9

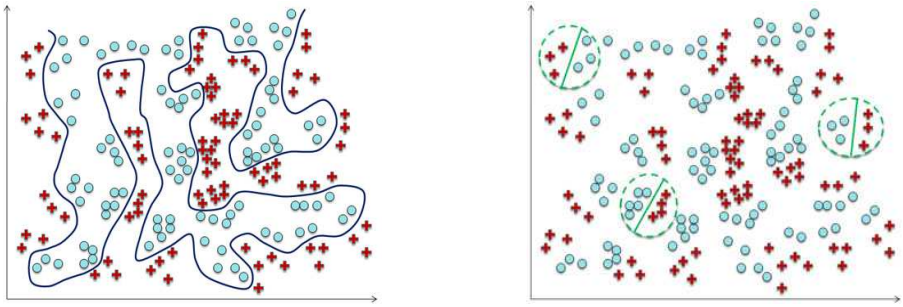
Ensemble Learning with LDA Topic Models for Visual Concept Detection

Sheng Tang[1], Yan-Tao Zheng[2], Gang Cao[3],
Yong-Dong Zhang[1] and Jin-Tao Li[1]
[1]*Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing,*
[2]*Google Inc, Mountain View, CA*
[3]*Beijing Software Testing & QA Center, Beijing,*
[1,3]*China*
[2]*USA*

1. Introduction

With the rapid growth of multimedia application technologies and network technologies, especially the proliferation of Web 2.0 and digital cameras, there has been an explosion of images and videos in the Internet. For example, the volume of videos uploaded to the YouTube every minute is amounting to 48 hours by May 2011, having doubled in the last two years. Such huge video collections hold useful yet implicit and nontrivial knowledge about various domains. To manage and utilize these resources effectively, video concept detection becomes a very important subject of intensive research by a large research community (Over et al., 2008). It is an integral part of visual data mining that is automatically extracting such knowledge from the huge unstructured visual data. It aims to automatically annotate video shots or keyframes with respect to a semantic concept (Tang et al., 2012). Ranging from objects like *airplane* and *car* to scenes like *urban street* and *sky*, semantic concepts serve as good intermediate semantic features for video content indexing and understanding, and thus, spurring much research attention (Jiang et al., 2010; Naphade & Smith, 2004; Snoek et al., 2006; Zheng et al., 2008). Essentially, concept detection is a classification task, in which a binary classifier is usually learned to predict the presence of a certain concept in a video shot or keyframe (image). Traditional concept detection methods are mainly global classification: use supervised machine learning techniques, such as single Support Vector Machine (SVM), etc., over whole training dataset.

Study on pedestrian classification (Munder & Gavrila, 2006) showed that the benefit of selecting the best combination of features and pattern classifiers was less pronounced than the gain obtained by increasing the training set, even though the base training set already involved many thousands of samples (Enzweiler & Gavrila, 2009). In other words, the data matters most (Enzweiler & Gavrila, 2009). For visual concept detection, this was also pointed out in (Huiskes et al., 2010), and made the authors simply use more data rather than design more intelligent classification algorithms and image representations since



(a) Single SVM (b) The proposed method

Fig. 1. Illustration of the differences of optimal separating hyperplanes between (a) and (b). For single SVM, each test instance will use the same holistic complex hyperplane (the blue curve), while for the proposed method, a test instance will trigger very fewer number of local classifiers, e.g. 3 green dashed circles in (b), with much simpler hyperplanes (green lines) to fire.

large-scale data can also directly benefit visual concept detection. Inspired by these studies, for multimedia data of high dimension and diversified patterns, it is necessary to construct large scale training dataset to reflect all sorts of patterns as much as possible. However, there exist some challenges for global classification methods trained on large scale dataset: huge intra-class variations, low training efficiency, and low testing efficiency resulting from complex classification hyperplane as shown in Figure 1(a). To address these difficulties, the focus of this work is to develop an ensemble learning method based on Latent Dirichlet Allocation (LDA) topic models for large scale concept detection.

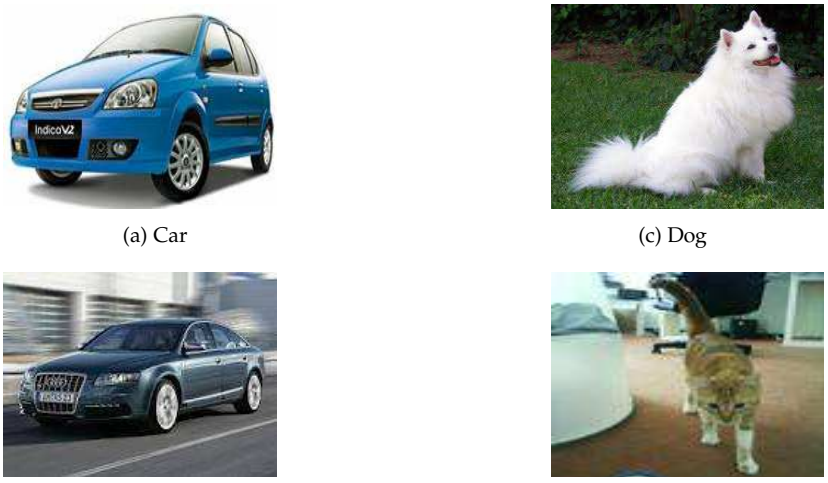
Ensemble learning refers to the process of combining multiple classifiers to provide a single and unified classification decision. Recent research have demonstrated, both theoretically (Krogh & Vedelsby, 1995) and empirically (Opitz & Shavlik, 1996a;b), that a good ensemble of localized classifiers can outperform a single (best) classifier learned over the entire dataset. Furthermore, learning a set of "smaller" localized classifiers usually possesses more efficient algorithmic complexity than a global classifier. Additionally, the former localized classifiers are generally more effective since their optimal separating hyper-planes may be much simpler to discriminate the data as illustrated in Figure 1(b), hence have better generalization performance than the latter due to the aforementioned problem of the huge intra-class variation. This motivates us to adopt an ensemble learning approach for concept detection.

There are, in general, two essential ingredients in a good ensemble classifier, which are: (1) the diversity of classifiers in the ensemble (Kuncheva & Whitaker, 2005), and (2) the fusion of classifiers (Opitz & Maclin, 1999; Zhang & Zhou, 2011). Diversity means that classifiers in the ensemble should possess different decision knowledge and make uncorrelated errors. In this way, the error of individual classifiers will not be the same and propagated to the ensemble, ensuring that individual classifiers have different "inductive biases", and thus, complement each other. The fusion of the classifier in the ensemble, on the other hand, is regarding how to coordinate individual classifiers for the final classification decision in a unified and theoretically principled fusion.

The most common way to achieve diversity is to train individual classifiers by using different training data. For example, the well known Bagging and Boosting (Freund & Schapire, 1997)

adopt this approach by randomly selecting (via re-sampling) different sets of training data for each individual classifier. Despite of simplicity, this approach ignores the intrinsic structure of data exemplars. To achieve classifier diversity, intuitively, similar data exemplars should be grouped together to train a localized classifier, as the simple subspace complexity usually leads to more effective localized classifier. The challenge here is how to group the data effectively. In this chapter, we investigate an instance grouping method via topic modeling.

Topic modeling is a newly emerging approach to analyze large volumes of unlabeled text (Griffiths & Steyvers, 2004). It specifies a statistical sampling technique to describe how words in documents are generated based on (a small set of) hidden topics. Particularly, we investigate the semantic grouping method through estimating the topical structure of large visual data under the framework of latent Dirichlet allocation (LDA)(Blei et al., 2003).



(a) Car (c) Dog

(b) Car (d) Cat

Fig. 2. Illustration of the insight from psychophysical studies that humans can perform coarse categorization between left column cars (a, b) and right column animals (c, d) quite easily and quickly, followed by successive finer but slower discrimination between different types of cars (a) and (b), or between different kinds of animals (c) dog and (d) cat.

As shown in Figure 2, our proposed solution is motivated by the insight from psychophysical studies that humans can perform coarse categorization of visual objects quite easily and quickly, followed by successive finer but slower discrimination (Kuncheva, 2004; S. Thorpe & Marlot., 1996). Specifically, since all the pictures of the same category often have some local parts in common, such as all the cars have common wheels and animals have common eyes and legs as shown in the Figure 2, we propose an ensemble learning with LDA topic models due to their great advantages in exploiting the co-occurrences of local features or visual words to discover intrinsic common or similar structures of data. The proposed ensemble learning can scale up to large data sets through combination of unsupervised semantic grouping and supervised learning. First, we use generative LDA model (Blei et al., 2003) to mine the hidden topical structure of large visual data, and then perform coarse categorization by grouping the large-scale training data set with huge variations into many diversified small topic localities. Second, we perform the successive finer discrimination by training each topic

locality to generate multiple small effective classifiers. Thereby, the data exemplars within one topic are deemed to be similar in part with respect to the hidden topic structure. The corresponding individual classifier then holds the decision knowledge mostly with respect to the topic. This ensures that the individual classifiers have reasonable diversity in varying regions of expertise. More importantly, for the fusion of classifiers, we propose to utilize the topic mixture coefficients in a generative probabilistic manner. For a given test sample, we adaptively select the most probable classifiers with large topic mixture coefficients for detection, which ensures that a sample is projected to only a few topic with top ranked non-zero coefficients. The resulting ensemble model is, therefore, sparse, in the way that only a small number of classifiers in the ensemble will fire on a testing sample as illustrated in Figure 1(b). Consequently, the efficiencies of both training and testing resulting can be greatly improved.

In summary, the main contribution of this chapter is that we propose a novel ensemble learning method for video concept detection by LDA topic modeling. Our preliminary results on the TREC Video Retrieval Evaluation (TRECVid) benchmark can be found in (Tang et al., 2008), and preliminary results on pornography detection for online videos can be found in (Tang et al., 2009). This chapter is an extension of both conference papers, and more detailed results of extensive tests on the TRECVid 08 benchmark and pornography detection will be provided to show that the proposed approach achieves promising results and outperforms existing approaches.

In the rest of the chapter, we first review the related work on concept detection, ensemble learning and LDA topic models in Section 2. Then, we elaborate on the details of the proposed ensemble learning algorithm in Section 3, which includes ensemble construction with LDA topic models and coordination of individual classifiers. Two systems based on the proposed ensemble learning algorithm, TRECVid concept detection system and online pornography filtering system are introduced in Section 4, and experimental results of the two systems are also given in Section 4. Finally, we present the conclusive remarks along with discussion for future work in Section 5.

2. Related work

2.1 Concept detection

Concept detection is a challenging yet useful task that has attracted attentions of many researchers. Early work on concept detection focuses on concept-specific handcrafted rules for tailor-made solution (A. Vailaya & Zhang, 1998; Smith & Chang, 1997; Szummer & Picard, 1998; Zhang et al., 1995). Distilling these rules automatically, machine learning approaches have then become the research focus, wherein a variety of classification techniques are explored. Majority of existing machine learning approaches are generally composed of five major steps (Jiang, 2009) as shown in Fig.3:

- **Preprocessing**: A video consists of a sequence of shots separated by shot boundaries including cuts and gradual transitions. Since video shots are often the basic unit for concept detection, videos are segmented into shots based on various shot boundary detection methods which uses different key frame features such as color histogram in (Yuan et al., 2007), SIFT in (Chang, Lee, Hong & Archibald, 2008), and similarity measurements. We refer readers to (Smeaton et al., 2010; Yuan et al., 2007) for a recent review on the subject. After shot boundary detection, either the middle I-frame or a set of

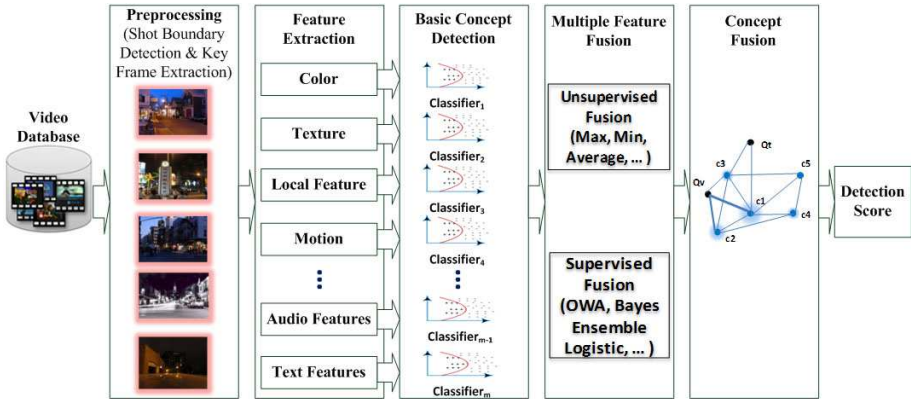


Fig. 3. General procedure of concept detection

key frames may be selected from each shot to represent the content for further detection (Borth et al., 2008).

- **Low-level feature extraction**: The purpose of feature extraction is to convert the shots or keyframes into a low dimensional feature vectors. Generally, two types of visual keyframe features are often used: global and local. Global features include (Chang, He, Jiang, Khoury, Ngo, Yanagawa & Zavesky, 2008; Ngo et al., 2009): color or edge histograms, correlograms, grid-based color moment and wavelet texture, histogram of oriented gradients (HOG) (Dalal & Triggs, 2005), local binary patterns (LBP) (Ojala et al., 1996), GIST (Siagian & Itti, 2007) and Gabor feature (Zhu et al., 2008), etc. Bag of Visual Words (BoVW) is the most widely adopted local feature which is based on a vocabulary of visual words clustered by a set of SIFT features (Lowe, 2004), and weighted by various schemes such as the traditional TF, TF-IDF, and the soft-weighting scheme which has been demonstrated to be more effective than the traditional ones (Jiang et al., 2010). Audio features such as in (Tang et al., 2007) are less frequently used while spatial temporal features including motion features, such as Space-Time Interest Points (STIP) (Laptev, 2005), are coming into use such as in (Jiang & *et al*., 2010) despite their expensive extraction.
- **Basic concept detection**: namely, uni-modality learning with a variety of classification techniques such as SVM (Cao et al., 2006), Gaussian Mixture Model (Amir et al., 2003), Hidden Markov Model (Pytlik et al., 2005), graph-based semi-supervised learning (Tang et al., 2010; 2011; Wang, Hua, Hong, Tang, Qi & Song, 2009), etc. The choice of different kernel functions for classification of BoW features was studied in (Jiang et al., 2010), and a novel neighborhood similarity measure beyond traditional distance measurement was proposed to explore the local sample and label distributions for concept detection in (Wang, Hua, Tang & Hong, 2009) recently.
- **Multiple feature fusion**: including both early fusion (i.e., vector concatenation) of multiple features, and late fusion of multiple classifiers (Snoek et al., 2005). Late fusion is widely used for efficiency and accuracy, and there are two common approaches to calculate the weight for each classifier: unsupervised, such as min, max, and average in (Amir et al., 2003), kernel fusion in (S. Ayache & Gensel, 2007), fusion with membership vector such as LDA topic mixture coefficients in (Tang et al., 2008), etc., and supervised, such as ordered weight averaging in (Tang et al., 2007).

- **Concept fusion**: Besides the above multi-modality fusion methods based on multiple features, contextual fusion techniques (Hauptmann et al., 2007; Jiang, 2009; Qi et al., 2007; Weng & Chuang, 2008) are emerging by exploiting inter-concept relationships, such as taking into account the detection scores of nearby objects in the scene, to improve the accuracy of detection.

The major challenge of concept detection lies primarily in the existence of the well-known semantic gap (Smeulders et al., 2000) between the low level visual features and the users's semantic interpretation of visual data, and the fact that different video shots w.r.t a certain concept often possess huge variations among different visual appearances, camera shooting and video editing styles, etc. This diversity renders video shots of the same semantic concept to have varying visual patterns. Therefore, the resultant huge intra-class variation hinders the performance of most machine learning approaches. Domain change caused by the mismatch between different domains (genres or sources (Borth et al., 2010)) may worsen this problem, which raises domain adaptation in concept detection, known as cross-domain learning techniques (Jiang et al., 2008; Ngo et al., 2009; Snoek et al., 2010), including our recent concept detection work on pseudo relevance feedback based domain transfer learning (Xu, Tang, Zhang & Li, 2011) and multi-modality transfer based on multi-graph optimization (Xu, Tang, Zhang, Li & Zheng, 2011), for transferring detectors trained in source domain to the target domain. For a comprehensive review on concept detection, refer to (Jiang, 2009; Snoek & Worring, 2009), and the high level feature extraction (or semantic indexing since the year 2010) task of TRECVid (Smeaton et al., 2006; 2009) as well as its workshop papers (NIST, 2001-2010) since TRECVid provides a large video data collection, uniform evaluation criteria, a workshop for active participants to discuss their approaches, and hence can be widely regarded as the actual standard for performance evaluation of concept based video retrieval systems (Snoek & Worring, 2009).

This chapter is an extension of our previous work (Tang et al., 2008), which attempts the concept detection in the framework of ensemble learning. In the proposed scheme, the individual classifiers and their fusion weights are learned in a unified framework without any additional classifier selection module.

2.2 Ensemble learning

Ensemble learning (Kuncheva, 2004; Rokach, 2010) coordinates the outputs of multiple classifiers using diversified data to improve the performance. Empirically, ensemble methods tend to yield better results when there is a significant diversity among the constituent classifiers (Kuncheva & Whitaker, 2005). The diversity of classifier outputs plays a critical role on the success of ensemble learning. Existing methods of constructing ensembles include Bayesian voting, manipulation on the training examples and input features, etc (Dietterich, 2000). Correspondingly, for concept detection in particular, Snoek and Worring (Snoek & Worring, 2009) identified three common approaches to achieve some form of independence (diversity) for coordination, which are using (1) separate features, (2) separate classifiers, and (3) separate set of labeled examples.

This work focuses on the manipulation on the training examples for classifier diversity, which can be generally classified into two schemes: (1) separate sampling of labeled instances, and (2) instance space partitioning. As for separate sampling of labeled examples, two most widely used strategies are Bagging and Boosting:

- **Bootstrap aggregating (Bagging):** Bagging (Breiman, 1996) aims at developing independent classifiers, and the diversity necessary to make the ensemble work is created by taking bootstrap replicates as the training sets. The samples are pseudo-independent because they are sampled with replacement from the same development set.
- **Boosting and AdaBoost**: Boosting (Freund & Schapire, 1997) is a general method for improving the performance of a weak classifier. Similar to bagging, Boosting develops the classifiers by resampling the training set, while contrary to bagging, the resampling mechanism in boosting focus on most useful sample in each consecutive iteration (Rokach, 2010). AdaBoost (Adaptive Boosting) (Y. & E., 1996) is a popular ensemble algorithm that improves the simple boosting algorithm via an iterative process. The main idea is to give more focus to patterns that are harder to classify (Rokach, 2010).

Parallel to the above sampling-based partitioning approaches, many space-based partition approaches have been developed for partitioning the training set into subsets according to their belonging to some part of the input space (Rokach, 2010). Particularly, inspired by the idea that similar instances should be assigned to the same subspace, researchers attempt to use some clustering method as a possible tool for partitioning the instance space recently. Lior Rokach proposed the naive decomposition method based on K-Means algorithm (Rokach, 2010). SVM-KNN was proposed in (Zhang et al., 2006) to train an SVM improvisedly by using the K nearest neighbors of the test sample, but for large-scale dataset such as TRECVid, it is too time-consuming to search for the K nearest neighbors and train an SVM for each test sample. Furthermore, it is evident that if a test image is only partially similar with the expected training images, the latter may not fall within the range of the K nearest neighbors if K is small, which turns in vain the subsequent training and testing. Recently, we proposed a localized multiple kernel learning method for realistic human action recognition based on multiple features (Song et al., 2011), and sparse ensemble learning for visual concept detection (Tang et al., 2012) by exploiting a sparse non-negative matrix factorization process to for ensemble construction and fusion.

In this chapter, we propose a novel space-based partitioning scheme by exploiting Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to partition the instance space into small topic subspaces (Tang et al., 2008).

2.3 LDA topic models

2.3.1 Vector space modeling

The first major progress in text processing was due to the vector space modeling (VSM) (Salton & McGill, 1986), in which "bag of words (BoW)" has been adopted to represent a document as a vector of frequency histogram where each dimension is associated with one term of the vocabulary and each entry is weighted by the term frequency (TF) or term frequency-inverse document frequency (TF-IDF) to reduce the importance of indiscriminant words that appear in many documents. Thereby, the whole corpus is represented as term-document matrix whose rows are indexed by the terms of the vocabulary and whose columns are indexed by the documents.

2.3.2 Latent Semantic Analysis

To address the inherent drawbacks of the VSM, such as the difficulty of capturing inter- and intra document statistical structure and the incompact description of the corpus (Alsumait

et al., 2010), Latent Semantic Analysis (LSA) (Deerwester et al., 1990) has been introduced to reduce the term-document matrix through singular value decomposition (SVD). However, the computation of the SVD is expensive, and the reduced feature space is very difficult to interpret (Alsumait et al., 2010).

2.3.3 Probabilistic Latent Semantic Analysis

To better understand LSA statistically, probabilistic Latent Semantic Analysis (pLSA) was proposed (Hofmann, 1999) as an alternative to LSA by applying Bayesian methods to document modeling. The pLSA model is a generative model which uses a probabilistic sampling process to generate words in documents based on the latent topics. It associates the documents d with a mixture of latent topics z weighted by the posterior $p(z|d)$, and represents each topic by a distribution over words w that appear in it $p(w|z)$. The graphical model of pLSA is shown in Fig.4(a). As shown by the figure, the joint probability of a document d and a word w_{di} can be given as:

$$p(d, w_{di}) = p(d)p(w_{di}|d) \tag{1}$$

Given that the observation pairs (d, w_{di}) are assumed to be generated independently, the conditional probability $p(w_{di}|d)$ can be computed by marginalizing over topics z_k. Therefore, the joint probability $p(d, w_{di})$ can be computed as:

$$p(d, w_{di}) = p(d) \sum_{k=1}^{K} p(z_k|d)p(w_{di}|z_k) \tag{2}$$

where K is the total number of latent topics, $p(z_k|d)$ is the probability of topic z_k occurring in document d, and $p(w_{di}|z_k)$ is the probability of word w_{di} occurring in a particular topic $p(z_k)$.

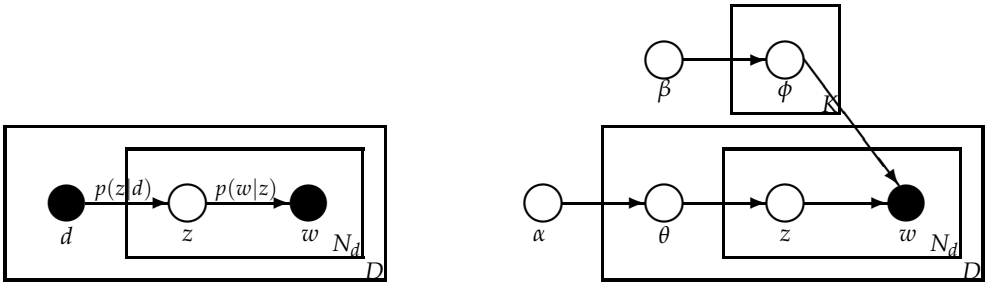


Fig. 4. A graphical model of pLSA (a) and LDA (b). Nodes are random variables. Darked ones are observed and other ones are unobserved. The plates indicate repetitions.

Actually, the pLSA is non-parametric pseudo-generative model since the document d is a dummy random variable indexed by the by the documents in a training set (Alsumait et al., 2010; Blei et al., 2003), and there is no natural way to use it to assign probability to a new testing observation (Li & Perona, 2005). Additionally, the model parameters grows linearly with the number of training examples (Li & Perona, 2005). Consequently, pLSA has the limitation of overfitting, hence poor generability for unseen documents (Blei et al., 2003). Despite its limitation, pLSA has invoked a huge amount of work in statistical machine learning and text

mining, which resulting in a class of probabilistic topic models aiming at discovering these hidden variables based on hierarchical Bayesian analysis (Alsumait et al., 2010).

2.3.4 LDA

LDA (Blei et al., 2003) is a three-level hierarchical Bayesian network, a truely generative probabilistic model for a corpus of documents. The basic idea of LDA is that documents are represented by a mixture of topics where each topic is a latent multinomial variable characterized by a multinomial distribution over a fixed vocabulary of words (Alsumait et al., 2010). A graphical model of LDA is shown in Fig.4(b). As shown by the figure, through introducing Dirichlet priors α on the document distributions over topics and β on the topic distributions over words, the generative model of LDA is complete and is capable of generalizing the topic distributions for generating unseen documents (Alsumait et al., 2010).

The generative process of the LDA is described as follows (Blei et al., 2003):

1. Generate K topic multinomials ϕ_k over a fixed vocabulary of words from a Dirichlet prior $Dir(\phi_k|\beta)$ given by β; $(p(\phi_k|\beta))$.
2. Generate D document multinomials θ_d over K topics from a Dirichlet prior $Dir(\theta_d|\alpha)$ given by α; $(p(\theta_d|\alpha))$.
3. For each document d in the corpus, and for each word w_{di} in the document d:
   (a) Sample a topic z_{di} from the document multinomial θ_d; $(p(z_{di}|\theta_d))$.
   (b) Sample a word w_{di} from the topic multinomial ϕ_z; $(p(w_{di}|\phi_{z_{di}}))$.

The joint distributions of the LDA model is:

$$p(\{w_{di}\},\{z_{di}\},\{\phi_k\},\{\theta_d\}|\alpha,\beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p(w_{di}|\phi_{z_{di}}) \quad (3)$$

where α and β are hyperparameters of Dirichlet priors, and ϕ_k , θ_d and z_{di} are hidden variables to be inferred.

2.3.5 LDA topic models in computer vision

Recently, inspired by its great success in finding useful structures in many kinds of documents in the field of text processing (Griffiths & Steyvers, 2004), LDA topic model, has been widely applied to computer vision problems such as object segmentation (Wang & Grimson, 2007), scene categorization (Li & Perona, 2005), action recognition (Niebles et al., 2008; Wang, 2011; Wang & Mori, 2009) and event detection(Pan & Mitra, 2011).

Under topic models, analogous to BoW in text processing, "Bag of visual words (BoVW)" (Jiang et al., 2007; Jurie & Triggs, 2005; Sivic & Zisserman, 2003; Zhang et al., 2007) is usually used to represent visual contents (such as key frames) as visual words. BoVW has first been introduced by Sivic in the case of video retrieval (Sivic & Zisserman, 2003) and became very popular in the fields of image retrieval and categorization due to its efficiency and effectiveness. After extraction of visual features, such as local features SIFT (Lowe, 2004) or SURF (Bay et al., 2006), BoVW consists of two main steps: visual vocabulary construction and feature quantization. Generally, various clustering methods are used to build the visual vocabulary by clustering features in to visual words (centroids) which are analogous to stems in text processing. Then, visual features are quantized into visual words and visual contents are represented as the frequencies of visual words. Topic models will compute latent concepts by exploring the co-occurrence of visual words to learn the models of different

patterns without manual annotation of training samples (Wang, 2011). Compared with other approaches, one of the major advantages of topic models is their unsupervised nature which is very important for discovering different patterns from large volumes of video data (Wang, 2011).

3. Algorithm

3.1 Preliminaries and problem formulation

In the task of concept detection, a video shot keyframe is processed to detect the presence of a set of predefined concepts. Let x denote the visual feature for the keyframe.

For each concept, we have a training set $X = \{x_i, i = 1, 2, ..., N\}$ with label $Y = \{y_i \in \pm 1, i = 1, 2, N\}$. The concept detection is thus naturally formulated as a classification task. Here, we adopt the binary classification in the framework of SVM (Vapnik, 1995) and aim to learn an ensemble discriminant function $F(x_t)$ for a test sample x_t

$$F(x_t) = \sum_{k=1}^{K} \Psi_k(x_t) \cdot (\langle \omega_k, \Phi(x_t) \rangle + b_k). \tag{4}$$

The discriminant function $F(x_t)$ is an ensemble of K localized classifiers that are built on instance localities π_k respectively, where $\Psi_k(x_t)$ are the gating functions that governs how localized classifiers $f_k(x_t)$ are coordinated for the final classification of test sample x_t.

Solving the primal SVM problem, we obtain $\omega = \sum_i \beta_i y_i \Phi(x_i)$. As plugging ω into Eq.(5), the ensemble discriminant function $F(x_t)$ becomes:

$$F(x_t) = \sum_{k=1}^{K} \Psi_k(x_t) \cdot \left(\sum_{i \in \pi_k} \beta_i y_i \langle \Phi(x_t), \Phi(x_i) \rangle + b_k \right). \tag{5}$$

Learning the ensemble discriminant function $F(x_t)$ can be decomposed into two steps: (1) computing the instance locality model and (2) estimating the localized kernel classifier parameters. The first step learns the instance localities π_k and gating function $\Psi_k(x_t)$.

In the next subsections, we describe the proposed data instance partitioning approach based on LDA topic model , and the coordination of individual classifiers based on LDA coefficients. Fig.5 shows the overall framework of the proposed ensemble learning method.

3.2 Ensemble construction with LDA topic models

We employ LDA to model the relationship between images to discover the hidden structures and perform coarse categorization for ensemble construction and fusion.

3.2.1 LDA topic modeling

To construct the ensemble, the first step is learn the instance locality π. Suppose we have a set of $D(d = 1, \ldots, D)$ keyframes (or shots) containing words from a vocabulary of size V. Each instance (a shot or keyframe) d is represented as a sequence of N_d visual words $w = (w_1, \ldots, w_{N_d})$. We set the number of latent topics K to the number of the above instance localities, i.e., the number of localized classifiers in the ensemble. Then, the LDA process that generates each instance d in the corpus is:

1. Choose the number of visual words N_d from $Poisson(\xi)$.

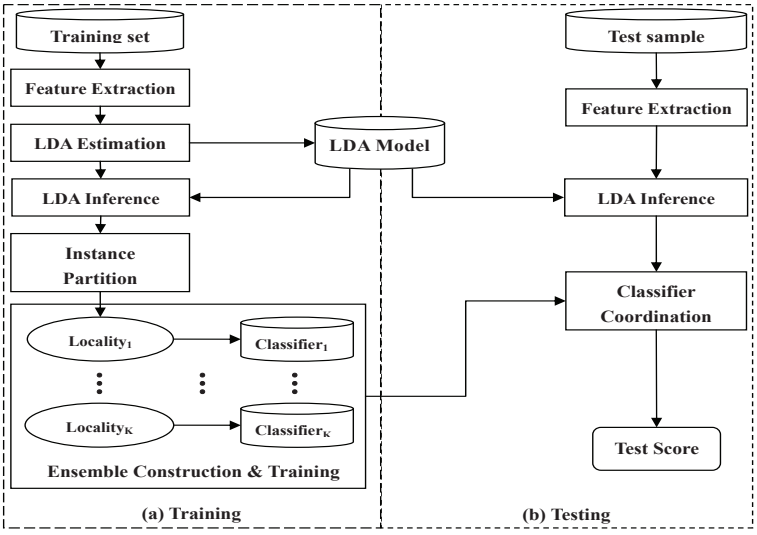


Fig. 5. The overall framework. The proposed ensemble first exploits LDA to represent data instances as a mixture of hidden topics and partition the data space into topic localities, and then coordinates the individual classifiers in each topic locality for final classification based on the topic mixture of LDA topics which is naturally achieved during LDA inference without additional classifier selection.

2. Choose the mixing proportions θ of the current instance d over K topics from $Dir(\alpha)$.
3. For each of the N_d visual words w_i:
   (a) Choose a topic z_i from $Multinomial(\theta)$.
   (b) Choose a visual word w_i from the multinomial distribution $p(w_i|z_i, \beta)$.

Here, the topic mixture θ is a multinomial distribution which is generated by K-dimensional Dirichlet distribution parameterized by the Dirichlet priors α. Additionally, the matrix β of size $K \times V$ is the parameter of the distribution of visual words conditioned on each topic locality, and each element of β corresponds to the probability $p(w_i|z_k)$.

The joint distributions of the LDA model is:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N_d} p(z_i|\theta) p(w_i|z_i, \beta) \tag{6}$$

where w is the set of words observed in the current instance, and z is their corresponding topic.

As mentioned in the previous section, under LDA topic models, BoVW feature (Sivic & Zisserman, 2003) (including other frequency-based features such as color histogram and edge histogram etc.) is usually used to represent keyframe. Therefore, the vocabulary size V is equal to the size M of the feature for LDA estimation. In some practice, $V = M + 1$ since a dummy word irrelevant to all other words should be included in the vocabulary. Additionally, in our proposed ensemble learning method, LDA is used only for partition

ensemble construction and individual classifier coordination. The features used for LDA can be different from the features for SVM.

In the proposed ensemble construction, we need to know how an instance is mixed over K hidden topic localities. So we must infer the posterior distribution of the hidden variables for a set of words w observed in the instance (shot or keyframe):

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{7}$$

where θ is specific to each instance and represents its latent topics distribution.

Generally, it is computationally intractable to perform above inference and parameter estimation for the LDA model. Up to now, two main approximation algorithms have been proposed to solve the problem: (1) variational inference Expectation-Maximization (EM) adopted in (Blei et al., 2003); and (2) gibbs sampling adopted in (Griffiths & Steyvers, 2004) which is easier to compute than the former method.

Once the topic mixture θ is inferred, we can know how topic localities are mixed in the current instance. There for we can exploit it to determine which localities the instance should be partitioned into as shown in the next subsection.

3.2.2 Adaptive instance-locality assignment

After LDA inference of the topic mixture θ, we can allocate the instance x_i to a few L localities, according to the top L large elements of $\theta = \{\theta_1, \ldots, \theta_K\}$. The greater the element θ_i is, the

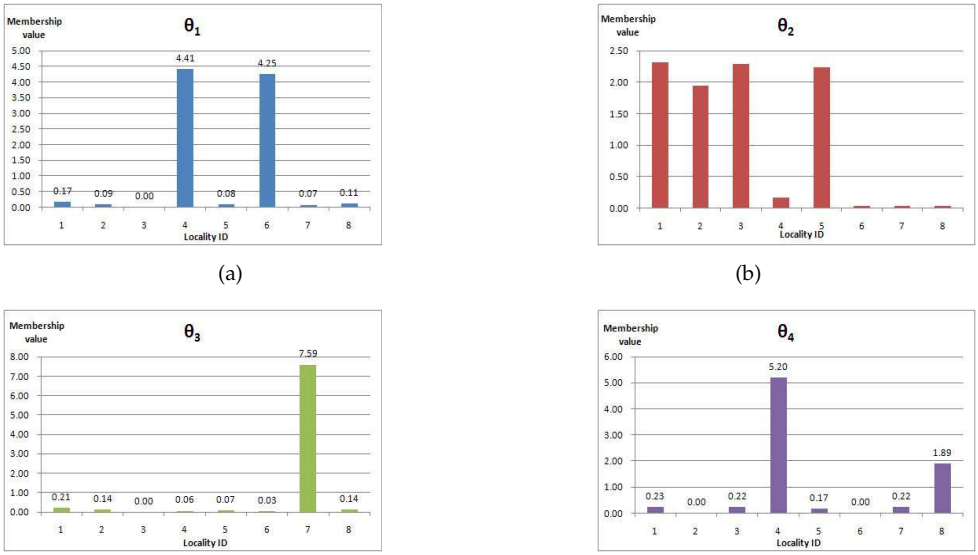


Fig. 6. Illustration of various distributions of topic mixture vector θ. Each bar denotes the value of one data instance in a topic locality. Intuitively, the data instance should be ideally assigned to a small number of localities with relatively large values only.

Algorithm: Adaptive Instance Assignment

Input: Topic mixture $\theta = \theta_1, \ldots, \theta_K \in R^K$;
Thresholds: $Th_{adj}, Th_{sum} \in [0, \ldots, 1]$;
Replication parameter: L.

Output: Locality index set: U; Normalized topic mixture: θ.

l_1-Normalize θ: $\theta_i \leftarrow \theta_i / \parallel \theta \parallel_1$, $i = 1, \ldots, K$;
Sort: [θ, $Index$] = sort(θ,'descend');
$S \leftarrow 0, Len \leftarrow min(K, L)$, $U \leftarrow \emptyset$;
$U \leftarrow U \cup \{Index_1\}$;
for ($i = 1$; $i < Len$; i++)
  if ($\theta_{i+1}/\theta_i < Th_{adj}$)
    break;
  $S \leftarrow S + \theta_{i+1}$;
  if ($S > Th_{sum}$)
    break;
  $U \leftarrow U \cup \{Index_{i+1}\}$;
end for;
if ($i + 1 < K$)
  reset: $\theta_j \leftarrow 0$, $j = i + 2, \ldots, K$;
l_1-Normalize θ: $\theta_j \leftarrow \theta_j / \parallel \theta \parallel_1$, $j = 1, \ldots, i + 1$;
Return U, θ.

Fig. 7. Adaptive instance assignment algorithm

more probably the instance is related to the corresponding i^{th} topic locality. Here, L is a replication parameter to control the maximum number of localities that a data instance can be assigned. This effectively controls the replication degree of instance.

One challenge here is that the topic mixture value of θ_i in θ may vary greatly, and it is not reasonable to assign the data instance to localities with very small values. For example in Fig.6, the data instance should be ideally assigned to a small number of localities with relatively large mixture values only. To do so, we leverage an ordered operator to select valid localities in an adaptive manner.

The main idea is to detect the abrupt decrease in two adjacent elements in the normalized and sorted (in descending order) topic mixture θ. If the ratio of the mixture value θ_{i+1} to the former θ_i is greater than the a given adjacent threshold (Th_{adj}) and the accumulated sum of the vector values is below a given accumulation threshold (Th_{sum}), then assign the data instance to the corresponding locality. Finally, we reset all the elements θ_j after the abrupt decrease in θ the to zero, and re-normalize the topic mixture vector θ. The adaptive instance-locality assignment algorithm is shown in Fig.7.

After grouping all the instances in the training set to localities, we finish coarse categorization by partitioning large-scale training data set into K small topic localities according to topic mixture θ of instances. Then, we train a linear discriminative classifier for each locality to learn the instance localities π_k. Once all the local classifiers are trained, we finish the ensemble construction and training process.

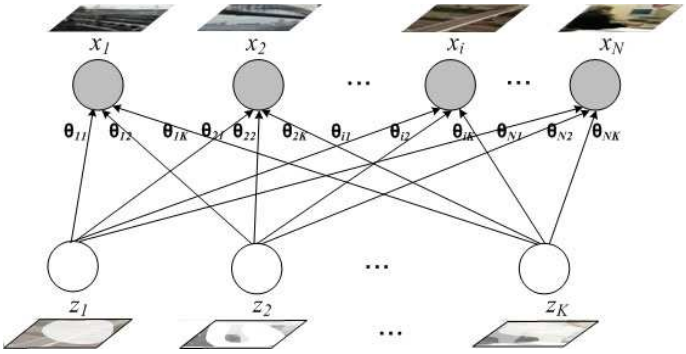


Fig. 8. In the probabilistic generative process of LDA, the observed data instances (documents) are represented as a mixture of hidden topics, and the topic mixture vector θ determines how an instance x is mixed over K hidden topic localities. Note that after reseting of θ in the above adaptive instance assignment, most connections between x and z will be of zero coefficients.

3.3 Coordination of individual classifiers

As shown in Fig.5, after LDA inference of test instances during testing process, we need to coordinate the learned local classifiers for the final classification.

The coordination of classifiers in the ensemble concerns learning the gating function $\Psi_k(x_t)$ in the Equation 5 for a test sample x_t. As shown in Fig.8, in the probabilistic generative process of LDA, the observed data instances (documents) are represented as a mixture of hidden topics, and the topic mixture vector θ determines how an instance x is mixed over K hidden topic localities. Therefore, in the LDA generative model, the influence of a topic locality z_k on x_t is represented by the connection strength θ_{tk}. We, therefore, utilize this influence to coordinate individual classifiers by setting $\Psi_k(x_t) = \theta_{tk}$. Note that after resetting of θ in the above adaptive instance assignment, most weak connections between x and z will be of zero coefficients. In other words, for a testing sample x_t, as only a few number ($T \leq L$ is the resultant size of the locality index set U returned by the algorithm) of large topic mixture coefficients are non-zero which trigger the corresponding number T of local classifiers to fire.

3.4 Analysis and discussion

3.4.1 The number of localities

The optimal tuning of the number of localities K is not necessary, since the instance space partitioning is a coarse process to group data. K can be roughly determined by the ratio of total number of training samples N to the desired average size of topic set which should be generally no more than ten thousands for the consideration of classifier training efficiency.

3.4.2 Computational complexity

The separate training of individual classifiers on each locality gives rise to the parallelism of SVM training. Assume that each locality possesses $O(N \cdot T/K)$ data instances. Since the theoretical computational complexity of SVM training (without considering space problem) is between $O(n^2)$ and $O(n^3)$ (n denotes the number of training samples) depending on the value of the hyper-parameter C (Bordes et al., 2005). Then the complexity of learning

its classifier is between the lower bound $O((N \cdot T/K)^2)$ and upper bound $O((N \cdot T/K)^3)$. The total complexity of the ensemble is between $K \cdot O((N \cdot T/K)^2) = \frac{T^2}{K} \cdot O(N^2)$, and $K \cdot O((N \cdot T/K)^3) = \frac{T^3}{K^2} \cdot O(N^3)$, which is much more efficient than the complexity of the single SVM that is learned over the entire dataset since $T \leq L \ll K$. Furthermore, if we take the space problem into consideration, as the number of training instances N grows, the large kernel matrix cannot be stored in memory and the cost of computing each kernel value is relatively high because Kernel values must be computed on the fly or retrieved from a cache of often accessed values (Bordes et al., 2005), which makes single SVM impractical. On the other hand, it means that the ensemble can scale up to large scale training dataset.

Similarly, the testing speed can be considerably improved since a testing instance belongs to only a few T localities and invokes the corresponding local classifiers only. The test complexity of the ensemble is $O(N_{sv})$, where N_{sv} is the number of support vectors (SV) which is proportional to number of training samples. Assume the testing data instance are assigned to only T localities as illustrated in Figure 1(b), then the number of training samples in all firing classifiers can be estimated to be $N \cdot T^2/K$ in average. Therefore, the testing efficiency can be considerably improved. This makes it practical for online detection in spite of large training data set.

3.4.3 Cross validation for classifier parameter optimization

Cross validation is widely used for parameter optimization of classifiers, such as the cost parameter C in soft-margin SVMs and the width parameter g of the Gaussian kernel for SVM classifiers. It has great influence on video classification performance (van Gemert et al., 2006). However, it is very time consuming for large scale training set. For the individual SVM classifier training on each locality, we do not adopt SVM cross validation due to the two facts: (1) the diversity of the ensemble makes uncorrelated decisions for each classifier thus complement each other, which makes cross validation less important compared with the case of single SVM training; (2) for the unbalanced data sets such as TRECVid, traditional accuracy-based SVM cross validation may not be good for model selection. Perhaps AP or InfAP based cross validation is more preferable.

4. System and evaluation

Based on the proposed LDA ensemble learning method, we developed two systems to test its effectiveness: (1)TRECVid concept detection system; (2) Online pornography filtering system. We will introduce them briefly as follows.

4.1 TRECVid concept detection system

To evaluate the performance of our proposed method, we developed a video concept detection system based on the TRECVid 08 video benchmark collection (Over et al., 2008). The preliminary results have been reported in our recent papers (Tang et al., 2008).

4.1.1 Datasets and experimental setup

In TRECVid 08, 20 concepts are used for evaluation as listed in Table 1. Its development set consists of 109-hour documentary videos of 43,616 keyframes(shots), and testing set of 109-hour videos of 35,766 keyframes(shots).

There are two kinds of the annotation efforts for the development set (Snoek & Worring, 2009): one is our manual annotation (Tang et al., 2008) and the other is collaboration annotation

ID	Concept	#Pos	#Hit	ID	Concept	#Pos	#Hit
1001	Classroom	241	64	1011	Harbor	217	35
1002	Bridge	186	30	1012	Telephone	203	106
1003	Emergency-Vehicle	103	22	1013	Street	1799	458
1004	Dog	136	94	1014	Demonstration-Or-Protest	159	87
1005	Kitchen	289	124	1015	Hand	1879	630
1006	Airplane-flying	80	64	1016	Mountain	265	140
1007	Two-people	4140	1090	1017	Nighttime	490	316
1008	Bus	106	47	1018	Boat-Ship	506	210
1009	Driver	302	364	1019	Flower	620	319
1010	Cityscape	331	337	1020	Singing	441	133

Note: The column “#Pos" denotes the number of positive training samples in the development set, and the column “#Hit" denotes the number of hits in the groundtruth of the test set provided by TRECVid.

Table 1. The list of 20 concepts in TV08.

Proposed (L=1)	Proposed (L=6)	Single-SVM	Bagging
0.116	0.138	0.130	0.132

Table 2. MAP Comparison of the proposed method, Single-SVM and Bagging on TV 08.

launched by Laboratory of Informatics of Grenoble (LIG) (Ayache & Quénot, 2008). In our experiments, we used the combination of both ours and LIG's annotation. The number of positive training samples and number of hits in the groundtruth are also shown in Table 1. The evaluation criteria used here is the inferred average precision (InfAP) (Yilmaz & Aslam, 2006) or inferred mean average precision (Inf MAP). InfAP is a very good estimate for average precision (AP). AP is the average of precisions computed at the point of each of the relevant documents for considering the order in the ranked sequence of documents, and it is one of the most commonly used system-oriented measures of retrieval effectiveness (Smeaton et al., 2009). InfAP was adopted to replace AP in TRECVid since 2005 to save large amount of judging effort as verified by Yilmaz and Aslam (Yilmaz & Aslam, 2006).

4.1.2 Features and ensemble parameters

We use the released VIREO-374 features (Jiang et al., 2010) to train and test our system. The primary visual feature we adopt is the local BoVW) features, due to its widely reported effectiveness. The BoVW representation is a histogram based on a visual vocabulary of 500 visual words clustered by a set of about 500,000 SIFT features (Lowe, 2004), and weighted by a soft-weighting scheme for taking into account the significance of each visual word in the keyframe, which has been demonstrated to be more effective than the traditional TF/TF-IDF weighting schemes (Jiang et al., 2010).

There are three principal types of parameters in the LDA ensemble construction and SVM training phase of the proposed ensemble:

(1) The number of localities K: we set $K = 100$ empirically;

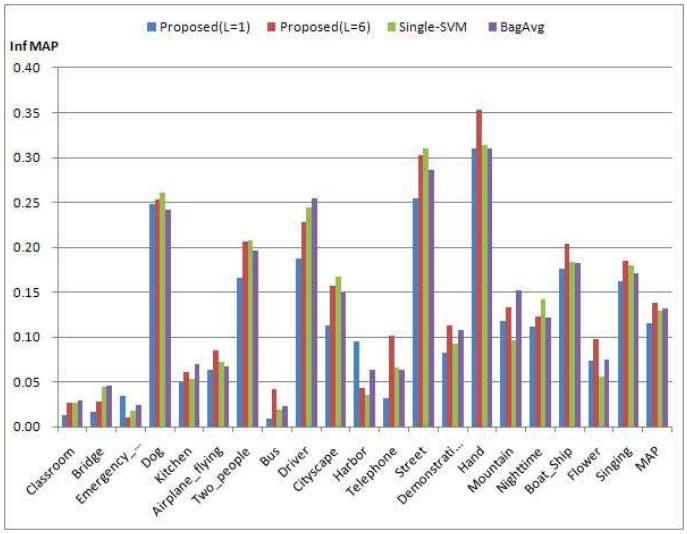


Fig. 9. Comparison of AP for each concept on TV 08 by different runs. As shown, the proposed method outperforms single SVM and Bagging.

(2) Parameters for the adaptive instance-locality assignment, and our recommendation is: the replication parameter $L = 6$, and the two thresholds: $Th_{adj} = 0.2$, $Th_{sum} = 0.95$. Th_{adj} and Th_{adj} is not so important since they can be determined empirically;

(3) For the individual SVM classifier of each locality, we utilize the default RBF kernel, and we do not adopt SVM cross validation according to the aforementioned reasons.

4.1.3 Experimental results

To investigate the effectiveness of the proposed ensemble, we compare its performances with the widely used single SVM method over the entire data set, and the well known ensemble learning approach - Bagging by imitating the random generation of $K = 100$ training subsets through sampling with replacement from the development set (Efron & Tibshirani, 1993).

The proposed ensemble learning ($L = 6$) gives rise to an MAP of 0.138, which is 6.2% relatively higher than the single SVM of MAP 0.130, and 4.5% higher than Bagging of MAP 0.132 as shown in Table 2. Fig.9 also shows the comparison of AP for each concept. As shown, the proposed method outperforms single SVM on evidently on 11 out of 20 concepts, which are object-oriented concepts like "Airplane-flying", "Bus", "Telephone", "Hand", "Boat-Ship", "Flower". Most of the rest concepts in which single SVM performs better are scene-oriented concepts like "Cityscape", "Street" and "Nighttime". As compared with Bagging, we can also see that the proposed method are better for object-oriented concepts.

Our conjecture is that the object-oriented concepts have intrinsic structures consisting of different subcategories (such as different kind of canoes and steamers associated with the concept "Boat-Ship") of common objects (such as "wheels" for the concept "Car"), which have similar local features, and hence makes LDA exhibit obvious advantages in capturing visual contents by exploiting the co-occurrences of visual words for instance space partitioning

during ensemble construction. On the other hand, scene-oriented concepts are too diversified to have common parts such as "Street" and "Nighttime".

Additionally, for the proposed method with the case of $L = 1$, its performances are greatly reduced as shown in Fig.9, even worse than the single SVM method. We attribute this to the insufficient positive training samples in most of topic localities due to the over-partition without replication of samples.

Besides the better accuracy, the proposed ensemble also enables much more efficient training than the single SVM. According to the previous computational complexity analysis, the lower bound complexity of the proposed ensemble ($L = 6$) is $\frac{L^2}{K} \cdot O(N^2)$, which is 0.36 of the single SVM. This is verified from the actual ratio (0.077) of the proposed training time (6.0 hours) to that of single SVM (78.3 hours), while the actual total number of training samples of the proposed amounts to about 250,000, approximately $L = 6$ times as that of single SVM.

4.2 Online pornography filtering system

Due to the explosion of images and videos in the Internet, the chances of individuals encountering adult-oriented contents such as pornographic images and videos increase dramatically, which has become a serious global socio-cultural problem. Therefore, it is of great importance to detect and filter these harmful contents to provide a cleaner internet environment for the sake of young adolescents' healthy growth.

Most existing methods for pornography filtering attempt to exploit text contents to classify web pages (Rowley et al., 2006). However, the textual approaches suffer from significant limitations such as dependence of languages, and unavailability of texts. Previous work on pornographic image detection can be divided into two broad categories (Hu et al., 2007; Rowley et al., 2006): skin-based methods (Zeng et al., 2004; Zheng, 2004) that are based primarily on skin color or texture, and model-based methods(Forsyth & Fleck, 1999) which analyze the shapes of skin colored regions to determine their similarity to human figures. Both categories rely heavily on skin detection which lacks sufficient robustness against significant variations in races, lighting conditions, textures, sex-positions, and other factors.

Due to the importance of data (Enzweiler & Gavrila, 2009), we first established a large-scale training image dataset to include all the kinds of possible variations aforementioned as much as possible. Then, in order to handle large-scale dataset both efficiently and effectively, we used the proposed LDA ensemble learning framework to develop an online pornography filtering system for detecting and monitoring images and video keyframes in the Internet, and the system is being used by governments and companies in real application. The details of the system are introduced as follows. The preliminary results have been published in our recent paper (Tang et al., 2009).

4.2.1 Construction of large-scale training dataset

We established a large-scale training image (including key-frames extracted from videos) dataset for pornography detection. Thanks to the proliferation of digital images and videos, it was no longer a difficult task to establish a large database with totally 420,615 training image samples collecting from a wide variety of diverse origins. We collected 1,108 pornographic videos from off-line VCD sources. We also captured about 20,000 short pornographic video clips from online media streams by the skin-based detection method (Zheng, 2004) from Dec 2007 to Dec 2008. We downloaded about 65,000 non-pornographic videos from YouTube, Tudou, YouKu and other websites. The non-pornographic images were mainly from Corel database while the pornographic images were downloaded from Pinkworld. After collection,

Samples	Images	Video Keyframes	Total
Positive	21,699	44,128	65,827
Negative	51,680	303,108	354,788

Table 3. Sample distribution of training dataset

we annotated all the images and keyframes after data collection and keyframe extraction. During annotation, in order to distinguish true pornographic images from non-pornographic bodies, we regarded only the images with exposed woman breast, anus, genital organs, or sexual intercourse scenes as positive samples, while others as negative samples regardless of exposed skin area. Details of the sample distribution are listed in Table 3. Up to now, few pornographic image detection systems are based on such a large-scale database with more than 10^5 images. To our best knowledge, the reported number of pornographic positive training samples is usually less than 10^4 images.

4.2.2 Features and ensemble parameters

We extracted the following three kinds of keyframe features in the system:

1. Color Histogram (CH) (Amir et al., 2005): This is a 166-dimensional histogram, a global representation of keyframe which is based on the distribution of pixels in an uniformly partitioned HSV color space.

2. Color Moments (CM) (Stricker & Orengo, 1995): To further incorporate spatial relationship into the color content, a keyframe is partitioned into a 5×5 grid and each patch is represented using the first three moments of the color distribution in LAB color space, i.e. the mean, standard deviation and the third root of the skewness of each color channel. The color moments for each patch are then concatenated to form a 255-dimensional feature vector. In our implementation (Chua et al., 2009), we pre-compute the transformation coefficients for color moment feature extraction which can provide up to five times speed up over the traditional extraction method.

3. Edge Histogram (EH) (Amir et al., 2005): It is localized edge histograms from a 5-region layout consisting of four corner regions and a center overlapping region, represented as a 320-dimensional vector with 8 edge direction bins and 8 edge magnitude bins based on a Sobel filter (64-dimensional) for each grid.

Through our hierarchical combination of unsupervised clustering and supervised learning, we used CH for LDA categorization at the top layer; and the prior fusion (concatenation) of the CM and EH for finer discrimination at the bottom layer (CM+EH, 545-dimensional). We used CH for coarse categorization due to its relatively lower dimension and faster extraction, while the concatenation of the two sets of features for finer discrimination for further removal of false detection caused by many existing skin-based methods. To meet the online detection speed requirements, we did not use the effective bag-of-visual-word features based on local features such as SIFT. Although there is no special emphasis on detecting skin, skin detection is actually included in the latent semantic analysis of the color histogram and training of SVM models with color moment.

We set the number of localities $K = 40$, the replication parameter $L = 1$, and the two thresholds: $Th_{adj} = 0.2$, $Th_{sum} = 0.95$. For the individual SVM classifier of each locality, we utilize the default RBF kernel, and we do not adopt SVM cross validation.

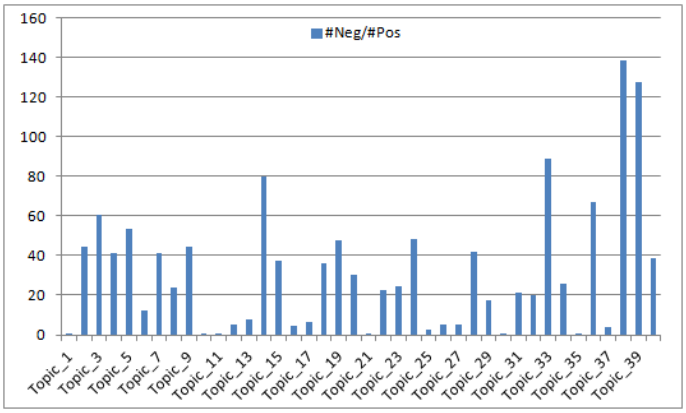


Fig. 10. Distributions of Negative/Positive samples over the $K = 40$ topics

4.2.3 Experimental results

After training all the $K = 40$ individual SVM classifiers, we collected the test image/keyframe samples from some companies independently. The total 7110 test samples include 1695 pornographic samples and 5415 non-pornographic samples.

Figure 10 show the distributions of the Negative/Positive samples over the $K = 40$ topics. We can see that the Negative/Positive ratio varies greatly from 0.29 in the topic 11 to 138.42 in the largest topic 38, which means there are more pornographic samples than non-pornographic ones in the topics (1, 10, 11, 21, 30 and 35) where the ratio is less than 1, while there are relatively much fewer pornographic samples in the topics (such as 38 and 39) where the ratio is very large. Compared with the nearly uniform distribution of Negative/Positive samples generated by random sampling, this imbalanced distribution of positive samples over topics indicates that LDA can mine the hidden structures of images effectively.

For comparison, we implemented the skin-based method in (Zheng, 2004) and used it for collecting our training database as mention above. We also implemented single SVM method with the same feature CM+EH over randomly selected 120, 000 samples from our training data set. We did not use the whole training dataset for single SVM due to the impractical amount of training computation.

The ROC curves for our proposed ensemble learning (LDA-SVM), SVM, and the skin-based method are shown in Figure 11, which indicates that the LDA-SVM is much more effective than other two methods. Particularly, when we set the detection score threshold to 0.95, the false positive rate can reach as low as 0.11% (only 6 out of 5415 non-pornographic samples are recognized pornographic ones) while keeping the recall rate still around 50% (840 out of 1695 pornographic samples are correctly recognized). On the other hand, when we set the detection score threshold to 0.5, the precision and recall rates can reach as high as 95.12% (corresponding to false positive rate of 4.88%) and 90.09% respectively.

To test the effectiveness of our coordination method with topic mixture θ, we use the average fusion method for comparison, and its ROC curve is shown in Figure 11 (the cyan one marked as LDA-SVM(AVG)). Since all the 40 SVM models is used for average fusion, its test time is 667ms about 13.6 times slower than that (49 ms) of our coordination method. So we can conclude that using the topic mixture coefficients for adaptive fusion is effective and efficient.

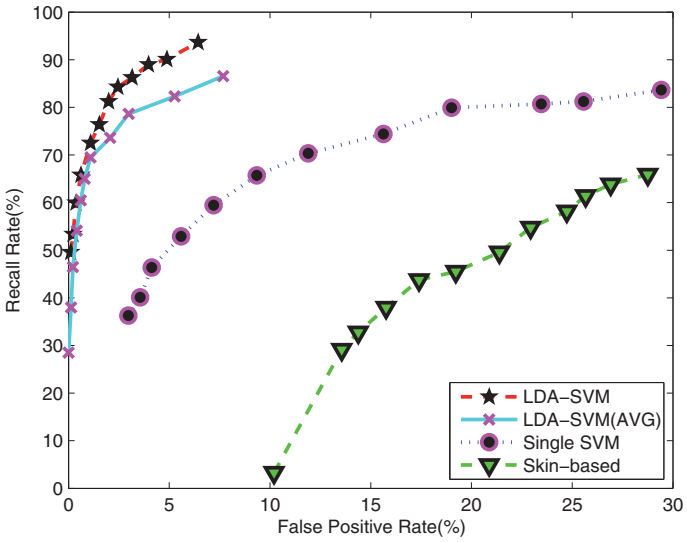


Fig. 11. ROC Curves for Pornography Detection

Methods	SVM	LDA-SVM
Training samples	120,000	420,615
Number of SVs	24,112	1,842 per topic
Training time	72 hours	6 hours
Testing time(320×240)	667 ms	49 ms

Table 4. Training & testing time of SVM methods

The training time, testing time and the numbers of samples and SVs of the single SVM method and LDA-SVM are shown in Table 4, which indicates the high training and testing efficiency of the proposed method.

5. Conclusion and future work

In this chapter, motivated by the insight from psychophysical studies, we propose a novel ensemble learning framework in LDA topic models for large scale concept detection through combination of unsupervised semantic grouping and supervised learning. Classifier diversity is achieved by digging the intrinsic topic structure of large visual data under the framework of LDA topic modeling. For the ensemble fusion, the individual classifiers are then coordinated based on the large LDA topic mixture coefficients in a generative probabilistic manner, which is naturally achieved without any additional classifier selection module. As the individual classifiers are often more compact due to their training on the smaller topic localities, and only a small number of classifiers in the ensemble will fire on a testing sample, the testing efficiency can be considerably improved. This makes it practical for online concept detection despite of large training data set. Extensive tests on the TRECVid 08 benchmark and pornography detection show that the proposed ensemble learning achieves promising results and outperforms existing approaches.

Several issues are worthy of further investigation. First, optimal feature, kernel selection and removal of redundant samples along with high-dimensional indexing should be taken into

consideration to further improve the performance. Then, the individual classifiers trained in each locality can be further explored for cross-domain concept detection. Finally, it is of great importance to use tags of web images to avoid laborious annotation of training samples.

6. Acknowledgments

This work was supported by National Nature Science Foundation of China (60873165).

7. References

A. Vailaya, A. K. J. & Zhang, H.-J. (1998). On image classification: City images vs. landscapes, *Pattern Recognition* 31: 1921–1936.

Alsumait, L., Wang, P., Domeniconi, C. & Barbarĺć, D. (2010). *Embedding Semantics in LDA Topic Models*, John Wiley & Sons, Ltd.

Amir, A., Argillander, J., Campbell, M. & *et al.* (2005). IBM Research TRECVID-2005 Video Retrieval System, *NIST TRECVID Workshop*.

Amir, A., Berg, M., Chang, S.-F. & *et al.* (2003). Ibm research trecvid-2003 video retrieval system, *NIST TRECVID Workshop*.

Ayache, S. & Quénot, G. (2008). Video corpus annotation using active learning, *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pp. 187–198.

Bay, H., Tuytelaars, T., Gool, V. & L. (2006). Surf: Speeded up robust features, *9th European Conference on Computer Vision*.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research* .

Bordes, A., Ertekin, S., Weston, J. & Bottou, L. (2005). Fast kernel classifiers with online and active learning, *J. Mach. Learn. Res.* 6: 1579–1619.

Borth, D., Ulges, A. & Breuel, T. (2010). *Adapting Web-based Video Concept Detectors for Different Target Domains*, Bentham Science Publishers, chapter 6.

Borth, D., Ulges, A., Schulze, C. & Breuel, T. (2008). Keyframe extraction for video tagging & summarization., *Informatiktage*, Vol. S-6 of *LNI*, GI, pp. 45–48.

Breiman, L. (1996). Bagging predictors, *Mach. Learn.* 24: 123–140.

Cao, J., Lan, Y., Li, J. & *et al.* (2006). Intelligent multimedia group of Tsinghua University at trecvid 2006, *NIST TRECVID Workshop*.

Chang, S.-F., He, J., Jiang, Y.-G., Khoury, E. E., Ngo, C.-W., Yanagawa, A. & Zavesky, E. (2008). Columbia university/vireo-cityu/irit trecvid2008 high-level feature extraction and interactive video search, *NIST TRECVID Workshop*.

Chang, Y., Lee, D. J., Hong, Y. & Archibald, J. (2008). Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature transform descriptor, *J. Image Video Process.* 2008: 9:1–9:10.

Chua, T.-S., Tang, S., Trichet, R., Tan, H. K. & Song, Y. (2009). Moviebase: A movie database for event detection and behavioral analysis, *Proceedings of ACM Multimedia 2009 Workshop on Web-Scale Multimedia Corpus*.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, IEEE Computer Society, Washington, DC, USA, pp. 886–893.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science* 41: 391–407.

Dietterich, T. G. (2000). Ensemble methods in machine learning, *Multiple classifers systems* 1857: 1–15.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.

Enzweiler, M. & Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31: 2179–2195.

Forsyth, D. A. & Fleck, M. M. (1999). Automatic detection of human nudes, *Int. J. Comput. Vision* 32: 63–77.

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55(1): 119–139.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics., *Proceedings of the National Academy of Sciences* pp. 5228–5235.

Hauptmann, A., Yan, R. & Lin, W.-H. (2007). How many high-level concepts will fill the semantic gap in news video retrieval?, *Proceedings of the 6th ACM International Conference on Image and video retrieval*, ACM, pp. 627–634.

Hofmann, T. (1999). Probabilistic latent semantic analysis, *Proceedings of Uncertainty in Artificial Intelligence, UAI*.

Hu, W., Wu, O., Chen, Z., Fu, Z. & Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29: 1019–1034.

Huiskes, M. J., Thomee, B. & Lew, M. S. (2010). New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative, *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, ACM, New York, NY, USA, pp. 527–536.

Jiang, W., Zavesky, E., Chang, S.-F. & Loui, A. (2008). Cross-domain learning methods for high-level visual concept classification, *International Conference on Image Processing*.

Jiang, Y.-G. (2009). *Large Scale Semantic Concept Detection, Fusion, and Selection for Domain Adaptive Video Search*, PhD thesis, City University of Hong Kong.

Jiang, Y.-G. & *et al*. (2010). Columbia-UCF TRECVID 2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching, *NIST TRECVID Workshop*.

Jiang, Y. G., Ngo, C. W. & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval, *Proceedings of ACM Conference on Image and video retrieval*, pp. 494–501.

Jiang, Y.-G., Yang, J., Ngo, C.-W. & Hauptmann, A. G. (2010). Representations of keypoint-based semantic concept detection: A comprehensive study, *IEEE Transactions on Multimedia* 12: 42–53.

Jurie, F. & Triggs, B. (2005). Creating efficient codebooks for visual recognition, *Proceedings of International Conference on Computer Vision*.

Krogh, A. & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning, *Proceedings of Advances in Neural Information Processing Systems*, pp. 231–238.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc.

Kuncheva, L. I. & Whitaker, C. J. (2005). Controlling the diversity in classifier ensembles through a measure of agreement, *Pattern Recognition* 38(11): 2195 – 2199.

Laptev, I. (2005). On space-time interest points, *Int. J. Comput. Vision* 64: 107–123.

Li, F.-F. & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories, *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pp. 524–531.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60(2): 91–110.

Munder, S. & Gavrila, D. (2006). An experimental study on pedestrian classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28: 1863–1868.

Naphade, M. R. & Smith, J. R. (2004). On the detection of semantic concepts at trecvid, *Proceedings of ACM International Conference on Multimedia*, pp. 660–667.

Ngo, C.-W., Jiang, Y.-G., Wei, X.-Y., Zhao, W., Liu, Y., Wang, J., Zhu, S. & Chang, S.-F. (2009). Vireo/dvmm at trecvid 2009: High-level feature extraction, automatic video search, and content-based copy detection, *NIST TRECVID Workshop*.

Niebles, J., Wang, H. & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words, *International Journal of Computer Vision* 79: 299–318.

NIST (2001-2010). TRECVID workshop papers, Website. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

Ojala, T., Pietikäinen, M. & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions, *Pattern Recognition* 29(1): 51–59.

Opitz, D. & Maclin, R. (1999). Popular ensemble methods: An empirical study, *Journal of Artificial Intelligence Research* 11: 169–198.

Opitz, D. W. & Shavlik, J. W. (1996a). Actively searching for an effective neural network ensemble, *Connect. Sci.* 8(3): 337–354.

Opitz, D. W. & Shavlik, J. W. (1996b). Generating accurate and diverse members of a neural-network ensemble, *Proceedings of Advances in Neural Information Processing Systems*, pp. 535–541.

Over, P., Awad, G., Rose, R. T., Fiscus, J. G., Kraaij, W. & Smeaton, A. F. (2008). Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics, *NIST TRECVID Workshop*.

Pan, C.-C. & Mitra, P. (2011). Event detection with spatial latent dirichlet allocation, *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pp. 349–358.

Pytlik, B., Ghoshal, A., Karakos, D. & Khudanpur, S. (2005). TRECVID 2005 Experiment at Johns Hopkins University: Using Hidden Markov Models for Video Retrieval, *NIST TRECVID Workshop*.

Qi, G. J., Hua, X. S., Y. Rui, J. T., Mei, T. & Zhang, H. J. (2007). Correl-ative multi-label video annotation, *ACM International Conference on Multimedia*.

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*, World Scientific Publishing Company.

Rowley, H., Jing, Y. & Baluja, S. (2006). Large scale image-based adult content filtering, *International Conference on Computer Vision Theory and Applications*.

S. Ayache, G. Q. & Gensel, J. (2007). Classifier fusion for svm-based multimedia semantic indexing, *European Conference on Information Retrieval*.

S. Thorpe, D. F. & Marlot., C. (1996). Speed of processing in the human visual system, *Nature* 381: 520–522.

Salton, G. & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA.

Siagian, C. & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention, *IEEE Trans. Pattern Anal. Mach. Intell.* 29: 300–312.

Sivic, J. & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos, *Proceedings of ICCV*, p. 1470.

Smeaton, A. F., Over, P. & Doherty, A. R. (2010). Video shot boundary detection: Seven years of trecvid activity, *Comput. Vis. Image Underst.* 114: 411–418.

Smeaton, A. F., Over, P. & Kraaij, W. (2006). Evaluation campaigns and trecvid, *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, ACM Press, pp. 321–330.

Smeaton, A. F., Over, P. & Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements, *in* A. Divakaran (ed.), *Multimedia Content Analysis, Theory and Applications*, Springer Verlag, Berlin, pp. 151–174.

Smeulders, A. W. M., M.Worring, Santini, S., Gupta, A. & Jain, R. (2000). Content-based image retrieval at the end of the early years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12): 1349–1380.

Smith, J. R. & Chang, S.-F. (1997). Visually searching the web for content, *IEEE MultiMedia* 4: 12–20.

Snoek, C. G. M. & Worring, M. (2009). Concept-based video retrieval, *Found. Trends Inf. Retr.* 2(4): 215–322.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M. & Smeulders, A. W. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia, *Proceedings of ACM Conference on Multimedia*, pp. 421–430.

Snoek, C. G., van de Sande, K. E., de Rooij, O. & *et al.* (2010). The mediamill trecvid 2010 semantic video search engine, *NIST TRECVID Workshop*.

Snoek, C., Worring, M. & Smeulders, A. (2005). Early versus late fusion in semantic video analysis, *Proceedings of the ACM International Conference on Multimedia*, Singapore, pp. 399–402.

Song, Y., Zheng, Y.-T., Tang, S. & *et al*. (2011). Localized multiple kernel learning for realistic human action recognition in videos, *IEEE Transactions on Circuits and Systems for Video Technology* 21(9).

Stricker, M. A. & Orengo, M. (1995). Similarity of color images, *Storage and Retrieval for Image and Video Databases (SPIE)*, pp. 381–392.

Szummer, M. & Picard, R. W. (1998). Indoor-outdoor image classification, *IEEE International Workshop on Content-based Access of Image and Video Databases, in Conjunction with ICCVąŕ98*.

Tang, J., Hong, R., Yan, S., Chua, T.-S., Qi, G.-J. & Jain, R. (2010). Image annotation by graph-based inference with integrated multiple/single instance representations, *IEEE Transactions on Multimedia* pp. 131–141.

Tang, J., Hong, R., Yan, S., Chua, T.-S., Qi, G.-J. & Jain, R. (2011). Image annotation by knn-sparse graph-based label propagation over noisily tagged web images, *ACM Trans. Intell. Syst. Technol.* **2**: 14:1–14:15.

Tang, S., Li, J.-T., Li, M. & *et al.* (2007). Trecvid 2007 high-level feature extraction by MCG-ICT-CAS, *NIST TRECVID Workshop*.

Tang, S., Li, J.-T., Li, M., Xie, C., Liu, Y., Tao, K. & Xu, S.-X. (2008). Trecvid 2008 high-level feature extraction by MCG-ICT-CAS, *NIST TRECVID Workshop*.

Tang, S., Li, J., Zhang, Y., Xie, C., Li, M., Liu, Y., Hua, X., Zheng, Y.-T., Tang, J. & Chua, T.-S. (2009). Pornprobe: an lda-svm based pornography detection system, *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pp. 1003–1004.

Tang, S., Zheng, Y.-T., Wang, Y. & Chua, T.-S. (2012). Sparse ensemble learning for concept detection, *IEEE Transactions on Multimedia* 14(1).

van Gemert, J. C., Snoek, C. G. M., Veenman, C. J. & Smeulders, A. W. M. (2006). The influence of cross-validation on video classification performance, *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, ACM, New York, NY, USA, pp. 695–698.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer-Verlag.

Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J. & Song, Y. (2009). Unified video annotation via multi-graph learning, *IEEE Transactions on Circuits and Systems for Video Technology* 19(5).

Wang, M., Hua, X.-S., Tang, J. & Hong, R. (2009). Beyond distance measurement: Constructing neighborhood similarity for video annotation, *IEEE Transactions on Multimedia* 11(3).

Wang, X. (2011). Action recognition using topic models, *in* T. B. Moeslund, A. Hilton, V. KrÍźger & L. Sigal (eds), *Visual Analysis of Humans*, Springer London, pp. 311–332.

Wang, X. & Grimson, E. (2007). Spatial latent dirichlet allocation, *Proceeding of Neural Information Processing Systems Conference (NIPS)*.

Wang, Y. & Mori, G. (2009). Human action recognition by semilatent topic models, *IEEE Trans. Pattern Anal. Mach. Intell.* 31: 1762–1774.

Weng, M.-F. & Chuang, Y.-Y. (2008). Multi-cue fusion for semantic video indexing, *ACM International Conference on Multimedia*.

Xu, S., Tang, S., Zhang, Y. & Li, J. (2011). A pseudo relevance feedback based cross domain video concept detection, *Proceedings of the Third International Conference on Internet Multimedia Computing and Service*, ICIMCS '11.

Xu, S., Tang, S., Zhang, Y., Li, J. & Zheng, Y.-T. (2011). Multi-modality transfer based on multi-graph optimization for domain adaptive video concept annotation, *Neurocomputing* 74.

Y., F. & E., S. R. (1996). Experiments with a new boosting algorithm, *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 325–332.

Yilmaz, E. & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments, *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pp. 102–111.

Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F. & Zhang, B. (2007). A Formal Study of Shot Boundary Detection, *IEEE Transactions on Circuits and Systems for Video Technology* 17.

Zeng, W., Gao, W., Zhang, T. & Liu, Y. (2004). Image guarder: An intelligent detector for adult images, *Asian Conference on Computer Vision*, ACCV '04, Jeju Island, Korea, pp. 1080–1084.

Zhang, H., Berg, A. C., Maire, M. & Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition, *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2126–2136.

Zhang, H.-J., Tan, S. Y., Smoliar, S. W. & Gong, Y. (1995). Automatic parsing and indexing of news video, *Multimedia Systems* 2: 256–266.

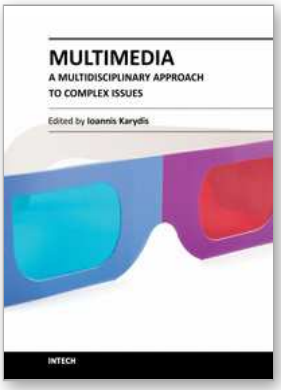
Zhang, J., Marsza, M., Lazebnik, S. & Schmid, C. (2007). Local features and kernels for cassification of texture and object categories: a comprehensive study, *International Journal of Computer Vision*, 73(2): 213–238.

Zhang, L. & Zhou, W.-D. (2011). Sparse ensembles using weighted combination methods based on linear programming, *Pattern Recogn.* 44: 97–106.

Zheng, H. (2004). *Maximum entropy modeling for skin detection: with an application to Internet filtering*, PhD thesis, UniveristÍę des Sciences et Technologies de Lille, France.

Zheng, Y.-T., Neo, S.-Y., Chua, T.-S. & Tian, Q. (2008). Probabilistic optimized ranking for multimedia semantic concept detection via RVM, *Proceedings of ACM Conference on Image and Video Retrieal (CIVR)*, pp. 161–168.

Zhu, J., Hoi, S. C., Lyu, M. R. & Yan, S. (2008). Near-duplicate keyframe retrieval by nonrigid image matching, *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, ACM, New York, NY, USA, pp. 41–50.



Multimedia - A Multidisciplinary Approach to Complex Issues
Edited by Dr. Ioannis Karydis

ISBN 978-953-51-0216-8
Hard cover, 276 pages
Publisher InTech
Published online 07, March, 2012
Published in print edition March, 2012

The nowadays ubiquitous and effortless digital data capture and processing capabilities offered by the majority of devices, lead to an unprecedented penetration of multimedia content in our everyday life. To make the most of this phenomenon, the rapidly increasing volume and usage of digitised content requires constant re-evaluation and adaptation of multimedia methodologies, in order to meet the relentless change of requirements from both the user and system perspectives. Advances in Multimedia provides readers with an overview of the ever-growing field of multimedia by bringing together various research studies and surveys from different subfields that point out such important aspects. Some of the main topics that this book deals with include: multimedia management in peer-to-peer structures & wireless networks, security characteristics in multimedia, semantic gap bridging for multimedia content and novel multimedia applications.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sheng Tang, Yan-Tao Zheng, Gang Cao, Yong-Dong Zhang and Jin-Tao Li (2012). Ensemble Learning with LDA Topic Models for Visual Concept Detection, Multimedia - A Multidisciplinary Approach to Complex Issues, Dr. Ioannis Karydis (Ed.), ISBN: 978-953-51-0216-8, InTech, Available from: http://www.intechopen.com/books/multimedia-a-multidisciplinary-approach-to-complex-issues/ensemble-learning-with-lda-topic-models-for-visual-concept-detection

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821