Pattern Recognition 68 (2017) 66–81

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Multi-layer multi-view topic model for classifying advertising video

Sujuan Hou[a,e], Ling Chen[b], Dacheng Tao[b], Shangbo Zhou[c], Wenjie Liu[d], Yuanjie Zheng[a,e,f,*]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China
[b] Centre for Quantum Computation and Intelligent Systems, FEIT,University of Technology, Sydney, Australia
[c] College of Computer Science, Chongqing University, Chongqing 400030, China
[d] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, PR China
[e] Institute of Life Sciences, Shandong Normal University, Jinan 250014, China
[f] Key Laboratory of Intelligent Information Processing, Shandong Normal University, Jinan 250014, China

ARTICLE INFO

Article history:
Received 9 August 2016
Revised 30 November 2016
Accepted 1 March 2017
Available online 2 March 2017

Keywords:
Video representation
Ad video classification
Multi-layer
Multi-view
Topic model

ABSTRACT

The recent proliferation of advertising (ad) videos has driven the research in multiple applications, ranging from video analysis to video indexing and retrieval. Among them, classifying ad video is a key task because it allows automatic organization of videos according to categories or genres, and this further enables ad video indexing and retrieval. However, classifying ad video is challenging compared to other types of video classification because of its unconstrained content. While many studies focus on embedding ads relevant to videos, to our knowledge, few focus on ad video classification. In order to classify ad video, this paper proposes a novel ad video representation that aims to sufficiently capture the latent semantics of video content from multiple views in an unsupervised manner. In particular, we represent ad videos from four views, including bag-of-feature (BOF), vector of locally aggregated descriptors (VLAD), fisher vector (FV) and object bank (OB). We then devise a multi-layer multi-view topic model, *mlmv_LDA*, which models the topics of videos from different views. A topical representation for video, supporting category-related task, is finally achieved by the proposed method. Our empirical classification results on 10,111 real-world ad videos demonstrate that the proposed approach effectively differentiate ad videos.

1. Introduction

Recent years have witnessed a fast and consistently growing online advertising market. Video advertising associates ads with online videos, and has become a key online monetization strategy. According to [1], digital advertising revenue accounted for $42.6 billion in 2013, and video advertising represented around 9.7% of that amount at about $4.15 billion. The Interactive Advertising Bureau (IAB) [2] also reported that more than two-thirds (68%) of marketers and agency executives expect to see their digital video advertising budgets increase in the next 12 months. According to the second annual Digital Content NewFronts 'Digital Video Spend Study,'[1] a survey of 305 buy-side professionals conducted by Advertiser Perceptions and the IAB, shows an increasing number of video websites, such as *YouTube, Dailymotion, Hulu* and *Youku*, are becoming profitable by providing effective video advertising services. There are two basic ways of placing advertising in videos: embedding advertising (e.g., in the front of the video) or overlaying text on certain frames (e.g., at the bottom of the videos).

Due to the surge in the amount of online ad videos, designing effective video ad management tools and systems is becoming imperative. This paper focuses on developing a fundamental tool for organizing ad videos by classifying them. Classifying ad videos is potentially invaluable in practice. First, classifying video ads with respect to the advertised products or services (e.g., automobiles, cosmetics etc.) can enable *contextual advertising*. This is defined as embedding relevant ad video in target videos. For example, it may be more effective to display an ad of sports wear, rather than an ad for furniture, in a video of a football match based on the assumption that users who are interested in football games would have a similar interest in sports products. Second, grouping ad videos by type of advertised product is helpful for *personalizing advertising services*. Much research has been conducted to model user interests into topics from their online behaviors (e.g., browsing and tagging). For a user who is interested in automobile products, embedding car ads is preferred over ads that are randomly selected. Third, categorizing ad videos makes browsing and searching for video advertising more efficient, and facilitates the development of related application, such as advertising library catalogues. For

* Corresponding author at: School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China.
E-mail address: zheng.vision@gmail.com (Y. Zheng).

[1] http://www.iab.com/.

http://dx.doi.org/10.1016/j.patcog.2017.03.003
0031-3203/

example, when creating a video advertisement for a new product promotion, it would be useful to be able to search through ads of similar products for inspiration.

The two main issues to overcome in addressing contextual ads are *contextual relevance* and *contextual intrusiveness. Contextual relevance* studies the problem of associating the most relevant ads to an online video, while *contextual intrusiveness* aims to seamlessly embed the ads at the most appropriate positions within the target video. To the best of our knowledge, we are the first to design a topical representation for ad video to further support category-related application, which can be regarded as the fundamental research into above two issues.

In the field of videos/images, many automatic classifying or indexing approaches have been proposed for types of video/image (e.g., movie, sports, news reports) [3–10]. Additionally, various strategies have been adopted to perform the classification tasks. However, according to [11], ad videos have at least two unique characteristics compared with generic videos (e.g., news and sport), caused by the special requirement of an ad video on drawing more audience's attention in the allotted amount of time. First, ad videos incline to produce special effects by taking advantage of combined versatile techniques related to color, rhythm, motion, object/scene variation, cartoon and involvement of a spokesperson. By contrast, generic videos are relatively consistent in using one or two techniques in the same time period of a typical ad video. Second, the storyline and characters in ad videos are designed specifically to express more *semantic*-like concepts in order to achieve effective perceptual impact. For example, a good ad video usually inspires modesty, self-confidence, action and adventure in order to promote the sale of a car to middle-aged people. Conversely, to promote a delicious food, the video will inspire relaxation, quietude, and calmness. Instead, generic videos prone to present more *object*-like concepts to provide objective facts of the story. Most of the existing work on effective video classification, as studied in [12], primarily focuses on detecting effective video content using low-level features (e.g., text features, audio features, visual features, or the combination of them). Although these low-level features have demonstrated adequate discriminative capability in some video classification tasks, their representation are limited in videos with complex semantics. Therefore, existing classification algorithms for non-ad videos may not be effective for classifying ad videos.

We observe that, in the field of image-related processing, image representation is becoming high-level oriented [13–16], because high-level representations captures more semantic information, which benefits from powerful local descriptors. At present, semantic analysis techniques have been adopted in multimedia processing for various applications [17–19], and show superior performance over approaches based on direct features, especially when the concept of interest is complex and the number of examples is limited. In addition, multi-view learning is a popular research topic in recent years [20] and widely used in classification field [21–23]. The unique characteristics of ad videos motivated us to design a video representation algorithm that captures the semantics of ad videos from a multi-view perspective, by combining these two types of popular techniques. We devise a model, named *multi-layer multi-view topic model* (namely *mlmv*_LDA), to comprehensively describe the video frames. It integrates high-level representations from several different views using topic modelling techniques. Fig. 1 illustrates the *mlmv_LDA* learning process.

An important aspect of the proposed model is that each frame can be represented from multiple views (e.g., view 1, view 2, etc.), which in turn are combined to respond to higher-level representation. This is very natural, since an image can be described from different perspectives. We know that the topic model is a method of dimension reduction, originally used in the field of text processing (from words to topics). Here, our aim is to obtain a concise video representation based on the features from multiple views. However, original one-layer topic models cannot achieve the goal of multi-view fusion. To fuse different data representations from multiple views, additional layer is designed in the proposed method, where the feature vector U^i $(1 \leq i \leq n)$ is learned for each view combined with the first-layer LDA. Through an iterative process in the second-layer LDA, the learned vector U^i $(1 \leq i \leq n)$ and its combination weight w_i $(1 \leq i \leq n)$ are utilized to obtain the final high-level topical representation Z, which serves as the discriminant information of ad videos to train classifiers, such as *kNN* and *random forest*. We collected real-world online ad videos and conducted extensive experiments to evaluate the performance of the proposed *mlmv_LDA* model. Our experimental results demonstrate that the proposed model, which integrates multiple views through multi-layer topic models, outperforms not only classifiers trained on mono-view but also those approaches that directly integrates multi-views.

The remainder of this paper is organized as follows: Section 2 discusses existing work related to our research; Section 3 provides a detailed description of the proposed *mlmv_LDA* model; the experimental results are presented in Section 4; and finally, Section 5 concludes this paper.

2. Related work and background

2.1. Existing research on multimedia advertisements

Conventional advertising treats multimedia ads as general text ads by displaying ads that are either relevant to queries or the web pages. However, media, like image and video, is now used almost as much as text in web pages, and has become a powerful and effective information carrier for online ads. A survey of the trends in Internet multimedia ads, and the methodologies for ads driven by the rich content of images and videos is provided in [24]. In this study, multimedia ads are grouped into three main categories according to medium: (a) text or keyword ad, (b) image ad, and (c) video ad.

Most reviews of text ads focus on two key issues: ad keyword selection, and ad relevance matching. Typical advertising systems analyze a web page or query to find prominent keywords or categories, and then sell those keywords or categories to advertisers. *TermsNet* [25] is a typical keyword generation framework for generating non-obvious yet relevant keywords. Yih et.al [26] propose a system that learns how to extract keywords from web pages for targeted ads. In terms of specific applications, eBay created a contextual ad platform [27] to solve the problem of keyword extraction using in contextually relevant links to *eBay* assets on third party sites. The platform uses linear and logistic regression models learned from manually labeled data, as well as documents, texts and eBay-specific features. Furthermore, Border et al. [28] present a semantic approach that matches ads to web pages.

In terms of image ads, contextual relevance deals with the selection of relevant ads according to a given image. *ImageSense* [29], the first attempt towards contextual in-image ads, is capable of automatically selecting suitable images from web pages not only based on textual relevance but also visual similarity. Another exemplary study on context-aware image ads is the advertising system in [30]. The system chooses a semantically relevant ad and blends with an enlarged thumbnail image based on visual consistency, allowing ads to be embedded into images in a non-intrusive and visually pleasant manner.

As mentioned in the Introduction, existing research on video ads mainly focuses on contextual ads, which aim to embed contextually relevant ad videos at proper positions within the video stream. *VideoSense* [31] is a contextual in-video advertising system.

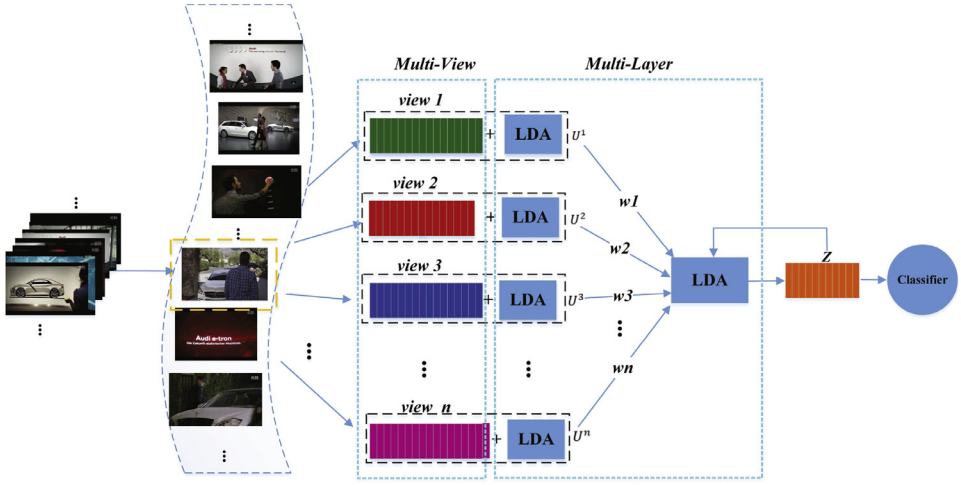


Fig. 1. An illustration of the *mlmv*_LDA learning process.

It automatically associates the relevant video ad and then seamlessly inserts the ad, at the appropriate positions, into each video. Several other studies [32–35] focus on context-aware video advertising technologies that can automatically insert video ads into video clips. In [36], real-time ads inserted into in baseball video are studied for their promotion effect. Many studies focus on video advertisement overlay techniques. These technique detect a set of spatio-temporal non-intrusive positions within a consecutive series of video frames and then overlay contextually relevant ads at these points. Some exemplars are reflected in [37–39]. Most recently, an annotation approach specifically for ad videos was proposed from the perspective of multi-label learning [40].

2.2. Mid/high-level image representation overview

In this section, we review several popular mid-level representations that produce a vector representation of an image from a set of local descriptors, and a high-level representation of image, i.e., object bank (OB) representation.

Mid/high-level representation becoming popular with large-scale datasets in the field of video/image, and provides a supplement for low-level features. There are two reasons why these mid/high-level representations are popular for categorization and indexing applications. First, these representations benefit from powerful local descriptors. Second, these vector representations are comparable with standard distances, and subsequently used by robust classification methods such as *k*NN and *random forest*. In this paper, we employ four different mid-level representations: bag-of-features (BOF) [41], vector of locally aggregated descriptors (VLAD) [16,42], improved fisher vector (IFV) [14,43] and OB [13].

Several local descriptors have been adopted to support high-level representation in real-world videos/images. Among these, SIFT [44] acts as a typical local visual descriptor with the ability to capture sufficiently discriminative local elements that have some invariant properties of geometric or photometric transformations and is robust to occlusions. In this study, we use a set of SIFT descriptors to represent the key frames. Suppose there are N feature points detected in a video keyframe, F can be represented as,

$$F = \{\boldsymbol{M}_k x_k, y_k, s_k, d_k\}, \quad for\ k \in \{1, 2, \ldots, N\},$$

where $\boldsymbol{M}_k$ is a 128-dimensional local edge orientation vector of the SIFT point, x_k, y_k are the x- and y-positions, and s_k, d_k are the scale and the dominant direction of the kth feature point.

With respect to the BOF representation [41], visual vocabularies are built from training descriptors using hierarchical k-means clustering. We set the size of vocabulary to w (w=10,000) in the experiments. Descriptors from keyframes are assigned to certain visual words and aggregated into a w-dimensional histogram, denoted as the BOF representation.

Before introducing two extensions to FV [43], we first review the FV representation, which models the code-words with Gaussian mixture model (GMM). Much more information is stored in the FV visual word compared to BOF, because FV stores the first and second order moments of the patches assigned to each visual word. Let $X = \{x_1, \ldots, x_T\}$ be a set of T local descriptors (e.g., SIFT) extracted from keyframes. The FV assumes the generation process of X can be modeled by the GMM with the parameters $\Theta = (u_k, \sum k, \pi_k : k = 1, \ldots, K)$. The GMM associates each local descriptor in X to a model k, in the mixture with a strength given by the posterior probability:

$$q_{ik} = \frac{exp[-1\frac{1}{2}(x_i - u_k)^T \sum_k^{-1}(x_i - u_k)]}{\sum_{t=1}^{K} exp[-\frac{1}{2}(x_i - u_k)^T \sum_k^{-1}(x_i - u_k)]} \tag{1}$$

For each model k, the mean and covariance deviation vectors are formulated as:

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji-u_{jk}}}{\sigma_{jk}} \tag{2}$$

$$v_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik}((\frac{x_{ji} - u_{jk}}{\sigma_{jk}})^2 - 1) \tag{3}$$

where $j = 1, \ldots, T$ spans the vector dimensions.

One recent extension to the FV is the VLAD [16], which aggregates the local SIFT descriptors into a compact vector representation. The VLAD can be regarded as a simplification of the FV. Specifically, it first learns a codebook $C = \{c_1, c_2, \ldots, c_k\}$ of k visual words (k=10,000 in our experiments) with hierarchical k-means. Then each local descriptor is associated to its nearest visual word. The main idea of the VLAD is to accumulate the difference between each local descriptor and its nearest visual word.

Another extension to the FV is the improved Fisher Vector (IFV). The improvements [14] are twofold : first, IFV adopts the L2 normalization strategy in the Fisher Kernel; while the second is it uses a non-linear additive kernel. Hellinger's kernel (or the Bhattacharya coefficient) can be used instead of the linear kernel at no cost by signed squared root.

Object Bank (OB) is a type of high-level image representation which is constructed from the responses of many object detectors

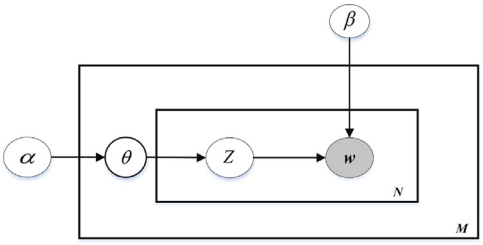


Fig. 2. Original LDA.

Table 1
The generative model for the original LDA.

1. Choose $N \sim$ Poisson (ξ).
2. Choose $\beta \sim$ Dir (α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$ Multinomial(θ).
   (b) Choose a word w_n from p ($w_n \mid z_n$, β),
A multinomial probability conditioned on the topic z_n.

Table 2
Parameter notations for the LDA.

Parameter	*Notation*
α	A prior topic distribution over all the documents with K-dimensional
β	A prior distribution, it is a matrix with $K\times V$ (V is the total number of words, and β_{ij}=$p(w_j\|z_i)$ is the probability of word w_j in topic z_i)
θ	The topic distribution over each document
z	The topic vector with N-dimensional
$\boldsymbol{w}$	The vector with N words

(it can also deemed as the responses of a 'generalized object convolution'). It is obtained from each key frame in ad videos by employing a large number of pre-trained generic object detectors described by the SIFT features. Each object detector is trained at 6 scales, 2 components and 3 spatial pyramid levels.

2.3. Classical works on topic model

When discussing topic models, we must mention Latent Dirichlet Allocation (LDA) [45], a popular technique in the scope of semantic analysis. It is a generative way to model a corpus, which assumes there is a given process to produce a document, and predicts the generation process based on the observed documents. Specifically, it assumes that words are generated by K topics (distribution over words), prior to producing a document, a topic distribution over the document needs to be produced, and the words are then produced. The graphical representation is shown in Fig. 2.

Given the parameters α and β, the joint probability of generating a document with a topic mixture θ, and a set of N topics z and a set of N words can be formed as:

$$p(\theta, z, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \tag{4}$$

For a document $\boldsymbol{w}$, Table 1 shows the generative process.

Table 2 lists the notations for the parameters used in the LDA.

The parameters θ and z are latent variables in Eq. (1). The marginal distribution of a document can be obtained by integrating over θ and summing over z:

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta \tag{5}$$

For a corpus D with M documents, we obtain the probability of D by taking the product of the marginal probabilities of M documents:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d \tag{6}$$

The purpose of the training process in the LDA is to find the parameters α and β that can maximize the value of $p(D \mid \alpha, \beta)$.

After deriving the optimal α and β, for a new document d_{new}, we can determine the topic distribution over d_{new} and predict which topic each word belongs to.

In recent years, topic models have been widely studied to solve semantic-related tasks. Several probabilistic topic models, such as [46,47], have been proposed to learn topic representations of document corpus. Improved topic models have also been studied in a variety of applications, such as topic learning from documents with dependencies between words [48], object discovery from images [49], latent topic learning from videos [50], and topic detection from online heterogenous data [51].

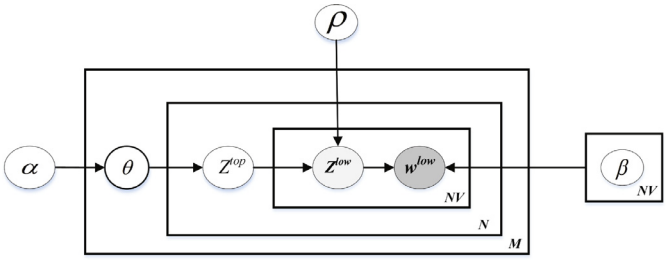


Fig. 3. Graphical model representation of the *mlmv_LDA* model.

Table 3
The generative procedure of the *mlmv*_LD A.

For each video $\boldsymbol{d}$ in $\boldsymbol{D}$= $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M\}$

1. Draw a topic proportion $\theta|\alpha \sim Dir(\alpha)$;
2. For each of $N\times$ NV words in $\boldsymbol{d}$
   (1) Draw top-level topic assignment $Z_n^{top}|\theta \sim Multi(\theta)$
   (2) For each view $\boldsymbol{p} \in$ {*VLAD, BOF, IFV, OB*}
      1) Draw low-level topic assignment $Z_n^{low}|Z_n^{top}, \rho_{1:K} \sim Multi(\rho)$
      2) For each *word* from the perspective $\boldsymbol{p}$ view
         ▲Draw a word $w_n^p|Z_n^{low}, \beta_{1:K^p} \sim Multi(\beta^p_{Z_n^{low}})$

3. Multi-layer multi-view topic model

3.1. mlmv_LDA

In this section, we introduce the proposed multi-layer multi-view topic model, *mlmv_LDA* for short. We know that different topic themes can be derived based on different lexicon sets. In *mlmv_LDA*, each ad video is viewed as the result of weighted topical themes based on *NV* (*NV*=4) different mid/high-level-representation views, i.e., VLAD, BOF, IFV and OB, as described in Section 2.2.

Fig. 3 illustrates the graph model of the proposed *mlmv_LDA*. The collection of M ad videos is denoted by $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M\}$. It assumes that in the *corpus* $\boldsymbol{D}$, each video $\boldsymbol{d}_i$ ($1 \le i \le M$) arises from a mixture distribution over numerous latent topics.

The generative process of the proposed *mlmv_LDA* model is listed in Table 3.

In the proposed model, video-level topic distribution is expressed as $\theta \sim Dir(\alpha)$, $\theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$, where θ_i ($1 \le i \le K$) represents the probability of selecting the ith topic. When certain topic is selected from θ, there are two levels of topic for words: Z^{top} (top-level topic) and Z^{low} (low-level topic). The Z^{top}

Table 4
Parameter notations for the *mlmv_LDA*.

Parameter	*Notation*
α	Hyperparameter on the mixing proportions, Dirichlet prior of θ
M	The number of ad videos
K	The number of topics/mixture components
N	The number of top words in current $\boldsymbol{d}$
NV	The number of views for each video (NV=4 here), they are {VLAD→ 1, BOF→ 2, IFV→ 3, OB→ 4}
K^1	Topics learned based on V^1 from the VLAD view
K^2	Topics learned based on V^2 from the BOF view
K^3	Topics learned based on V^3 from the IFV view
K^4	Topics learned based on V^4 from the OB view
β^1	Hyperparameter on the VLAD word probabilities, a matrix with $K^1 \times V^1$
β^2	Hyperparameter on the BOF word probabilities, a matrix with $K^2 \times V^2$
β^3	Hyperparameter on the IFV word probabilities, a matrix with $K^3 \times V^3$
β^4	Hyperparameter on the OB word probabilities, a matrix with $K^4 \times V^4$

can be deemed as the weighted combination of NV Z^{low}. We assign VLAD→ 1, BOF→ 2, IFV→ 3 and OB→ 4, therefore, there are four types of Z^{low}, i.e., Z_1, Z_2, Z_3 and Z_4 toward the four corresponding views. They are also treated as the elements of top-level topic. Four types of dictionary words, namely V^1, V^2, V^3, V^4 corresponding to four views, constitute the low-level lexicon set V^{low}, where $V^{low} = \{V^1, V^2, V^3, V^4\}$.

The number of topics and words in these four views are represented as $|K^1|$, $|K^2|$, $|K^3|$, $|K^4|$ and $|V^1|$, $|V^2|$, $|V^3|$, $|V^4|$, respectively. The parameter ρ is a matrix with $K\times$ NV, each row is a normalized vector with NV_dimension and follows the multinomial distribution.

Several notations and illustrations in the above model are given in Table 4.

For each video $\boldsymbol{d}$, given the parameters α, ρ, β^1, β^2, β^3, β^4, the joint distribution of a topic mixture θ, a set of topics Z^{top}, and a set of N words is given by:

$$\begin{aligned} &p(\theta, z^{top}, \boldsymbol{d}|\alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4) \\ &= p(\theta|\alpha)\prod_{n=1}^{N} p(z_n^{top}|\theta)\prod_{\eta=1}^{NV}(p(z_\eta^{low}|z_n^{top}, \rho)p(w_\eta^{low}|z_\eta^{low}, \beta^\eta)) \end{aligned} \tag{7}$$

The marginal distribution of a video $\boldsymbol{d}$ can be obtained by integrating over θ and summing over z^{top} joint distribution of a topic mixture θ, a set of top topic z^{top}, and a set of N top words is given by:

$$\begin{aligned} p(\boldsymbol{d}|\alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n^{top}} p(z_n^{top}|\theta)\right. \\ \left.\times\prod_{\eta=1}^{NV}\sum_{z_\eta^{low}}(p(z_\eta^{low}|z_n^{top}, \rho)p(w_\eta^{low}|z_\eta^{low}, \beta^\eta))\right)d\theta \end{aligned} \tag{8}$$

The probability of the corpus $\boldsymbol{D}$ with M ad videos can be obtained by taking the product of the marginal probabilities of single videos.

$$\begin{aligned} p(\boldsymbol{D}|\alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{dn=1}^{N_d}\sum_{z_{dn}^{top}} p(z_{dn}^{top}|\theta_d)\right. \\ \left.\times\prod_{\eta=1}^{NV}\sum_{z_\eta^{low}}(p(z_\eta^{low}|z_{dn}^{top}, \rho)p(w_\eta^{low}|z_\eta^{low}, \beta^\eta))\right)d\theta_d \end{aligned} \tag{9}$$

Using the variational expectation-maximization (EM) algorithm, we can solve the proposed model. The *E*-step approximates the probability distribution $p(\theta, z^{top}|\boldsymbol{d}, \alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4)$ while the *M*-step estimates the parameters by maximizing the lower bound of the log likelihood. The following section will give the details.

3.2. Initialization

More complicated probabilistic models are always accompanied by an explosion in required training time. Thus, we suggest computing a decent initial estimation of the conditional probabilities in a divide-and-conquer procedure. This procedure first computes the independent topic model for each view. Next, the computed topics of all the views are taken as the observed words at the next higher level. The procedure can continue until the top-most topic vector is learned. The finial representation, the top-level topic distribution for each document describes each video as a "distribution over topic distributions" and thereby fuses the multi-view.

3.3. Parameter estimation

This section gives the Bayes method for parameter estimation in the proposed model. Given a corpus composed of M documents $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M\}$, our goal is to find the parameter $\Theta = (\alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4)$ that maximizes the log-likelihood of corpus. The objective function is given in Eq. (10).

Objective function:

$$\begin{aligned} &\mathcal{L}(\boldsymbol{D}|\alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4) \\ &= \underset{\Theta}{argmax}\sum_{d=1}^{M} log(p(\boldsymbol{d}_d)|\Theta) \end{aligned} \tag{10}$$

This equation can not be computed tractably. Thus, we adopt the variational expectation-maximization (EM) algorithm to maximize the log likelihood.

E-step: the aim is to approximate the probability distribution $p(\theta, z^{top}|\boldsymbol{d}, \alpha, \rho, \beta^1, \beta^2, \beta^3, \beta^4)$. The solution is to find a simple variational distribution to approximate the true posterior distribution, by minimizing the *KL* (*Kullback–Leibler*) divergence between them. Here, the simple variational distribution is supposed to be,

$$q(\theta, z^{top}|\gamma, \phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n^{top}|\phi_n) \tag{11}$$

We know that,

$$\begin{aligned} &log(\boldsymbol{d}|\Theta) \\ &= log\int\sum_{z^{top}} p(\theta, z^{top}, \boldsymbol{d}|\Theta)d\theta \\ &= log\int\sum_{z^{top}} \frac{p(\theta, z^{top}, \boldsymbol{d}|\Theta)q(\theta, z^{top}|\gamma, \phi)}{q(\theta, z^{top}|\gamma, \phi)}d\theta \end{aligned}$$

$$\geq \int \sum_{z^{top}} q(\theta, z^{top}|\gamma,\phi) log \frac{p(\theta, z^{top}, \boldsymbol{d}|\Theta)}{q(\theta, z^{top}|\gamma,\phi)} d\theta$$
$$= \int \sum_{z^{top}} q(\theta, z^{top}|\gamma,\phi) log p(\theta, z^{top}, \boldsymbol{d}|\Theta) d\theta$$
$$- \int \sum_{z^{top}} q(\theta, z^{top}|\gamma,\phi) log q(\theta, z^{top}|\gamma,\phi) d\theta$$
$$\doteq \mathcal{L}(\gamma,\phi;\Theta) \tag{12}$$

And that is

$$log p(\boldsymbol{d}|\Theta) - \mathcal{L}(\gamma,\phi;\Theta)$$
$$= KL(q(\theta, z^{top}|\gamma,\phi)||p(\theta, z|\boldsymbol{d},\Theta)) \tag{13}$$

This shows that minimizing the *KL* divergence value is equivalent to maximizing the $\mathcal{L}(\gamma,\phi;\Theta)$.

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

$$(\gamma^*,\phi^*) = \underset{\gamma,\phi}{argmax}\, \mathcal{L}(\gamma,\phi;\Theta) \tag{14}$$

This maximization could be achieved via an iterative fixed-point method. For learning over plentiful videos, the variational updates of Θ are iterated until convergence for each video.

By using factorizations, we expand the $\mathcal{L}(\gamma,\phi;\Theta)$ as:

$$\mathcal{L}(\gamma,\phi;\Theta)$$
$$= log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K} log\Gamma(\alpha_i)$$
$$+ \sum_{i=1}^{K}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right))$$
$$+ \sum_{n=1}^{N}\sum_{i}^{K}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)$$
$$+ \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{ni} log(\rho_{i1}\sum_{m=1}^{|V^1|}\delta 1_n^m \beta_{i,m}^1$$
$$+ \rho_{i2}\sum_{m=1}^{|V^2|}\delta 2_n^m \beta_{i,m}^2 + \rho_{i3}\sum_{m=1}^{|V^3|}\delta 3_n^m \beta_{i,m}^3$$
$$+ \rho_{i4}\sum_{m=1}^{|V^4|}\delta 4_n^m \beta_{i,m}^4) - log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right)$$
$$+ \sum_{i=1}^{K} log\Gamma(\gamma_i) - \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{ni} log\phi_{ni}$$
$$- \sum_{i=1}^{K}(\gamma_i - 1)\left(\Psi\left(\gamma_i - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\right) \tag{15}$$

Where $\delta 1_n^m = 1$ when the nth word comes from the mth of VLAD lexicon set, $\delta 1_n^m = 1$ when the nth word comes from the mth of BOF lexicon set, $\delta 3_n^m = 1$ when the nth word comes from the mth of IFV lexicon set and $\delta 4_n^m = 1$ when the nth word comes from the mth of OB lexicon set.

We then respectively maximize Eq. (15) with respect to ϕ_{ni}, γ_i, set the derivatives to zero and find:

$$\phi_{ni} \propto exp\left(\Psi\left(\gamma_i - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\right)\left(\rho_{i1}\sum_{m=1}^{|V^1|}\delta 1_n^m \beta_{i,m}^1 + \rho_{i2}\sum_{m=1}^{|V^2|}\delta 2_n^m \beta_{i,m}^2 + \rho_{i3}\sum_{m=1}^{|V^3|}\delta 3_n^m \beta_{i,m}^3 + \rho_{i4}\sum_{m=1}^{|V^4|}\delta 4_n^m \beta_{i,m}^4\right) \tag{16}$$

Table 5
The number of details in the ad video data set.

Ad type	*Cars*	*Drinks*	*Dig-devices*	*Cosmetics*	*Footwear*	*Bags*
#Vides	624	343	4649	1855	1093	1547
					Totals	**10,111**

$$\gamma_i = \alpha_i + \sum_{n-1}^{N}\phi_{ni} \tag{17}$$

If we set $v = arg_m(\delta q_n^m = 1), q = 1,2,3,4$ in Eq. (16),

$$\phi_{ni} \propto exp\left(\Psi\left(\gamma_i - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\right)$$
$$\times(\rho_{i1}\beta_{i,v}^1 + \rho_{i2}\beta_{i,v}^2 + \rho_{i3}\beta_{i,v}^3 + \rho_{i4}\beta_{i,v}^4) \tag{18}$$

After the variational parameters ϕ and γ are fixed, we estimate the Θ adopting *MLE* based on a variational EM procedure. For the corpus $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_M\}$, the overall log-likelihood of $\boldsymbol{D}$ is the sum of the log-likelihood for the individual ad video. Furthermore, the overall variational lower bound is the sum of the individual variational bounds.

$$\mathcal{L}(\gamma,\phi;\Theta) = \sum_{n=1}^{M} log p((d)_n||\Theta)$$
$$\geq L(\gamma_n,\phi_n;\Theta) \doteq L(D) \tag{19}$$

M-step: the aim is to estimate the parameters by maximizing the lower bound of the log-likelihood. We first take the derivative with respect to ρ_{ij}, and set it to zero and find:

$$\rho_{ij} \propto \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\beta_{i,m}^j} \tag{20}$$

We then take the derivative with respect to β^1, β^2, β^3, and β^4 set it to zero and find:

$$\beta_{iu^{1,m}}^1 \propto \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\rho_{i1}} \tag{21}$$

$$\beta_{iu^{2,m}}^2 \propto \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\rho_{i2}} \tag{22}$$

$$\beta_{iu^{3,m}}^3 \propto \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\rho_{i3}} \tag{23}$$

$$\beta_{iu^{4,m}}^4 \propto \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dni}}{\rho_{i4}} \tag{24}$$

Using the Newton–Raphson method, the update formula of α is the same as the basic LDA model by maximizing the lower bound with respect to α, $\alpha' = \arg\max_{\alpha}\mathcal{L}(\gamma,\phi;\Theta)$.

4. Experiments

4.1. Video data set

As there is no currently publicly available ad video data set, we collected our data set to perform the experimental analysis. To evaluate the proposed approach, 10,111 real-world ad videos from the internet were used in all experiments. Six categories were considered from the perspective of the promoted product. They are: cars, drinks, digital devices, cosmetics, footwear and bags. The number of videos and details about each data set is given in Table 5.

4.2. Baseline methods

To validate the effectiveness of the proposed *mlmv_LDA*, we compared the proposed model with other eight baselines, including: mono-view topic model approaches, non-topic-model methods, and one classical multi-view framework. They are:

- **BOF_LDA**: a latent Dirichlet allocation paradigm integrated with the BOF representation.
- **VLAD_LDA**: a latent Dirichlet allocation paradigm integrated with the VLAD representation.
- **OB_LDA**: a latent Dirichlet allocation paradigm integrated with the OB representation.
- **IFV_LDA**: a latent Dirichlet allocation paradigm integrated with the IFV representation.
- **Majority voting view with topic model (MVV)**: trains the classifier in the latent Dirichlet allocating unified learning paradigm by employing four mono-views separately, then gathers their decisions, and outputs the final result through a voting process.
- **Feature_Merged (FM)**: integrates four types of image representations (i.e., VLAD, BOW, IFV and OB) into one vector representation towards each key frame in an ad video. Specifically, four types of feature representations are merged to represent a key frame in video before employing the latent Dirichlet allocation learning paradigm.
- **CNN [52]**: a popular convolutional neural network (CNN) architecture. We adopted the parameter setting suggested in the paper [52]. The training was performed using stochastic gradient decent with image batch size of 128 images and the learning rate was reduced by hand after 20 K iterations from an initial setting of $1e^{-3}$.
- **MCCA (Multi-view canonical correlation analysis) [53]**: a widely used multi-view CCA-based framework that first extracts the BOF using hierarchical *k*-means, and then accumulates these features into a histogram for each view. After every histogram is normalized to 1, the MCCA [53] framework is employed to fuse multi-view features, and finally obtains the video representation.

4.3. Preprocessing and experimental setting

To obtain mid/high-level representations from the videos, we first segmented all the videos and extracted the key frames from all the shots. The SIFT was then extracted from each key frame and treated as the low-level feature. When carrying out single-view representation, the local features were quantized into V=10,000 visual words using hierarchical *k*-means clustering technology for the BOF and VLAD representations. To generate the OB representation, we adopted 177 pre-trained general object detectors [13], each of which was trained at 6 scales, 2 components and 3 spatial pyramid levels. With respect to the IFV representation, as in [14], we applied *power* and *L2* normalization and 256 Gaussians with SIFT features to compute the improved variant of the FV.

In the experiments of the LDA-based and MCCA approaches, all results are conducted based on 10-*fold* cross validation by adopting two typical classification algorithms, including *k*NN and *random forest* classification. We verified the different classification performance for the LDA-based approaches, when Z_k (the number of topics) and *nTree*(the number of trees) differed, where $Z_k = \{10, 30, 50, 70, 90, 120, 150\}$ and $nTree = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200\}$. In total, 105 groups of experiments were conducted with the *k*NN classifier, and 140 groups conducted with the *random forest* classifier. In each run, we randomly sampled 90% of the data set for training and reserved the rest for testing.

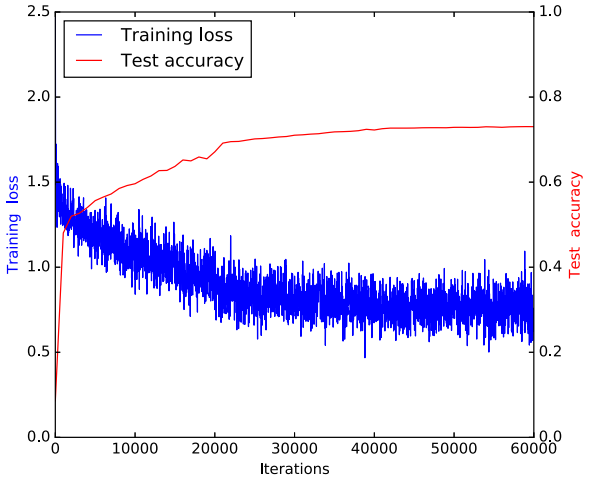


Fig. 4. The learning rate with iterations using CNN on the Ad video data set. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

The performance was measured by the evaluating the *mean average accuracy* (MAA) and *stand deviation*(SD) of 10-*fold* cross validation.

4.4. Experimental results and discussion

(1) **Results by CNN [52]**: the learning rate curves for the test accuracy and training loss are reported in Fig. 4. The blue curve indicates the training loss rate, and the red curve indicates the test accuracy.

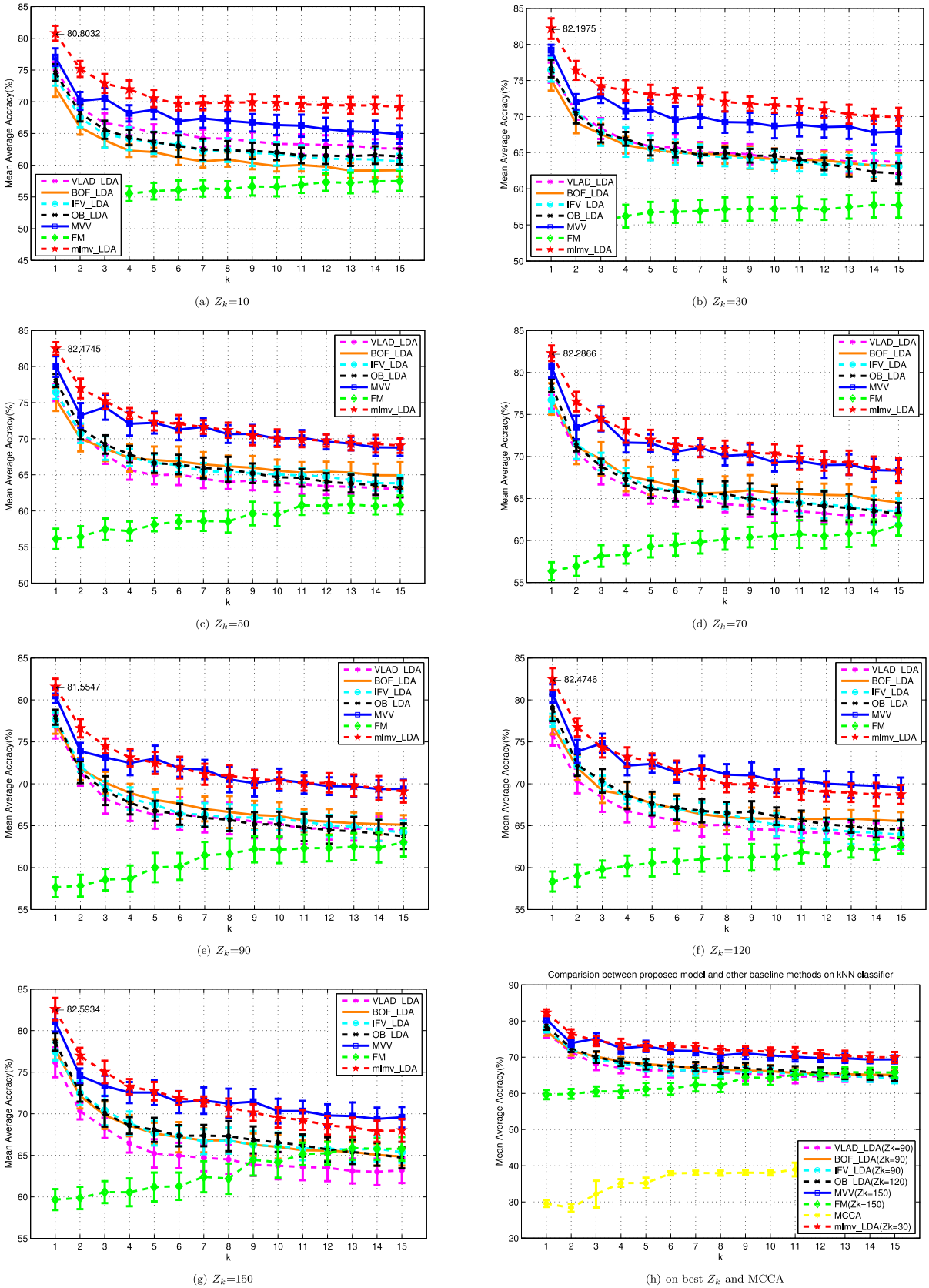
The CNN algorithm achieved convergence after about 60,000 iterations. In terms of test accuracy, there was an initial dramatic increase when the number of iterations increased from 0 to 220,000, and then a slight increase before reaching 40,000. The accuracy leveled off at 73.04%. By contrast, almost the opposite occurred with training loss.

(2) **Results by *mlmv*_LDA and other baselines**

Results by adopting *k*-NN classifier: we conducted the *k*-NN classification tasks with respect to BOF_LDA, VLAD_LDA, OB_LDA, IFV_LDA, MVV, FM, MCCA, and *mlmv_LDA* in terms of 15 different values of *k* (the number of the nearest-neighbours). For the LDA-based strategies, the experiments were designed to find a trade-off between Z_k and *k*. Theoretically speaking, the greater the value of Z_k, the better the performance, as larger values of Z_k can carry more elaborated discriminative semantic information.

Fig. 5 provides a graphic display of the *k*NN classification results with several strategies under different parameter settings. Fig. 5(a)–(g) show the performance trend when Z_k increases from 10 to 150, and Fig. 5(h) shows the best performance of the different LDA-based methods in their respective best Z_k. The performance of MCCA is also shown. Both the MAA and SD of the accuracy are applied to the results.

The results of Fig. 5(a)–(g) demonstrate that: (1) the performance gets worse as *k* increases for most strategies, so we can achieve the best point at k=1 for these methods. (2) As an exception, the performance of the FM strategy becomes better with *k* increases. (3) the proposed *mlmv_LDA* strategy produces dominant performance compared to other approaches when Z_k differs from 10 to 90. However, the MVV method becomes competitive when Z_k is greater than 90 and *k* is greater than 3. Nevertheless, the best performance point drops at *mlmv_LDA* with regard to the different Z_k. We observe from Fig. 5(h) that: (1) the proposed *mlmv_LDA* method is superior to other baselines on almost all classification tasks. (2) the MVV method is comparable to *mlmv_LDA* in a few



(a) Z_k=10 (b) Z_k=30

(c) Z_k=50 (d) Z_k=70

(e) Z_k=90 (f) Z_k=120

(g) Z_k=150 (h) on best Z_k and MCCA

Fig. 5. *k*NN classification performance comaprison. (a)–(g) Different LDA-based strategies (including the proposed *mlmv*_LDA)when Z_k varies from 10 to 150. (h)The best Z_k for LDA-based approaches and MCCA strategy.

Table 6
The highest performance *k*NN classifier with different strategies.

%	VLAD _LDA	BOF _LDA	IFV _LDA	OB _LDA	MVV	FM	MCCA	CNN	*mlmv* _LDA
k	1	1	8	1	1	13	12	–	2
Z_k	90	150	90	120	150	150	–	–	150
MAA	76.83	77.75	77.67	79.06	81.12	65.88	40.42	73.04	**82.59**

Table 7
The highest performance on *random forest* classifier with different strategies.

%	VLAD _LDA	BOF _LDA	IFV _LDA	OB _LDA	MVV	FM	MCCA	CNN	*mlmv* _LDA
nTree	150	180	140	160	120	100	190	–	180
Z_k	90	120	150	30	150	150	–	–	30
MAA	80.11	80.60	81.16	82.15	82.22	70.00	81.61	73.04.	**85.08**

cases, by assigning k as 3 or 5. (3) the accuracy of VLAD_LDA, BOF_LDA, IFV_LDA and OB_LDA tend to be insensitive to the value of k as it increases. (4) the VLAD_LDA did not perform as well as several other methods (e.g., IFV_LDA, BOF_LDA) in some cases. This may be because the VLAD representation is just a simple version of Fisher Vector at the expense of some discriminative ability. (5) the FM strategy becomes more competitive compared to other mono-view strategies (i.e., VLAD_LDA, BOF_LDA, IFV_LDA and OB_LDA) when k is greater than 13. (5) the MCCA method does not consider semantic information, and this may account for its last-place perform.

In addition, we present the parameter combinations achieving the best performance for different strategies with *k*NN classifier in Table 6. After combining the four views and semantic theme, the proposed *mlmv_LDA* respectively improves the classification MAA by 5.8%, 4.8%, 4.9%, 3.5%, 1.5%, 16.7%, 42.2% and 9.6% compared with VLDA_LDA, BOF_LDA, IFV_LDA, OB_LDA, MVV, FM, MCCA and CNN.

Results by adopting a random forest classifier: this section provides experimental results conducted on a *random forest* classifier with different strategies to further verify the effectiveness of proposed approach. We performed a total of 140 group of experiments to verify the best *nTree* and Z_k for LDA-based approaches hoping to find the trade-off between *nTree* and Z_k.

We performed different strategies on the *random forest* classifier and show a graphic display of them in Fig. 6. Fig. 6(a)–(g) show the performance trend when Z_k increases from 10 to 150, and Fig. 6(h) shows the best classification performance for all methods in their respective best Z_k. The performance result of MCCA is also compared. Both the MAA and SD for accuracy are were applied to the reported results.

We notice that: (1) the proposed *mlmv_LDA* strategy consistently outperforms other baseline approaches on all Z_k. (2) With respect to all the strategies, the performance apparently tends to be better when *nTree* increases. (2) Nearly all strategies are not sensitive to the value of *nTree*, which indicates that even small *nTree* still could capture sufficiently discriminative power. In a word, according to the overall classification accuracy ranking in part (h), we can draw the conclusion as follows: *mlmv_LDA*> MVV> OB_LDA> MCCA> IFV_LDA> BOF_LDA> VLAD_LDA> FM.

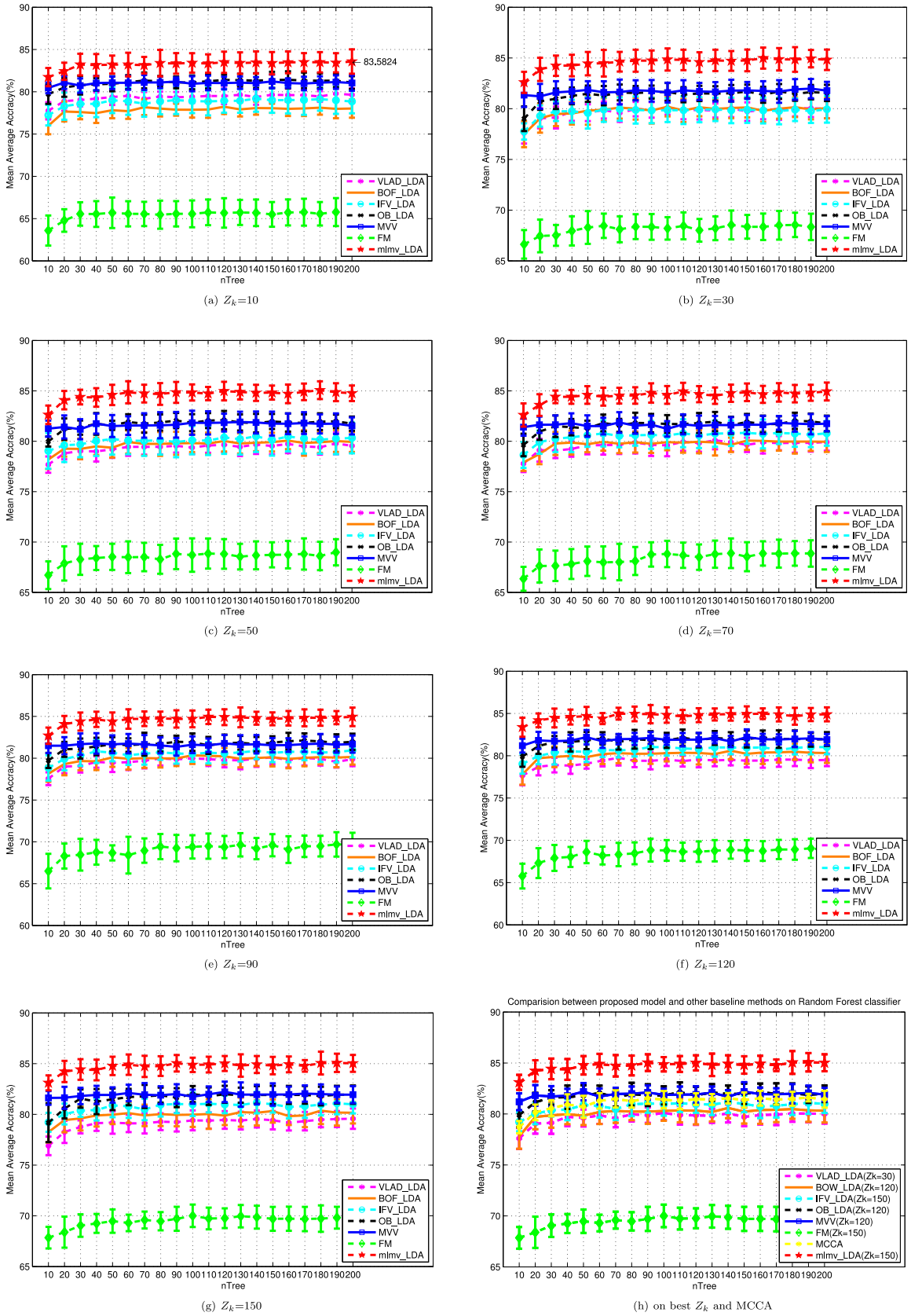
Furthermore, the parameter combinations leading to the best performance for different strategies with *random forest* classifier are listed in Table 7, which also demonstrates the superiority of proposed model. We have the similar observation that *mlmv_LDA* performs the best, and obtains 5.0%, 4.5%, 3.9%, 2.9%, 2.9%, 15.1%, 3.5%, and 12.0% improvement on the average performance compared with VLAD_LDA, BOF_LDA, IFV_LDA, OB_LDA, MVV, FM, MCCA and CNN, respectively.

It is worth mentioning that in the reported results, the FM performed poorly in most cases, even worse than the methods using a single view. This was possibly caused by feature incomparability among different views. First, dimensions differ greatly among these features. For example, the IFV feature has 65,536 dimensions, while the VLAD feature only has 128. In this case, the representation ability of the VLAD would be severely weakened when combined with the IFV. Second, the real values differ greatly among these features. Most real values in the OB and part of the IFV are negative, while those in the VLAD and the BOF are positive. Although all the single-view features were normalized before merging, they are still inevitable lopsidedness. Last, the sparsity degree differs greatly among different feature spaces. For example, only a few dimensions have meaningful real values in a BOF with more than 10,000 dimension. By contrast, each dimension has a meaningful real-value in a VLAD with 128 dimensions. The representation abilities of these types of feature may weaken each other when simply merging together. In this sense, simply merging multi-view features into a vector representation is not a good method to represent data with multi-view features, when the difference among features is large.

Though the number of topics (i.e., Z_k in our experiments) in *mlmv*_LDA needs to be predetermined and the learned concepts may not have explicit semantic meaning that is identifiable by people, the proposed *mlmv*_LDA is an excellent representation learning model that is able to automatically discover category-related latent concepts, and learn the underlying topical representation in an unsupervised manner. Fig. 7 illustrates an example of the general process of obtaining a topical video representation when setting the dimension of semantic space to 10 (i.e., Z_k=10) using the proposed *mlmv*_LDA to support further ad video classification.

The right part of Fig. 7 gives three examples of the learned topical representation using *mlmv*_LDA, where subfigure (a) shows the topical representation of "*Audi*" car promotion, subfigure (b) indicates the topical representation of "*BMW*" car promotion, while subfigure (c) illustrates the topical representation of the "*LV*" bag. Intuitively, subfigures (a) and (b) are more similar to each other than to subfigure (c). Because both (a) and (b) belong to the category of *car*, while (c) belongs to the class of *bag*. If Euclidean distance is used to measure the distance between the videos in latent semantic space, the topical distance between subfigures (a) and (b) would be 0.3854, while the distance between subfigures (a) and (c) is 0.6259. That is, the distance measured in the learned latent semantic space properly reflects the proximity between the three subfigures, which verifies that our *mlmv*_LDA model effectively captures the semantic content of ad videos.

Lastly, in order to facilitate classifier selection, we summarize the best results for these two classifiers as in Table 8.



(a) Z_k=10 (b) Z_k=30

(c) Z_k=50 (d) Z_k=70

(e) Z_k=90 (f) Z_k=120

(g) Z_k=150 (h) on best Z_k and MCCA

Fig. 6. *Random forest* classification performance comparison. (a)-(g) Different LDA-based strategies (including the proposed *mlmv*_LDA)when Z_k varies from 10 to 150. (h)The best Z_k for LDA-based approaches and MCCA strategy.

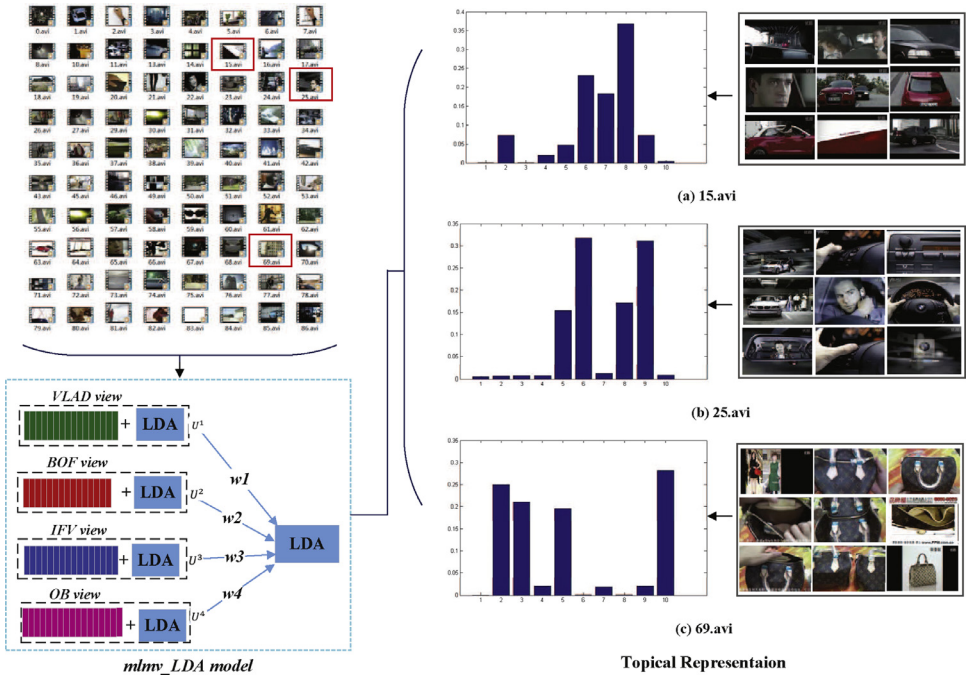
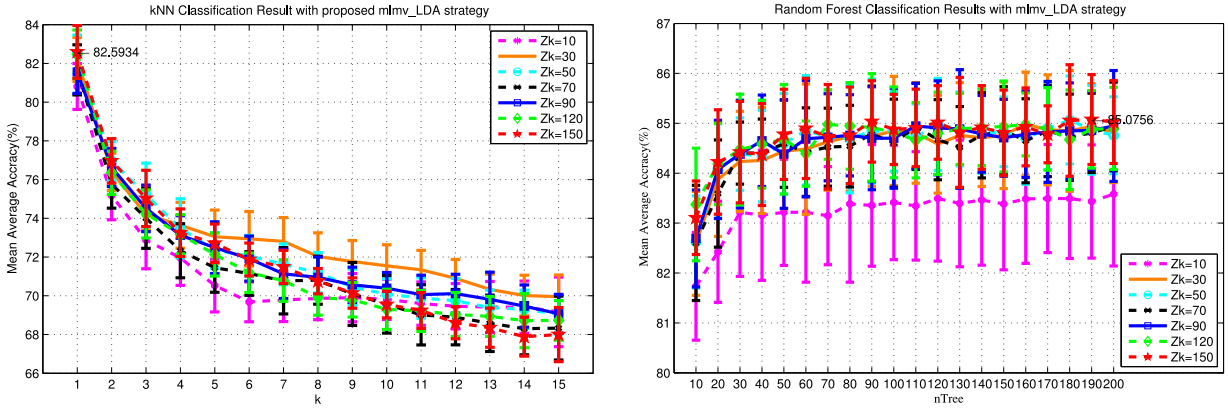


Fig. 7. Representation ad videos in latent semantic space employing *mlmv*_LDA.



(a) Results on *k*NN classifier

(b) Results on *random forest* classifier

Fig. 8. Classification results with *mlmv_LDA* strategy when Z_k differs.

Table 8
The best results for these two classifier s.

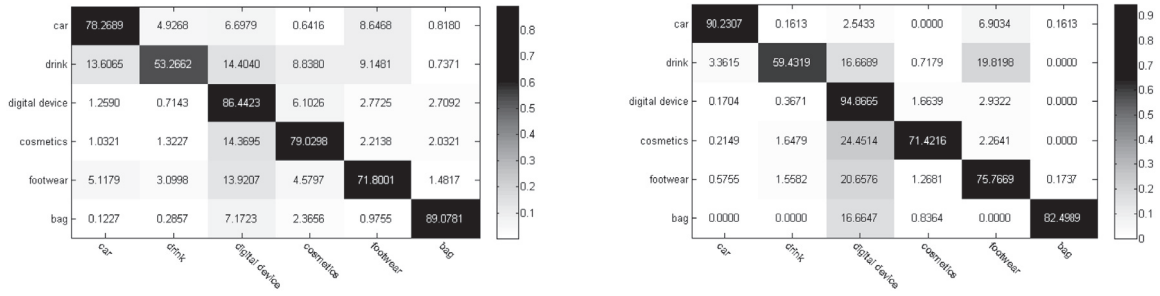
The best result(%)	*k*NN	*Random forest*
VLAD_LDA	76.83	**80.11**
BOF_LDA	77.75	**80.60**
IFV_LDA	77.67	**81.16**
OB_LDA	79.06	**82.16**
MVV	81.12	**82.22**
FM	65.88	**70.00**
MCCA	40.42	**81.61**
mlmv_LDA	82.59	**85.08**

The performance of the *random forest* classifier is significantly superior to the *k*NN classifier with regard to each strategy. This is due to the *random forest* acting as a typical ensemble classifier, can make full use of the samples, benefit from its sampling method, and integrate the decision ability from multiple decision trees.

4.5. Computational cost

This section reported the results from a computational cost perspective. Results for CNN [52] method were implemented in CNN architecture on a Linux workstation NVIDIA GPU with 640 CUDA cores and 4 GB of video memory, using a Caffe [54] framework. For the remaining methods, the code was written in MatLab and run on an Intel i7 core with 32 GB of RAM. Table 9 lists the details of the feature learning time, classification(CLS) time with a *k*NN classifier, and the classification(CLS) times with a *random forest* classifier for all methods.

From the results we conclude the following: (1) the classification efficiency of the proposed method outperforms the other baselines with the *random forest* classifier. (2) when performing *k*NN classification tasks, the proposed method still demonstrates its comparative superiority compared to most baseline methods. (3) with respect to time costs in the process of training the model (i.e., feature learning), the proposed *mlmv*_LDA model takes more time than the mono-view methods and FM. This is reasonable be-



(a) kNN results with Z_k=70 and k=1

(b) *random forest* results with Z_k=150 and $nTree$=19

Fig. 9. Confusion matrix of *mlmv*_LDA.

(a) Part of keyframes from an ad video promoted by "Audi" car

(b) Part of frames from a sport video

Fig. 10. Visual difference between ad video and sports video.

Table 9
The time cost comparison between *mlmv*_LDA and other baselines.

(Sec)	*Feature learning time*	*kNN CLS time*	*Random forest CLS time*
BOF_LDA	238.92	135.78	277.16
VLAD_LDA	572.72	494.94	308.09
IFV_LDA	404.37	206.03	126.93
OB_LDA	917.68	69.66	288.48
MVV	-	435.52	1160.68
FM	289.96	135.80	403.78
MCCA	1260.00	23495.57	7473.10
CNN	65400.00	12631.54	2522.00
mlmv_LDA	1250.64	109.25	126.31

cause the proposed *mlmv*_LDA method is a two-layer topic model. However, its performance is superior to the other two popular baseline methods(i.e., CNN and MCCA).

Overall, the proposed *mlmv*_LDA is an optimal option for ad video representation when both accuracy and time overhead are taken into account.

4.6. A self-comparison of proposed mlmv_LDA

In this section, we first report the performance trend of the proposed *mlmv_LDA* when Z_k differs from 10 to 150 on *k*NN and *random forest* classifier in Fig. 8.

In evaluating *k*NN classification tasks, we find from Fig. 8(a) that: (1) the performance of *mlmv_LDA* decreases with an increase in *k*. (2) increasing the number of Z_k results in a correspondingly improves the performance when *k* is smaller than 3. (3) the best Z_k drops at 30 when *k* varies from 4, which implies that the classification performance is not as sensitive to Z_k as except. However, with respect to *random forest* classification task, we can draw the conclusions from Fig. 8(b) that: (1) the performance of *mlmv_LDA* tends to be better when *nTree* increases; and (2) the classification performance becomes insensitive to Z_k when it varies from 30.

Then, the confusion matrices obtained with the proposed *mlmv_LDA* model on *k*NN classifier and *random forest* classifier are reported at their best performance points in Fig. 9. As shown, there is a clear separation in nearly all categories. Most confusion in *k*NN classification tasks occur between *cosmetics* and *digital devices*, and between *drinks* and *digital devices*. We observe that similar confu-

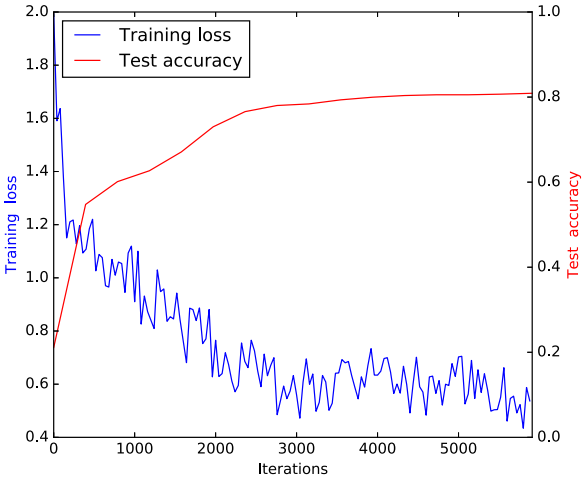


Fig. 11. Learning rate with iterations by CNN on UCF101. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

sion appears in *random forest* classification tasks between *footwear* and *digital devices*.

4.7. Discussion about proposed mlmv_LDA

The results of our analysis on the ad video data set indicate that the proposed model can learn powerful video representations. The proposed method is suitable for videos that have characteristics similar to ad videos (e.g., with large diversity in terms of scenes and objects), like a miniature movie with a specific theme. However, we cannot ensure it also works well for videos with static scenes and constant objects, like sports video. In order to illustrate this issue, we chose a markedly different non-ad benchmark data set to conduct similar experiments.

UCF101 [55] consists of 13,320 videos belonging to 101 categories which are divided into five groups: human-object interaction (HOI), body-motion (BM), human-human interaction (HHI), playing instruments (PI), and sports. Table 10 presents the number details of each category. The "#categories" means the number of categories in each group and the "#videos" is the number of videos belonging to every video group.

Unlike time-varying scenes which often occur in ad video, UCF101 has almost the same visual information in each frame of each video. An illustrative example is shown in Fig. 10. Fig. 10(a) shows part frames of an ad video and Fig. 10 (b) shows part frames of a sports video from UCF101.

Table 10
Number details of UCF101 data set.

Video group	*HOI*	*BM*	*HHI*	*PI*	*Sports*	*Totals*
#Categories	20	16	5	10	50	**101**
#Videos	2619	1910	690	1428	6673	**13,320**

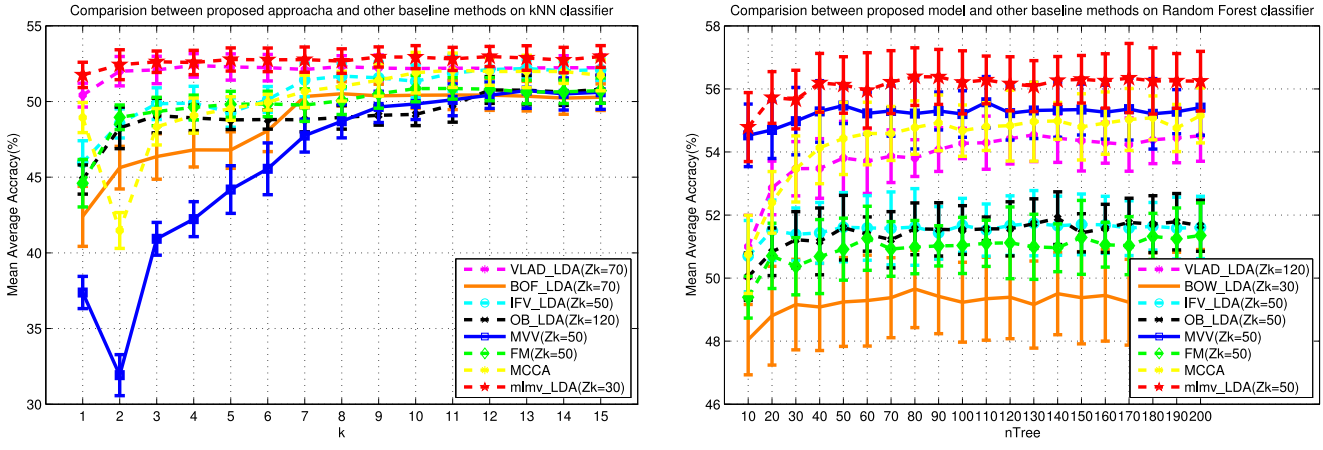
We first report results using CNN [52]. Fig. 11 shows the learning rate curves for test accuracy and training loss on UCF101 data set. The blue curve indicates the training loss rate, and the red curve shows the test accuracy. The algorithm reaches convergence after only 6000 iterations compared with 60,000 iterations for the ad video data set.

Similarly, we conducted the *k*NN and *random forest* classification tasks, with respect to BOF_LDA, VLAD_LDA, OB_LDA, IFV_LDA, MVV, FM, MCCA [53] and *mlmv_LDA*. For these LDA-based approaches, including the proposed *mlmv*_LDA, we conducted the experiments under multiple parameter combinations by varying the values of Z_k, k and *nTree*, and chose the one with the best performance for each method. The comparison result is provided in Fig. 12

Though the proposed *mlmv*_LDA model performs much better than the most baseline methods on UCF101 data set, classification performance was far from being satisfactory. One way to explain this result is that UCF101 lacks sufficient diversity in scenes/objects. This is exactly treated as the discriminative information in vision-based models, including our *mlmv*_LDA model and other baselines.

5. Conclusion and future work

Mining High-level semantic concepts in video classification is a challenging problem due to possible object variations. In this paper, we present a novel video representation model, called *mlmv_LDA*, which not only fuses several different representations from a multi-view, but also integrates topical theme information. The proposed model effectively avoids the lopsidedness of single-view and simultaneously supports ad categorization in latent semantic space. Experiments on real-world data set using two typical classifiers, *k*NN and *random forest*, demonstrate the proposed representation approach achieves better classification performance than other mono-view representations (i.e., VLAD_LDA, BOF_LDA, IFV_LDA and OB_LDA), MCCA, CNN, majority voting view (MVV)



(a) *k*NN classification results (b) *random forest* classification results

Fig. 12. Classification performance comparison of various strategies on UCF101.

and feature_merged (FM) strategies, even other LDA-based baseline methods that also consider topical information.

We also provide a discussion on the choice of classifiers given the analysis of our experimental results.

We conclude that the proposed *mlmv*_LDA is an effective and efficient tool for classifying videos with largest diversity in scenes and objects.

In future work, we hope to incorporate broader media information, such as motion and sound, to obtain more powerful and generic video representations.

Acknowledgment

S. Hou would like to thank Maoying Qiao for helpful discussions about the proposed model, Jianwei Lin for help in designing the CNN architecture in the revision. We are also grateful to Jia Wu for suggestions about the revision. This work was made possible through support from the major project of Natural Science Foundation of Shandong Province (ZR2016FQ20), Youth Cultivation Foundation of Shandong Normal University, Fundamental Science and Frontier Technology Research of Chongqing CSTC (cstc2015jcyjBX0124), Natural Science Foundation of China (NSFC) (61572300), Natural Science Foundation of Shandong Province in China (ZR2014FM001) and Taishan Scholar Program of Shandong Province in China (TSHW201502038).

References

[1] http://www.199it.com/archives/289664.html.
[2] http://www.iab.com/insights/iab-internet-advertising-revenue-report-conducted-by-pricewaterhousecoopers-pwc-2/.
[3] F. Cricri, M.J. Roininen, J. Leppänen, S. Mate, I.D. Curcio, S. Uhlmann, M. Gabbouj, Sport type classification of mobile videos, IEEE Trans. Multimed. 16 (4) (2014) 917–932.
[4] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, P. Oittinen, Content-based prediction of movie style, aesthetics, and affect: data set and baseline experiments, IEEE Trans. Multimed. 16 (8) (2014) 2085–2098.
[5] S.U. Maheswari, R. Ramakrishnan, Sports video classification using multi scale framework and nearest neighbor classifier, Indian J. Sci. Technol. 8 (6) (2015) 529–535.
[6] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[7] M. Rouvier, S. Oger, G. Linarès, D. Matrouf, B. Merialdo, Y. Li, Audio-based video genre identification, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (6) (2015) 1031–1041.
[8] M. Hayat, M. Bennamoun, S. An, Deep reconstruction models for image set classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 713–727.
[9] D. Tao, L. Jin, Y. Wang, X. Li, Person reidentification by minimum classification error-based KISS metric learning, IEEE Trans. Cybern. 45 (2) (2015) 242–252.
[10] A.A. Hamed, R. Li, Z. Xiaoming, C. Xu, Video genre classification using weighted kernel logistic regression, Adv. Multimed. 2013 (2013) 2.
[11] C. Colombo, A. Del Bimbo, P. Pala, Retrieval of commercials by video semantics, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1998, pp. 572–577.
[12] D. Brezeale, D.J. Cook, Automatic video classification: a survey of the literature, IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 38 (3) (2008) 416–430.
[13] L.-J. Li, H. Su, L. Fei-Fei, E.P. Xing, Object bank: a high-level image representation for scene classification & semantic feature sparsification, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.
[14] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: Proceedings of the European Conference on Computer Vision, Springer, 2010, pp. 143–156.
[15] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: Proceedings of the European Conference on Computer Vision, Springer, 2006, pp. 490–503.
[16] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 3304–3311.
[17] X. Cao, X. Wei, Y. Han, X. Chen, An object-level high-order contextual descriptor based on semantic, spatial, and scale cues, IEEE Trans. Cybern. 45 (7) (2015) 1327–1339.
[18] J. Yi, Y. Peng, J. Xiao, Exploiting semantic and visual context for effective video annotation, IEEE Trans. Multimed. 15 (6) (2013) 1400–1414.
[19] M. Wang, W. Li, D. Liu, B. Ni, J. Shen, S. Yan, Facilitating image search with a scalable and compact semantic mapping, IEEE Trans. Cybern. 45 (8) (2015) 1561–1574.
[20] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, Neural Comput. Appl. 23 (7–8) (2013) 2031–2038.
[21] Y. Luo, T. Liu, D. Tao, C. Xu, Multiview matrix completion for multilabel image classification, IEEE Trans. Image Process. 24 (8) (2015) 2355–2368.
[22] M. Liu, Y. Luo, D. Tao, C. Xu, Y. Wen, Low-rank multi-view learning in matrix completion for multi-label image classification., in: Proceedings of the AAAI Conference on Artificial Intelligence, 2015, pp. 2778–2784.
[23] J. Yu, D. Tao, Y. Rui, J. Cheng, Pairwise constraints based multiview features fusion for scene classification, Pattern Recognit. 46 (2) (2013) 483–496.
[24] T. Mei, X.-S. Hua, Contextual internet multimedia advertising, Proc. IEEE 98 (8) (2010) 1416–1433.
[25] A. Joshi, R. Motwani, Keyword generation for search engine advertising, in: Proceedings of the Sixth IEEE International Conference on Data Mining-Workshops (ICDMW), IEEE, 2006, pp. 490–496.
[26] W.-t. Yih, J. Goodman, V.R. Carvalho, Finding advertising keywords on web pages, in: Proceedings of the 15th International Conference on World Wide Web, ACM, 2006, pp. 213–222.
[27] J. Liu, C. Wang, W. Yao, Keyword extraction for contextual advertising, Chin. Commun. 7 (4) (2010) 51–57.
[28] A. Broder, M. Fontoura, V. Josifovski, L. Riedel, A semantic approach to contextual advertising, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2007, pp. 559–566.
[29] T. Mei, X.-S. Hua, S. Li, Contextual in-image advertising, in: Proceedings of the 16th ACM International Conference on Multimedia, ACM, 2008, pp. 439–448.
[30] Z. Li, L. Zhang, W.-Y. Ma, Delivering online advertisements inside images, in: Proceedings of the 16th ACM International Conference on Multimedia, ACM, 2008, pp. 1051–1060.
[31] T. Mei, X.-S. Hua, S. Li, Videosense: a contextual in-video advertising system, IEEE Trans. Circ. Syst. Video Technol. 19 (12) (2009) 1866–1879.
[32] G. Kastidou, R. Cohen, An Approach for Delivering Personalized Advertisements in Interactive TV Customized to Both Users and Advertisers, in: Interactive Digital Television: Technologies and Applications, 2008, pp. 52–73.
[33] W.-S. Liao, K.-T. Chen, W.H. Hsu, Adimage: video advertising by image matching and ad scheduling optimization, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 767–768.
[34] S.H. Sengamedu, N. Sawant, S. Wadhwa, vadeo: video advertising system, in: Proceedings of the 15th ACM International Conference on Multimedia, ACM, 2007, pp. 455–456.
[35] K. Yadati, H. Katti, M. Kankanhalli, Cavva: computational affective video-in-video advertising, IEEE Trans. Multimed. 16 (1) (2014) 15–23.
[36] Y. Li, K.W. Wan, X. Yan, C. Xu, Real time advertisement insertion in baseball video based on advertisement effect, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, ACM, 2005, pp. 343–346.
[37] J. Beaton, System and Method for Overlay Advertising and Purchasing Utilizing On-line Video or Streaming Media, 2016. US Patent 9336528.
[38] T. Mei, X.-S. Hua, S. Li, J. Guo, Intelligent Overlay for Video Advertising, 2013. US Patent 8369686.
[39] J. Guo, T. Mei, F. Liu, X.-S. Hua, AdOn: an intelligent overlay video advertising system, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2009, pp. 628–629.
[40] S. Hou, S. Zhou, L. Chen, Y. Feng, K. Awudu, Multi-label learning with label relevance in advertising video, Neurocomputing 171 (2016) 932–948.
[41] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2, IEEE, 2006, pp. 2169–2178.
[42] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1704–1716.
[43] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[44] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[45] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[46] C. Chemudugunta, P. Smyth, M. Steyvers, Modeling general and specific aspects of documents with a probabilistic topic model, in: Proceedings of the Conference and Workshop on Neural Information Processing Systems(NIPS), 19, 2006, pp. 241–248.
[47] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[48] D. Inouye, P.D. Ravikumar, I.S. Dhillon, Admixture of poisson MRFs: a topic model with word dependencies., in: Proceedings of the International Conference on Machine Learning, 2014, pp. 683–691.
[49] W. Ou, Z. Xie, Z. Lv, Spatially regularized latent topic model for simultaneous object discovery and segmentation, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2015, pp. 2938–2943.
[50] R. Fernandez-Beltran, F. Pla, Latent topics-based relevance feedback for video retrieval, Pattern Recognit. 51 (2016) 72–84.
[51] Z. Wang, L. Li, Q. Huang, Cross-media topic detection with refined CNN based image-dominant topic model, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 1171–1174.

[52] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[53] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data, IEEE Trans. Image Process. 11 (3) (2002) 293–305.
[54] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[55] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, Comput. Sci. (2012).

Sujuan Hou received the Ph.D. degree from Chongqing University, Chongqing, China, in 2015. She is currently a lecture with Shandong Normal Univertisy. Prior to that, she was a visiting Ph.D. student with Faculty of Engineering and Information Technology (FEIT), University of Technology, Sydney, from 2013 to 2015. Her research interests include video data mining, multimedia retrieval and pattern recognition.

Ling Chen received her Ph.D. in 2008, in Computer Engineering, from Nanyang Technological University, Singapore. She is currently a Senior Lecturer with Faculty of Engineering and Information Technology (FEIT), University of Technology, Sydney. Prior to that, she was a Postdoc Research Fellow with L3S Research Center, Leibniz University Hannover, Germany. Ling's main research interests include data mining, machine learning, social media etc.

Dacheng Tao is Professor of Computer Science with the Centre for Quantum Computation andIntelligent Systems and the Faculty of Engineering and Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics for data analysis problems in computer vision, data mining, machine learning, multimedia, and video surveillance. He has authored and co-authored more than 100 scientific articles at top venues including IEEE T-PAMI, T-IP, T-NNLS, AISTATS, ICDM, CVPR, ECCV, ACM SIGKDD and Multimedia, with the best theory/algorithm paper runner up award in IEEE ICDM'07.

Shangbo Zhou received the B.Sc. degree from Guangxi National College in 1985, the M.Sc. degree from Sichuan University in 1991, both in Mathematics, and Ph.D. degree in Circuit and System from Electronic Science and Technology University. From 1991 to 2000, he was with the Chongqing Aerospace Electronic and Mechanical Technology Design Research Institute. Since 2003, he has been with the Department of Computer Science and Engineering of Chongqing University, where he is now a Professor. His current research interests include artificial neural networks, physical engineering simulation, visual object tracking and nonlinear dynamical system.

Wenjie Liu is an associate professor of Computer Sciences at Nanjing University of Information Science and Technology, China. He received his Bachelor (HZAU, China, 1997), Master (WHU, China, 2004), Ph.D. (SEU, 2011). His research interests include quantum machine learning, quantum secure multiparty computation, quantum secure communication, etc.

Yuanjie Zheng is currently a professor in the School of Information Science and Technology at Shandong Normal University and a Taishan Scholar of People's Government of Shandong Province of China. He is also acting as a vice-dean at both the School of Information Science and Technology and the Institute of Life Sciences of Shandong Normal University. He used to be a Senior Research Investigator in Perelman School of Medicine of University of Pennsylvania. His research is in the fields of medical image analysis, translational medicine, computer vision and computational photography. His ultimate research goal is to enhance patient care by creating algorithms for automatically quantifying and generalizing the information latent in various medical images for tasks such as disease analysis and surgical planning, through the applications of computer vision and machine learning approaches to medical image analysis tasks and development of strategies for image-guided intervention/surgery.