Visual Concept Modeling Scheme Using Early Learning of Region-based Semantics for Web Images

(ウェブ画像に対する領域ベースのセマンティックマイニングによるビジュアルコンセプトのモデリングに関する検討)

Yongqing Sun†, Satoshi Shimada (member)†, Masashi Morimoto††, Yukinobu Taniguchi (member)‡

Abstract In this paper we present a novel approach to modeling visual concepts effectively and automatically using web images. The selection of training data (positive and negative samples) is strongly related to the quality of learning algorithms and is an especially crucial step when using noisy web images. In this scheme, first, images are represented by regions from which training samples are selected. Second, region features effectively representing a semantic concept are determined, and on their basis, the representative regions corresponding to the concept are selected as reliable positive samples. Third, high quality negative samples are determined using the selected positive samples. Last, the visual model associated with a semantic concept is built through an unsupervised learning process. The presented scheme is completely automatic and performs well for generic images because of its robustness in learning from diverse web images. Experimental results demonstrate its effectiveness.

Key words: Web image mining, visual concept model, image learning.

1. Introduction

Semantic model-based indexing and searching has been proposed as an effective method of managing large multimedia archives at the semantic level. As part of this general problem, image learning and visual concept-based categorization has been actively studied in industry and academia.

Modeling semantic concepts within a limited and static database has been comprehensively investigated and much success has been achieved. The basic framework for building a visual model can be summarized as the following three steps: (1) Defining a category or vocabulary including desired semantic concepts, (2) Manually collecting a large amount of training images associated with these concepts, and (3) Training a classifier based on the training data. The semantic concept of a new instance is recognized based on the classifier. This framework has produced a rich body of work. For example, Maron proposes a supervised method based on Multiple Instance Learning (MIL) for learning common color and spatial properties of a set of images. In the learning process, one image is viewed as a positive sample if there is at least one positive instance in it, and it is viewed as a negative sample if no instance in it is positive. In recent years, learning within a static database performs well in semantic concept detection on the TRECVID evaluation benchmark that was established in 2001 to evaluate and promote research in video retrieval.

However, the challenging problems that still exist in such learning methods limit performance of visual concept modeling in the real world. Manually collecting and labeling training data is very time- and cost-demanding work. As a result, these methods are ill-suited to practical use such as large-scale web image search and real-world image annotation since such semantic concept models are built based on limited amounts of data and a static database.

Since there is a huge amount of images available in current search engines such as Yahoo! and Goo, directly exploring the rich resources offers a promising solution for image learning and clustering. Because visual concepts in the real world, especially on the web, are characterized both as textual and visual informa-
tion, using both kinds of information to build a visual concept model is more semantically meaningful. Therefore, learning from web images to build visual concept models has recently attracted a lot of interest as a topic of research.

For example, the work in proposes an approach to unifying textual and visual statistics information in a single index vector for web image clustering. By building possible statistical links between the web document and images, the approach shows its validity in image retrieval. However, simply combining visual and textual information in a single index raises the issue of appropriate characterization of web images.

In more recent work, web images are first gathered using an object name such as “airplane” and are all regarded as positive samples, then the proposed model called TSI-pLSA is used to learn an object category from the noisy web images. However, directly selecting positive samples from such noisy data and only using them to build model impedes the robustness of visual concept building.

A method proposed in builds probabilistic model for extracting visual patterns using web images. It first selects positive and negative samples from the web images gathered by a query word, then builds a probabilistic model for visual concepts based on linear Gaussian distribution.

Appropriate selection of training data (positive and negative samples) is strongly related with the quality of visual concept modeling, which is especially crucial and challenging step when using noisy contaminated web images. Although the existing image learning methods using web images have shown somewhat effectiveness in building visual models, many of them still have limitations, which have been mentioned in the above.

In this paper we propose a novel approach that utilizes web image resources effectively and automatically to build visual models. Our goal is to tackle the above mentioned problems and improve the quality of image learning from web images.

In our work, images are segmented and represented by a set of regions from which training samples are selected.

It is observed that among images related with a semantic concept, usually there is a distinctive object in almost every image. It is called representative object and expected to describe the semantic meaning. It is ideal to segment representative objects from these images and use them for model building, but precise object segmentation still keeps challenging problem. However, it is also observed that employing current segmentation methods such as the work, representative object in one image may be separated into several parts (called as regions), most of which share similar features such as color, shape and texture. In other words, a representative object consist of assembled multiple regions, which contribute representing power to the concept. Based on this fact, we detect such regions and regard them as positive sample in our work. For example, for “tiger” concept, in one image its representative object—tiger may be separated into regions like head, legs, body. But from all the semantic-related images, such regions are detected and gathered together as positive samples. As a result, in some sense, the whole positive samples consisting of different parts of tiger are expected to represent semantic meaning completely at last.

Regarding the selection of negative samples, compared with either generating them randomly or treating them as a single distribution, we make an effort to select them more reliably utilizing the property of positive samples.

After the determination of training samples, the visual model associated with a semantic concept is built through a learning process. The proposed method performs well on generic images because of its enhancement in learning from noisy web sources, and it needs no human invention.

In summary, the main feature of this approach is that it challenges the problem of selection of proper training samples from noise web images for building reliable visual concept models. In term of using positive and negative samples, our approach is kind of similar to the work. However, since it did not put much effort on the selection of training samples, its coarse determination of training data may affect the convergence and robustness property of the model building process. In the positive samples irrelevant samples such as various “background” regions are always included during the model building process. Besides, generating negative samples is conducted randomly. Compared with the work, we make an effort to select regions corresponding to “representative object” of a concept as the positive samples and carefully select the negative samples by utilizing the property of positive samples. Therefore, compared with the work, good quality of visual concept model is expected to be obtained from noisy web images by our method.

The remainder of this paper is organized as follows. In
section 2, we give a brief overview of the whole scheme. In section 3 a visual concept is built through a learning approach automatically, which is based on the reliable selection of positive and negative samples. The experimental results are discussed in section 4, and the last section is the conclusion.

2. System Overview

The proposed method is depicted in Fig.1.

Given a query keyword, a set of images are automatically gathered from the web search engines. Note that because the image gathering is conducted in a text-based search, not by image content matching, positive images relevant to the query word and negative images irrelevant to the query word are always mixed within the retrieved images. Therefore, it is not so easy to divide them directly into purely positive or negative images.

However we have an empirical observation for the distribution property of the gathered web images as the following. Due to the expertise of current web engines, images retrieved in the top few pages tend to be “positive” although there are a few irrelevant images included. Here we call them as “positive candidate images”. Images retrieved in the latter pages tend to be “negative” although a few irrelevant images are unavoidably mixed. Here we call them as “unlabeled images”. Based on this observation, our scheme directly utilizes these two groups as a rough but quick start because it is very easy to obtain them.

In this work, since training samples (positive and negative) for model building indicate image regions, by the following processing steps in our scheme, positive and negative regions will be carefully selected from the regions of “positive candidate image” and “unlabeled images”, respectively. Therefore, even “positive candidate images” or “unlabeled images” are not purely at the beginning, our whole scheme will push inappropriate regions away from the selection of training samples at last.

Then, all images in this method are segmented into regions and represented by region-based features.

Next, in the positive candidate group, good features with discriminant representing power for a semantic concept are determined, and based on them the representative regions corresponding to the concept are selected as positive samples. Next, high quality of negative samples are determined utilizing the properties of the selected positive sample group and the unlabeled group. At last, the visual model associated with a semantic concept is built through a learning process.

3. Building Visual Model

In this section, we introduce the building process of the visual model related to a semantic concept.

Images in the proposed scheme are segmented into regions using the approach presented in. Then based on the consideration that different image categories may have different characteristics, after segmentation, for each region various kinds of features are extracted to characterize region content, which are listed in Table 1. For example, color histogram is used to describe color distribution in one region, and according the work, 218 bins desirably reflect most important color information. Regarding color moments with , it is adopted to give an efficient color description of one region by statistically describing color distribution by average, variance and skewness in 3 color channels.

All features in this paper are normalized to put equal importance on each feature component using the Gaussian normalization method.

To facilitate the following presentation, denotations used are described in Table 2.

3.1 Selection of Positive Samples

As for the selection of positive samples, we have the following considerations.

First, for most image categories, it is observed that the semantic meaning of an image is often represented with a large-size region that is corresponding to a distinctive object in real world under ideal image segmen-
Table 1: Image features in the proposed scheme.

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Color Histogram(^{(2)})</td>
<td>218-d</td>
</tr>
<tr>
<td>2</td>
<td>Color moment(^{2})</td>
<td>9-d</td>
</tr>
<tr>
<td>3</td>
<td>Shape( 3-d Normalized Inertia and Region Size(^{(18)})</td>
<td>4-d</td>
</tr>
<tr>
<td>4</td>
<td>Tamura Texture (coarseness, contrast, directionality)</td>
<td>3-d</td>
</tr>
<tr>
<td>5</td>
<td>Distance Histogram(^{(1)})</td>
<td>10-d</td>
</tr>
<tr>
<td>6</td>
<td>Angle Histogram(^{(1)})</td>
<td>36-d</td>
</tr>
</tbody>
</table>

Table 2: Denotation in the proposed scheme.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_{i}^{m,pc})</td>
<td>feature vector of region (i) in positive candidate group in (m) feature space</td>
</tr>
<tr>
<td>(f_{i}^{m,u})</td>
<td>feature vector of region (i) in negative group in (m) feature space</td>
</tr>
<tr>
<td>(f_{i}^{m,p})</td>
<td>feature vector of region (i) in positive group in (m) feature space</td>
</tr>
<tr>
<td>(f_{i}^{m,u})</td>
<td>feature vector of region (i) in unlabeled group in (m) feature space</td>
</tr>
<tr>
<td>(\mu_{i}^{m,pc})</td>
<td>the centroid of positive candidate group in (m) feature space</td>
</tr>
<tr>
<td>(\mu_{i}^{m,p})</td>
<td>the centroid of positive group in (m) feature space</td>
</tr>
<tr>
<td>(\mu_{i}^{m,u})</td>
<td>the centroid of unlabeled group in (m) feature space</td>
</tr>
</tbody>
</table>

Then, according to \(\{D_{j}^{m,pc}\}\), the mean squared deviation of positive candidates in \(m\) space is obtained, which is denoted as \(D_{average}^{m,pc}\).

In similar ways, in the unlabeled group, one sample has feature \(f_{j}^{m,u}\). Let its distance to the centroid of the group be \(D_{j}(f_{j}^{m,u}, \mu_{j}^{m,u})\). Here, \(j = 1, 2, \ldots, U\), \(U\) is the number of regions in the unlabeled group. Among \(\{D_{j}^{m,u}\}\), the maximum and minimum values \(D_{min}^{m,u}, D_{max}^{m,u}\) are achieved.

Next in \(m\) feature space, the sparse degree (SD) of the positive candidate group is estimated by integrating the distribution information of the positive candidate group and the unlabeled group, using the following strategy:

\[
SD(m) = \frac{D_{average}^{m,pc}}{D_{max}^{m,u} - D_{min}^{m,u}}
\]  \(1\)

With this function, the sparse degree of positive candidates in \(m\) feature space is measured utilizing the ratio of the feature distances between the positive candidate group and the unlabeled group. It also can be seen that the feature distance in \(m\) space is normalized by the distance range of the unlabeled group.

The idea behind the above function is that if the points of a particular feature have a dense distribution in the positive candidate group, and the distribution is relatively highly biased to the positive candidate group compared with the unlabeled group, the feature is regarded as good feature, which has good discriminating power for semantically related images in the positive candidate group. The above heuristic strategy computes \(O(N)\) (\(N\) is the number of training samples) quantities in determining good feature because it effectively and simply utilizes the distribution of positive candidate samples and unlabeled samples. This is much efficient
compared with those complex algorithms for feature selection in paper\textsuperscript{22(23)}. For example, feature selection in\textsuperscript{21} is conducted using complex mixture models where their parameters are estimated iteratively by EM. As a result the amount of computation is demanding with $O(NDK)$, where $N$ is the number of samples, $D$ is the dimension of features, and $K$ is the number of mixture models.

All $SD(m) \cdot m = 1, 2, \cdots, M$ is sorted by ascending order, top $G$ features are selected as good feature and are used to build the visual model associated with a semantic concept in the following process. Here, the selected good features are denoted as $GF = \{f^g, g = 1, 2, \cdots, G\}$. For instance, if they are color moments and shape features, here $G = 2$, $f^1$ is color moments, and $f^2$ is shape features. Note that both the representative features and their order are saved in $GF$ with the consideration that if a feature has a higher rank in $GF$, it shows that the feature with a higher order has more discriminating power in characterizing a particular visual pattern. According to our experimental experience, $M$ is set as 2 in the implementation.

Although images in the positive candidate group are highly semantically relevant, there are still “noise” regions that do not contribute to the description of a semantic concept. Therefore, a fine filtering process is carried out to explicitly extract positive samples (regions) from the positive candidate group in $GF$ feature space.

Currently, an efficient K-means clustering algorithm\textsuperscript{20} is used in the proposed scheme. The K-means method clusters the regions in the positive candidate group into several groups, in which the clustered feature vector is the combination of good features in $GF$. The stopping rule for the iterative process in K-means algorithm is determined when the overall sum of the distances between all feature points to corresponding cluster centers is lower than a threshold, or when the size of the biggest group (regions) is lower than $\Gamma$ of the size of the whole positive candidate group.

We have two considerations for determining positive samples. First, it is observed that semantically related images should include one common object (region) with similar visual characteristics, and as a result, in the positive candidate group containing images with highly semantic relevance, the samples of the biggest cluster should reflect the common region referred as visual pattern. The second consideration is that the purpose of the fine filtering process is just to extract high quality positive samples for the building process of the visual model, but not all positive samples. As a result, in the proposed scheme, the biggest clustered group is selected as the positive sample group. $\Gamma$ is set as $1/3$ based on our experimental experience. This is reasonable since one image is usually segmented into 3-5 regions in this scheme.

Here, the proposed method for selecting positive samples has the advantages that the selection of positive samples and good features are unified in the same framework efficiently and that using high quality positive samples can lead to a more effective learning process.

### 3.2 Selection of Negative Sample

Basically, the more reliable the selected information is, the better the performance a learning process for a semantic concept is. After positive samples are selected, another important issue for an effective learning process is the reliable selection of negative samples. Random generation of negative samples in previous work\textsuperscript{10,11} may still introduce positive samples to the negative group. In addition the efficiency of convergence for the learning process and the effectiveness of the learned visual model are not satisfactory.

Here negative samples are selected carefully from the unlabeled group using the property of the positive group achieved in the previous section.

For one sample in the unlabeled group, let its features in $GF$ feature space be $f_i^g, g = 1, 2, \cdots, G$. Its representing power to the positive group is estimated using the following function, which calculates the weighted distance between one sample and the centroid of the positive group.

$$
\rho_i = \sum_{g=1}^{G} e^{-\Delta^{g, p}} \cdot D_i(f_i^g, \mu^{g, p})
$$

(2)

In the previous section, we mentioned that in $GF$, different good features make different contributions to the description of a semantic concept. Assuming that there are two features in $GF$, color moments($f^1$) and texture feature($f^2$), their importance to representing power is reflected by setting the weight for color moments to a higher value than the other weight for texture in the equation 2.

Here $\Delta^{g, p}$ is defined as the following, which reflects the standard deviation of a positive group in $g$ feature space.
\[ \delta^{\theta,p} = \frac{1}{\dim} \frac{1}{N'} \left( \sum_{i=1}^{N'} D_i(f_i^{\theta,p}, \mu^{\theta,p}) \right)^{\frac{1}{2}} \]  

(3)

where \( N' \) is the number of positive samples, and \( \dim \) is the dimension of \( \theta \) feature.

If \( \rho_i \) is larger than a threshold value \( \tau \), then the region \( i \) is considered not to have the representing power to the positive group, and is taken as a reliable negative sample. Here, \( \tau \) is defined as the maximum of the value of the representing power of all samples in positive group, which is defined as:

\[ \tau \leftarrow \max_{j, \rho \in \rho} \left( \sum_{g=1}^{G} e^{-\delta^{p,g} \cdot D(f_j^{p,g}, \mu^{p,g})} \right) \]  

(4)

### 3.3 Building a Visual Model

Now the selected positive samples and negative samples are used to build a visual model.

In previous work, usually during the learning process, the negative samples are treated as the same distribution as in a Gaussian model. However, irrelevant samples often come from heterogeneous categories and they are too often too sparse to represent their true distribution. Any attempt to cluster them is not only unnecessary, but also potentially damaging. It is thus desirable to treat the positive and negative samples differently.

This kind of asymmetry is addressed in biased discriminant analysis (BDA) algorithm\(^{24}\), which is proposed in order to cluster only positive samples and pull the negative samples far away from the positive ones.

This idea fits our goal, which is only to mine the intrinsic visual model of a semantic concept underlying the web images, and does not care about how the distribution of negative samples is.

Let \( f^G \) indicate the region feature vector in combined GF feature space. Then let one positive sample in GF feature space be \( f_i^{G,p} \), the centroid of the positive group is denoted as \( C \), and one negative sample is denoted as \( f_j^{G,n} \). According to the BDA algorithm, the objective function is modeled by finding an optimal projection \( W \) that maximizes the following ratio in this work:

\[ W = \arg \max \left[ \frac{W^T S_y W}{W^T S_x W} \right] \]  

(5)

Here,

\[ S_y = \sum_{i=1}^{N'} (f_i^{G,p} - C) \cdot (f_i^{G,p} - C)^T \]  

(6)

\[ S_x = \sum_{j=1}^{K} (f_j^{G,n} - C) \cdot (f_j^{G,n} - C)^T \]  

(7)

where \( N' \) is the number of regions in the positive group and \( K \) is the number of regions in negative group.

The columns of the optimal \( W \) are the generalized eigenvectors \( V \) corresponding to the largest eigenvalue \( \Omega \) which is defined as the following function:

\[ W = \text{eigen}(S_x^{-1} S_y) \]  

(8)

Let \( \Omega = \text{diag}\{\Omega_1, \Omega_2, \cdots, \Omega_D\} \) be the eigenvalues, where \( \{\Omega_1, \Omega_2, \cdots, \Omega_D\} \) is sorted in descending order, and \( V = \{v_1, v_2, \cdots, v_D\} \) are the corresponding eigenvectors.

For a particular semantic concept \( S \), its corresponding discriminating space is constructed using only the first \( m \) eigenvalues and eigenvectors as the orthonormal basis vectors.

\[ L(S) = V' \cdot (\Omega')^{\frac{1}{2}} \]  

(9)

In this new space, positive samples related to a semantic concept are clustered tightly, and negative samples are separated from the centroid of the positive group. In addition, in the discriminating space feature dimensions are also reduced.

Finally, in this scheme, for a semantic concept \( S \), the visual model is characterized by its discriminating space \( L(S) \), and the centroid of positive samples in the \( L(S) \) space that is denoted as \( C(S) \).

### 3.4 Identification of a Semantic Concept in Images

Using the visual model associated with one semantic concept, we now present the process of identifying whether an image contains the semantic concept.

We measure whether one image \( I \) contains semantic concept \( S \) using the following procedure.

First, after the image is segmented into regions, only large regions with number \( P \) are used to estimate a semantic concept, based on the consideration that not all regions are important to the determination of a semantic concept, and larger regions are more likely to attract human attention than smaller ones. In our implementation, it is set that the total area of \( P \) regions should cover above 50% area of the total image.

Let the feature vectors of the extracted regions in GF feature space be \( f_i^G, i = 1, 2, \cdots, P \). Projecting \( f_i^G \) into the corresponding \( L(S) \) discriminating space, its new representation is defined in the following equation:

\[ f_i^G = L^T(S) \cdot f_i^G \]  

(10)

Then the similarity distance of the semantic concept in this image is estimated as:
\[ P(S|I) = \frac{1}{P} \sum_{i=1}^{P} D(f_i^G, C(S)) \] (11)

If \( P(S|I) \) is below a threshold value, image \( I \) is determined to contain \( S \) semantic concept.

4. Experiment

In this section, the proposed method is evaluated through experiments. The experimental system was developed in \( C++ \) and perl programming languages. According to the statistics\(^3\) that reflects popular user queries, we selected ten semantic concepts from “pet”, “natural scene”, “sports”, “sightseeing places” and “animal” and used the proposed approach to build visual models of them independently. In the following, due to the space limitation, five concepts “tiger”, “soccerball”, “greatwall”, “cat”, and “sunset” are used as examples to illustrate the results in details.

In our work, Yahoo! was selected as the image search engine because it provides convenient image search web services. First, each semantic concept was submitted to Yahoo! as a query. The referred images were gathered by an image crawler that was developed in perl. Here the image format was limited to JPEG since images in GIF format are often animated pictures. In addition, very small images less than 60 pixels in width or height were removed from the set of gathered images due to their low quality. As a result, in the experiment, about 1000 web images were gathered as training data for each concept.

Next the gathered training data was divided into positive candidate group and unlabeled image group. Here, we have two considerations on the size of the positive candidate group. First, BDA learning algorithm is designed to be robust under different size of training data. Second, an empirical statistics shows that semantically relevant images often appear among the top retrieved 100 \( \sim \) 200 images\(^2\). As a result, of the 1000 training data, 100 images from the first web pages were taken as the positive candidate group to tradeoff between learning efficiency and effectiveness. Then, with the similar size as positive candidate group, the last 100 images of 1000 web images were taken as the unlabeled group. In total, 200 images were used to build the visual model. Next, from these two groups positive and negative samples were selected with high quality.

Finally, the visual model associated with one semantic concept \( S \), which is characterized using its discriminating space \( L(S) \), and the centroid \( C(S) \) of positive samples in the \( L(S) \) space, was built through the BDA learning process.

The effectiveness of the model was tested by utilizing it to re-rank all the 1000 web images, which made an attempt to make the semantic-related images had higher rank than the noise images.

4.1 Illustrative Examples of Visual Model Building

To give a more vivid impression of the visual model building of a semantic concept by the proposed method, some positive region samples and negative region samples are shown as in Fig.2.

For positive region samples of each semantic concept, the second row shows the regions close to the \( C(S) \) in corresponding \( L(S) \) discriminant space, which are highlighted in a black color background. The first row corresponds to their original images. In similar fashion, for negative region samples, the second row shows the negative regions and their corresponding original images are in the first row.

Especially, when observing the results of positive regions in Fig.2, it can be seen that the proposed approach can extract meaningful regions from web images to effectively represent the visual model of a semantic concept. For instance, with regard to the “tiger” concept, semantically meaningful parts of “tiger” such as “head” and “body” are extracted to characterize the visual model, even if they have different colors and are embedded in different backgrounds.

Besides the visualized examples above, the distributions of the positive regions in \( L(S) \) discriminant spaces are estimated using an empirical probability function and is shown in Fig.3, where the \( x – axis \) indicates the distance-bins between positive samples and the \( C(S) \) in the discriminant space, and \( y – axis \) is the probabilities corresponding to the distance-bins. From Fig.3, it can be seen that most positive regions are close to the center \( C(S) \), which verifies that the selected positive group reflects visual model related to a semantic concept.

4.2 The Application of the Visual Model to Image Retrieval

The visual model built by our method is applied for image retrieval. For each query word, as mentioned in the beginning of “Experiment” section, dataset used here consists of 1000 images retrieved from the web. Within the dataset each image is manually labeled as semantic-related or not to the query word, and used as the ground truth.

To conduct image retrieval, first the similarity dis-
Fig. 2 The visualization of the visual model building for semantic concepts. In positive region samples, the second row is the positive regions close to $C(S)$ in corresponding $L(S)$ discriminant feature space; the first row is the original images containing these regions. Negative region samples are shown in similar ways.

\begin{equation}
\text{Precision} = \frac{n(\text{related image according to groundtruth})}{n(\text{top retrieved images})}
\end{equation}

To demonstrate the effectiveness of our approach, we compared the retrieval precision using our visual models with those using the other two methods, referred as method $A$ and $B$.

Method $A$ indicates the current search engines.
Fig. 3  Empirical probability density function of the distances from positive regions to the $C(S)$ in $L(S)$

discriminant space.

Note that it is difficult to compare our approach directly to other current methods because learning visual concepts from web images is still in its research infancy and there are few related methods, if any, the datasets used are very different due to the heterogeneous property of web images. We implemented a reference algorithm by our own over our experimental dataset. In method $B$, a one-class SVM classifier was built based on the positive samples obtained in section 3.1, and used to retrieve images.

The considerations of conducting method $B$ are as the following: (1). In contrast with some approaches such as\cite{12} that build visual models directly from noise contaminated web images, method $B$ is based on positive samples of good quality. (2). In previous work\cite{20}, one-class SVM shows that it has promising capability in nonlinear estimation of probability distribution compared with the Gaussian model that is adopted in other related research. (3). Therefore, in the sense of methodology, method $B$ shows more potential in terms of discriminating power and may provide a reasonable reference.

In fact, referring to the proposed scheme in Fig.1, the comparison of three methods indicates the retrieval re-
results in (1) the text-based step (Method A), that is, precision is calculated within 1000 images retrieved by a query word, in (2) the positive sample only step (Method B), that is, precision is calculated within the re-ranked 1000 images whose rank are determined by the similarity between the visual model built by one-class SVM and each image, and in (3) retrieval with our visual model, that is, precision is calculated within the re-ranked images set whose rank are determined by the similarity between the visual model built by our method and each image. As can be seen, it also indicates how retrieval precision is improved by our method step by step.

Fig. 4 and 5 summarize the experiment results. As can be seen in Fig. 4 and Fig. 5, method B provides better retrieval results than method A for many semantic concepts due to its utilization of image features. But it also can be seen that the performance improvement of method B is just a little, which implies that building a visual model based only on positive samples has lim-
Fig. 5 The average retrieval precisions of 10 visual concepts in the proposed approach, in method A and in method B. They are computed as the average values of precisions in top 50, 100, 150, ..., 400 retrieved images. From left to right, they are corresponding to method A, method B and the proposed approach.

It is apparent that our approach has much better retrieval accuracy than methods A and B. This is because our method utilizes the information of both text and image content during the visual modeling process compared with method A being text-based, and compared with method B only using positive samples, our method enhances the discriminating power of visual models significantly by including carefully selected negative samples in the modeling process. As can be seen in Fig.5, retrieval precision can be improved by an average of 13% compared with method B.

Another observation is noticed. According to current search engine, within the web images gathered by a query word, semantic-related images tend to gather intensely in the top retrieved images (about 100–200 images) and distribute sparsely in the latter retrieved images. As a result, three methods basically show the tendency that the more number of images are retrieved, the lower the precision becomes according to the equation 12, in which numerator (the related images) becomes less following denominator (the top retrieved images) becomes larger. However, as can be seen in Fig.4, compared with the retrieval results of method A and method B in which the precision fall down sharply, our method tries to makes the tendency of the precision (green line) become smoother. It indicates that our method can grasp those semantically related images which are sparsely distributed in the low ranked retrieval images and give an intense overview of similar images in the top ranked retrieval results.

This experiment demonstrates that the visual model learned by the proposed method is effective in boosting or optimizing retrieval results. In fact, it is not very appropriate to compare our algorithm with current search engines in this way due to the different retrieval methods and the different scale of images used in the retrieval systems. We have presented the retrieval results merely to provide an insight into the effectiveness of our approach.

5. Conclusion and Discussion

In this paper a novel approach is proposed for building visual models associated with a semantic concept. With reliable selection of positive samples and negative samples from noisy web image resources, an effective visual model associated with a semantic concept is built using learning algorithm. The proposed scheme is completely automatic, needing no human intervention, and is robust and applicable for generic images.

The experimental results demonstrate the effectiveness of the visual model learned by our method on image clustering and retrieval. The proposed approach yields better retrieval accuracy with the image categories that include distinctive objects inside them. This is because the regions of the objects in such images can be segmented more semantically and the representative objects of the image categories would cluster much densely according the method described in section 3.1.

In general, experimental verification of the learning visual model using web images is a very difficult task. There is no standard dataset, and there are also no widely agreed upon evaluation metrics. Therefore, in the future, we need to more exhaustively investigate
evaluation methods to demonstrate the soundness of our system.

[References]

2) http://yahoo.com
3) http://www.google.jp
4) http://www.ilt.nist.gov/iaiui/894.02/projects/trecvid
9) K. Barnard and D. Fornyth, “Learning the Semantics of Words and Pictures”,Proc.8th Int. Conf.on Computer Vision, 2, pp. 408-415, 2001
23) A. Blum, P.Langley, “Selection of relevant features and examples in machine learning”, Artificial Intelligence, 97(1-2), pp.245 -271,(1997)

Yongqing Sun received the B.E. and M.E. degrees in Computer Science from Xi’an Jiaotong University (XJTU), China in 1998 and 2001 respectively, and the Ph.D. degree in Information and Computer Science from Keio University, Japan in 2005. She is currently a researcher in NTT Cyber Solution Laboratories. Her research interests include image processing, pattern recognition, machine learning, data mining and multimedia retrieval.

Satoshi Shimada received his B.E., M.E. and Ph.D. degree in engineering from Kanazawa University, Kanazawa, Japan in 1984, 1987 and 2004, respectively. He is currently a senior research engineer in NTT Cyber Solution Laboratories. His research interests include image and video processing, pattern recognition, multimedia handling and retrieval. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan.

Masashi Morimoto received his B.E. and M.E. degrees in information engineering and the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan in 1986, 1988 and 2006, respectively. He is currently a senior research engineer, supervisor in NTT Cyber Space Laboratories. His research interests include image and video processing, pattern recognition, multimedia handling and retrieval. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Information Processing Society of Japan (IPSJ).

Yukinobu Taniguchi Senior Research Engineer, Supervisor, Promotion Project 2,NTT Cyber Solution Laboratories. He received the B.E., M.E., and Dr. Eng. degrees in Mathematical Engineering from the University of Tokyo in 1990, 1992 and 2002, respectively. In 1992, he joined NTT Human Interface Laboratories Japan. He received the best paper award of the Institute of Electronics, Information and Communication Engineers (IEICE), in 2000. He is a member of IEICE, Information Processing Society of Japan (IPSJ), and ACM.