Topical Video Object Discovery from Key Frames by Modeling Word Co-occurrence Prior

Gangqiang Zhao, Junsong Yuan
School of EEE, Nanyang Technological University
Singapore, 639798
{gqzhao,jsyuan}@ntu.edu.sg

Gang Hua
Stevens Institute of Technology
Hoboken, NJ 07030
ghua@stevens.edu

Abstract

A topical video object refers to an object that is frequently highlighted in a video. It could be, e.g.*, the product logo and the leading actor/actress in a TV commercial. We propose a topic model that incorporates a word co-occurrence prior for efficient discovery of topical video objects from a set of key frames. Previous work using topic models, such as Latent Dirichelet Allocation (LDA), for video object discovery often takes a bag-of-visual-words representation, which ignored important co-occurrence information among the local features. We show that such data driven co-occurrence information from bottom-up can conveniently be incorporated in LDA with a Gaussian Markov prior, which combines top down probabilistic topic modeling with bottom up priors in a unified model. Our experiments on challenging videos demonstrate that the proposed approach can discover different types of topical objects despite variations in scale, view-point, color and lighting changes, or even partial occlusions. The efficacy of the co-occurrence prior is clearly demonstrated when comparing with topic models without such priors.*

1. Introduction

With the prevalence of video recording devices and the far reach of online social video sharing, we are now making more videos than ever before. The videos usually contain a number of topical objects, which refer to objects that are frequently highlighted in the video, e.g., the leading actor/actress in a film. It is of great interests to automatically discover topical objects in videos efficiently as they are essential to the understanding and summarization of the video contents.

One potential approach to automatically discover video objects is using frequent pattern mining [6]. Although significant progress has been made along this path [21] [19], it is still a challenge to discover topical objects in videos automatically using frequent pattern mining methods. As a bottom-up approach, frequent pattern mining requires the predefined items and vocabularies. However, different instances of the same video object may endure significant variabilities due to viewpoint, illumination changes, scale changes, partial occlusion, etc. This makes the frequent item set mining with video data to be very difficult with the ambiguity of visual items and visual vocabularies.

To mitigate this challenge, several methods have been proposed to discover topical objects in images and videos [21] [17] [14] [15] [23]. Notwithstanding their successes, these methods are limited in different ways. For example, Zhao and Yuan [23] have proposed to discover topical objects in videos by considering the correlation of visual items via cohesive sub-graph mining. It has demonstrated effectiveness in finding one topical object, but it can only find multiple video objects one by one.

Russell *et al*. have proposed to discover objects from image collections by employing Latent Dirichlet Allocation (LDA) model [14] [2]. It can discover multiple objects simultaneously while each object is one topic discovered by LDA model in a top-down manner. However, the computational cost will be too high if LDA model is directly leveraged to discover video object, as one second video contains dozens of frames. One possible mitigation is to discover the video object from selected key frames only. As a consequence, dense motion information can no longer be exploited, and any model needs to address the problem with learning of limited number of training examples to avoid overfitting.

To effectively address the issue of limited training samples, we propose a new topic model which explicitly incorporates a word co-occurrence prior using a Gauss-Markov network over the topic-word distribution in LDA. We call our model as LDA with Word Co-occurrence prior (LDA-WCP). This data-driven word co-occurrence prior can effectively regularize the topic model to learn effectively from limited samples.

In our model, a video sequence is characterized by a number of key frames and each frame is composed of a

collection of local visual features. Each key frame is segmented at multiple resolutions [14]. After clustering the features into visual words, we obtain the bag-of-words representation for each segment. The parameter of the word co-occurrence prior is obtained by analyzing the spatial-temporal word co-occurrence information. After that, the topical objects are discovered by the proposed LDA-WCP model.

By combining data-driven co-occurrence prior from bottom-up with top-down topic modeling method, the benefits of our method are three-fold. First, by using the multiple segmentation and the bag-of-words representation, our method is able to cope with the variant shapes and appearance of the topical video objects. Second, through the proposed LDA-WCP model, our method can discover multiple topical objects simultaneously. Last but not least, by incorporating the word co-occurrence prior, the proposed LDA-WCP model can successfully discover more instances of the topical video objects. Experimental results on challenging video datasets demonstrated the efficacy of our model.

2. Related Works

Most existing visual object discovery methods fall into one of the two categories: bottom-up based methods that find visual objects from bottom-up, and generative model based methods that find objects through top-down reasoning. One type of bottom-up methods depend on the frequent pattern mining algorithms [6]. These methods first translate each image into a collection of "visual words" and then discover the common object through frequently co-occurring words mining [21] [19] [20] [22]. 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 [21] [22]. Han *et al*. summarize the frequent pattern mining algorithms in [6].

Another type of bottom-up methods discover visual objects by graph or tree matching or mining. Traditional subgraph matching methods characterize an image as a graph or a tree composed of visual features. Then, the visual object is discovered by graph or tree matching [17]. Liu and Yan use sub-graph mining to find the common patterns between two images [10]. Zhao and Yuan characterize all key frames using an affinity graph of all visual features and find the topical object by cohesive sub-graph mining [23]. However, the sub-graph mining algorithms are not naturally designed for multiple object discovery.

To consider the inter-image or intra-image statistics for video object discovery, one possible way is using latent semantic embedding structures to describe the image or image region. Most notable methods include probabilistic LSA (pLSA) model [7] and Latent Dirichlet Allocation (LDA) model [2]. Russell et al. [14] discover visual object categories based on LDA model. They first segment the images multiple times and then use LDA to discover object topics form a pool of segments. Liu and Chen [8] show promising video object discovery results by combining pLSA with Probabilistic Data Association (PDA) filter based motion model.

Variants of LDA models have been applied in many different applications. Among them, there are works related to exploring the order or the spatial correlation of words in each document. Gruber et al. propose to model the topics of words in the document as a Markov chain [5]. Wang and Grimson propose a Spatial Latent Dirichlet Allocation (SLDA) model which encodes the spatial structure among visual words [18]. Cao and Li propose a spatially coherent latent topic model [4] for recognizing and segmenting object and scene classes. Philbin *et al*. propose a Geometric Latent Dirichlet Allocation (gLDA) model for unsupervised particular object discovery in unordered image collections [13]. It is an extension of LDA, with the affine homography geometric relation built into the generative process. Some other works explore different priors over the topic proportion such as using logistic normal prior [3] or Dirichlet tree prior [1] to develop correlated topic models.

Another theme of research, such as [9], engages human in the loop for video object discovery. Their model takes weakly supervised information from the user, which is suitable for targeted object discovery that is tailored to users' interests.

3. LDA with Word Co-occurrence Prior

To discover topical objects from videos, visual features are extracted from key frames and clustered into visual words first. Then each video frame is segmented at different resolutions to obtain the bag-of-words representation for each segment. After that we obtain the word co-occurrence prior by analyzing the spatial-temporal word co-occurrence information. Finally, video objects are discovered by the proposed LDA-WCP model. This section describes details about LDA-WCP model while the word-occurrence prior is introduced in Sec. 4.

3.1. Preliminaries

Our method first extracts a set of local visual features from key frames, e.g., SIFT feature [11]. Each visual feature in key frame I_l is described as a feature vector $\phi_l(\mathbf{u}) = [\mathbf{u}, \mathbf{h}]$, where vector $\mathbf{u}$ is its spatial location in the frame, and high-dimensional vector $\mathbf{h}$ encodes the visual appearance of this feature. Then, a key frame I_l is represented by a set of visual features $I_l = \{\phi_l(\mathbf{u}_1), ..., \phi_l(\mathbf{u}_p)\}$. Clustering algorithms, such as k-means, group the features in all F frames $\{I_l\}_{l=1}^{F}$ according to the similarity between their appearance vectors, yielding V visual words $\mathit{\Pi} = \{w^1, w^2, ..., w^V\}$.

To consider the spatial information of visual objects, each key-frame is segmented multiple times using normal-

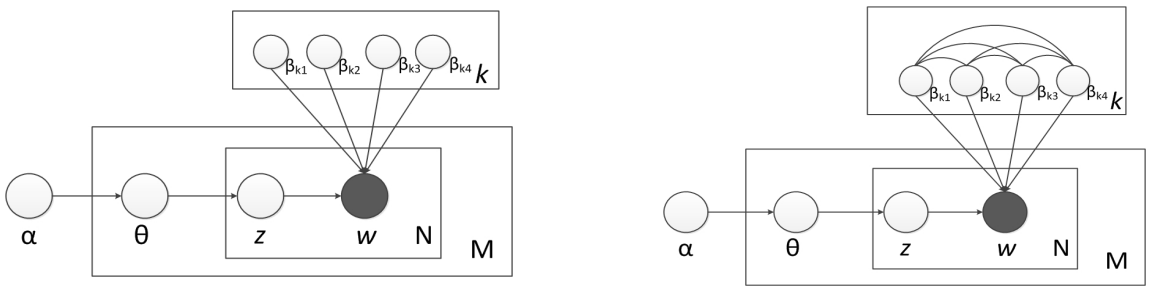


Figure 1. Graphical model representation for (a) original LDA, and (b) the proposed LDA-WCP. Here we set the number of words to be four for illustration convenience.

ized cut [16] to generate segments at different resolution levels.. Then each segment is represented by its corresponding visual words and denoted by $\mathbf{w} = \{w_n | n = 1, ..., N\}$, which is considered as one document. All segments of one video are collected as a corpus denoted by $D = \{\mathbf{w}_m | m = 1, ..., M\}$. In the following, we also use d to represent one specific document.

3.2. Original LDA

Before describing the proposed model, we first briefly introduce original LDA model [2]. LDA shown in Figure 1(a) assumes that in the corpus, each document d arises from a mixture distribution over latent topics [2]. Each word w_{dn} is associated with a latent topic z_{dn} according to the document specific topic proportion vector θ_d, whose prior is Dirichlet with parameter α. The word w_{dn} is sampled from the topic word distribution parameterized by a $K \times V$ matrix β, where each row $\beta_i, 1 \leq i \leq K$, satisfies the constraint $\sum_{j=1}^{V} \beta_{ij} = 1$. Here K and V denote the number of topics and the vocabulary size, respectively.

The generative process for original LDA is as follows:

1. For each document d, $\theta_d \sim \text{Dirichlet}(\alpha)$;
2. For each of the N_d word in document d:
   Choose a topic $z_{dn} \sim \text{Multinomial}(\theta_d)$;
   Choose a word $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$.

For each document d, the joint distribution of a topic mixture θ_d, a set of N_d topics $\mathbf{z}$, and a set of N_d words $\mathbf{w}$ is given by:

$$p(\theta_d, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta_d | \alpha) \prod_{n=1}^{N_d} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta),$$

where $p(z_{dn}|\theta_d)$ is simply θ_{di} for unique i such that $z_{dn}^i = 1$. Integrating over θ_d and summing over $\mathbf{z}$, the marginal distribution of document d, is obtained as:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta_d | \alpha) (\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta)) d\theta_d.$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus [2]:

$$p(D | \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d | \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d.$$

Figure 2. Illustration of visual features for several frames of one video. The blue cross + shows visual words which have strong co-occurrence with each other, while the green star ⋆ shows words which have weak co-occurrence with each other.

3.3. LDA-WCP Model

LDA model is computationally efficient which can also capture the local words co-occurrences via the document-word information. However, the assumption that the words in the document are independent may be an oversimplification. Take the video data as an example, a topical object may contain unique patterns composed of multiple co-occurrence features. Besides, the video objects may be small and hidden in the cluttered background, these co-occurrence features can provide highly discriminative information to differentiate the topical object from the background clutter.

The above consideration motivates us to propose LDA-WCP model, which impose *a priori* constraints on different visual words to encourage co-occurrence visual words in the same topic, as shown in Figure 1(b). This is technically achieved by placing a Markovian smoothness prior $p(\beta)$ over the topic-word distributions β, which encourages two words to be categorized into the same topic if there is strong co-occurrence between them. In video object discovery, the visual words belonging to the same object co-occur frequently in the video, as shown in Figure 2. With the help of prior $p(\beta)$, these words are more likely to be clustered to the same topic. Therefore, more instances of this object will be categorized to the same topic even when some instances contain the noisy visual words from other objects or the background.

A typical example of prior $p(\beta)$ is the Gauss-Markov random field prior [12], expressed by:

$$p(\beta) \propto \prod_{i=1}^{K} \sigma_i^{-V} \exp\left[-\frac{1}{2} \frac{\sum_{j=1}^{V} E(\beta_{ij})}{\sigma_i^2} \right], \qquad (1)$$

$$E(\beta_{ij}) = \sum_{m \in \Pi^j} \epsilon_m (\beta_{ij} - \beta_{im})^2, \quad (2)$$

where Π^j represents the words which have co-occurrence with word w^j and ϵ_m is the co-occurrence weight between word w^m and word w^j. $E(\beta_{ij})$ is the co-occurrence evidence for word j within topic i. The parameter σ_i captures the global word co-occurrence smoothness of topic i and enforces different degrees of smoothness in each topic in order to better adapt the model to the data. The larger the parameter σ_i is, the stronger word co-occurrence is incorporated in topic i. The estimation of word co-occurrence prior is introduced in Sec. 4. Considering the Gauss-Markov random field prior, the probability of a corpus becomes:

$$p(D|\alpha, \beta, \Pi) = p(\beta) p(D|\alpha, \beta). \quad (3)$$

In this way, the prior term incorporates the interaction of different co-occurrence words and forces them to co-occur in the same topic.

The proposed LDA-WCP model can be solved using a variational expectation-maximization (EM) algorithm similar to the one for LDA. The E-step approximates the posterior distribution $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ while the M-step estimates the parameters by maximizing the lower bound of the log likelihood. The following two subsections describe the E-step and M-step, respectively.

3.4. Variational Inference for LDA-WCP

The inference problem for LDA-WCP is to compute the posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$, which is intractable due to the coupling between θ and β, as shown in Figure 1. The basic idea of variational inference is to use a tractable distribution q to approximate the true posterior distribution p, by minimizing the Kullback-Leibler divergence between the two distributions. Here we approximate the posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ by $q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \sum_{n=1}^{N} q(z_n|\phi_n)$, where the Dirichlet parameter γ and the multinomial parameters $\phi_1, ..., \phi_N$ are free variational parameters. Since the Gauss-Markov random field prior does not couple with other variables, we directly use $p(\beta)$. After this approximation, the lower bound of log likelihood of the corpus (Eq.3) is obtained:

$$L(\gamma, \phi; \alpha, \beta) \leq \log p(D|\alpha, \beta). \quad (4)$$

The specific formulation of $L(\gamma, \phi; \alpha, \beta)$ can be found in [2]. The values of varitional parameters ϕ and γ can be obtained by maximizing this lower bound with respect to ϕ and γ:

$$(\gamma^*, \phi^*) = \arg\max_{\gamma, \phi} L(\gamma, \phi; \alpha, \beta). \quad (5)$$

This maximization can be achieved via an iterative fixed-point method. For learning with LDA-WCP over multiple documents, the variational updates of ϕ and γ are iterated until the convergence for each document. This section is presented to make the description complete. Further details can be found in the appendix of [2].

3.5. Parameter Estimation for LDA-WCP

The influence of word co-occurrence prior is adjusted through the strength parameter σ when estimating values of β. Considering the lower bound of log likelihood with respect to β:

$$L_{|\beta|} = L'_{|\beta|} + \sum_{i=1}^{K} \sum_{j=1}^{V} \left(-\log(\sigma_i^V) - \frac{1}{2} \frac{E(\beta_{ij})}{\sigma_i^2} \right), \quad (6)$$

where $L'_{|\beta|}$ is the lower bound of log likelihood without the Gauss-Markov random field prior:

$$L'_{|\beta|} = \Phi(\beta) + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \beta_{ij} - 1 \right), \quad (7)$$

where $\Phi(\beta) = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{dni} w_{dn}^j \log \beta_{ij}$; ϕ_{dni} is the topic i proportion for item n in document d; w_{dn}^j indicates the occurrence of word w^j of item n in document d; and λ_i is the Lagrange multipliers for constraint $\sum_{j=1}^{V} \beta_{ij} = 1$.

The word co-occurrence prior is included in the objective function of Eq.6 and it is more challenging to solve this problem. So we first obtain the solution of β_{ij} by solving $L'_{|\beta|}$. Take the derivative $L'_{|\beta|}$ with respect to β_{ij}, set it to zero, and find:

$$\beta'_{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j. \quad (8)$$

Then, the solution for parameters σ_i^2 are obtained by setting $\partial L_{|\beta|} / \partial \sigma_i^2 = 0$:

$$\sigma_i^2 = \frac{1}{V} \sum_{j=1}^{V} E(\acute{\beta}_{ij}), \quad (9)$$

where $E(\acute{\beta}_{ij}) = \sum_{m \in \Pi^j} \epsilon_m (\beta'_{ij} - \beta'_{im})^2$. After that, we add the Gauss-Markov smooth information back by solving the following problem:

$$L_{|\beta|} = \Phi(\beta) + \sum_{i=1}^{k} \sum_{j=1}^{V} \left(-\log(\sigma_i^V) - \frac{1}{2} \frac{E(\hat{\beta}_{ij})}{\sigma_i^2} \right), \quad (10)$$

where $E(\hat{\beta}_{ij}) = \sum_{m \in \Pi^j} \epsilon_m (\beta_{ij} - \beta'_{im})^2$ and β'_{im} is obtained by Eq.8. To simplify the formulation, we will consider the constraint $\sum_{j=1}^{V} \beta_{ij} = 1$ later. Let $\psi_w^{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j$ and by changing the order of summation, we obtain:

$$L_{|\beta|} = \sum_{i=1}^{K} \sum_{j=1}^{V} \left(\psi_w^{ij} \log \beta_{ij} - \log \sigma_i^V - \frac{1}{2} \frac{E(\hat{\beta}_{ij})}{\sigma_i^2} \right). \quad (11)$$

To compute parameter β_{ij}, we have to maximize $L_{|\beta|}$ with respect to β_{ij}, that is, to compute its partial derivative and set it to zero. Considering a neighborhood Π^j and setting $\partial L_{|\beta|}/\partial\beta_{ij} = 0$ gives a second degree polynomial equation with respect to β_{ij}:

$$\psi_w^{ij}\frac{1}{\beta_{ij}} - \frac{\sum_{m\in\Pi^j}\epsilon_m\beta_{ij} - \sum_{m\in\Pi^j}\epsilon_m\beta'_{im}}{\sigma_i^2} = 0. \quad (12)$$

Multiply by both sides with β_{ij} and σ_i^2, we obtain the following second degree polynomial equation:

$$\left(\sum_{m\in\Pi^j}\epsilon_m\right)\beta_{ij}^2 - \left(\sum_{m\in\Pi^j}\epsilon_m\beta'_{im}\right)\beta_{ij} - \psi_w^{ij}\sigma_i^2 = 0. \quad (13)$$

This equation has two solutions for β_{ij}. It is easy to check that there is only one non-negative solution for the β_{ij} and we select it as the final solution. As β'_{im} is initialized by solving Eq.8 without using the Gauss-Markov prior, we apply a fixed point iteration to estimate β_{ij}. We can see that parameter σ controls the weight of smooth.

After obtaining the solutions for all β_{ij}, β is normalized such that $\sum_{j=1}^{V}\beta_{ij} = 1$. The estimation for parameter α is the same as the basic LDA model by maximizing the lower bound with respect to α, *i.e.*, $\alpha^* = \arg\max_\alpha L(\gamma, \phi; \alpha, \beta)$. The overall algorithm is summarized in algorithm 1:

Algorithm 1 The EM algorithm for LDA-WCP model

input : The corpus D and word co-occurrence prior $p(\beta)$.
output : The topic document matrix γ and the topic word matrix β.

repeat
  /* E-step: variational inference */
  for *d = 1 to D* **do**
    $(\gamma^*, \phi^*) = \arg\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta)$
  end
  /* M-step: parameter estimation */
  estimate β' using Eq.8
  estimate topic smoothness parameter σ using Eq.9
  update β with word co-occurrence prior by solving Eq.13
  normalize β to satisfy the constraint $\sum_{j=1}^{V}\beta_{ij} = 1$
  $\alpha^* = \arg\max_\alpha L(\gamma, \phi; \alpha, \beta)$
until *convergence* ;

4. The Word Co-occurrence Prior for Videos

For video corpus, we can obtain the word co-occurrence prior by considering the spatial-temporal co-occurrence of words. In a typical video, a number of visual words may have strong co-occurrence while others have weak co-occurrence. To estimate the word co-occurrence prior, we find the k nearest neighbors for each visual word w^j in each video frame I_l first. The neighbors in frame I_l are selected according to their spatial distances with word w^j and denote the selected nearest neighbor set in frame I_l as Π_l^j. Second, we obtain the global neighbor set Π^j for each visual word w^j by assembling the nearest neighbors of word w^j in all frames:

$$\Pi^j = \{\Pi_1^j, \cdots, \Pi_l^j, \cdots, \Pi_F^j\}. \quad (14)$$

Then we count the number of occurrence of each visual word w^min the neighbor set Π^j:

$$\mathcal{N}(w^m) = |\{w^m : w^m \in \Pi^j\}|. \quad (15)$$

After that, we select the top k visual words according to the numbers of their occurrences in the neighbor set Π^j and denote the selected visual word set as $\widehat{\Pi^j}$. The co-occurrence weight ϵ_m between the selected neighbor visual word w^m and w^j is calculated as $\epsilon_m = \frac{\mathcal{N}(w^m)}{|\widehat{\Pi^j}|}$. The Gauss-Markov random field prior $p(\beta)$ is build using the neighbor set of each word as Eq. 1.

In this section, we show how to obtain the word co-occurrence prior by simply checking their co-occurrence frequency in the whole video. It is important to note that the word co-occurrence prior can also be obtained by considering the frequent pattern mining algorithms [6], or by employing the human knowledge.

5. Evaluation

To evaluate our approach, we test it on challenging videos for topical object discovery. In addition, we compare the proposed approach with the state-of-the-art methods [14] [23].

5.1. Video Datasets

In the first experiment, we discover video objects from fourteen video sequences downloaded from YouTube.com. We test our method on the video sequences one by one, and try to find one topical object from each video. Most of the videos have the well-defined topical objects, e.g., the product logo. In the second experiment, we test our method on another ten video sequences which have well defined multiple video objects. It is possible that one video frame contains multiple objects and some video frames contain only one topical object.

5.2. Experimental Setting

To obtain the segment representation for videos, we first sample key-frames from each video at two frames per second. SIFT features are extracted from each key-frame. For each sequence, the local features are quantized into $V = 1000$ visual words by the k-means clustering. The number of visual words is selected experimentally. The top 10% frequent visual words that occur in almost all key

Figure 3. Sample results of single object discovery. Each row shows the discovery result of a single video. The segment with normal color contains the discovered topical object, while the segments highlighted by the green color correspond to the background region. The red bounding boxes indicate the ground truth position of the topical objects and the frames without bounding boxes contain non instances of topical objects.

frames are discarded in the experiments. Then each key-frame is segmented at multiple levels using normalized cut [16]. In our implementation, each key-frame is segmented into 3, 5, 7, 9, 11 and 13 segments, respectively. We perform normalized cut in both original key-frames as well as the down-sampled key-frames of half size of the original key-frames. After the segmentation, each segment is described by the bag-of-words representation. To employ LDA-WCP model, the word co-occurrence prior is estimated by using the top 10 neighbors for each visual word as shown in Sec. 4.

To quantify the performance of the proposed approach, we manually labeled the ground truth bounding boxes of the instances of topical objects in each video frame. The bounding boxes locate $2D$ sub-images in each key frame. Let DR and GT be the discovered segments and the bounding boxes of ground truth, respectively. The performance is measured by $F\text{-}measure$. To calculate the $F\text{-}measure$ value for one video, the $F\text{-}measure$ value is first calculated for each key frame and then the average value of all key frames is used to evaluate the whole video.

In the following experiments, we set the topic number $K = 8$ for LDA-WCP model. After obtaining a pool of segments from all key frames, object topics are discovered using the proposed LDA-WCP model. Then the most supportive topics are selected. The supportiveness of one topic is measured by using the ground truth. The more instances of the ground truth object in one topic, the higher the supportiveness of this topic is. One topic is selected in the first experiment for the performance evaluation, while two are selected in the second experiment.

5.3. Single Video Object Discovery

Many videos contain a single topical object, *e.g.*, the Starbucks logo in a commercial video of Starbucks coffee. Such a topical object usually appears frequently. Figure 3 shows some sample results of video object discovery by LDA-WCP model. In the video sequences, the topical objects are subject to variations introduced by partial occlusions, scale, viewpoint and lighting condition changes. It is possible that some frames contain multiple instances of video objects and some frames do not contain any video objects. On average, each video have 42 keyframes and the proposed method can correctly discover 16 instances from total 19 instances of topical object. These results show that the proposed approach performs well for discovering single object from video sequences.

Table 1. Numbers of topical frames and instances of two topical objects in each video sequence.

Seq.	1	2	3	4	5	6	7	8	9	10
FNo.	24	33	31	30	28	19	46	49	40	27
INo.	37	39	40	27	32	31	45	48	33	27
CNo.	29	29	28	24	32	27	45	35	30	27

5.4. Multiple Video Objects Discovery

Many videos contain a number of objects which have comparable importance for video understanding. Such objects can be the objects that are frequently highlighted in the video, or the persons that appear commonly, e.g., the leading actor/actress in the same video. In this experiment, multiple objects are discovered for each video.

Figure 4 shows sample results of multiple topical object discovery. For one video, we show two discovered topical objects and each row shows the result of one topical object. It can be seen that the proposed approach can categorize the instances of different topical object to different topics, even when one video frame contains multiple types of topical objects. Table 1 summarizes the information of ten video sequences. For each sequence, the number of key frames(*FNo.*), the ground truth number of topical object instances (*INo.*) and the corrected detected number of topical object instances (*CNo.*) are shown in three rows, respectively. The instance numbers of two discovered topical objects are considered together. Averagely, the proposed method can correctly discover 31 instances from total 36 instances of two topical objects. These results show that the proposed approach performs well for discovering multiple topical objects from videos simultaneously.

5.5. Comparison with Other Approaches

We compare our video object discovery method with two other methods: (1) LDA based approach and (2) sub-graph mining approach. The LDA based approach [14] is the

Figure 4. Sample results of multiple object discovery in one video. The instances of two discovered topical objects are given. (a) and (c) show the object discovery results of the proposed LDA-WCP model while (b) and (d) show the results of LDA. The LDA-WCP model categorized more instances of one topical object to the same topic than the LDA model. Two frames of (b) and (d) do not have the red bounding boxes as LDA clustered two background segments into the discovered topics.

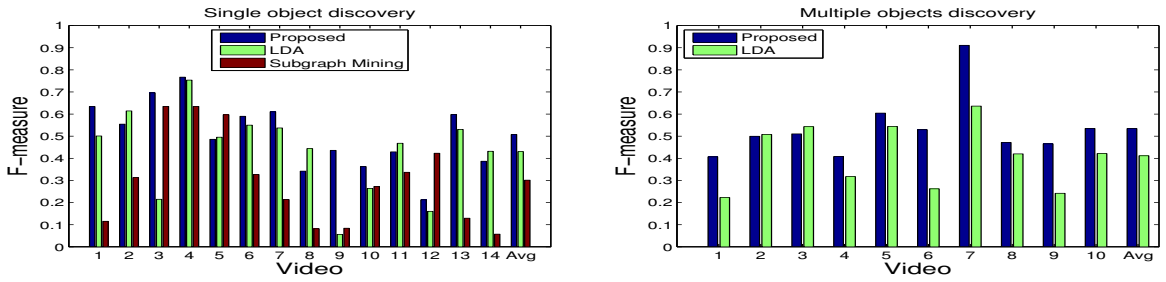
state-of-the-art approach for object categorization and object discovery. To find the video object, each key frame is segmented multiple times with varying number of segments and scales. After obtaining a pool of segments from all key frames, object topics are discovered using LDA following the work in [2]. The visual words and other settings are same as our method for a fair comparison. In the second method, we use the sub-graph mining approach as described in [23]. To find the topical object using sub-graph mining approach, each key frame is segmented multiple times as our method first. Then the affinity graph is built to represent the relationships of all segments. After that, by cohesive sub-graph mining, the instances of topical object are selected from the segments which have strong pair-wise affinity relationships. As this method only obtains the maximum sub-graph each time, we compare it with two other methods for single object discovery only.

As shown in Figure 5(a), our proposed approach outperforms both LDA approach and sub-graph mining approach in terms of the *F-measure* for single topical object discovery, with an average score of 0.51 (Proposed) compared to 0.43 (LDA) and 0.30 (Subgraph Mining), respectively. LDA approach does not consider the co-occurrence prior of visual words and its results only depend on the bag-of-words information. The topics of segments may be affected by the words of the background as the segmentation is not always reliable. On the contrary, the proposed method can achieve a much better result. The same conclusions can be obtained for the experiment of multiple object discovery, as shown in Figure 5(b). It is interesting to note that LDA-WCP performs worse than LDA in some videos. This is because the average F-measure of all discovered object instances is used to evaluate the whole video. As some segments may occupy a small part of the object, more discovered instances of one object may sometimes lead to a lower average F-measure.

We further compare the number of discovered topical object instances by LDA model and the proposed LDA-WCP model. Figure 6(a) shows the discovered instance numbers of single video object and Figure 6(b) shows the discovered instance numbers of multiple objects. It can be seen that LDA-WCP model can categorize more instances of one object to the same topic than LDA model. By incorporating the word co-occurrence prior, LDA-WCP model encourages the words to be categorized to the same topic if there is strong co-occurrence prior between them. This implies that LDA-WCP model makes the learned topics more interpretable by considering both the bag-of-words information and the word co-occurrence prior. These comparisons clearly show the advantages of the proposed video object discovery technique.

6. Conclusion

Video object discovery is a challenging problem due to the possibly large object variations, the complicated dependencies between visual items and the prohibitive computational cost to explore all the candidate set. We first propose a novel Latent Dirichlet Allocation with Word Co-occurrence Prior (LDA-WCP) model, which naturally in-



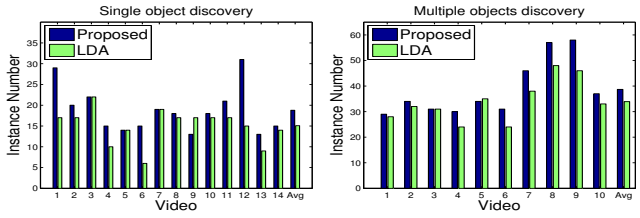
(a) (b)

Figure 5. The performance comparison of three approaches using two video datasets. (a) shows the single object discovery performance of our approach (Proposed), LDA approach (LDA) [14] and sub-graph mining approach (Subgraph Mining) [23]. (b) shows the multiple object discovery performance of two approaches.

tegrates the word co-occurrence prior and the bag-of-words information in a unified way. Then we apply the LDA-WCP model to discover multiple objects from videos simultaneously. Experiments on challenging video datasets show that our method is efficient, robust and accurate.

7. Acknowledgement

This work is supported in part by the Nanyang Assistant Professorship M4080134.040 and GH is partially supported by the start-up funds from Stevens Institute of Technology.



(a) (b)

Figure 6. The number of discovered video object instances by LDA (LDA) and the proposed LDA-WCP (Proposed). (a) shows the number of single video object and (b) shows the number of multiple video objects.

References

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, 2009. 2

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. 1, 2, 3, 4, 7

[3] M. Blei and J. D. Lafferty. A correlated topic model of science. *Ann. Appl. Stat.*, 2007. 2

[4] L. Cao and F.-F. Li. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007. 2

[5] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden topic markov models. *Artificial Intelligence and Statistics (AISTATS)*, 2007. 2

[6] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. In *Data Mining and Knowledge Discovery*, 2007. 1, 2, 5

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999. 2

[8] D. Liu and T. Chen. A topic-motion model for unsupervised video object discovery. In *CVPR*, 2007. 2

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

[10] H. Liu and S. Yan. Common visual pattern discovery via spatially coherent correspondences. *CVPR*, 2010. 2

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[12] C. Nikou, N. P. Galatsanos, and A. Likas. A class-adaptive spatially variant mixture model for image segmentation. *TIP*, 16(4):1121–1130, 2007. 3

[13] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *BMVC*, 2008. 2

[14] 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 *CVPR*, 2006. 1, 2, 5, 6, 8

[15] M.-K. Shan and L.-Y. Wei. Algorithms for discovery of spatial co-orientation patterns from images. *Expert Syst. Appl.*, 37:5795–5802, August 2010. 1

[16] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22:888–905, 2000. 3, 6

[17] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *TPAMI*, 2008. 1, 2

[18] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007. 2

[19] Y. Xie and P. S. Yu. Max-clique: A top-down graph-based approach to frequent pattern mining. In *ICDM*, 2010. 1, 2

[20] J. Yuan and Y. Wu. Spatial random partition for common visual pattern discovery. In *ICCV*, 2007. 2

[21] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *KDD*, 2007. 1, 2

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

[23] G. Zhao and J. Yuan. Discovering thematic patterns in videos via cohesive sub-graph mining. In *ICDM*, 2011. 1, 2, 5, 7, 8