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

Gangqiang Zhao, Junsong Yuan, *Senior Member, IEEE,* and Gang Hua, *Senior Member, IEEE,* Jiong Yang

***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.**

***Index Terms*—LDA, word co-occurrence prior, video object discovery, Gaussian Markov, Top-down, Bottom-up**

I. 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 [1]. Although significant progress has been made along this path [2], it is still a challenge to automatically discover topical objects in videos 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, and 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 [2] [3] [4] [5] [6]. Notwithstanding their demonstrated successes, these methods are limited in different ways. For example, Zhao and Yuan [6] have proposed to discover topical objects in videos by considering the correlation of visual items via cohesive sub-graph mining. It has been shown to be effective 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 the Latent Dirichlet Allocation (LDA) [5] [7]. It can discover multiple objects simultaneously while each object is one topic discovered by the LDA model in a top-down manner. However, the computational cost will be too high if the 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. In addition, the LDA model requires to segment each image to acquire the word-document representation. As a perfect image segmentation is not always achievable, the topical objects may be hidden in the segments of cluttered background. Therefore, the instances of topical objects may not be discovered if considering only the word-document information.

To effectively address the issue of limited training samples and the imperfect segment based representation, 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 from limited samples and imperfect segmentation, as illustrated in Figure 1.

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 [5]. 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-

This work is supported in part by Nanyang Assistant Professorship SUG M4080134 and Singapore Ministry of Education Tier-1 Grant M4011272 to Dr. J. Yuan

G. Zhao is with Morpx Inc., Hangzhou, China, 310051. e-mail: gangqiangzhao@gmail.com. The major part of this research was carried out at the Nanyang Technological University, Singapore

J. Yuan and J. Yang are with the School of EEE, Nanyang Technological University, Singapore, 639798. e-mail: jsyuan@ntu.edu.sg, yang0374@e.ntu.edu.sg

G. Hua is with the visual computing group of Microsoft Research Asia, Bejing, China, 100080. e-mail:ganghua@gmail.com

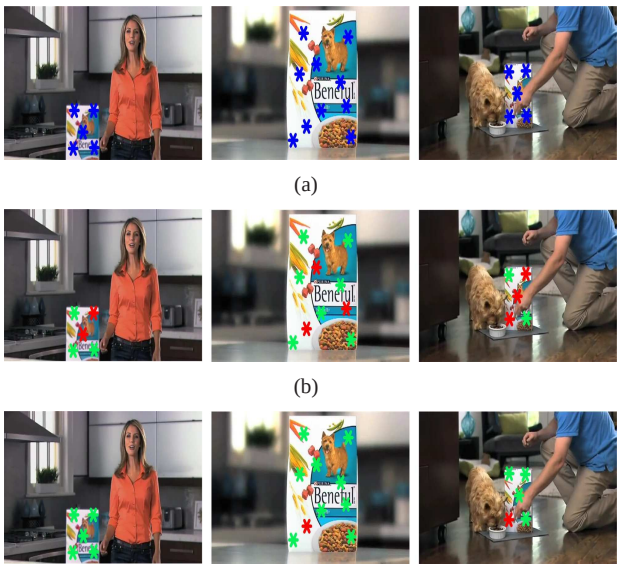


Fig. 1. Illustration of the importance of visual word co-occurrence for video object discovery. (a) shows three keyframes of one video. Several visual words of the topic video object are shown in blue color. (b) shows the topics of each visual word estimated by LDA model. The green color represents the topic of video object while the red color represents other topics. With the help of word co-occurrence prior, the proposed LDA-WCP model can adjust the topics of visual words, as shown in (c). Therefore, more instances of one topical video object will be categorized to the same topic.

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 simultaneously discover multiple topical objects. 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 the proposed unsupervised topical video object discovery method.

A preliminary version of this paper was described in [8]. The current version described here differs from the former in several ways, including: the description of one new word co-occurrence prior estimation algorithm; the introduction of the temporal documents co-occurrence prior of adjacent frames; comprehensive evaluation of the method with more datasets; further analysis and discussion of the whole approach; as well as the introduction of more related works about visual object discovery. While the preliminary version in [8] focuses on the incorporating of word co-occurrence, the current version will provide more details on the word and document co-occurrence estimation techniques, too.

This paper is organized as follows. In Section 2, we briefly survey the visual object discovery. After giving the overview for LDA model, we describe the proposed LDA with word co-occurrence in Section 3. The temporal co-occurrence of documents is presented in Section 4. Thorough experiments are conducted in Section 5 for evaluation. We conclude our paper in Section 6.

II. Related Works

Over the past decade, visual object discovery has received increasing attention in the computer vision community. Most existing visual object discovery methods fall into one of two categories: bottom-up methods and top-down methods [9]. The bottom-up object discovery methods start with basic visual units (e.g., features and their nearest neighbors [10]) and then merge these basic units until the visual objects are found. In contrast, the top-down method start with the modeling of visual objects (e.g., topic-word distribution [5] or attributed relational graphs [11]) and then infer the pattern discovery result for the given data. In this section, we first give an overview on the literature of visual object discovery and then briefly introduce the variants of topic model techniques.

A. Bottom-up methods

For bottom-up methods, it is essential to estimate the repetitiveness of visual features. Different bottom-up methods use different ways to address this issue.

The first type of bottom-up methods discovers the visual object by directly matching the local features. Heath *et al.* extract the visual objects by merging the corresponding matched features between different images [12]. Chum and Matas rely on the min-Hash algorithm for fast detection of pairs of images with spatial overlap [13]. Yuan *et al.* propose to detect visual object by speeding up the local feature matching through LSH-Hash [14] [15]. Cho *et al.* [16] propose a multi-layer match-growing method for visual object discovery. The geometric relations between object instances are estimated through local feature matching and represented by the object correspondence networks.

The second type of bottom-up methods depends on the frequent pattern mining algorithms. These methods first translate each image into a collection of visual words and then discover the visual object through frequently co-occurring words mining [10] [2]. To represent each image using the transaction data, Sivic and Zisserman build spatial configurations of individual visual words on their k-nearest neighbors in the image space [10]. Yuan *et al.* consider the spatial k-nearest neighbors of each local features as a transaction record [2]. Wang *et al.* [17] [18] integrate the transaction building and visual object discovery in a uniform solution. However, it is difficult to select the size k of the nearest neighbors as there is no *a priori* knowledge about the visual object scale.

Another type of bottom-up methods formulates the visual object discovery as a sub-graph mining problem. The graph can be used to model the affinity relationship of visual features in the same image. With this kind of graph, Gao *et al.* propose a frequent sub-graph pattern mining algorithm

to discover high-order geometric patterns which occur in a single image [19]. Chu and Tsai also employ the sub-graph mining algorithm to discover visual object [20]. By building this kind of graph, Liu and Yan discover the visual objects by the maximum sub-graph mining algorithm [21]. The similar affinity graph is used in [22]. To handle multiple images simultaneously, Zhao and Yuan propose another kind of graph which describes the affinity relationship of all visual features of one video and they find the video object by a cohesive sub-graph mining algorithm [6]. Recently, J. Liu and Y. Liu [23] propose a joint assignment algorithm for visual object mining which considers visual words and instances of visual object simultaneously in the process of sub-graph matching.

In general, the bottom-up methods discover the visual objects by linking the basic visual units together. However, the bottom-up methods do not provide a model of the visual object and it is also very difficult to incorporate the prior knowledge of visual objects into them.

B. Top-down methods

For top-down methods, it is essential to model the visual objects. According to the modeling approaches, we categorize the top-down methods to three types.

The first type of top-down methods employs the topic model for visual object discovery. The topic model, such as Latent Dirichlet Allocation (LDA) [7] and probabilistic Latent Semantic Analysis (pLSA) [24], discovers semantic topics from a corpus of documents. The bag of words representation is used to model the documents. Sivic *et al.* introduce the topic model to discover the objects in images [25]. Following this idea, Russell *et al.* [5] discover the visual object categories based on the LDA and pLSA models. They first segment the images multiple times and then discover object topics from a pool of segments. To discover the hierarchical structure of visual objects, Sivic *et al.* investigate the hierarchical Latent Dirichlet Allocation (hLDA) model [26]. Liu and Chen show promising video object discovery results by combining pLSA with Probabilistic Data Association (PDA) filter based motion model [27].

The second type of top-down methods uses sub-space projection for visual object discovery. Inspired by the success of topic model based visual object discovery, Tang and Lewis [28] propose to use non-negative matrix factorization (NMF) to approximate the semantic structure of visual objects. The results of NMF are comparable with that of LDA on the same dataset. Sun and Hamme [29] further integrate spectral cluster and NMF for visual object discovery.

The third type of top-down methods explicitly use a graph or tree to model the spatial structure of visual objects. Hong and Huang [11] model the visual object as a mixture of attributed relational graphs whose nodes represent the basic primitives. They also propose an expectation-maximization (EM) algorithm to learn the parameters of visual object model. Tan and Ngo also represent each image as one attributed relational graph of image segments while Earth Movers Distance (EMD) is used to estimate the segment similarity [30]. Todorovic and Ahuja model the spatial layout of primitive regions in a tree structure to learn object category [4].

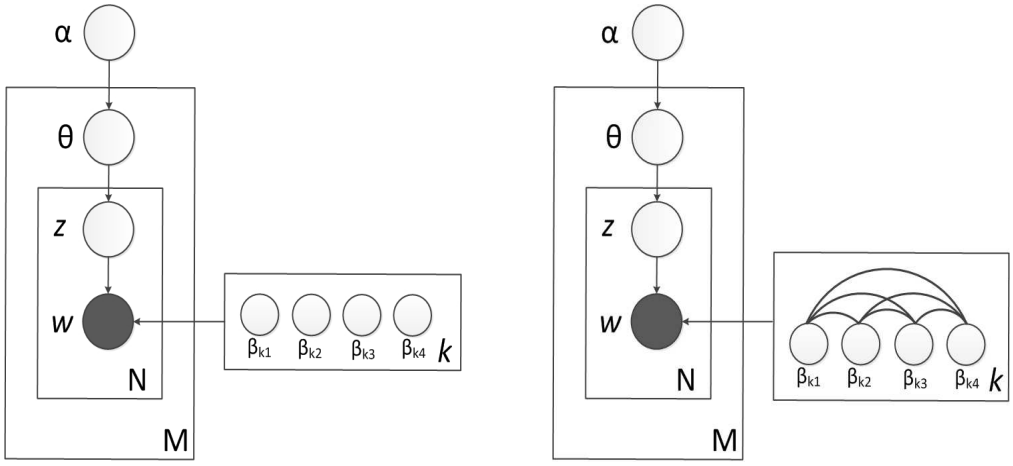
In general, the top-down methods can capture the overview of the visual objects. However, for most top-down methods, there are only approximated solutions for the inference and parameter estimation. This may affect the performance of top-down methods.

C. Variants of Topic Models

As our work most closely builds upon the topic model based object discovery method [5], we further give a brief overview on the variants of topic models. 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 [31]. Wang and Grimson propose a Spatial Latent Dirichlet Allocation (SLDA) model which encodes the spatial structure among visual words [32]. The word-document assignment is not fixed *a priori* but depends on the generative procedure which assigns visual words into the same documents if they are close in space. Philbin *et al.* propose a Geometric Latent Dirichlet Allocation (gLDA) model for discovering a particular object in unordered image collections [33]. It is an extension of LDA, with the affine homography geometric relation built into the generative process. Cao and Li propose a spatially coherent latent topic model for recognizing and segmenting object and scene classes [34]. Andreetto *et al.* propose a new Affinity-Based Latent Dirichlet Allocation (A-LDA) which considers the affinities between pixels to improve the segmentation performance [35].

The temporal properties of documents are also helpful for several applications. Levent *et al.* employs the temporal ordering of the documents to discover the topic propagation between different time segments [36]. Wang *et al.* use the temporal ordering to capture the causal relationship of social media event. Different with these applications, the temporal ordering can not be applied for the topical video object discovery problem. This is because different instances of one same topical object have no casual relationship. Therefore, we only model the temporal co-occurrence between segments of adjacent frames and do not consider the temporal ordering of different segments.

As for Markov Random Field model, zhao *et al.* defines a Markov Random Field over hidden labels of an image to enforce the spatial coherence between topic labels for neighboring regions [37]. Verbeek *et al.* improve the performance of PLSA by introducing an image-specific Markov Random Field to enforce the spatial coherence on the labels of the fine-grained local patches [38]. Wallach *et al.* relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word [39]. Different with these methods, we define the Markov word co-occurrence prior over the whole video for the topical object discovery.



(a) (b)

Fig. 2. Graphical model representation for (a) the original LDA, and (b) the proposed LDA-WCP. The curves between the items of topic-word distribution β in (b) imply the incorporation of word co-occurrence prior. Here we set the number of words to be four for illustration convenience.

III. LDA-WCP MODEL

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 and the word co-occurrence prior estimation.

A. Preliminaries and LDA Model

Our method first extracts a set of local visual features from key frames, e.g., SIFT feature. 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 T frames $\{I_l\}_{l=1}^{T}$ 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 normalized cut 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.

Before describing the proposed model, we first briefly introduce the original LDA model [7][1]. LDA shown in Figure 2(a) assumes that in the corpus, each document d arises from a mixture distribution over latent topics [7]. 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 the original LDA is as follows:

1. For each document d, $\theta_d \sim$ Dirichlet(α);
2. For each of the N_d word in document d:
   Choose a topic $z_{dn} \sim$ Multinomial(θ_d);
   Choose a word $w_{dn} \sim$ 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 an 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 all documents, we obtain the probability of a corpus [7]

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

LDA model is computationally efficient which can also capture the local words co-occurrences via the document-word information. However, the video level feature co-occurrence information is not considered in LDA model. Take the video data as an example, a topical object

[1]Here we consider the un-smoothed version of LDA model for explanation convenience.

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.

B. LDA-WCP Model

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 2(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 a strong co-occurrence between them. In video object discovery, the visual words belonging to the same object co-occur frequently in the video. 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 [40], expressed by

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

$$E(\beta_{ij}) = \sum_{h\in\Xi^j} \mathcal{E}_{jh}(\beta_{ij} - \beta_{ih})^2, \quad (2)$$

where Ξ^j represents the words which have co-occurrence with word w^j and $\mathcal{E}_{jh}$ is the co-occurrence weight between word w^h 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. Considering the Gauss-Markov random field prior, the probability of a corpus becomes

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

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

C. Inference and Learning in LDA-WCP

The Gauss-Markov prior couples the parameter σ and β, which makes direct estimation of them intractable. Therefore we propose a new variational expectation-maximization (EM) algorithm to solve LDA-WCP model. The E-step approximates posterior distribution $p(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$ using variational inference similar to the one for LDA. The M-step estimates the parameters in a closed form by maximizing the lower bound of the log likelihood.

1) Variational Inference: 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 2. 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

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

The values of variational parameters ϕ and γ can be obtained by maximizing this lower bound with respect to ϕ and γ

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

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 supplementary material.

2) Parameter Learning: 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

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

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 $\sum_{j=1}^{V}\beta_{ij} = 1$.

Algorithm 1 The EM algorithm for LDA-WCP model

input : The corpus D and word co-occurrence prior $\mathcal{E}$.
output : The topic document matrix γ and the topic word matrix β.

repeat

/* E-step: variational inference */
for *d = 1 to D* **do**
$(\gamma^*, \phi^*) = \arg\max_{\gamma,\phi} L(\gamma, \phi; \alpha, \beta, \sigma)$
end

/* M-step: parameter learning */
estimate $\beta^{'}$ using Eq.8
estimate topic smoothness parameter σ using Eq.9
update β with word co-occurrence prior by solving Eq.11
normalize β to satisfy the constraint $\sum_{j=1}^{V} \beta_{ij} = 1$
$\alpha^* = \arg\max_{\alpha} L(\gamma, \phi; \alpha, \beta)$

until *convergence* ;

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

$$\beta^{'}_{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w^{j}_{dn}. \tag{8}$$

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

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

where $E(\beta^{'}_{ij}) = \sum_{h \in \Xi^j} \mathcal{E}_{jh} (\beta^{'}_{ij} - \beta^{'}_{ih})^2$. After that, we add the Gauss-Markov smooth information back by solving the following problem

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

where $E(\hat{\beta}_{ij}) = \sum_{h \in \Xi^j} \mathcal{E}_{jh} (\beta_{ij} - \beta^{'}_{ih})^2$ and $\beta^{'}_{ih}$ is obtained by Eq.8. To simplify the formulation, we will consider the constraint $\sum_{j=1}^{V} \beta_{ij} = 1$ later. Let $\psi_w^{ij} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w^{j}_{dn}$ and by changing the order of summation, we obtain $L_{|\beta|} = \sum_{i=1}^{K} \sum_{j=1}^{V} \left(\psi_w^{ij} \log \beta_{ij} - \log \sigma_i^V - \frac{1}{2} \frac{E(\hat{\beta}_{ij})}{\sigma_i^2} \right)$. 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 Ξ^j and setting $\partial L_{|\beta|}/\partial \beta_{ij} = 0$, we obtain a second degree polynomial equation with respect to β_{ij}, i.e. $\psi_w^{ij} \frac{1}{\beta_{ij}} - \frac{\sum_{h \in \Xi^j} \mathcal{E}_{jh} \beta_{ij} - \sum_{h \in \Xi^j} \mathcal{E}_{jh} \beta^{'}_{ih}}{\sigma_i^2} = 0$. Multiply both sides with β_{ij} and σ_i^2, we obtain the following second degree polynomial equation

$$\left(\sum_{h \in \Xi^j} \mathcal{E}_{jh} \right) \beta_{ij}^2 - \left(\sum_{h \in \Xi^j} \mathcal{E}_{jh} \beta^{'}_{ih} \right) \beta_{ij} - \psi_w^{ij} \sigma_i^2 = 0. \tag{11}$$

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 $\beta^{'}_{ih}$ 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. , $\alpha^* = \arg\max_{\alpha} L(\gamma, \phi; \alpha, \beta, \sigma)$. The overall algorithm is summarized in algorithm 1.

As shown in the evaluation section, our model can effectively incorporate the word co-occurrence to the LDA model. A theoretical analysis of the convergence properties of the EM algorithm is an interesting open problem for the statistical and machine learning community. In addition, we does not consider the collapsed Gibbs sampling inference [41] for LDA-WCP model due to its high computational cost and its integration out of variable β, which are used for estimation of smoothness parameter σ.

D. Word Co-occurrence Prior Estimation

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 may have weak co-occurrence. As shown in many vision tasks [23] [6] [21] [42], the frequency of word co-occurrence is an important criterion for estimating the affinity of visual words. However, due to the inherent complexity of a video object, the co-occurrence frequency of a word pair does not always suggest accurate and meaningful affinity relationship. Even if a word pair co-occurs frequently, it is not clear whether such co-occurrence is statistically significant or just by chance. Therefore, inspired by mutual information criterion, we employ the following criterion to estimate the co-occurrence prior $\mathcal{E}_{jh}$ of two words w^j and w^h

$$\mathcal{E}_{jh} = \frac{\mathcal{P}(w^j, w^h)}{\mathcal{S}(w^j) \times \mathcal{S}(w^h)}, \tag{12}$$

where $\mathcal{P}(w^j, w^h)$ represents the effective co-occurrence number of a pair of visual word w^j and w^h, and $\mathcal{S}(w^j)$ is the effective occurrence number of visual word w^j. To capture the most important word co-occurrence, for each visual word w^j, we select its top C co-occur visual words according to the estimated co-occurrence prior values. All the selected co-occur visual words for word w^j is denoted as Ξ^j.

Algorithm 2 Word Co-occurrence Prior Estimation

input : The video with V visual words and T frames.
output : The word co-occurrence prior $\mathcal{E}$.

initialize $\mathcal{S}$ and $\mathcal{P}$ to zero
/* estimate the effective occurrence of words and word pairs */
for j = 1 *to* V **do**
  for l = 1 *to* T **do**
    $\mathcal{S}(w^j) = \mathcal{S}(w^j) + \Omega_s(w_l^j)$
  end
  for h = 1 *to* V **do**
    for l = 1 *to* T **do**
      if j *and* h *co-occurs in frame* l **then**
        $\mathcal{P}(w^j, w^h) = \mathcal{P}(w^j, w^h) + \Omega_p(w_l^j, w_l^h)$
      end
    end
  end
end
/* estimate the word co-occurrence prior */
for j = 1 *to* V **do**
  for h = 1 *to* V **do**
    $\mathcal{E}_{jh} = \mathcal{P}(w^j, w^h)/(\mathcal{S}(w^j) \times \mathcal{S}(w^h))$
  end
end

To estimate the effective co-occurrence frequency of word pairs, we need to decide the co-occurrence of two visual words in each frame is effective or not by checking their repetitiveness in the whole video. Therefore, we find the k nearest neighbors for each visual word instance in each video frame according to the spatial distances, as illustrated in Figure 3. Assume the instance number of visual word w^j in frame I_l is $\mathcal{M}$ and w^{j_m} is the m^{th} instance of visual word w^j in frame I_l. The nearest neighbor set of all instances of word w^j in frame I_l is denoted as $\Pi_l^j = \{\pi_l^{j_1}, \pi_l^{j_2}, \cdots, \pi_l^{j_\mathcal{M}}\}$, where $\pi_l^{j_m}$ is the neighbors of visual word instance w^{j_m}. The repetitiveness of word instance w^{j_m} is obtained by comparing $\pi_l^{j_m}$ with the nearest neighbor set of word w^j in all other frames. Take frame I_t as an example and assume the instance number of word w^j in frame I_t is $\mathcal{M}'$. The repetitiveness of visual word instance w^{j_m} at frame t is estimated by $s_t^{j_m} = \max\{|\pi_l^{j_m} \cap \pi_t^{j_{m'}}|\}_{m'=1}^{\mathcal{M}'}$, which is the maximum number of intersections between $\pi_l^{j_m}$ and the nearest neighbor set Π_t^j of visual word w^j in frame I_t.

With the help of the estimated repetitiveness of each visual word, we can calculate the effective occurrence of each word and each pair of words. Assume the instance number of word w^h in frame I_l is $\mathcal{N}$. For two word instance $w_l^{j_m}$ and $w_l^{h_n}$, the pairwise repetitiveness is estimated as $R(w_l^{j_m}, w_l^{h_n}) = \sum_{t=1}^{T} \min(s_t^{j_m}, s_t^{h_n})$ where T is the total number of frames in one video. The final effective co-occurrence of word pair w^h and w^j in frame I_l is estimated by

$$\Omega_p(w_l^j, w_l^h) = \max\{R(w_l^{j_m}, w_l^{h_n})\}_{m=1}^{\mathcal{M}}{}_{n=1}^{\mathcal{N}}. \quad (13)$$

The final occurrence of word w^j in frame I_l is estimated by selecting the maximum effective co-occurrence of word w^j

$$\Omega_s(w_l^j) = \max\{\Omega_p(w_l^j, w_l^h)\}_{h=1}^{V}. \quad (14)$$

The overall algorithm is summarized in algorithm 2.

Fig. 3. Illustration of nearest neighbors for two visual words. Five nearest neighbors are shown for both word 1 (red circle) and word 2 (green diamond) in two frames.

Although the computational complexity of this algorithm is proportional to $V^2 \times T$, the algorithm is still efficient due to the low frequency of words co-occurring in the same frame.

IV. Incorporate Temporal Document Co-occurrence to LDA-WCP

The proposed LDA-WCP model can effectively incorporate the visual word co-occurrence prior for topical video object discovery. However, the co-occurring phenomenon exists not only among visual words. Because of the temporal dependence of video frames, the co-occurrence of segments of adjacent key frames might also provide beneficial prior information for object discovery. As the video objects may be small and hidden in the cluttered background, this temporal co-occurrence can provide discriminative information to differentiate the topical object from the background clutter.

A. Temporal Document Co-occurrence Modeling

To incorporate the temporal co-occurrence to the LDA-WCP model, we model the temporal co-occurrence between segments of adjacent frames as variables $y_{1:M,1:M}$, where $y_{d,d'}$ represents the temporal co-occurrence prior between document d and document d', which are two segments of adjacent key frames. Inspired by Relational Topic Model (RTM) [43], we assume the distribution probability of temporal co-occurrence $y_{d,d'}$ as

$$p(y_{d,d'}|\eta, \nu) = \exp(\eta^T(\bar{z}_d \circ \bar{z}_{d'}) + \nu), \quad (15)$$

where $\bar{z}_d = \frac{1}{N_d}\sum_n z_{d,n}$ is the average topic assignment of all N_d words in document d; $\circ$ is the element-wise product (Hadamard product); η and ν are two parameters of the probability function. Different with Relational Topic Model (RTM) [43] which uses binary $y_{d,d'}$ to describe the link of network data, we assume $0 \leq y_{d,d'} \leq 1$ in order to handle the temporal document co-occurrence of adjacent key frames.

When considering the temporal co-occurrence and LDA-WCP model simultaneously, the generative process for our model is as follows:

1. For each document d:
   a. $\theta_d \sim$ Dirichlet(α);
   b. For each of the N_d words in document d:
      Choose a topic $z_{dn} \sim$ Multinomial(θ_d);
      Choose a word $w_{dn} \sim$ Multinomial($\beta_{z_{dn}}$).
2. For each pair of documents of adjacent frames:

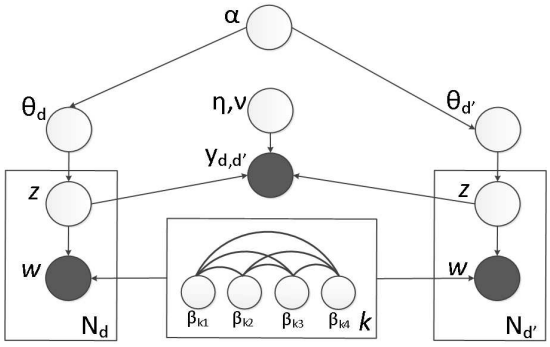


Fig. 4. Graphical model representation for LDA-WCP with temporal document co-occurrence. Here we show a single pair of documents and set the number of words to be four for illustration convenience.

Choose a temporal co-occurrence $y_{d,d'} \sim p(y_{d,d'}|\eta,\nu)$ where $z_d = z_{d,1}, z_{d,2}, \ldots, z_{d,n}$

By comparing with the generative process of the original LDA, it can be seen that the temporal co-occurrence is generated for pairs of documents. Figure 4 illustrates this process for a single pair of documents using the graphical model. In this way, our model incorporates both the word level co-occurrence and the document level co-occurrence. The inference and parameter learning of this new model can still be handled by the proposed EM algorithm, as shown in the supplementary material. By modeling the temporal dependence, the topic of each document depends on the visual appearance of the document as well as that of its adjacent document. Therefore, two documents which have temporal co-occurrence will be more likely to be clustered to the same topic.

B. Temporal Document Co-occurrence Estimation

The temporal co-occurrence prior of two segments d and d' of adjacent frames can be decided using the appearance similarity obtained by matching the instances of visual words between two frames. Denote the number of matched word instances between two segments d and d' as $N_{d,d'}$, the temporal co-occurrence is defined by

$$y(d,d') = \frac{N_{d,d'}}{|\mathbf{w}_d \cup \mathbf{w}_{d'}|}, \tag{16}$$

where $|\mathbf{w}_d \cup \mathbf{w}_{d'}|$ is the number of all words of two segments. If all word instances of segment d and segment d' can match with their corresponding instances, $y(d,d')$ is 1. Otherwise, $y(d,d')$ is less than 1.

To count the number of matched word instances $N_{d,d'}$, we estimate the matching score of one word instance in segment d using its nearest neighbor. Denote $\pi_d^{j_m}$ as the k neighbors for word instance w^{jm} in segment d, and $\Pi_{d'}^{j} = \{\pi_{d'}^{j_1}, \pi_{d'}^{j_2}, \cdots, \pi_{d'}^{j_{\mathcal{M}'}}\}$ as the nearest neighbor set of all $\mathcal{M}'$ instances of visual word w^j in segment d'. The matching score of visual word w^{j_m} is $s^{j_m} = \max\{|\pi_d^{j_m} \cap \pi_{d'}^{j_{m'}}|\}_{m'=1}^{\mathcal{M}'}$, which is the maximum number of intersections between $\pi_d^{j_m}$ and elements of $\Pi_{d'}^{j}$. With the estimated matching score of visual word, the word instance w^{j_m} is considered as one matched word instance if s^{j_m} is larger than half of k.

V. 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 [5][6][8][43].

A. Video Datasets and Experimental Setting

To evaluate the proposed method, two video datasets are collected. Dataset 1 contains twenty-four video sequences downloaded from YouTube.com. Most of the videos in Dataset 1 are the commercial videos and and length of the video in this dataset ranges from 20 seconds to 40 seconds. Dataset 2 contains 101 videos from diverse categories such as news, commercials, documentary, etc. The length of the video in this dataset ranges from 20 seconds to 4 minutes. Several videos are shared by both Dataset 1 and Dataset 2.

In the first experiment, we discover video objects using Dataset 1 and Dataset 2. Most of the videos have the well-defined primary topical objects, e.g., the product logo. We test our method on the video sequences one by one, and try to find one primary topical object from each video. Besides a primary topical object, many videos contain a number of other 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 frequently, e.g., the leading actor/actress in the commercial video. Therefore, in the second experiment, we test our method on Dataset 1 to discover multiple video objects from each video.

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 $V = 400$ visual words by the k-means clustering. The number of visual words is selected experimentally. Then each key-frame is segmented at multiple levels using normalized cut. In our implementation, each key-frame is segmented into 3, 5, 7, 9, 11, 13 and 15 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 $C = 30$ co-occurring words for each word as shown in Sec. III-D. We use ten nearest neighbors for each visual word for both word co-occurrence prior and temporal document co-occurrence estimation.

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. The most supportive topics are selected 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 for single topical

object discovery, while two are selected for multiple topical objects discovery. For the selected topic, all segments of the same key frame are sorted based on the topic assignment values. The segment with the highest rank is selected as the instance of the topical object.

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. One segment is considered as discovered by our method only when the overlap between the discovered segment and the ground truth is larger than 50 pixels. Let DR and GT be the discovered segments and the bounding boxes of ground truth of one frame, respectively. The performance of each object instance is measured by two criteria: $precision$ and $recall$. By combining $precision$ and $recall$, we obtain F-$measure$ as the metric for performance evaluation [4]. To evaluate the performance of one video, the $precision$ and $recall$ is first estimated for each video by averaging the results of all discovered instances. Then we normalize the average $precision$ and $recall$ value by multiplying them with the discovered instance number weight $\frac{N_c}{N_g}$, where N_g is the ground truth instance number of topical objects and N_c is the corrected detected instance number of topical objects. After that, the average F-$measure$ value of one video is estimated using the normalized average $precision$ and $recall$. For each dataset, the performance is measured by the average results obtained after running the LDA-WCP algorithm 5 rounds.

B. Video Object Discovery using LDA-WCP

To demonstrate the advantage of the proposed LDA-WCP model, we evaluate it with the challenging video datasets for topical video object discovery.

Many videos in Dataset 1 contain a primary topical object, e.g. , the Starbucks logo in a commercial video of Starbucks coffee. Such a topical object usually appears frequently. Figure 6 shows some sample results of video object discovery by the 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 has 42 keyframes and the proposed method can correctly discover 19 instances from a total of 23 instances of topical object. We further evaluate the proposed approach using Dataset 2. Figure 7 shows some sample results of video object discovery using Dataset 2 by the LDA-WCP model.

Besides a primary topical object, many videos of Dataset 1 also contain several other objects which are important for video understanding. The proposed approach can categorize the instances of different topical objects to different topics, even when some video frames contain multiple types of topical objects. On average, the proposed method can correctly discover 37 instances from a total of 44 instances of two topical objects.

These results show that the proposed approach performs well for discovering both single and multiple topical objects from videos.

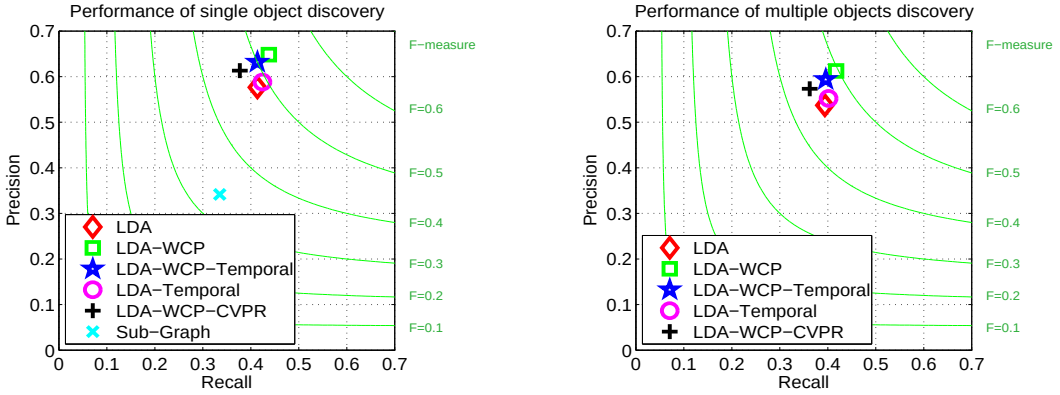
C. Comparison with LDA and sub-graph mining approach

We compare our video object discovery method (LDA-WCP) with two state-of-the-arts methods: LDA based approach (LDA) and sub-graph mining approach (Sub-Graph). The LDA based approach (LDA) is one of the state-of-the-art approaches for object discovery [5]. 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 [7]. 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 (Sub-Graph) as described in [6]. To find the topical object using sub-graph mining approach, each key frame is again first segmented multiple times in the same way as our method. 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 other methods for single object discovery only.

To evaluate the effect of the word co-occurrence prior and temporal document co-occurrence prior, we report the results of three variants of our method: LDA-WCP-Tempral method which incorporates the proposed temporal document co-occurrence prior to LDA-WCP method; LDA-Tempral method which incorporates the proposed temporal document co-occurrence prior to LDA method [5]. LDA-Tempral can be considered as our implementation of Relational Topic Model (RTM) [43] for topical video object discovery; LDA-WCP-CVPR method which uses our previous word co-occurrence prior estimation algorithm as described in [8].

As shown in Figure 5(a) and Figure 9, 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 0.52 (LDA-WCP) compared to 0.44 (LDA) and 0.34 (Sub-Graph), respectively. LDA approach does not consider the co-occurrence prior of visual words and its results only depend on the word occurrence frequency. 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 multiple objects discovery, as shown in Figure 5(b).

By incorporating the word co-occurrence prior, LDA-WCP model encourages the words to be categorized to the same topic if there is a strong co-occurrence prior between them. This implies that LDA-WCP model makes the learned topics more interpretable by considering both the word occurrence frequency and the word co-occurrence



(a) (b)

Fig. 5. Performance comparison of different methods using Dataset 1. (a) and (b) show the precision/recall results of different methods for single and multiple video object discovery, respectively. The precision/recall values are the average results of all videos. The green curves shows the corresponding precision/recall values of the same $F\text{-}measure$ value.

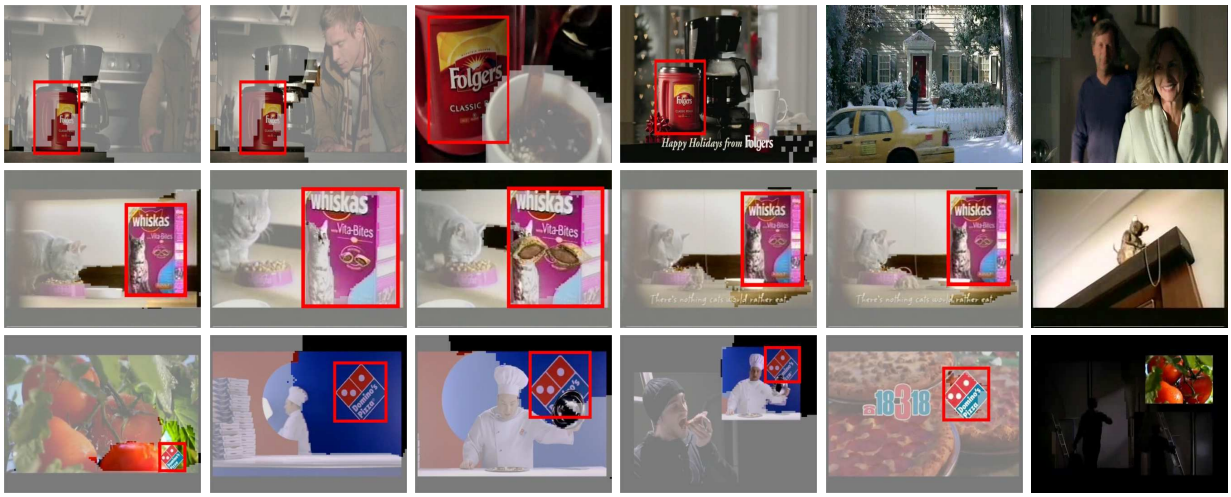


Fig. 6. Sample results of single object discovery using Dataset 1. Each row shows the discovery result of a single video. The segment with normal color contains the discovered topical object, while the segments overlaid by a transparent filter correspond to the background region. The red bounding boxes indicate the ground truth position of the topical objects and the frames without bounding boxes do not contain any instances of topical objects.

prior. These comparisons clearly show the advantages of the proposed video object discovery technique.

The comparison between LDA-WCP and LDA-WCP-CVPR [8] demonstrates the advantages of the proposed word co-occurrence prior estimation algorithm. In LDA-WCP-CVPR, the word co-occurrence prior is estimated by only considering the co-occurrence frequency between two visual words as described in [8]. On the contrary, we first estimate the repetitiveness of each visual word and then use the mutual information criterion to obtain the word co-occurrence prior for LDA-WCP.

We also observe that, by incorporating the temporal document co-occurrence to LDA, LDA-Temporal model outperforms LDA model for both single and multiple objects discovery. Although incorporating the temporal co-occurrence does not improve the overall performance of LDA-WCP for all videos, the temporal co-occurrence boosts the performance of LDA-WCP for about one third videos. By analyzing the video contents, we find that the temporal co-occurrence does not work for other two thirds videos as these videos do not have a strong temporal document co-occurrence due to the information loss in the process of keyframe sampling. We expect that if the key-frame is extracted more densely, then modeling the temporal co-occurrence may show more benefits.

D. Comparison with the CNN (convolutional neural nets) method

Since 2012, CNN (convolutional neural nets [44]) becomes the state-of-the-art methods in problems such as object detection, face recognition. To using the CNN for video object discovery, we employ the pretrained 1000 classes imagenet-caffe-alex model [44].

Specifically, for one image segment s in keyframe I, its circumscribed rectangle region is selected firstly. Then, this rectangle region is used as input of the pretrained imagenet-caffe-alex model [44] and the class $C(s)$ of segment s is decided. After that, one segment s_{max} with the largest classification score is selected and its class $C(s_{max})$ is counted as one instance of class $C(s_{max})$ in the video.

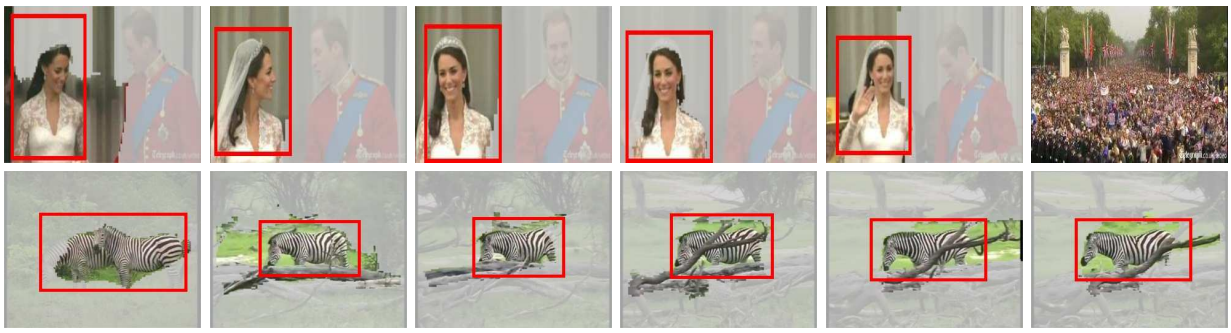


Fig. 7. Sample results of single object discovery using Dataset 2. Each row shows the discovery result of a single video.

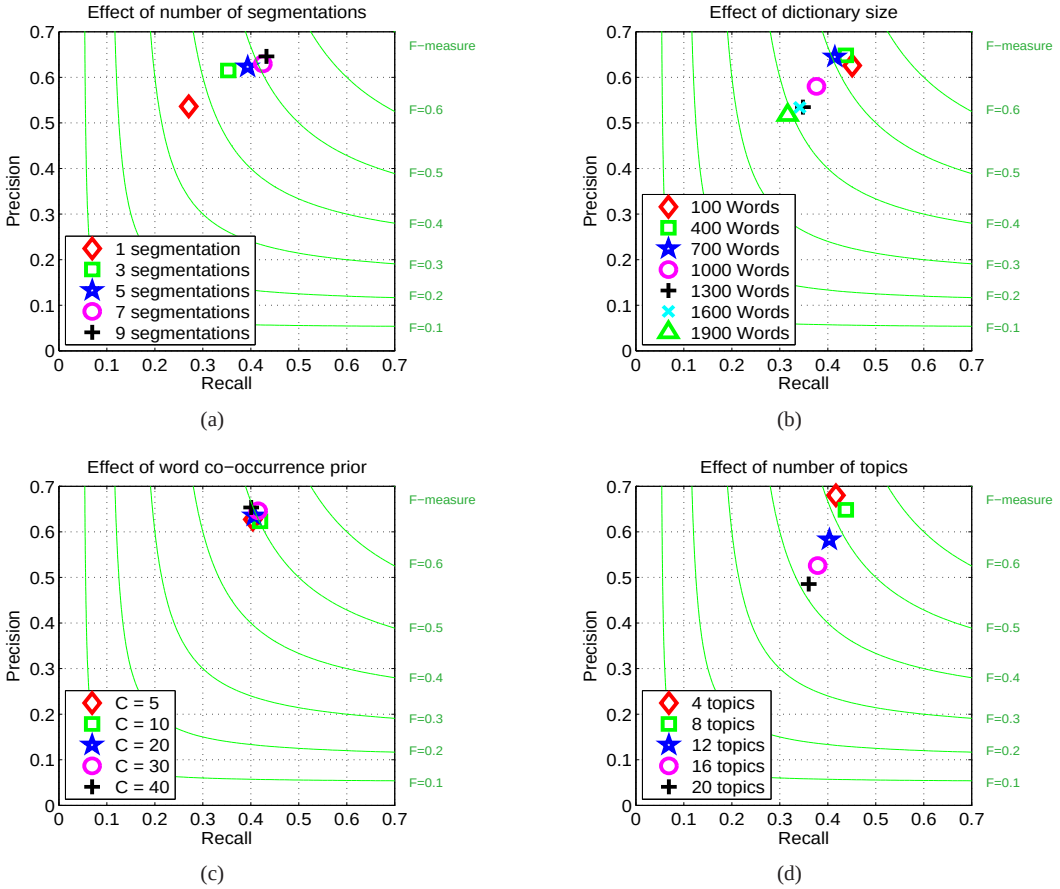


Fig. 8. Precision/recall plots for our approach with different parameters using Dataset 1. (a) shows the performance of LDA-WCP model when using different segmentations of each frame. (b) shows the performance of LDA-WCP model with different dictionary sizes. (c) shows the performance of LDA-WCP model with different word co-occurrence priors. (d) shows the performance of LDA-WCP model with different number of topics.

Finally, we select the top 3 classes by counting their instance numbers in the whole video. Each selected class is considered as one discovered topical video objects. Figure 10 shows the discovered results of CNN method for one video.

From this evaluation, we can draw two conclusions. First, sometimes CNN can assign the same class label to the instances of the same ground truth topical video object, e.g., the class ”plate rack”. However, only a very limited number of topical object instances are assigned the same class label as the topical object instances are subject to variations introduced by partial occlusions, scale, viewpoint and lighting condition changes. Second, the pretrained CNN model describes the category-level information and it classifies the image regions based on the learned category-level features. Therefore, it may classify the instances of the topical objects and the regions of the background to the same topic, e.g., the class ”sweatshirt”. In summary, the CNN method with the pretrained model is not able to improve the topical video object discovery performance. However, the CNN approach is still a promising research direction as it can provide the semantic label for the discovered object.

Fig. 10. Sample results of CNN based topical video object discovery of one video. The discovered top 3 classes are shown. For each discovered class, 4 input image regions are shown. Sometimes, CNN method can assign the same class label to the instances of the same ground truth topical video object, e.g., the class "plate rack". However, it may classify the instances of the topical objects and the regions of the background to the same topic, e.g., the class "sweatshirt".

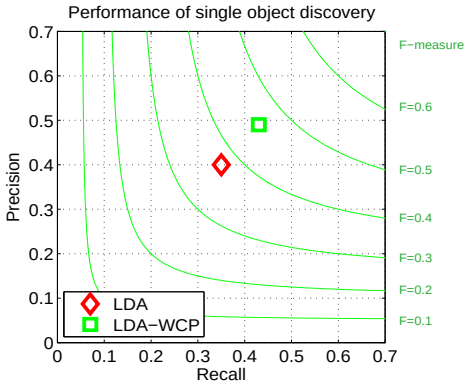


Fig. 9. Performance comparison of different methods using Dataset 2. The precision/recall values are the average results of all videos. The green curves shows the corresponding precision/recall values of the same F-$measure$ value.

E. Evaluation of Parameter Selection

To further evaluate the proposed approach, we discuss the influence of key-frame segmentation, dictionary size of bag-of-words representation, size of co-occurring set of each word and the number of topics.

1) Performance versus segmentation: The proposed approach requires to segment each image to acquire the document-word representation. However, as a perfect image segmentation is not always achievable, the topical objects may be hidden in the segments of cluttered background. Inspired by [5], we segment the frames multiple times and expect that each object instance is correctly segmented by at least one segmentation. We expect that the more number of segmentation we perform for each frame, the better the performance of our method will be. To verify this intuition, we obtain nine segmentations of each frame using normalized cut while each key-frame is segmented into 3, 5, 7, 9, 11, 13, 15, 17 and 19 segments, respectively. We test our method several rounds by using different numbers of segmentations for each frame. In the first round, only the first segmentation of each frame is used. In the following rounds, we gradually increase the number of segmentations for each frame.

Figure 8 (a) shows the performance of our method in different rounds. The average F-$measure$ value of all video sequences is shown. We observe that the performance of LDA-WCP gets better when more segmentations of each frame are used. The F-$measure$ is 0.35 when using one segmentation while the F-$measure$ is 0.53 when using all nine segmentations. We also observe that when more than seven segmentations of each frame are used, the performance does not change significantly. Therefore, seven segmentations of each frame is used for our approach.

2) Performance versus dictionary size: Figure 8 (b) illustrates the performance of LDA-WCP when the dictionary size varies. We observe that the advantage is gained by the dictionary size of 400 visual words. The overall performance of LDA-WCP does not change significantly when the dictionary size is between 100 and 700 visual words. The appropriate number of visual words is helpful to capture the repetitiveness of video objects and handle the variabilities of topical video object due to viewpoint, illumination changes, scale changes, and partial occlusion, etc.

3) Performance versus number of co-occurrence visual words: For the testing videos, we obtain the word co-occurrence prior by considering the spatial-temporal co-occurrence of words in the whole videos. The Gauss-Markov random field prior $p(\beta)$ is built using the co-occurring set of each word as described by Eq.12. Figure 8 (c) shows the performance of our methods for tested videos when using different numbers of co-occurring visual words. It can be seen that the overall performance of LDA-WCP does not change significantly with the size of co-occurring word set, with a minor advantage being gained by using the top $C = 30$ co-occurring visual words for each word. This demonstrates that the small number of co-occurring words is able to capture the important co-occurrence prior information.

4) Performance versus number of topics: Figure 8 (d) illustrates the performance of LDA-WCP when the number of topics varies. We observe that the advantage is gained by the smaller number of topics. The overall performance of LDA-WCP did not change significantly when the number of topics is between 4 and 8. However, the larger number of topics reduces the performance of LDA-WCP as the instances of one topical video objects might be clustered to multiple topics.

These evaluation results demonstrate that it is convenient

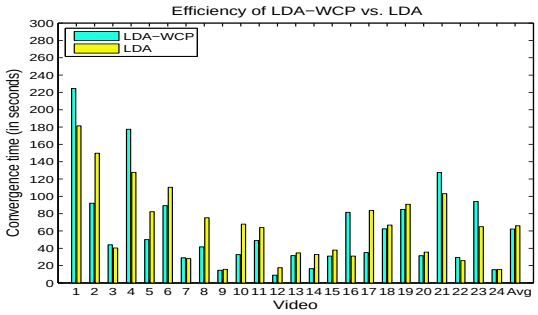


Fig. 11. Comparison of computational cost of LDA-WCP and LDA using Dataset 1. For each video, it shows the convergence time of EM based inference and learning algorithm.

to set the parameters of the proposed LDA-WCP model.

F. The Computational Cost of LDA-WCP

In this section, we report the computational cost of LDA-WCP model. After obtaining the document-word representation and the word co-occurrence prior for video clips, the un-optimized LDA-WCP implementation in Matlab requires about 60 seconds on average to discover topical video objects from one video using one CPU core on a Xeon 2.67GHz PC. The convergence time of EM algorithm for all videos of Dataset 1 are shown in Figure 11. Due to the low frequency of words co-occurrence in the same frames, the estimation of word co-occurrence prior requires only about 10 more seconds on average for each video. To process one video, the original LDA requires about 65 seconds on average.

VI. CONCLUSION

Video object discovery is a challenging problem due to the potentially 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 integrates the word co-occurrence prior and the bag-of-words information in a unified way. Then we propose a new variational expectation-maximization (EM) algorithm to solve the LDA-WCP model. This EM algorithm makes the problem tractable and allows for an elegant iterative solution. Experiments on challenging video datasets show that our method is superior to LDA for topical video object discovery.

There are several directions that could be further explored in the future. Currently, we estimate the word co-occurrence prior by checking their effective co-occurrence frequency in the whole video. An alternative approach that can be pursued is leveraging the weakly supervised information about the visual objects [45][46]. This is suitable for targeted object discovery that is tailored to users' interests. In addition, our model can be combined with co-segmentation algorithms [47] and visual saliency discovery [48].

REFERENCES

[1] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Min. Knowl. Discov.*, vol. 15, pp. 55–86, August 2007.
[2] J. Yuan, Y. Wu, and M. Yang, "From frequent itemsets to semantically meaningful visual patterns," in *Proc. KDD*, 2007.
[3] J. Yuan and Y. Wu, "Mining visual collocation patterns via self-supervised subspace learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 2, pp. 334–346, 2012.
[4] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2158–2174, Dec. 2008.
[5] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using multiple segmentation to discover objects and their extent in image collections," in *Proc. CVPR*, 2006.
[6] G. Zhao and J. Yuan, "Discovering thematic patterns in videos via cohesive sub-graph mining," in *Proc. ICDM*, 2011.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[8] G. Zhao, J. Yuan, and G. Hua, "Topical video object discovery from key frames by modeling word co-occurrence prior," in *Proc. CVPR*, 2013.
[9] H. Wang, G. Zhao, and J. Yuan, "Visual pattern discovery in image and video data: a brief survey," *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, vol. 4, no. 1, pp. 24–37, 2014.
[10] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. CVPR*, 2004.
[11] P. Hong and T. S. Huang, "Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs," *Journal of Discrete Applied Mathematics*, vol. 139, pp. 113–135, 2004.
[12] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. J. Guibas, "Imagewebs: Computing and exploiting connectivity in image collections," in *Proc. CVPR*, 2010.
[13] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, Feb. 2010.
[14] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2207–2219, 2012.
[15] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," in *Proceedings of the IEEE Conference on Computer Vision*, ser. ICCV07, 2007.
[16] M. Cho, Y. M. Shin, and K. M. Lee, "Unsupervised detection and segmentation of identical objects," in *Proc. CVPR*. IEEE, 2010.
[17] H. Wang, J. Yuan, and Y. Tan, "Combining feature context and spatial context for image pattern discovery," in *Proc. ICDM*, Vancouver, Canada, 2011.
[18] H. Wang, J. Yuan, and Y. Wu, "Context-aware discovery of visual co-occurrence patterns," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1805–1819, 2014.
[19] J. Gao, Y. Hu, J. Liu, and R. Yang, "Unsupervised learning of high-order structural semantics from images," in *Proc. ICCV*, Kyoto, Japan, 2009.
[20] W.-T. Chu and M.-H. Tsai, "Visual pattern discovery for architecture image classification and product image search," in *Proc. ICMR*, 2012.
[21] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. CVPR*, 2010.
[22] H. Xie, Y. Zhang, K. Gao, S. Tang, K. Xu, L. Guo, and J. Li, "Robust common visual pattern discovery using graph matching," *J. Visual Communication and Image Representation*, vol. 24, no. 5, pp. 635–646, 2013.
[23] J. Liu and Y. Liu, "Grasp recurring patterns from a single view," in *Proc. CVPR*, 2013.
[24] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.
[25] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. ICCV*, 2005.
[26] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. CVPR*, 2008.
[27] D. Liu and T. Chen, "A topic-motion model for unsupervised video object discovery," in *Proc. CVPR*, 2007.

[28] J. Tang and P. H. Lewis, "Non-negative matrix factorisation for object class discovery and image auto-annotation," in *Proc. CIVR*, 2008.
[29] M. Sun and H. V. Hamme, "Image pattern discovery by using the spatial closeness of visual code words," in *Proc. ICIP*, 2011.
[30] H.-K. Tan and C.-W. Ngo, "Localized matching using earth mover's distance towards discovery of common patterns from small image samples," *Image Vision Comput.*, vol. 27, no. 10, pp. 1470–1483, 2009.
[31] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden topic markov models," in *Proc. AISTATS*, 2007.
[32] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Proc. NIPS*, 2007.
[33] J. Philbin, J. Sivic, and A. Zisserman, "Geometric lda: A generative model for particular object discovery," in *Proc. BMVC*, 2008.
[34] L. Cao and F.-F. Li, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. ICCV*, 2007.
[35] M. Andreetto, L. Zelnik-Manor, and P. Perona, "Unsupervised learning of categorical segments in image collections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1842–1855, 2012.
[36] L. Bolelli, c. Ertekin, and C. L. Giles, "Topic and trend detection in text collections using latent dirichlet allocation," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ser. ECIR '09, 2009, pp. 776–780.
[37] B. Zhao, L. Fei-Fei, and E. P. Xing, "Image segmentation with topic random field," in *Proc. ECCV*, 2010.
[38] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *Conference on Computer Vision & Pattern Recognition*, jun 2007, pp. 1–8.
[39] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proc. ICML*, 2006.
[40] C. Nikou, N. P. Galatsanos, and A. Likas, "A class-adaptive spatially variant mixture model for image segmentation," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1121–1130, 2007.
[41] J. Chang, *Uncovering, Understanding, and Predicting Links*. Princeton University, 2011. [Online]. Available: http://books.google.com.sg/books?id=VHs2MwEACAAJ
[42] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. ICCV*, 2005.
[43] J. Chang and D. M. Blei, "Relational topic models for document networks," *Journal of Machine Learning Research - Proceedings Track*, vol. 5, pp. 81–88, 2009.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *26th Annual Conference on Neural Information Processing Systems 2012.*, 2012, pp. 1106–1114.
[45] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, 2010.
[46] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *Proc. CVPR*, June 2013.
[47] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," *Proc. CVPR*, June 2013.
[48] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, and J. Brandt, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT)*, 2016.

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

From September 2003 to December 2009, he was a Research Assistant in the Pervasive Computing Lab at ZJU. From March 2010 to September 2014, he was a Research Fellow at Nanyang Technological University. Since October 2014, he has been a Senior Research Scientist at Morpx Inc., Hangzhou, China. His current research interests include computer vision and machine learning.

Junsong Yuan is currently an Associate Professor and program director of video analytics at School of EEE, Nanyang Technological University, Singapore. He received Ph.D. from Northwestern University, USA, and M.Eng. from National University of Singapore. Before that, he graduated from Special Class for the Gifted Young of Huazhong University of Science and Technology, China. His research interests include computer vision, video analytics, action and gesture analysis, large-scale visual search and mining, etc. He has authored and co-authored 3 books, and 130 conference and journal papers. He serves as Program Co-Chair of IEEE Visual Communications and Image Processing (VCIP'15), Organizing Co-Chair of Asian Conf. on Computer Vision (ACCV'14), Area chair of IEEE Winter Conf. on Computer Vision (WACV'14), IEEE Conf. on Multimedia Expo (ICME'14 15), and Asian Conf. on Computer Vision (ACCV'14). He also serves as guest editor for International Journal of Computer Vision (IJCV), associate editor for The Visual Computer journal (TVC), IPSJ Transactions on Computer Vision and Applications (CVA), and Journal of Multimedia (JMM). He co-chairs workshops at SIGGRAPH Asia14, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'12 13 15), IEEE Conf. on Computer Vision (ICCV'13), and gives tutorials at ACCV'14, ICIP'13, FG'13, ICME'12, SIGGRAPH VRCAI'12, and PCM'12. He received Nanyang Assistant Professorship from Nanyang Technological University, Outstanding EECS Ph.D. Thesis award from Northwestern University, Best Doctoral Spotlight Award from CVPR'09, and National Outstanding Student from Ministry of Education, P.R.China.

Gang Hua received the B.S. degree in automatic control engineering and the M.S. degree in control science and engineering from Xian Jiaotong University (XJTU), Xian, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA, in 2006. He was enrolled in the Special Class for the Gifted Young of XJTU in 1994. He is currently a Research Manager in visual computing group of Microsoft Research Asia, Bejing China. He was a research staff member of the IBM Research Thomas J. Watson Center, Hawthorne, NY, USA, from 2010 to 2011, a Senior Researcher with the Nokia Research Center, Hollywood, CA, USA, from 2009 to 2010, and a Scientist with the Microsoft Live Labs Research, Redmond, WA, USA, from 2006 to 2009. He has authored over 50 peer-reviewed publications in prestigious international journals and conferences. He holds three U.S. patents and 17 more patents pending. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IAPR Journal of Machine Vision and Applications, and a Guest Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the International Journal on Computer Vision. He was the Area Chair of the 2011 IEEE International Conference on Computer Vision and the 2011 ACM Multimedia, and the Workshops and Proceedings Chair of the 2011 IEEE Conference on Face and Gesture Recognition. He is a member of the Association for the Computing Machinery.

Jiong Yang is currently a PhD candidate in Rapid Rich Object Search Lab, College of Engineering, Nanyang Technological University, Singapore. He received the B.Eng degree with first class honor in School of Electrical Electronic Engineering, Nanyang Technological University, Singapore in 2013. His research interests include computer vision and machine learning.