Discovering Thematic Patterns in Videos via Cohesive Sub-graph Mining

Gangqiang Zhao, Junsong Yuan
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore
gqzhao@ntu.edu.sg, jsyuan@ntu.edu.sg

Abstract—One category of videos usually contains the same thematic pattern, e.g., the spin action in skating videos. The discovery of the thematic pattern is essential to understand and summarize the video contents. This paper addresses two critical issues in mining thematic video patterns: (1) automatic discovery of thematic patterns without any training or supervision information, and (2) accurate localization of the occurrences of all thematic patterns in videos. The major contributions are two-fold. First, we formulate the thematic video pattern discovery as a cohesive sub-graph selection problem by finding a sub-set of visual words that are spatio-temporally collocated. Then spatio-temporal branch-and-bound search can locate all instances accurately. Second, a novel method is proposed to efficiently find the cohesive sub-graph of maximum overall mutual information scores. Our experimental results on challenging commercial and action videos show that our approach can discover different types of thematic patterns despite variations in scale, view-point, color and lighting conditions, or partial occlusions. Our approach is also robust to the videos with cluttered and dynamic backgrounds.

Keywords—thematic pattern; unsupervised; cohesive sub-graph; mining;

I. INTRODUCTION

One category of videos usually shares the same thematic pattern. Such a thematic pattern can be the visual object that is frequently highlighted in the video, e.g., the product logo in a commercial video, or a specific event that appears commonly, e.g. spin action in skating videos. It is of great interests to discover thematic patterns in videos as they are essential to the understanding and summarization of the video contents.

Even though tremendous progress has been made for the frequent pattern mining over a decade [3], there are three major challenges for mining thematic visual pattern in videos. First, visual patterns generally exhibit large variabilities in their visual appearances. Different instances of the same thematic pattern may vary significantly due to viewpoint and illumination changes, scale changes, partial occlusion, not to mention the large variations of videos events. Second, visual patterns have complex structures. Different visual items (image patches) can be correlated due to temporal and spatial dependency and co-occurrences. Finally, it is challenging to locate the thematic pattern accurately in the cluttered and dynamic video scenes. Therefore, it is difficult and time consuming, even for human beings, to find the thematic patterns in videos accurately.

To deal with the above challenges, some recent approaches have been proposed to discover common objects in images [14] [11] [9]. Despite a moderate success, these image-based methods cannot be extended to video data directly. For example, to represent the images using the transaction data, Russell et al. have proposed to segment the image multiple times [9]. However, it is very difficult, if not impossible, to segment the video sequences multiple times due to the high computational cost. Besides, the thematic video object may be small and hidden in the cluttered background, while existing object discovery approaches usually assume the object dominates the whole image [11]. Furthermore, for thematic event, it is important to take its spatio-temporal characteristics into consideration, while previous image-based methods only deal with spatial patterns [15].

We propose a novel thematic video pattern discovery method that addresses the challenges mentioned above. A video sequence is characterized by a number of key frames and each frame is composed of a collection of local visual features. After clustering the features into visual words, we measure the pairwise relationship between two words using the mutual information criterion and represent the relationship of all words as an affinity graph. The thematic pattern becomes a cohesive sub-graph which has the maximum overall mutual information score and this sub-graph can be obtained efficiently by the proposed mining approach. In addition, to identify the thematic pattern, we perform an efficient bounding box search to locate all of the instances of the thematic pattern. Consequently, the following contributions have been made in this paper:

1) To the best of our knowledge, this is the first attempt to unsupervised extraction of meaningful video themes.
2) A new formulation for thematic pattern discovery is proposed. The cohesive sub-graph formulation is robust with respect to visual pattern variations.
3) A novel algorithm is proposed to efficiently find the cohesive sub-graph by solving the corresponding binary quadratic programming problem.

The experimental results on challenging commercial videos and action datasets show that our approach can discover the thematic pattern despite its variations due to scale, view-point, color and lighting condition changes.
II. RELATED WORK

Most existing common pattern discovery methods fall into one of the four categories: local feature matching based methods, frequent pattern mining based methods, topic model based methods, and sub-graph matching based methods. Local feature matching based methods discover the common pattern by directly matching the local features through RANSAC algorithm [4] or matching the local features through hash indexing [15]. These methods can obtain the good mining performance but result in a very high computational cost. Frequent pattern mining based methods first translate each image into a collection of “visual words” and then discover the common pattern through frequently co-occurring words mining [14] [12] [10]. To represent each image using the transaction data, Yuan et al. consider the spatial K-nearest neighbors (K-NN) of each local features as a transaction record [14]. However, it is difficult to select the size K of the nearest neighborhood as there is no a priori knowledge about the thematic pattern scale.

Topic model based methods discover the common pattern through topic discovery [9] [2]. To represent the images using the transaction data, Russell et al. proposed to segment the image multiple times [9]. After obtaining a pool of segments from all images, object topics are discovered using Latent Dirichlet Allocation model (LDA) and the most supportive topic is selected as the common topic. The segments corresponding to the common topic are selected as the common objects. Traditional sub-graph matching methods characterize an image as a graph or a tree composed of visual features. Then, the common pattern is discovered by graph or tree matching [11]. However, the existing approaches require the training step and have high computational cost.

III. THEMATIC PATTERN DISCOVERY

To discover the thematic patterns from videos, visual features are extracted from key frames and clustered into visual words first. Then, we build an affinity graph to capture the pairwise relationships of the visual words in all key frames. Next, the cohesive sub-graph corresponding to the thematic pattern is discovered by the proposed mining method. Finally, instances of the thematic pattern are located using the branch-and-bound method. Figure 1 illustrates the main steps of our method. The following sub-sections describe details about these steps.

A. Preliminaries

Our method first extracts a set of local visual features from the key frames, e.g., the SIFT features [6] or spatial-temporal interest point (STIP) features [5]. Each visual feature in key frame $I_m$ is described as a feature vector $\phi_m(u) = [u, d]$, where the 3D vector $u$ is its spatial location and temporal order, and the high-dimensional vector $d$ encodes the visual appearance of this feature. Then, a key frame $I_m$ is represented by a set of visual features $I_m = \{\phi_m(u_1), ..., \phi_m(u_p)\}$. Clustering algorithms, such as k-means, group the features in $\{I_m\}_{m=1}^M$ according to the similarity between their appearance vectors, yielding $N$ visual words $\Pi = \{W_1, W_2, ..., W_N\}$. The features clustered to word $W_i$ are named as instances of word $W_i$. We denote this representation as $\{\Phi_{mi}\}_{m=1}^M_{i=1}^N$, where $\Phi_{mi} = \{\phi_{mi}\}$ represents all instances of word $W_i$ in key frame $I_m$, 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 between two words, as shown in Figure 1(a). As is customary, we represent the affinity graph with the corresponding affinity matrix $A$. Given two words $W_i$ and $W_j$, the affinity value $A_{i,j}$ describes the chance of two words $W_i$ and $W_j$ belonging to the same thematic pattern. While $A_{i,j} \geq 0$ implies that the two words have a high probability of being in the same thematic pattern, and $A_{i,j} \leq 0$ otherwise.

B. Pairwise Affinity Measurement

Given the visual words representation, we can estimate the affinity $A_{i,j}$ between each wordpair $P = \{W_i, W_j\}$. Existing works demonstrate that word co-occurrence is an important criterion for common pattern discovery [9]. However, due to the inherent complexity of a visual pattern, visual words that co-occur frequently do not always suggest an accurate and meaningful affinity relationship. Even if a wordpair appears frequently, it is not clear whether such co-occurrences are statistically significant or just by chance. Therefore, inspired by the work [14], we employ the pointwise mutual information criterion to estimate the affinity relationship of two words:

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

(1)
where \( Pr(\mathcal{P}) \) is the joint probability of wordpair \( \mathcal{P} \) and \( Pr(W_i) \) is the individual probability of word \( W_i \). This quantity can take on both negative and positive values and is zero if \( W_i \) and \( W_j \) are independent. Its value is positive if \( W_i \) and \( W_j \) are positively correlated, while it is negative if they are negatively correlated.

To estimate the probability \( Pr(\mathcal{P}) \) and \( Pr(W_i) \), we assume the features belong to the same thematic pattern should be close in the same key frame. Therefore, we count the number of effective occurrence of \( \mathcal{P} \) as:

\[
N(\mathcal{P}) = |\{m : D_{\min}(\phi_{mi}, \phi_{mj}) < \Lambda_D\}|, \tag{2}
\]

where \( D_{\min}(\phi_{mi}, \phi_{mj}) \) represents the minimal distance between instances of \( W_i \) and \( W_j \) in key frame \( I_m \), and \( \Lambda_D \) is a threshold. For spatio-temporal features, \( D_{\min}(\phi_{mi}, \phi_{mj}) \) should also consider the temporal information. The number of effective occurrence of each word \( N(W_i) \) is also obtained as \( N(W_i) = |\{m : \exists j, D_{\min}(\phi_{mi}, \phi_{mj}) < \Lambda_D\}| \). Instances of \( W_i \) and \( W_j \) that satisfy \( D_{\min}(\phi_{mi}, \phi_{mj}) < \Lambda_D \), will be added to \( \Phi_{mi} \) and \( \Phi_{mj} \), respectively. The set \( \Phi_{mi} \subseteq \Phi_{mi} \) represents the visual features of key frame \( I_m \) which contribute to the estimation of \( N(W_i) \).

Based on the point-wise mutual information criterion, the wordpair affinity value is:

\[
A_{i,j} = \log \frac{N(\mathcal{P})/M}{N(W_i)/M \times N(W_j)/M}, \tag{3}
\]

where \( M \) is the number of frames. If \( W_i \) and \( W_j \) do not have any effective co-occurrence, \( A_{i,j} \) is set to a fixed negative value \( \tau \).

### C. Thematic Pattern Discovery Formulation

In order to discover thematic patterns from videos, we consider the integrity and uniqueness of the visual pattern’s representation. In other words, the thematic pattern is composed of a specific group of words. Therefore, following this intuition, we represent the thematic pattern as the cohesive sub-graph and denote this sub-graph using its vertices set \( \Omega \), where elements of \( \Omega \) are the words belong to the same thematic pattern. We define the affinity potential function of the sub-graph \( \Omega \) as \( f(\Omega) = \sum_{W_i, W_j \in \Omega} A_{i,j} \) and the solution to the following optimization problem gives the maximum cohesive sub-graph:

\[
\Omega^* = \arg\max_{\Omega \subseteq \mathcal{H}} f(\Omega), \tag{4}
\]

i.e., the sub-graph that has the largest affinity potential is the maximum cohesive sub-graph. As each thematic pattern is presented by a sub-graph, we can discover them one by one. After one pattern is discovered, the features belong to this pattern will be removed and another pattern can be found.

When obtaining the affinity matrix \( A \) for all wordpairs, the subset optimization problem in Eq. 4 is converted to a binary optimization problem. Given a sub-graph \( \Omega \), let \( x = \{x_i\}_{i=1}^N \) with \( x_i \in \{-1, 1\} \) represents its indicator vector. When \( x_i = 1 \), word \( W_i \) belongs to sub-graph \( \Omega \), and vice versa. As the indicator vector \( x \) and the sub-graph \( \Omega \) correspond to each other, Eq. 4 can be rewritten as:

\[
x^* = \arg\max_x f(x) = \frac{1}{4}(1 + x)^T A (1 + x), \tag{5}
\]

s.t. \( x_i \in \{-1, 1\}, i = 1, \ldots, N \),

where \( f(x) = \frac{1}{4}(1 + x)^T A (1 + x) \) is the objective function. Eq. 5 is a binary quadratic programming (BQP) problem. Since \( A \) may not be the positive definite matrix, the objective function \( f(x) \) can be non-convex, thus it is a NP problem. Section III-D describes the proposed solution of this problem.

### D. Cohesive Sub-graph Mining

To solve Eq. 5, we observe that a binary constraint \( x_i \in \{-1, 1\} \) is always equivalent to an equilibrium constraint, i.e., \(-1 \leq x_i \leq 1, (1 + x_i)(1 - x_i) = 0 \). Furthermore, this equilibrium constraint is implied by the nonlinear complementarity problem (NCP) function \( \psi(1 + x_i, 1 - x_i) = 0 \) [1]. In the implementation, we select the popular Fischer-Burmeister function \( \psi(a, b) = \sqrt{a^2 + b^2} - (a + b) \) and obtain the differentiable constraint functions:

\[
\psi(1 + x_i, 1 - x_i) = \sqrt{2 + 2x_i^2} - 2 = 0. \tag{6}
\]

To simplify the representation, we denote the Fischer-Burmeister function of Eq. 6 as \( \psi(x_i) \).

To deal with the \( \psi(x_i) \) constraint, we introduce the quadratic penalty \( \sum_{i=1}^N \psi^2(x_i) \) into the objective function:

\[
x^* = \arg\max_x F(x, \beta) = f(x) - \frac{\beta}{2} \sum_{i=1}^N \psi^2(x_i), \tag{7}
\]

s.t. \( x^T x = N \),

where \( \beta > 0 \) is a penalty parameter. To solve the optimization problem of Eq. 7 with a specific penalty parameter \( \beta \), we look at its Lagrangian: \( L(x, \beta, \lambda) = f(x) - \frac{\beta}{2} \sum_{i=1}^N \psi^2(x_i) - \lambda(x^T x - N) \). By taking the derivative and setting \( \frac{\partial L(x, \beta, \lambda)}{\partial x} = 0 \), we obtain:

\[
\frac{1}{2} A(1 + x) - \beta \sum_{i=1}^N \psi(x_i) \psi'(x_i) = 2 \lambda x, \tag{8}
\]

\[
x^T x = N.
\]

Solving Eq. 8 explicitly is difficult. Therefore, as suggested by [16], we employ a fixed point iteration procedure to obtain the solution \( x \). By adding the \( -\frac{\beta}{2} \sum_{i=1}^N \psi^2(x_i) \) into the objective function, it can not only incorporate the constraint but also obtain a concave objective function when the penalty parameter \( \beta \) is large enough. This advantage greatly contributes to the searching for an optimal solution or a favorable suboptimal solution of Eq.5 via a sequence of maximization procedures with an increasing penalty parameter.
Algorithm 1: Cohesive sub-graph mining

\textbf{input}: Matrix \( A \), number of iterations \( K \), penalty parameter \( \beta_0 \), perturbation threshold \( \delta \).

\textbf{output}: the maximum cohesive sub-graph \( \Omega^* \).

1. construct random solution \( x \)
2. \( x_{best} = x \), \( A = A \)
3. for \( k = 1 \) to \( K \) do
4. \( \beta = \beta_0 \)
5. find the solution \( x^a \) with original matrix \( A \)
6. \( x^a = \arg \max_x f(x) = \frac{1}{N} \sum_{i=1}^{N} \psi^2(x_i) \)
7. if \( f(x^a) > f(x_{best}) \) then
8. \( x_{best} = x^a \)
9. end
10. end
11. obtain sub-graph \( \Omega^* \) based on \( x_{best} \).

To explore the neighborhood of the solution for different \( \beta \), we also perturb the matrix \( A \) by a small quantity and hope that this change could lead to a better solution. The perturbation of matrix \( A \) is obtained by adding an \( N \times N \) perturbation matrix \( P \), where entries of the matrix \( P \) (\( 0 \leq P_{ij} \leq 1 \)) are randomly generated based on the values of the \( x \) vector. The proposed sub-graph mining algorithm is summarized in Alg 1. The maximum cohesive sub-graph is obtained according to the best solution found after all iterations.

After obtaining the cohesive sub-graph \( \Omega^* \), we can locate the thematic patterns in videos via the occurrences of their corresponding visual features [16]. To speed up the localization process for thematic patterns, we apply the branch-and-bound search proposed in [13].

IV. Evaluation

To evaluate our approach, we test it on challenging commercial videos and action video collections for thematic pattern discovery. In addition, we compare the proposed approach with the state-of-the-art methods [9] [7].

A. Video Dataset

In the first experiment, we discover thematic objects from twenty video sequences downloaded from YouTube.com. We test our method on the video sequences one by one. Each video sequence is one advertisement, where the length of videos range from 7 to 41 seconds. In the second experiment, we apply our method to five action video collections to discover the thematic actions. Two of them (i.e., Hand Clap and Hand Wave) come from MSR action dataset [13] and the other two (i.e., Jumping Jack and Golf Swing) come from UCF action dataset [8]. The last one (Figure Skating Spin) is downloaded directly from YouTube.com. We test our method on the video collections one by one. Both datasets ([13], [8]) for action discovery are available publicly.

B. Experimental Setting

Several parameters should be set first. The distance threshold \( \Lambda_D \) is set according to the size of each key frame, i.e. we set \( \Lambda_D = 0.33 \times W \), where \( W \) denotes the width of the video frame. If two visual words do not have any co-occurrence in the datasets, their affinity is set to be -3, i.e., \( \tau = -3 \). For the cohesive sub-graph mining algorithm, the number of iterations is set to be 20, i.e., \( K = 20 \), the penalty parameter is set to be 1.1, i.e., \( \beta_0 = 1.1 \) and the perturbation threshold is set to be a very small number, i.e., \( \delta = 0.01 \). All these parameters are fixed in our experiments. All the experiments are performed on a Xeon 2.67GHz PC and our approach is implemented in Matlab.

To quantify the performance of the proposed approach, we manually labeled the ground truth bounding boxes of the instances of thematic patterns in each dataset. The discovered bounding boxes are decided by the branch-and-bound search method [13]. Let \( DR \) and \( GT \) be the discovered bounding boxes and the bounding boxes of ground truth, respectively. The performance is measured by two criteria: \( \text{precision} = \frac{|GT\cap DR|}{|DR|} \) and \( \text{recall} = \frac{|GT\cap DR|}{|GT|} \). By combining \( \text{precision} \) and \( \text{recall} \), we use a single \( F \)-measure as the metric for performance evaluation [11].

C. Thematic Object Discovery from Videos

Many commercial videos contain the thematic objects, e.g., the Starbucks logo in a commercial video of Starbucks coffee. Such a thematic object usually appears frequently, and the discovery of it is essential for understanding and summarizing the video contents. The employed videos are 24 frames per second and we sample key frames from each video at two frames per second, and discover the instances of thematic objects from these extracted key frames. For each video sequence, we only discover the most significant thematic object, i.e., the thematic object corresponding to the maximum cohesive sub-graph. In our visual words representation, SIFT local features are extracted from each key frame. For each sequence, the local features are quantized into 1000 visual words by the \( k \)-means clustering. The top 10% frequent visual words that occur in almost all key frames are discarded in the experiments. Table I summarizes the information of twenty video sequences. For each sequence, the number of key frames (\( F_{No.} \)) and the ground truth number of thematic object instances (\( I_{No.} \)) are shown in the first and second rows, respectively.

The first row of Figure 2 shows some sample results of thematic object discovery. In the video sequences, the
thematic objects are subject to variations introduced by partial occlusions, scale, viewpoint and lighting condition changes. It is possible that one video sequence contains multiple thematic objects and some frames do not contain any thematic objects. We also count the number of correctly detected instances of thematic object CNo and the number of falsely detected instances of thematic object WNo for each sequence. Their ratios to the ground truth number of thematic object instances are calculated as CorrectRatio = CNo \over TNo and FalseRatio = WNo \over TNo. Figure 3(a) illustrates the Correct Ratio and False Ratio of all twenty videos and the average Correct Ratio of twenty videos is about 94% while the average False Ratio is about 3%.

Moreover, our method is also able to find thematic objects from video collections, as shown in the second row of Figure 2. These results show that the proposed approach performs well for discovering identical thematic objects from video sequences.

D. Thematic Action Discovery from Video Collections

Discovering actions in the video space is much more complicated than discovering objects in the image space. In this experiment, we discover the thematic action from five different action video collections. For each video collection, we only discover the first thematic action, i.e., the thematic action corresponding to the maximum cohesive sub-graph.
The third row of Figure 2 shows some sample results of thematic action discovery. In the video sequences, the actions are subject to the intra-pattern variations of actions such as scale and speed variations, and dynamic and cluttered backgrounds and even partial occlusions. Figure 3(c) illustrates the Correct Ratio and False Ratio of all five video collections and the average Correct Ratio of five video collections is about 90% while the average False Ratio is about 9%. Figure 3(d) illustrates the F-measure of all 5 video collections. These mining results show that the proposed approach performs well for mining thematic actions from video collections.

E. Comparison with Other Approaches

We compare our thematic pattern discovery method with two other methods: (1) topic discovery approach and (2) dominant set mining approach. The topic discovery approach [9] is the state-of-the-art approach for object categorization and object discovery. To discover thematic actions from action video collections, we employ the LDA model [9] directly and select the most supportive topic as the thematic action topic. In the second method, we use the dominant set mining approach as described in [7]. As this method only provides the probability of each word that belongs to the dominant set, we have to set a probability threshold to decide whether one word is selected or not. Therefore, we select the same number of words as the proposed method.

As shown in Figure 3(b), our proposed approach outperforms both topic discovery approach and dominant set mining approach in terms of the F-measure for thematic object discovery, with an average score of 0.63 (Proposed) compared to 0.47 (Dominant set mining) and 0.32 (Topic discovery), respectively. The topic discovery approach does not consider the spatial relationship of the visual features and its results highly depend on the performance of the key frame segmentation. Due to multiple objects and the cluttered background per key frame, performing a reliable segmentation is not a trivial task. In this case, topic discovery approach only obtains a very coarse discovery of the thematic object, which is far from being satisfactory. The dominant set mining can obtain good results in several video collections and the average Correct Ratio of five video collections is about 90% while the average False Ratio is about 9%. Figure 3(d) illustrates the F-measure of all 5 video collections. These mining results show that the proposed approach performs well for mining thematic actions from video collections.

VI. CONCLUSION

Thematic pattern discovery in videos is a challenging problem due to the possibly large visual pattern variations and the prohibitive computational cost to explore the candidate set without a priori knowledge. By representing the affinity relations of all words as a graph, we formulate the thematic pattern discovery problem as a novel cohesive sub-graph mining problem and obtain its solution by solving the binary quadratic programming problem. Our approach has the ability to identify the thematic pattern and accurately locate its instances in the cluttered and dynamic video scenes. Experiments on challenge commercial videos and action video collections show that our method is efficient, robust and accurate. Future work can be carried out to test the cohesive sub-graph mining algorithm on other applications such as video clustering and categorization. In addition, our method can be extended to consider more cues like video saliency and feature distinctiveness information.

REFERENCES