\lambda_D$.

Comparison with other approaches

We compare our method with two other methods: topic discovery with multiple segmentations, and a PageRank-based approach. The topic-discovery-based method is proposed in Russell et al. First, each keyframe is segmented multiple times using a normalized cut, with a different number of segments $K$. Given multiple segments, we expect that one of the segments will cover the thematic object. By clustering visual features into a visual vocabulary, each segment is further represented as a bag of words. After obtaining a pool of segments (visual documents) from all of the keyframes, common topics are discovered using text-mining methods. Finally, for each discovered topic, all segments of the same keyframe are sorted by how well they are explained by that topic. The segment at the highest rank is selected as the thematic object. To obtain the bag-of-words presentation, both SIFT and maximally stable extremal regions (MSER) local features are extracted from each keyframe and characterized by the SIFT descriptor of 128 dimensions.

For each sequence, the same types of local features are quantized into 1,000 visual words by $k$-means clustering. The second approach is the well-known PageRank method, which is an efficient method for searching rank. As this method can only assign each word a numerical weight, which represents the probability of one word belonging to the thematic word group, we have to decide on the size of thematic word group. Therefore, we test different thresholds: select the same number of words as the proposed method and select the top 10, 20, 30, and 40 percent of the words.

Figure 5 presents the quantitative comparison of the different approaches. Overall, our proposed approach outperforms both the topic discovery and PageRank approaches in terms of the intersection-over-union, with an average score of 0.38 (for our proposed method) compared to 0.20 (for topic discovery) and 0.31 (for PageRank). In addition, when the size of the thematic word group is set as 10, 20, 30, and 40 percent of the total number of visual words, the average intersection-over-union scores for the PageRank method are about 0.31, 0.34, 0.35, and 0.33, respectively. Table 2 compares the running time of the proposed

---

**Table 1. Evaluation of our method with different sequences.**

<table>
<thead>
<tr>
<th>Sequence</th>
<th># frame</th>
<th># instances</th>
<th># correct</th>
<th># false</th>
<th>CorrectRa</th>
<th>FalseRa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59</td>
<td>30</td>
<td>26</td>
<td>0</td>
<td>0.87</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>35</td>
<td>29</td>
<td>0</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>59</td>
<td>17</td>
<td>17</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
<td>22</td>
<td>22</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>49</td>
<td>21</td>
<td>21</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>60</td>
<td>13</td>
<td>13</td>
<td>2</td>
<td>1.00</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>19</td>
<td>18</td>
<td>17</td>
<td>0</td>
<td>0.94</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>13</td>
<td>10</td>
<td>0</td>
<td>0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>19</td>
<td>19</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>22</td>
<td>22</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>21</td>
<td>20</td>
<td>0</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
<td>15</td>
<td>8</td>
<td>4</td>
<td>0.53</td>
<td>0.27</td>
</tr>
<tr>
<td>14</td>
<td>49</td>
<td>19</td>
<td>17</td>
<td>0</td>
<td>0.89</td>
<td>0.00</td>
</tr>
<tr>
<td>15</td>
<td>59</td>
<td>21</td>
<td>15</td>
<td>1</td>
<td>0.71</td>
<td>0.05</td>
</tr>
<tr>
<td>16</td>
<td>59</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>1.00</td>
<td>0.11</td>
</tr>
<tr>
<td>17</td>
<td>82</td>
<td>33</td>
<td>33</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>18</td>
<td>57</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>19</td>
<td>58</td>
<td>57</td>
<td>57</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>20</td>
<td>59</td>
<td>16</td>
<td>16</td>
<td>5</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>21</td>
<td>30</td>
<td>22</td>
<td>21</td>
<td>0</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>22</td>
<td>40</td>
<td>18</td>
<td>14</td>
<td>0</td>
<td>0.78</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>32</td>
<td>17</td>
<td>17</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>24</td>
<td>33</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>19</td>
<td>18</td>
<td>0</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>26</td>
<td>19</td>
<td>14</td>
<td>14</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>27</td>
<td>19</td>
<td>19</td>
<td>18</td>
<td>0</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>16</td>
<td>11</td>
<td>9</td>
<td>1</td>
<td>0.82</td>
<td>0.09</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>0</td>
<td>0.95</td>
<td>0.00</td>
</tr>
<tr>
<td>30</td>
<td>37</td>
<td>22</td>
<td>19</td>
<td>2</td>
<td>0.86</td>
<td>0.09</td>
</tr>
<tr>
<td>31</td>
<td>34</td>
<td>21</td>
<td>20</td>
<td>6</td>
<td>0.95</td>
<td>0.28</td>
</tr>
<tr>
<td>32</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>33</td>
<td>59</td>
<td>20</td>
<td>14</td>
<td>13</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>34</td>
<td>59</td>
<td>43</td>
<td>12</td>
<td>0</td>
<td>0.28</td>
<td>0.00</td>
</tr>
<tr>
<td>35</td>
<td>31</td>
<td>18</td>
<td>18</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>36</td>
<td>59</td>
<td>22</td>
<td>8</td>
<td>7</td>
<td>0.36</td>
<td>0.31</td>
</tr>
</tbody>
</table>
The three approaches are implemented in Matlab. Our approach requires only about 200 seconds to process the 30-second commercial video clip and it’s about 10 times faster than the topic-discovery-based approach.

This comparison clearly shows the advantages of combining the spatially collocated word group mining technique and the branch-and-bound search algorithm. On the one hand, the topic-discovery-based approach doesn’t consider the spatial relationship among the visual features and its results highly depend on the performance of the key-frame segmentation. Even though each key-frame is segmented multiple times with a different number of segments, the thematic object is often not well extracted. Due to multiple objects and the cluttered background per keyframe, obtaining a reliable segmentation is not a trivial task. In this case, the topic-discovery approach only obtains a coarse discovery of the thematic object, which is far from satisfactory. On the other hand, the PageRank method can only obtain better results in several video data sets. However, in the thematic-pattern-mining applications, there is no a priori knowledge about the size of thematic word groups.

**Conclusion**

Thematic video object discovery is a challenging problem due to the possible visual pattern variations and the large computational cost when exploring the candidate set without a priori knowledge. We found that spatially collocated word groups can be discovered efficiently by using our proposed binary-optimization technique. We also found that such a spatially collocated word-group-mining procedure is efficient and robust, because the qualified candidates of visual words are always kept, while unqualified visual words are discarded.

Compared with previous approaches that use the bag-of-words model for thematic object discovery, our approach has the ability to capture the thematic objects globally and detect thematic objects that appear at random

---

**Figure 4.** Performance comparison when using different distance threshold $\Lambda_D$. The data sets are ordered according to the intersection-over-union values of the proposed method. Data set 37 shows the average intersection-over-union value of all data sets.

**Figure 5.** Performance comparison of our approach, topic-discovery approach, and PageRank approach. For the PageRank approach, we select the same number of words as the proposed method. The data sets are ordered according to the intersection-over-union values of the proposed method. Data set 37 shows the average intersection-over-union value of all data sets.

**Table 2.** Time costs to process a 30-second commercial video clip.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>205</td>
</tr>
<tr>
<td>Topic</td>
<td>2,084</td>
</tr>
<tr>
<td>PageRank</td>
<td>203</td>
</tr>
</tbody>
</table>

Method with that of the two other methods. The three approaches are implemented in Matlab. Our approach requires only about 200 seconds to process the 30-second
Related Work

To discover common objects from images, some previous approaches characterize an image as a graph or a tree composed of visual features, such as corners, interest points, and image segments.\(^1\) The common object discovery is formulated by mining the common subgraph or subtree from a collection of graphs\(^4\) or trees.\(^3\) However, in general, for the graph-based image representation, matching and mining subgraphs are computationally demanding. Several other approaches discover common objects through local feature matching. In Liu and Yan, spatially coherent constraints are employed for finding the common object from a pair of images.\(^5\) However, this solution is not designed for object discovery from a collection of images, or a video sequence.

In Heath et al., the random sample consensus (Ransac) algorithm is used to remove matching outliers and discover the common object from images.\(^6\) To make Ransac work, the approach has to provide the transform type information between different instances of a common object. Bagon et al. detects and sketches the common object from only a few images.\(^7\) However, as the employed self-similarity descriptor is not scale-invariant, this method has a limited ability in handling the scale variations of the common object.

Another category of approaches of object discovery applies the conventional topic-discovery approach in text documents, by translating each image into a collection of visual words through clustering visual features. To obtain the object boundary, Russell et al. segments the image multiple times and expects that one of the obtained segments will correspond to the target object.\(^8\) However, even though each image is segmented multiple times with a different number of segments, the method can’t guarantee that the common object is correctly extracted. The methods proposed in work by Yuan, Wu, and Yang perform spatial image data mining to search for frequently co-occurring words that are also spatially collocated.\(^9\) The approach in Gao et al. goes one step further by introducing the geometric relationships.\(^11\) However, the proposed graph-mining method can’t easily handle the scale changes of the common object as well. Besides mining objects from images, there is also recent work in discovering objects from video sequences.\(^12\) This method needs supervision information about the object; user labeling is required to initialize the search. Thus this approach is not fully unsupervised.

locations within video sequences, vary in either scales or shapes, or have missing features. Experiments on challenge video sequences show that our method is scalable to a video sequence, without requiring exhaustive checking for the scale, location, and number of thematic objects. In the future, we plan to further develop the method and employ it to process other types of data, for example, text and audio.

References


Acknowledgment

This work is supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Junsong Yuan.

References


Gangqiang Zhao is a research fellow at Nanyang Technological University. His research interests include computer vision, multimedia data mining, and content analysis. Zhao has a PhD from Zhejiang University, Hangzhou, China. Contact him at gqzhao@ntu.edu.sg.

Junsong Yuan is an assistant professor at Nanyang Technological University. His research interests include computer vision, image and video data mining and content analysis. Yuan has a PhD in electrical engineering and computer science from Northwestern University. Contact him at jsyuan@ntu.edu.sg.

Jiang Xu is a PhD candidate in the Department of Electrical Engineering and Computer Science at Northwestern University. His research interests include computer vision and pattern recognition. Contact him at jxu323@vision1.ece.northwestern.edu.

Ying Wu is an associate professor of electrical engineering and computer science at Northwestern University. His research interests include computer vision and multimedia data mining. Wu has a PhD in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, Illinois. Contact him at yingwu@northwestern.edu.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.