Multimodal Fusion using Learned Text Concepts for Image Categorization

Qiang Zhu Mei-Chen Yeh Kwang-Ting Cheng

Electrical & Computer Engineering Department
University of California, Santa Barbara, CA, 93106, USA

{qzhu,myeh,timcheng}@ece.ucsb.edu

ABSTRACT

Conventional image categorization techniques primarily rely on low-level visual cues. In this paper, we describe a multimodal fusion scheme which improves the image classification accuracy by incorporating the information derived from the embedded texts detected in the image under classification. Specific to each image category, a text concept is first learned from a set of labeled texts in images of the target category using Multiple Instance Learning [1]. For an image under classification which contains multiple detected text lines, we calculate a weighted Euclidian distance between each text line and the learned text concept of the target category. Subsequently, the minimum distance, along with low-level visual cues, are jointly used as the features for SVM-based classification. Experiments on a challenging image database demonstrate that the proposed fusion framework achieves a higher accuracy than the state-of-art methods for image classification.

Categories and Subject Descriptors: I.4.9 [Image Processing and Computer Vision]: Applications

General Terms: Algorithms, Experimentation

Keywords: Multimodal fusion, image categorization, text detection, multiple instance learning, image annotation

1. INTRODUCTION

Image categorization is an open problem in machine vision and has many attractive applications, such as image annotation, retrieval and scene understanding. The existing approaches are based on a wide range of technologies, ranging from the design of the features to the classification schemes. These approaches follow a similar principle: training an image classifier using low-level visual cues such as color, shape, texture and etc., as features. However, it has been proved that using low-level visual cues alone achieves a fairly low accuracy on unconstrained image sets.

Recently, Boutell and Luo [2] suggested the use of camera metadata cues for semantic scene classification. The camera metadata related to the capture conditions provides cues somewhat independent of the captured scene content and, thus, can be used to improve classification accuracy. More specifically, they proposed a Bayesian network to fuse content-based and metadata cues in the probability domain. In their work, only two problems were considered: indoor-outdoor classification and sunset detection. Wu and Chang [3] extended this idea for photo annotation. They proposed a probabilistic framework which uses influence diagrams to fuse metadata of multiple modalities among contextual information, visual content and semantic ontology in a synergistic way. Moreover, the image categories considered in their work were more diverse, and the photo semantic ontology has a hierarchical, multi-layer structure which requires finer categorization. Although the use of camera metadata somewhat improves the image classification accuracy, the main limitation of this idea is that the information derived from camera metadata works effectively only for some special image categories. For example, both approaches reported a significant improvement for classifying sunset images. Note that sunset scene is a very unique natural phenomenon which occurs within a fixed period of time of the day and is usually taken under very particular camera settings. However, considering a more general case such as classifying a "Highway" scene vs. a "Street" scene, there is unlikely any useful information contained in the camera metadata. Therefore, more information at the semantic level, other than the camera metadata, needs to be extracted for improving the classification accuracy.

Some recent studies result in a number of successful systems in detecting text in natural images. Gao and Yang [4] present a system for automatic extraction and interpretation of signs from natural scenes. The results can be further shown on a hand-held wearable display, or a head-mounted display. In [5], Chen and Yuille use the AdaBoost algorithm to learn a strong classifier for detecting text in unconstrained city scenes. The type of texts contained in their image database includes street signs, hospital signs, bus numbers, and etc. Both systems are designed mainly for helping visually handicapped people to increase their environmental awareness. Other potential applications include a driving assistant system [6] with detection of text on road signs from videos, or helping visual indexing for large photo and video databases. A detailed survey on text detection techniques will be given in Section 2.2. We note that most applications of text detection ignore the scene itself, i.e. they do not explore the information regarding the object that the detected text is attached to, nor the surrounding scene of the detected text.

After a thorough investigation of both topics, we identify several interesting connections between image/scene classification and text detection/recognition:

- For the purpose of scene understanding, the image database often contains some relevant text information, such as road signs, building names, bus numbers and etc.

MM'06, October 23-27, 2006, Santa Barbara, California, USA.

- For some scene/object categories, the texts embedded in images of the same category may exist some uniform patterns over text colors, text sizes, locations, and etc. This finding indicates that the text detected in an image may contain useful information for helping image classification.
- The detected text can be better interpreted if we know the associated scene/object. For example, we can develop a smart system which is able to provide both the OCR results and the objects/scenes to which these texts are attached.

Inspired by these observations, we propose a multimodal fusion framework using the information derived from the detected text as well as the low-level visual cues. The proposed framework follows a two-step process depicted in Figure 1.

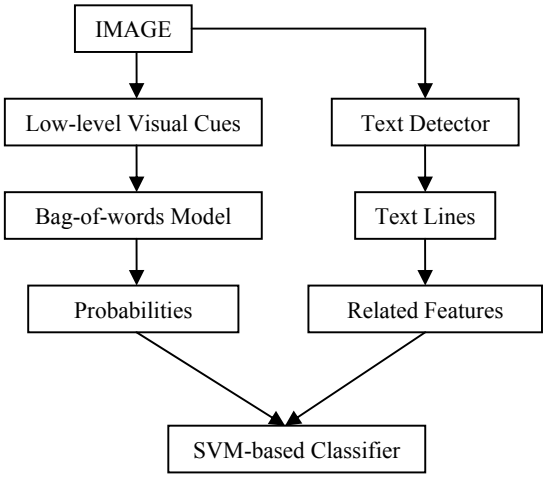


Figure 1: Multimodal fusion using visual cues and texts

For a given image, we first perform image classification using a bag-of-words model [7] which is based on the low-level visual cues. The outcomes of this step are the probabilities of this image belonging to each image category. In the meantime, a text detector is used to locate the text lines in this image. In Step 2, a SVM-based classifier is used to fuse the information of two different models together. As the flow presents, two specific problems need to be addressed. First, the specifics of integrating a bag-of-words model into a SVM model need to be developed. Our experiment indicates that different combinations between the two models could result in very different performance, which will be detailed in the experimental section. The second problem is the extraction of useful features from the detected texts. As one image may contain multiple text lines, some of which are relevant and some are not, it raises a nontrivial question: shall we take all the text lines into account or use some criteria to choose the most representative one? We will discuss the details of our solution to this question in Section 5.1.

As a generative approach, the bag-of-words model is proved to be very successful for object/scene classification. On the contrary, the SVM model is typically selected for a discriminative approach. Our proposed approach can be viewed as an implicit way of combining a generative model and a discriminative model. Note that promising results have been reported for prior approaches using such a hybrid strategy [8, 9] for image categorization.

In the rest of the paper, we first give some background on image categorization and text detection. We then describe the image database used in our work. Section 4 introduces the bag-of-words model, followed by the details of our fusion framework and experimental results. Finally, we discuss two potential applications and conclude with a short discussion.

2. RELATED WORK

2.1 Image Categorization

The term *image categorization* refers to the labeling of images into one of a number of predefined categories. A significant amount of research has been done to address this problem in the communities of Computer Vision [7, 10], and Image Retrieval [11]. The main difficulties for accurately solving this problem include variable image conditions (view point, illumination, scale, etc.), hard-to-describe objects in images, occlusion of objects, and most importantly, the gap between the low-level features (color, texture, shape, etc.) and the high-level semantic concepts. Most previous work employed a learning approach in order to fill the semantic gap. These methods differ in terms of how an image is represented, and how a classifier is formed given the training data.

Representation: For each image, we can extract features from the content, which can be categorized into global features and local features. For global features, frequency distribution, edge orientations and color histograms have been widely used. Recently, the work in [10] shows that local features are more powerful than global features in terms of robustness to occlusions and spatial variations. Local regions can be extracted by even or random sampling, by Kadir & Brady Saliency Detector [12], or by Lowe's DoG Detector [10]. In [13], Mikolajczyk et al. empirically demonstrate that the Scale Invariant Feature Transform (SIFT) descriptors [10] seem to be more robust than the pixel gray value representation, and dense sampling grid outperforms all other detectors.

Learning: Given the training data, how do we form a classifier to differentiate a certain image category from others? The design of the learning algorithms considers level of supervision, i.e. given a training sample, what kind of label is provided. Moreover, the training methods can be broadly characterized as either generative or discriminative, according to whether or not the distribution of the image features is modeled.

Recently the bag-of-words model [7, 14] receives attention in the computer vision community for object recognition and natural scene classification. This model originated from approaches categorizing the documents, and was later successfully generalized and extended for categorizing images. The main advantages of this method are its simplicity, its computational efficiency and its invariance to affine transformation, as well as to occlusion, lighting, and intra-class variations. Although it does not exploit the geometric information of the components, surprisingly, this method produces good categorization accuracy. In our framework, we apply the bag-of-words model as the first step to select candidate categories, and further fuse the score of each candidate category with learned text concepts in the second step to improve the overall categorization accuracy. Details of the bag-of-words model and its variations will be discussed in Section 4.

2.2 Embedded Text Detection

Embedded text detection is to find the location of texts in images or videos. Texts are usually designed to attract attention and to

reveal information. Therefore, they usually present strong coherence in space, geometry and color. Previous work, primarily relying on image processing techniques to extract these characteristics for the text detection task, can be roughly divided into two leading approaches: the connected component (CC)-based and the texture-based approaches. The CC-based methods [15] basically segment an image into a set of CCs, group them into potential text regions according to their geometric relations, and then examine these potential text regions by some rule-based heuristics which utilize characteristics such as the size, the aspect ratio and the orientation of the region. While these methods are efficient in finding texts, the CC-based methods encounter problems when the text is multi-colored, textured, with a small font size, or overlapping with other graphical objects. The texture-based methods [16] assume that texts have distinct textural properties, which can be used to discriminate them from the background. These methods perform well even for images with noisy, degraded, or complex texts and/or background. However, the texture-based methods are usually time-consuming as texture analysis is inherently computational-intensive. A comparison of various text detection algorithms is summarized in [17]. However, it is hard to make direct comparisons among these methods because different approaches have been reporting their results using different evaluation metrics and different datasets.

In this paper, we use the embedded-text detector developed in our previous work [18] which has a high detection rate even though the false positive rate might also be high. The requirement of the text detector for supporting the proposed framework is that at least one of the text regions which could represent the concept of the image (For example, the license plate in a vehicle image) shall be detected by the text detector. A slightly higher false positive rate is acceptable and can somewhat be tolerated by our text concept learning process which will be described in Section 5.1.

2.3 Multimodal Fusion

Multimedia data, such as images and videos, are represented by features from multiple sources. For example, videos are represented by features which can be extracted from visual, audio, and caption channels. In addition to visual features, the camera metadata, which are embedded in the digital image files, could provide cues for contextual information [2, 3]. However, as discussed earlier, the information contained in camera metadata is valuable only to a limited subset of image categories. It is also common that such camera metadata is missing in on-line image storage. Therefore, it will be highly beneficial to develop more cues, e.g. text information, with discriminative power for the image categorization problem.

It has been studied extensively in the past to derive text information to help manage the video database [19, 20]. The OCR from the captions provides rich cues for efficient indexing, searching and categorizing a video clip. One key reason that such clues can be robustly derived is because the artificial captions in a video frame can be localized and recognized relatively easily, On the contrary, performing OCR on natural scene images is a very challenging task. In [21], the author proposed to extract text-related features from images attached in an email and use such features to categorize spam emails. Note that the embedded texts in such an application are artificial, and thus reliable text detection is not as challenging as that for natural scene images. In their approach, they simply incorporate the information derived from all the text lines detected in the target image into a support vector machine without the step of learning a text-concept in each target category.

3. IMAGE DATABASE

We have collected an image database which contains around 2000 images, about half of which were collected from the web and the remaining were generated by ourselves - with pictures taken in local towns and on campus. The images are categorized into eight groups. Table 1 summarizes the image categories and statistics. The reason we built our own database is that the multimodal fusion framework involves both image categorization and text detection. None of the public databases available in both fields meets our specific requirement, i.e. a given image must contain some embedded text. We manually labeled the text lines in each image, and the average number of text lines per image in each category is reported in the 3rd column of Table 1. Among the eight categories, categories 'Highway' and 'Street' can be interpreted as scenes while categories 'Camera', 'Car' and 'Door' contain images mostly presenting a single object with a clear background. For categories 'Airplane', 'Building' and 'Bus', the images contain a fair amount of background clutters and scale variations. Therefore, these three categories can be interpreted as either a scene or an object. Figure 2 shows a few sample images, where labeled texts are highlighted as colored rectangles. For the experiments, we randomly chose 70% of the images for training and the remaining 30% for testing.

Table 1: Summary of image database

Category name	Number of images	Average number of text lines per image
Airplane	260	1.7
Building	207	1.2
Bus	294	5.2
Camera	260	2.1
Car	220	3.3
Door	200	1.6
Highway	205	3.4
Street	250	1.6

Observing from the images in the database, we noted the following: (I) For some pairs of categories such as categories 'Highway' vs. 'Street', the classification is inherently challenging; (II) For some image categories, the colors, the sizes or the locations of the text follow some uniform patterns for images in the same category. For example, one or multiple text lines in images of the 'Highway' category have white text on green background which are from the signs on the highway; (III) Most images contain multiple text lines. The average number of text lines is as high as 5.2 for the 'Bus' category. The inherent challenging problem addressed by this paper presents an example of a variable-length and unordered feature set: different images have different numbers of text regions and the detected text lines are unordered.

4. BAG-OF-WORDS MODEL

In this section we discuss the bag-of-words model for image categorization, and some variations of each component in this

model. We also describe the implementation and the results which are considered as the baseline performance, and will be further compared to our proposed method in the experimental section.

4.1 Overview

In the formulation of the bag-of-words model, each image is represented as a collection of patches, each of which corresponds to a *visual word* in a pre-defined large dictionary. An image is described by a histogram over the visual words. A generative model, such as a Naïve Bayesian classifier, is built based on given training images for each image category. For recognition, we calculate the histogram of the image under classification, feed it to each classifier and obtain a score (logarithmic posteriori probability). The image is then labeled as the category C_i with the highest score. Three key factors affect the performance of the bag-of-words model: (1) The choice of the image patch description; (2) Vocabulary construction; (3) The classifiers. In the following, we discuss these components in more details.

Figure 2: A few sample images from our database

4.2 Features & Vocabulary

Although there exist a number of descriptors to represent an image patch, the SIFT descriptors [10] are most popular for its effectiveness to extract distinctive and invariant features. Given an image patch, the SIFT algorithm computes Gaussian derivatives at 8 orientation planes over a 4x4 grid of spatial locations, resulting in a 128-D vector. These vectors have been shown to be robust in matching across an affine distortion, change in 3D viewpoint, addition of noise, and change of illumination.

Given a collection of SIFT descriptors from the training images of all categories, the next question is how to select the most representative ones and form a codebook? A conventional way is to *learn* a codebook by performing the k-means algorithm on the descriptor database, and choose the centers of clusters as the visual words. Recently, more sophisticated clustering and vector quantization algorithms have been proposed. In [22], Jurie and Triggs combined the advantages of on-line clustering and mean-shift into a unified framework. The algorithm produces an ordered list of centers, with the quantization rule that patches are assigned to the first center in the list that lies within a fixed radius of them. If the descriptors are distributed non-uniformly in the high-dimensional space, their algorithm might outperform the k-means. Winn and Minka [23] learned a universal visual dictionary by pair-wise merging of visual words from an initially large dictionary. More interesting, the size of the dictionary can be automatically determined during the learning process. However, this method requires very fine labeling for all training images. It might be as tedious and expensive as manual annotation for which each of the images is hand-annotated with a category label.

4.3 Classifiers

Naïve Bayes is a simple classifier commonly used in pattern recognition. While it has the assumption that the feature attributes are independent, the accuracy of the Naïve Bayes classification is typically high [24].

Consider again the bag-of-words model, each image I is represented by a histogram over the visual words $W = \{\mathbf{w}_i\}$, $i = 1$ to N, where N is the size of the dictionary. We denote $n(t, i)$ the number of visual word $\mathbf{w}_t$ occurs in image I_i. To categorize a testing image, the Naïve Bayes decides category c by the maximum a posteriori rule:

$$c^* = \arg\max_c \; p(c \mid I_i) \propto p(c)\,p(I_i \mid c)$$
$$= p(c)\prod_{t=1}^{N} p(w_t \mid c)^{n(t,i)}$$

The class-conditional probabilities $p(w_t \mid c)$ can be estimated from the training data:

$$p(w_t \mid c) = \frac{1 + \sum_{\{I_i \in c\}} n(t,i)}{N + \sum_{s=1}^{N}\sum_{\{I_i \in c\}} n(s,i)}$$

Here the estimates are calculated with Laplace smoothing in order to avoid the probabilities becoming a zero.

More complex hierarchical Bayesian models, such as the Probabilistic Latent Semantic Analysis (pLSA) model and the Latent Dirichlet Allocation (LDA) model, have been successfully applied to the problem of image categorization [7] from the statistical text literature.

4.4 Experimental Results – The Baseline

The implementation of the bag-of-words model is based on the Matlab code released from the ICCV 2005 short course on "*Recognizing and Learning Object Categories*" [25] with slight modification. We chose SIFT as the feature descriptors, and

constructed a dictionary with 200 visual words by the k-means. A Naïve Bayesian classifier was trained for each category by randomly selecting 70% of our image dataset as the training samples. Table 2 shows the confusion matrix on our testing dataset containing the remaining 30% images, where the overall categorization accuracy is **81.32%**. This implementation achieves a comparable accuracy using the same parameter setting on the image database used in [7]. Note that the confusion matrix in our work is slightly different from the definition given in [7]. Our classification is based on the criterion of choosing the most probable category, while in [7] they select an equal-error-rate point in the ROC curve of each category for classification.

Table 2: Confusion matrix with the bag-of-words model

	air	bldg	bus	cam	car	door	hwy	strt
Airplane	**84.6**	0.00	5.13	0.00	3.85	0.00	5.13	1.28
Building	1.61	**46.8**	1.61	1.61	9.68	12.9	0.00	25.8
Bus	0.00	0.00	**95.5**	0.00	3.41	0.00	0.00	1.14
Camera	0.00	3.03	0.00	**92.4**	4.55	0.00	0.00	0.00
Car	1.28	2.56	2.56	1.28	**84.6**	3.85	0.00	3.85
Door	1.67	6.67	1.67	1.67	6.67	**78.3**	1.67	1.67
Highway	1.64	1.64	3.28	0.00	3.28	4.92	**80.3**	4.92
Street	0.00	2.67	0.00	0.00	5.33	0.00	4.00	**88.0**

*Row: eight image categories and the sum of each row is 100%.
*Column: eight object classifiers trained for each category

Table 2 indicates that the bag-of-words model categorizes most images correctly among the categories of *airplane*, *bus*, *camera*, *car* and *street*, but not among the others, especially the "Building" category. Only 46.8% of the building images are correctly classified, and a good fraction of them, say 25.8%, are confused with the street images. The reason of its low accuracy is that some common objects, such as trees, are present in images of both categories. Thus, they have overlapped visual words which confuse the model and further lead to a poor classification accuracy.

5. MULTIMODAL FUSION

The bag-of-words model is used as the first step in our system. The confusion matrix, shown in Table 2, is constructed based on the criterion that the most probable category is chosen as the classification result. In Figure 3, we show the detection accuracy as the cumulative sum over the number of categories we choose, which is referred to as the Top-K strategy. Obviously, for our dataset which has 8 categories, the Top-8 strategy achieves a 100% detection rate while it does not provide any useful information about the true image category. The curve presented in Figure 3 indicates that we could use the Top-K strategy to eliminate a number of categories without compromising much on the detection rate. For example, the Top-3 strategy maintains a very high detection rate (97%) while eliminates 5 image categories from Step 2 of the classification process.

Step 2 uses an SVM-based classifier to refine the classification result (i.e. choosing one out of the K categories selected in Step 1) by incorporating the information derived from the detected texts. A fundamental question in designing the classifier for Step 2 is: should we use one multi-class SVM classifier, or K binary "one per class" SVM classifiers (OPC), or K × (K-1)/2 pair-wise binary SVM classifiers (PWC). In [26], a study was made on how the binary SVM classifiers can be effectively combined to tackle the multi-class image classification problem. Their experiments concluded that PWC achieves the best accuracy for image classification. The main disadvantage of PWC is the need of K×(K-1)/2 classifications for each testing image. As K increases, the computation cost becomes the major limitation. However, in our framework, this limitation is alleviated through the use of the Step 1 as a filtering process for Step 2. The experimental results show that the Top-3 strategy achieves the best performance, in terms of both accuracy and computation. More details will be provided in Section 5.2.

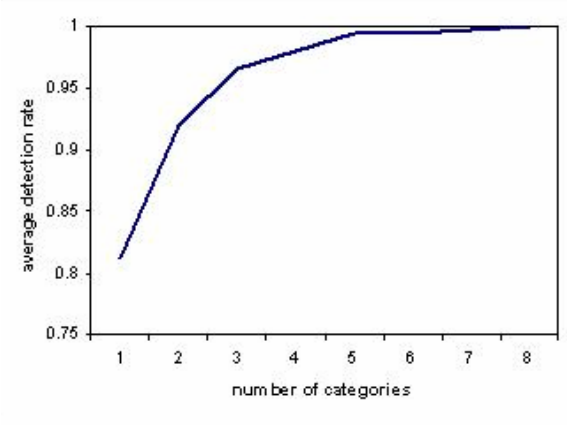


Figure 3: Accumulated detection rate over Top-K

In Step 2, a critical task is derivation and integration of text-related information. Specifically, we learn a text concept from the text lines labeled in the images of each category using *Multiple Instance Learning*. The learned concept is then used to extract features from the detected text lines from the image under classification. In the following, Section 5.1 discusses this approach in great details, followed by the specification of a pair-wise binary SVM classifier in Section 5.2.

5.1 Text Concept Learning

Text commonly appears as part of the scene/object in images. For example, a building often has its name on the wall. The picture of a street often contains road signs showing the street names. Typically, the text detection is used as a pre-processing step for OCR. From the human perception point of view, it might be interesting to directly utilize the text detection result to help image classification. Note that if the image classification accuracy is high, the knowledge of the image category can be further used to help improve the accuracy of the OCR results.

In this section, we address the following two problems: (I) designing a set of features to describe a text line detected in the image under classification; (II) learning a text concept for an image category to help image classification.

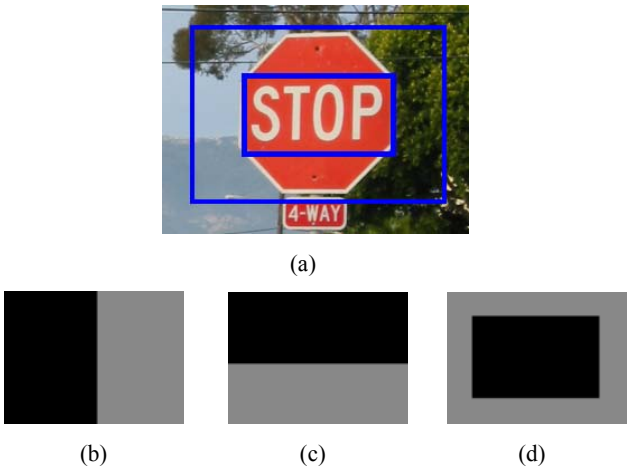
5.1.1 Feature

A text line in an image can be viewed as an image patch identified by a text detector. One critical task for utilizing the text line information for classification is to design a set of effective features to describe this image patch. Such features should capture common patterns within a wide range of text lines found in images of the same category. As the text patches are typically

small, instead of using a large set of complicated features, we design sixteen simple features only, which are categorized into five groups:

- Text color: we use the H value in the HSV color space as a feature. Segmentation is required to differentiate the foreground text from the background in a text line.
- Text size and location: the width, the height, the aspect ratio, location X and location Y of the text line are included in the feature set.
- Edge density: we use Canny edge detection to produce an edge map for computing the edge density.
- Global textures: this group of features includes brightness, contrast, mean and variance extracted from the text region.
- Local structures: a local window is defined around the text region (illustrated in Figure 4(a)), and three local structures based on the local window (illustrated in Figure 4(b-d)) are further defined. For each local structure, the intensity difference between the two sub-regions within the local window is evaluated as a feature.

One may argue that the OCR result could be a good cue for image classification. The reasons we exclude it from the feature set are twofold: (I) The OCR accuracy for texts embedded in natural scenes could be low due to the perspective deformations under different views. (II) The recognized words in the target category could be too diverse, e.g. restaurant names on the building or street names on the road sign, to be useful for classification.



(a)

(b) (c) (d)

(a) The small rectangle contains the text line, and the big rectangle defines its local window; (b) The horizontal structure; (c) The vertical structure; (d) The boundary structure

Figure 4: Local window and its three local structures

5.1.2 Learning

A straightforward way of incorporating the text information is directly using the 16-D feature vector for the SVM-based classification in Step 2. However, such an intuitive solution is not feasible because an image may contain multiple text lines. We could either choose a "representative" one from the multiple text lines or use all of them for computing the feature vector. However, choosing the most "representative" one is a nontrivial problem, while concatenating them together would result in a variable-size feature vector which cannot be handled by an SVM-based classifier. To address this problem, we propose a *text concept learning* algorithm.

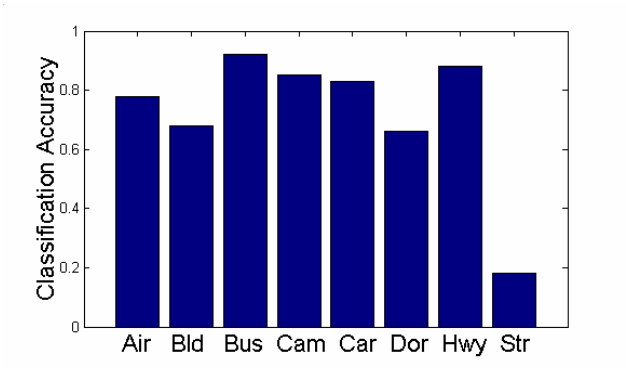
Text Concept: The text lines from images of the same category, each of which is represented by a 16-D feature vector, should follow some common patterns. Specifically, such a common pattern, referred to as a *text concept* in this paper, is defined as a 16-D vector $T : \{t_i\}_{i=1}^{16}$ associated with weights $W : \{w_i\}_{i=1}^{16}$. Thus, a weighted Euclidean distance between a text concept and any text line, represented by its 16-D feature vector, can be computed. Given a testing image containing N text lines and a text concept of an image category, we use the minimum distance, among the N distances, as the metric for measuring the similarity between the testing image and the text concept. Since one text concept is learned for each image category, such information can be effectively integrated with other information for improving image classification. The main advantages of learning a text concept are twofold: (1) the text line with the minimum distance to the text concept is chosen as the "representative" one among all text lines so we effectively remove the irrelevant text lines detected in an image; (2) the approach implies a significant reduction of dimensionality from a 16-D feature vector to a single value of the weighted Euclidean distance.

Multiple Instance Learning (MIL): The MIL [27] was proposed as a variation of supervised learning for the problem with ambiguous labels on training examples. The study on Multiple Instance Learning was first motivated by the problem of predicting the drug molecule activity level. Afterwards, this idea was refined and applied to a wide spectrum of applications. In particular, Maron and Lozano-Perez [1] applied this idea to image concept learning using the Diverse Density (DD) algorithm. In their formulation, an image is denoted as a positive bag if it is from the target category while a negative bag represents an image from other categories. Each image contains a number of small patches, where each patch is represented as a feature vector and called an instance in that bag. Thus each bag is a collection of instances. A bag is labeled negative if all the instances in it are negative, and positive if at least one of the instances in the bag is positive. Such a framework is used to model the ambiguity in mapping an image to many possible templates which describe the image. The key idea of the Diverse Density approach is to find a concept point in the feature space that is close to at least one instance from every positive bag and meanwhile far away from all instances in negative bags. The optimal concept point is located at the position with a maximum value of Diversity Density, which measures how many different positive bags have instances near the point, and how far the negative instances are away from that point. For our application, if text lines are treated as image patches, this image concept learning framework can be directly applied to finding the text concept for an image category. Note that the DD algorithm allows us to find the best weights for the initial feature set automatically. In our experiments, the implementation of the DD algorithm is taken directly from a Multiple Instance Learning Library (MILL) [28].

5.1.3 Learned Concepts

Using Multiple Instance Learning, we learn eight text concepts with respect to eight image categories in our database. When learning a text concept for an image category, all images in the

target category are treated as the positive samples/bags while we randomly sample an equal number of images from the other seven categories as the negative samples/bags. Each learned text concept is represented by the point with maximum Diversity Density in the 16-D feature space, where the located point is close to at least one instance, i.e. one text line, from every positive image and meanwhile far away from all instances in negative images.



** Average accuracy is 72.3%, or 80.1% if the "Street" category is excluded. The bag-of-words model achieves 81.3% in average*

Figure 5: Detection accuracy using the "text concept" alone

Given a testing image containing N text lines and a text concept of an image category, we use the minimum distance, among the N distances, as the metric for measuring the similarity between the testing image and the text concept. Based on this metric, we obtain eight minimum distances, one for each of the eight learned text concepts for the eight image categories. Therefore, we can directly apply this metric for image classification, i.e. choosing the category with the smallest value among the eight minimum distances. Figure 5 shows the classification accuracy for the eight categories entirely based on this simple metric. The resulting accuracy explicitly shows the discriminative power of the text information for image classification. The result shows that a reasonably high accuracy is achieved for most image categories. One exceptional case is the "Street" category, for which only 18% of the images are correctly classified. One explanation behind this low accuracy is that street scene usually contains a mixture of text types, such as road signs, banners, shop/building names and etc. Thus, learning of a uniform pattern for the "Street" category is very challenging, if not impossible. Although satisfactory classification results are achieved for most categories, using text information alone does not achieve a higher accuracy than using a bag-of-word model. Therefore, a framework of fusing both sources of information together is developed to further improve the accuracy.

Note that our text concept leaning involves a set of weights, which reflects the impact of each feature in the 16-D vector. Table 3 presents the Top-3 features with maximum weights for learning eight text concepts. We observed that the result is consistent with some simple findings discussed earlier. For example, color is the most important cue for the "Highway" category, which reveals the fact that the signs on the highway typically have white text on green background. Another example of the coherence is the importance of the aspect ratio in the "Car" category due to the existence of a license plate in a good fraction of the car images. In general, the edge density, the location and the local structures are among the most frequent features in Table 3.

Table 3: Most important text feature for learning a concept

Category	Top-3 features with maximum weights
Airplane	{location, contrast, edge density}
Building	{edge density, location, local structure}
Bus	{edge density, location, aspect ratio}
Camera	{text size, edge density, local structure}
Car	{aspect ratio, location, edge density}
Door	{aspect ratio, edge density, variance}
Highway	{color, mean, local structure}
Street	{local structure, location, edge density}

5.2 Pair-wise SVM

Based on the bag-of-words model, the Top-K strategy used in Step 1 eliminates the image categories with low probabilities from further analysis. That is, in Step 2, only the K categories with the highest probabilities, P_1, P_2..., P_K, are considered. We then calculate K minimum distances, D_1, D_2..., D_k, between the image under classification and the text concepts of the K categories.

Step 2 adopts pair-wise SVM classifiers using a linear kernel for classification, instead of using a single multi-class SVM classifier.

Specifically, for any pair of Categories i and j, we assemble seven features as the input for a pair-wise classifier SVM_{CiCj}: P_i, P_j, P_i/P_j, M, D_i, D_j and D_i/D_j, where M is the number of the text lines detected in the image. In addition to the two probabilities, P_i and P_j, and the two distances, D_i and D_j ,their ratios, P_i/P_j and D_i/D_j, are included as well. In total 8×(8-1)/2, i.e. 28, pair-wise SVM classifiers are trained for the eight image categories. Note that SVM_{CiCj} and SVM_{CjCi} in fact share the same hyper-plane in the feature space. Therefore, we remove the redundant information by limiting i<j for SVM_{CiCj}. For testing, as we only consider the Top-K categories, K×(K-1)/2 binary classifications will be performed, using the K×(K-1)/2 relevant SVM classifiers. After that, a majority voting is used to make the final decision. In the case of having equal voters, we take the confidence, provided by the scores reported by the SVM classifiers, into account.

6. EXPERIMENTS

6.1 Classification Results

We first examine the impact of the parameter K for the Top-K strategy. Although a larger K implies a lower accuracy loss in Step 1, it inevitably complicates Step 2 and may not necessarily improve the overall accuracy, in addition to a higher computation cost. In Figure 6, the curve of the average accuracy vs. K is plotted. The curve indicates that the Top-3 strategy results in the highest accuracy. In later experiments, we only present the classification accuracy using the Top-3 strategy.

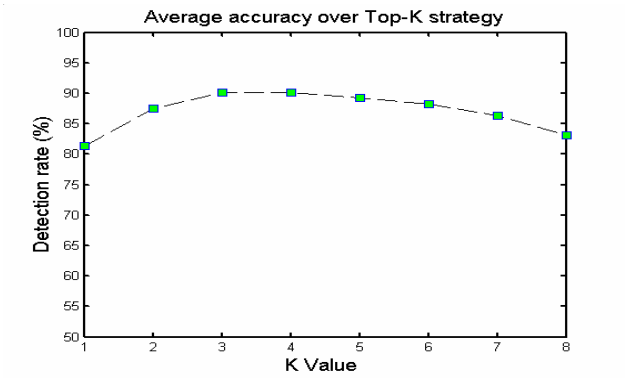


Figure 6: Average accuracy curve in terms of K

Table 4 shows the confusion matrix in the same format used in Section 4.4. The overall categorization accuracy is improved from **81.3%,** which uses a bag-of-words model, to **90.1%** by a multimodal fusion approach. Figure 7 shows a direct comparison between the bag-of-words model and the proposed two-step fusion framework for each of the eight categories. That is, Figure 7 compares the diagonal elements of the two confusion matrixes of Tables 2 and 4. The results indicate that the multimodal fusion improves the classification accuracy for 7 of the 8 categories, with the best improvement for the "Building" category (an improvement of 37.1%) and an average improvement of about 8.8% for the 8 categories. This level of improvement clearly demonstrates that the fusion of multimodal information would benefit image categorization. The accuracy of the "Street" category is slightly dropped by the proposed approach. We believe that this slight performance degradation is mainly caused by the failure of learning a concrete text concept for this category.

Table 4: Confusion matrix using the two-step fusion framework with the Top-3 strategy in Step 1

	air	bldg	bus	cam	car	door	hwy	str
Airplane	**96.2**	1.28	0.00	1.28	1.28	0.00	0.00	0.00
Building	3.23	**83.9**	0.00	1.61	3.23	1.61	1.61	4.84
Bus	0.00	1.14	**97.7**	0.00	1.14	0.00	0.00	0.00
Camera	1.52	1.52	0.00	**94.0**	0.00	0.00	0.00	3.03
Car	0.00	5.13	5.13	1.28	**88.5**	0.00	0.00	0.00
Door	1.67	1.67	1.67	0.00	6.67	**86.7**	0.00	1.67
Highway	0.00	0.00	0.00	1.64	4.92	3.28	**86.9**	3.28
Street	0.00	8.00	0.00	0.00	4.00	0.00	2.67	**86.7**

6.2 Discussions

In the following, we summarize and discuss our framework in terms of three design factors, proposed in [29], that would affect fusion performance.

Modality independence: Our fusion is based on the information derived from two different models: a bag-of-words model using low-level visual cues and a text concept learned from a set of labeled text lines using Multiple Instance Learning. These two models, which serve for completely different purposes and use distinct feature sets, are inherently independent.

Curse of dimensionality: A pair-wise SVM classifier, responsible for fusing information together, only takes seven features as the input. Therefore, curse of dimensionality should not be a problem for our scheme.

Fusion-model complexity: The super-kernel fusion, proposed in [29], used SVM classifiers in both steps, while our two-step fusion is based on a bag-of-words model in Step 1 and pair-wise SVM classifiers in Step 2. This combination takes full advantage of a successful generative approach, i.e. a bag-of-words model, as well as the powers of a discriminative approach, i.e. SVM-based classifiers, for image categorization. In addition, the pair-wise design in our approach allows a linear SVM classifier with seven features to effectively differentiate a pair of categories.

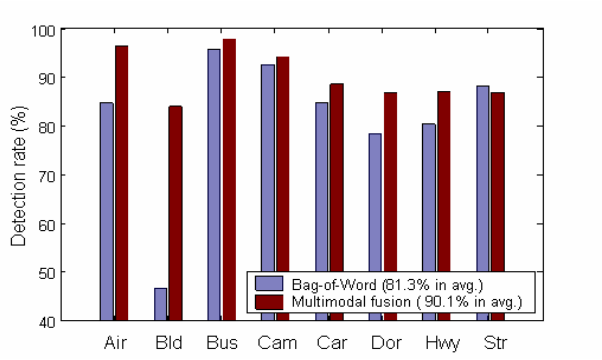


Figure 7: A direct comparison of the classification accuracy

7. APPLICATIONS

Inspired by several practical systems based on image categorization and text detection technologies, we believe this proposed work particularly benefits two applications.

First, in scene understanding, the OCR text can be better interpreted by incorporating the information regarding image category into the final result. Figure 8 shows a few typical examples from our dataset. The same text can be associated with different objects, or extracted from different scenes. For example, a number can be interpreted as an apartment number, a bus route, or a part of a car license plate depending on the associated object/scene. Figure 9 shows another example of applying our technique. We examined a number of sample images from our "Camera" category. Most texts in this category contain relevant information, such as camera brands and camera models. The integration of the text detection/OCR results into image categorization would lead to a powerful image annotation system, which in turn would allow us to retrieve cameras with a certain brand or a special model.

8. CONCLUSIONS

We demonstrate a multimodal fusion framework that follows a two-step process. Step 1 uses a bag-of-words model which fully utilizes the low-level visual features. In Step 2, we incorporate features derived from the detected, embedded text lines into pair-wise SVM classifiers to improve the initial result of Step 1. Central to the success of our approach is that the learning of *text concepts* from a large set of labeled text lines from the training images using Multiple Instance Learning, and a smart design of the pair-wise SVM classifier in Step 2. For our database

containing eight image categories, the overall categorization accuracy is improved from **81.3%,** which uses a bag-of-words model, to **90.1%** by the proposed multimodal fusion approach.

The key value of this work is a fusion framework which eliminates the ambiguity and bridges the information in two different fields, namely, image categorization and text detection. We demonstrated that, for many real-world applications, such a fusion system could produce more valuable information, e.g. a number associated with a car license plate, which in turn would result in considerable improvement to image categorization accuracy. However, we should point out that the database used in the experiments is specific and the size is limited. In addition, for those concepts/categories which contain no text, the gain of the proposed fusion framework would be limited.

9. REFERENCES

[1] Maron, A. L. Ratan. Multiple-Instance Learning for Natural Scene Classification. *Proc. 15th Int. Conf. on Machine Learning (ICML 98)*, Madison, USA, 1998.

[2] M. Boutell and J. Luo. Bayesian Fusion of Camera Metadata Cues in Semantic Scene Classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 04)*, Washington, DC, USA, 2004.

[3] Y. Wu, E. Y. Chang and B. L. Tseng. Multimodal Metadata Fusion Using Casual Strength. *ACM International Conference on Multimedia (MM 05)*, Singapore, 2005.

[4] J. Gao and J. Yang. An Adaptive Algorithm for Text Detection from Natural Scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 01)*, Hawaii, USA, 2001.

[5] X. Chen and Alan L. Yuille. Detecting and Reading Text in Natural Scenes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 04)*, Washington,, USA, 2004.

[6] Wen Wu, Xilin Chen and Jie Yang. Incremental Detection of Text on Road Signs from Video with Application to a Driving Assistant System, *ACM International Conference on Multimedia (MM 04)*, New York, USA, 2004.

[7] L. Fei-Fei and P Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)*, San Diego, CA, USA, 2005.

[8] Ulusoy, I. and C. M. Bishop. Generative Versus Discriminative Models for Object Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05)*, San Diego, CA, USA, 2005.

[9] Alex D. Holub, Max Welling, Pietro Perona. Combining Generative Models and Fisher Kernels for Object Recognition*, IEEE International Conference on Computer Vision (ICCV 05)*, Beijing, China, 2005.

[10] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.

[11] Y. Chen, J. Li and J. Z. Wang. Machine Learning and Statistical Modeling Approaches to Image Retrieval, Published by Kluwer Academic Publishers, 2004

[12] T. Kadir and M. Brady. Scale, Saliency and Image Description. *International Journal of Computer Vision*, 2001.

[13] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1615–1630, 2005.

[14] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray. Visual Categorization with Bags of Keypoints. *In ECCV'04 Workshop on Statistical Learning in Computer Vision*, 2004.

[15] S. Loncaric, A survey of shape analysis techniques. *Pattern Recognition*, vol. 31, pp. 983–1001, 1998.

[16] K. I. Kim, K. Jung, and J. H. Kim. Texture-based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm. *IEEE Transactions on PAMI*, 2003.

[17] Simon M. Lucas. ICDAR 2005 Text Location Competition Results. *In Proceedings of the IEEE Conference on Document Analysis and Recognition*, vol. 1, pp. 80-84, 2005.

[18] C. T. Wu, Embedded-text Detection and its Application to Anti-Spam Filtering, Master Thesis,, 2005 University of California, Santa Barbara, CA

[19] Rainer Lienhart and Wolfgang Effelsberg. Automatic Text Segmentation and Text Recognition for Video Indexing. *ACM Multimedia Systems*, Vol. 8. pp.69-81, January 2000

[20] D.Q. Zhang and S.F. Chang. A Bayesian Framework for Fusing Multiple Word Knowledge Models in Videotext Recognition, *IEEE Computer Vision and Pattern Recognition (CVPR 03)*, Madison, Wisconsin, June, 2003

[21] H. Aradhye, G. K. Myers, J. A. Herson: Image Analysis for Efficient Categorization of Image-based Spam E-mail. *ICDAR 2005*, Seoul, Korea, August, 2005

[22] F. Jurie and B. Triggs. Creating Efficient Codebooks for Visual Recognition. *IEEE International Conference on Computer Vision (ICCV 05)*, Beijing, China, 2005.

[23] J. Winn, A. C. and T. M. Object Categorization by Learned Universal Visual Dictionary. *IEEE International Conference on Computer Vision (ICCV 05)*, Beijing, China, 2005.

[24] P. Domingos and M. Pazzani. On the Optimality of Simple Bayesian Classifier Under Zero-one Loss. *Machine Learning*, 1997.

[25] Fei-Fei Li, Rob Fergus, Antonio Torralba, Recognizing and Learning Object Categories, ICCV 05 short course, Beijing

[26] K-S Goh, E. Chang and K-T Cheng. Support Vector Machine Pairwise Classifiers with Error Reduction for Image Classification. *The 3rd Intl Workshop on Multimedia information retrieval (MIR2001)*, Ottawa, Canada, 2001

[27] T. G. Dietterich, R. H. Lathrop, T. Lozano-Perez. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence Journal*, pp-89, 1997.

[28] Jun Yang, MILL: A Multiple Instance Learning Library, http://www.cs.cmu.edu/~juny/MILL

[29] Y. Wu, E. Y. Chang, K. C.-C. Chang, and John R. Smith. Optimal Multimodal Fusion for Multimedia Data Analysis. *ACM International Conference on Multimedia (MM 04)*, New York, USA, 2004

Figure 8: Some sample images of image categorization with the text detection output

Figure 9: A few examples from the "Camera" category with labeled texts of camera brands and models