Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 28 (2007) 1581–1591

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Integrated patch model: A generative model for image categorization based on feature selection

Feng Xu [*], Yu-Jin Zhang

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Received 22 January 2006; received in revised form 9 March 2007
Available online 11 April 2007

Communicated by M.-J. Li

Abstract

Image categorization could be treated as an effective solution to enable keyword-based image retrieval. In this paper, we propose a novel image categorization approach by learning semantic concepts of image categories. In order to choose representative features and meanwhile reduce noisy features, a three-step feature selection strategy is proposed. First, salient patches are detected. Then all the detected salient patches are clustered and the visual keyword vocabulary is constructed. Finally, the region of dominance and the salient entropy measure are calculated to reduce the similar and non-common noises of salient patches. Based on the selected visual keywords, the Integrated Patch (IP) model is proposed to describe and categorize images. As a generative model, the IP model represents the appearance of the combination of the visual keywords, considering the diversity of the object or the scene. The parameters are estimated by the EM algorithm. The experimental results on the Corel image dataset demonstrate that the proposed feature selection and the image description model are effective in image categorization.

Keywords: Salient patch; Visual keyword; Feature selection; Image categorization; Generative model; EM algorithm

1. Introduction

Although content-based image retrieval (CBIR) has been well studied over decades, it is still a challenging problem to search for images from a large-scale image database because of the well acknowledged semantic gap between low-level features and high-level semantic concepts. An alternative solution is to use keyword-based approaches, which usually associate images with keywords by either manually labeling or automatically extracting surrounding text. Although such a solution is widely adopted by most existing commercial image search engines, it is not perfect. First, manual annotation, though precise, is expensive and difficult to extend to large-scale databases. Second, automatically extracted surrounding text might be incomplete and ambiguous in describing images, and even more, surrounding text may not be available in some applications.

To overcome these problems, automated image categorization and annotation are considered two promising approaches in understanding and describing the content of images. Besides obtaining text annotation, a successful image categorization will significantly enhance the performance of the content-based image retrieval system by filtering out images from irrelevant classes during matching. At the cognitive semantic level, images can be categorized according to objects and scenes they contained. In some image categories, the general concept of images is about the object, such as “Horse”, “Bus”, “Ship”, and so forth. Some other categories focus on the scene, such as “Beach”, “Skiing”, “Surfing”, and so on. Image category labels are text keywords describing objects or scenes. Our problem is to categorize images based on object or scene concepts.

[*] Corresponding author. Tel.: +86 10 62781291; fax: +86 10 62770317.
E-mail address: f-xu02@mails.tsinghua.edu.cn (F. Xu).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.03.016

Many good results have been reported in two-class image classification tasks, such as city vs. landscape (Vailaya et al., 1998; Liu et al., 2005), indoor vs. outdoor (Szummer and Picard, 1998; Liu et al., 2005). However, we want to investigate more powerful methods to solve multiple-class categorization problem. Chen and Wang (2004) proposed an approach for image categorization by learning and reasoning with regions. In their work, images are viewed as bags, each of which contains a number of instances corresponding to regions obtained from image segmentation. Then Multiple-Instance Learning and SVM classifier are applied to image categorization. Csurka et al. (2004) proposed bags of key-point of objects as features. Based on that, the visual vocabulary is constructed by K-means clustering algorithm. Both Naïve Bayes and SVM classifiers are applied to categorization. Recently, many approaches for object class recognition have been proposed and demonstrated to be promising to solve multiple-class image categorization tasks. Fergus et al. (2003) proposed the constellation model, which is learned in a Bayesian manner, to recognize several classes of objects. This categorization scheme was further improved by Li et al. (2003) to classify more categories with less training samples. A good application of this scheme is filtering Google images (Fergus et al., 2004). Taking into account shape, appearance, occlusion and relative scale, the constellation model well describes an object in multiple semantic aspects with low-level features, and demonstrates promising potentials in image understanding.

In the object class recognition, generative models are successful. Compared with the discriminative models, the generative models exhibit advantages of incrementally learning the added classes, handling missing data, and incorporating prior information. However, most of the object class recognition models are too complex with a large quantity of parameters to be extended to larger scale datasets. Hence, for image categorization, generative models with low complexity are worthy to be investigated.

For image categorization, it is useful to have access to high-level information about objects and scenes contained in the images to manage image collections. The high-level information must be learned from low-level features. As low-level features are usually noisy and uninformative, feature selection is of great importance and needs to be conducted before modeling images. Although there are many feature selection approaches, few of them select features according to image content, such as the work by Vasconcelos and Vasconcelos (2004). Hence, an effective feature selection method based on image content is necessary.

Recent progresses in object class recognition have shown that local salient features are more informative in describing image content than global features (Csurka et al., 2004; Fergus et al., 2003). In image categorization, features are required to be common for the same class and discriminative for different classes. In semantic concept learning, the model should emphasize the object or the scene in an image category. Therefore it is essential to select the common features and meanwhile reduce noisy features contributed by various backgrounds. Here, noisy features are defined as the points in non-common parts of all the images in the same category. However, to the best of our knowledge, there is no related work explicitly eliminating noises. Thus a robust feature selection strategy based on the image category is crucial and worthy of investigation.

Considering these aspects, we propose a feature selection strategy and an image category description model to categorize images according to the cognitive semantics. First, a three-step feature selection strategy is explicitly conducted. Then considering the diversity of the image appearance in the same concept, a generative model, the Integrated Patch (IP) model is proposed and used for each image category. For a new test image, its posterior probability in each category is calculated, and then the label with the largest probability is assigned to it. Through the proposed feature selection and the IP model, the image category can be discriminated.

It is desired the proposed approach can learn the common parts of images in the same category, so that the general concept of images can be learnt. This is different from the object categorization. Object categorization will focus on the object and object recognition, while our approach will focus on the common parts of images. If the images in the same category contain the same object and the same background, the features from the object and the background are both reserved as the visual keywords, and the model learns the object and the background as a whole.

The main contributions of this paper can be highlighted as follows:

First, a novel feature selection strategy is proposed. Taking into account visual quantization, region of dominance and salient entropy, it is capable of selecting the most informative and common salient patches for one category, and excluding most similar and non-common noise points.

Second, a generative model, the IP model, is proposed to describe image categories. Compared with two-class classification models, this model is capable of being extended to larger scale image databases.

The rest of the paper is organized as follows. Section 2 presents the feature selection strategy. Section 3 presents the IP model and parameter estimation. Section 4 shows the experimental results to evaluate the performance of our approach. Section 5 gives concluding remarks.

2. Feature selection strategy

In most existing categorization methods, all the local salient features are used to train the classifier, such as works by Csurka et al. (2004) and Fergus et al. (2003). However, some salient features may be noises and are contributed by only a few images in an image category. To model the image category, the common features are the most important because the common features are considered to be capable of representing the object away from the different backgrounds. Thus it is essential to conduct feature selection before modeling image categories. In this section, we describe the proposed feature selection strategy in detail. Features used in the categorization model are generated from two stages: salient patch selection and feature extraction.

2.1. Salient patch selection

Salient patch selection consists of three steps: salient patch detection, visual keywords construction and noise exclusion.

2.1.1. Salient patch detection

In this step, salient patches are detected by the local salient feature detector proposed by Kadir and Brady (2001). This detector finds regions that are salient over both location and scale. Only intensity information is used to detect and represent features. Once the regions are identified, they are cropped from the image and rescaled to the size of a small pixel patch. Because a high dimensional Gaussian is difficult to manage, principal component analysis (PCA) is performed on the patches from all images. Then each patch is represented by a vector of the coordinates within the first 15 principal components.

2.1.2. Visual keyword construction

The 15-dimensional feature vectors are used to construct the visual keyword vocabulary. The vector quantization is performed on the vectors of all the images within one category, conducted by *K*-means clustering. Clusters of the vectors are the *visual keywords* for the category. Cluster histogram of salient patches shows its distribution, in which each bin corresponds to a visual keyword. Those visual keywords with large number of salient patches or over a predetermined threshold, regarded as the most important features for the image category, are selected. The visual keywords with a small quantity of salient patches are considered noises from different backgrounds. An example of cluster histogram is illustrated in Fig. 1. The shadowed bins denote the selected clusters.

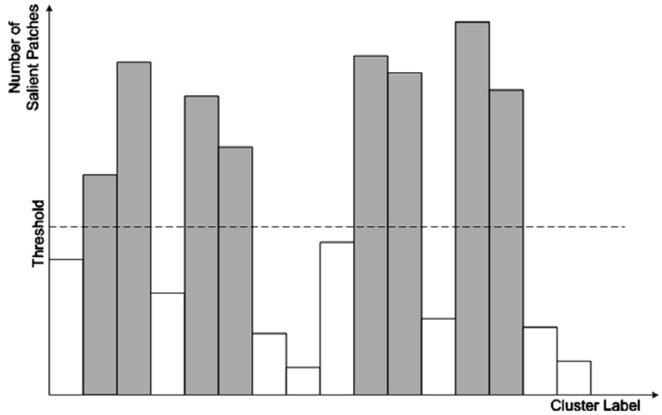

Fig. 1. Salient patch histogram.

2.1.3. Noise exclusion

Two types of noises can be excluded. First, the most similar noises can be excluded by the region of dominance (ROD) (Liu and Collins, 2000). ROD is defined as the maximal distance between two patch clusters in the feature space. The two patch clusters are determined by the histogram of salient patches, in which one is the current bin and the other is each of the detected local maximum bins. If the maximal distance is smaller than a preset threshold, the current local maximum is regarded as being similar to one of the detected local maxima and will not be preserved as a visual keyword.

Second, the most non-common noises can be excluded by the salient entropy. The salient entropy is defined as

$$H(n) = -\sum_{m=1}^{M} p_m(n)\log(p_m(n)) \tag{1}$$

where n denotes the index of the cluster, m denotes the index of the image and M denotes the total number of images in one category. $p_m(n)$ is the ratio between the number of salient patches in the nth cluster of the mth image and the total number of salient patches in this cluster of all the images.

The salient entropy reflects the distribution of a certain salient patch cluster in each image within the same category. If the distribution is more uniform, the selected feature is more common. So the visual keywords with larger entropies are preserved. According to the entropy measure, those visual keywords contributed by a few images are excluded, despite of large number of salient patches in the histogram.

2.1.4. Case study

This three-step feature selection strategy can be modeled as a feature filter shown in Fig. 2. Through this feature filter, the most important and common features are reserved while noisy patches on different backgrounds are removed as many as possible.

Some images from "Ship" category in the Corel image database are shown in Fig. 3 to demonstrate the feature selection results. The first row shows the salient patches without noise reduction while the second row shows the salient patches after feature filtering. For illustration, here,

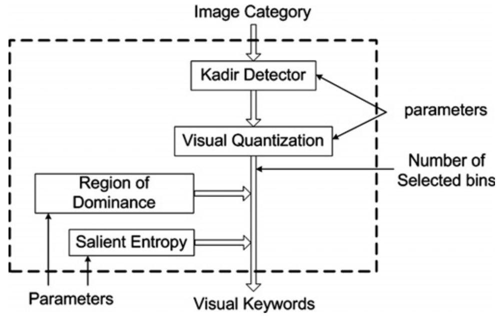

Fig. 2. Feature filter.

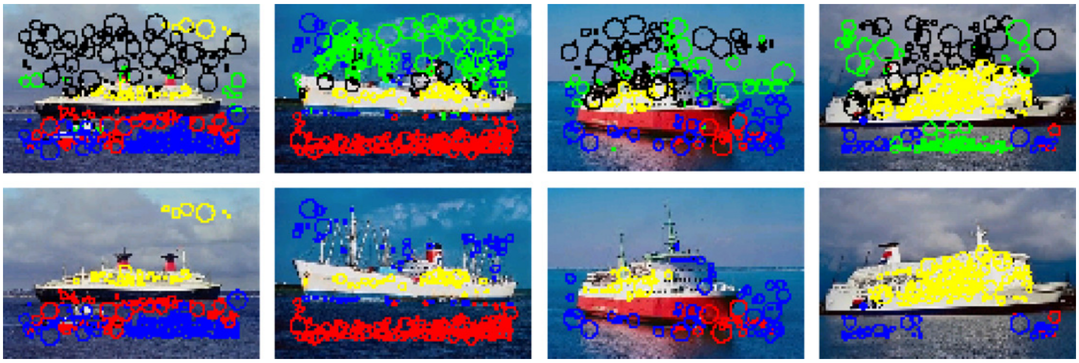

Fig. 3. Image examples with salient patches in “Ship” category in Corel image database.

five visual keywords are constructed and the different colors represent the different visual keywords. After feature filtering, three visual keywords are preserved. Each image presents a ship in water, in which “ship” is the object and “water” is the common background. But other parts of background are different (e.g. “mountain” or “sky”). After feature selection, the salient patches on “ship” and “water” are reserved as visual keywords, and patches from “mountain” or “sky” are removed. From these examples, it can be found that the preserved salient patches correspond to the visual keywords which are the common parts of all the images and relevant to the semantic concept of the image category, while the noises on the various backgrounds or irrelevant to the category concept are removed.

2.2. Feature extraction

For each selected salient patch, a 64-dimensional feature is extracted, in which the 44 dimensional elements are banded auto-color correlogram (Huang et al., 1997), the 14-dimensional elements are color texture moment (Yu et al., 2002) and the 6-dimensional elements are color moment (Deng et al., 2001; Vailaya et al., 2002).

2.3. Analysis of the feature selection

The proposed feature selection strategy can be explained by the Fisher Discriminant Analysis (FDA). Assume the salient feature set in class c $(c = 1, \ldots, M)$ is $X^c = \{x_j^c | j = 1, \ldots N\}$. We define the feature selection criterion function for image category c as the ratio between scatters of intra-class and inter-class:

$$R^c = \frac{\sum_{s_1=1}^{N_c}\sum_{s_2=1}^{N_c} d(x_{s_1}^c, x_{s_2}^c)}{\sum_{c_1=1}^{M}\sum_{c_2=1}^{M}\sum_{j_1=1}^{N_{c_1}}\sum_{j_2=1}^{N_{c_2}} d(x_{j_1}^{c_1}, x_{j_2}^{c_2})}, \quad s_1 \neq s_2, \quad c_1 \neq c_2 \tag{2}$$

where $d(x_{j_1}^{c_1}, x_{j_2}^{c_2})$ is a measure between two salient features in two different categories, and $d(x_{s_1}^c, x_{s_2}^c)$ is the same measure between two salient features in the same category. The measure can be any distance, discriminative information or other measures that can represent the similarity between two salient features. The ratio between the intra-class measure and the inter-class measure is the objective function to optimize, i.e.

$$\arg\min_{\Gamma}(R^c) \tag{3}$$

where Γ is defined as the indicator function of X^c, i.e.

$$\Gamma = \{\delta_j^c | j = 1, \ldots N, \delta_j^c \in \{0, 1\}\} \tag{4}$$

$$\delta_j^c = \begin{cases} 1 & \text{if} \quad x_j^c \text{ is selected} \\ 0 & \text{if} \quad x_j^c \text{ is removed} \end{cases} \tag{5}$$

Minimizing R^c will result in the optimal feature set $\{x_j^c\}$.

In order to select the most common features within one category, the salient features are clustered in the same category. However, this will lead to some extremely similar visual keywords in different classes, i.e. $\{x_{j_1}^{c_1} \simeq x_{j_2}^{c_2}\}$, so that $\{d(x_{j_1}^{c_1}, x_{j_2}^{c_2}) \simeq 0\}$. These visual keywords cannot distinguish different categories and are regarded as uninformative feature points. If these uninformative features can be removed, the sum of inter-class distances (the denominator in Eq. (2)) will almost be invariable. However, the sum of intra-class distances (the numerator in Eq. (2)) will decrease significantly, so that R^c will decrease. Until the minimal R^c is achieved, the reserved feature points are regarded as the important features for the categorization. Moreover, since the selected points are based on the analysis of the image content, they are regarded as the informative features for the category concept.

3. Image category description model

In order to categorize images at the cognitive semantic level, we model each image category based on the selected salient patches. The model could learn the semantic concept, and then categorize the new test images.

In the previous feature selection, the detector proposed by Kadir and Brady (2001) is used, which depends on the local intensity information. Then the 64-dimensional feature is extracted so that color and texture information are included. Considering the diversity of the newly included features, we use the finite mixture model. Each category is modeled as a combination of all the visual keywords, and the appearance of the visual keywords is defined as the Gaussian density distribution.

Suppose M images from the same category are given. Let K denote the optimal number of mixture components. In an image, there are N salient patch clusters corresponding to the selected visual keywords for the image category. Let $\boldsymbol{X}_{mn}$ denote the feature vector for cluster n in image m.

For an image category I, the model can be defined as

$$p(I|\Theta) = \prod_{m=1}^{M} p(I_m|\Theta) = \prod_{m=1}^{M}\left[\sum_{k=1}^{K}\prod_{n=1}^{N} p(\boldsymbol{X}_{mn}, c_k|\Theta)\right]$$
$$= \prod_{m=1}^{M}\left[\sum_{k=1}^{K}\prod_{n=1}^{N} p(\boldsymbol{X}_{mn}|c_k, \Theta)p(c_k)\right] \quad (6)$$

where $p(I_m|\Theta)$ is the probability of the mth image. $p(\boldsymbol{X}_{mn}|c_k,\Theta)$ is the probability of the nth patch in the mth image and the kth mixture component. For each component, there are N independent means and N covariance matrices corresponding to N clusters. $\Theta = \{\boldsymbol{\mu}_{kn}, \boldsymbol{\Sigma}_{kn}, k = 1,\ldots,K, n = 1,\ldots,N\}$ is the set of the parameters for these mixture components. $\boldsymbol{X}_{mn}$ is the 64-dimensional visual feature. $p(c_k)$ is the mixture weight subject to constraints:

$$0 \leqslant p(c_k) \leqslant 1, \quad \text{and} \quad \sum_{k=1}^{K} p(c_k) = 1 \quad (7)$$

It should be mentioned that the visual properties of a certain type of objects or scenes may look various at different lighting and capturing conditions. For example, the *ship* consists of various appearances, especially various colors, such as "red ship pattern", "white ship pattern" and "yellow ship pattern", which have very different properties. Thus, the data distribution for a certain type of images is approximated by using multiple mixture components to accommodate the variability of the same type of objects or scenes, i.e. presence/absence of distinctive parts, variability on overall shape, changing of visual properties due to the object patterns and viewpoints, etc.

The key of the proposed model is the multiplication of N cluster probabilities, each of which corresponds to a visual keyword selected by the algorithm in Section 2. It is assumed that the different visual keyword is independently identical distribution (i.i.d.), and can be defined as a Gaussian density distribution.

The optimal model structure and parameters $(\hat{c}_k, \widehat{\Theta})$ for an image category are determined by

$$(\hat{c}_k, \widehat{\Theta}) = \underset{c_k,\Theta}{\arg\max}\{L(\Theta|I)\} = \underset{c_k,\Theta}{\arg\max}\{\log p(I|\Theta)\} \quad (8)$$

The likelihood function is

$$L(\Theta|I) = \log p(I|\Theta) = \sum_{m=1}^{M}\log\left[\sum_{k=1}^{K} p(I_m, c_k|\Theta)\right]$$
$$\geqslant \sum_{m=1}^{M}\left[\sum_{k=1}^{K} q_m(c_k)\log\frac{p(I_m, c_k|\Theta)}{q_m(c_k)}\right] \triangleq B(\Theta; c_k) \quad (9)$$

where $B(\Theta;c_k)$ is the lower bound, and the inequality is deduced by Jesen Inequality.

The Maximum Likelihood Estimation (MLE) can be achieved by using the EM algorithm (Dempster et al., 1977). In E-step, the posterior distribution of c_k is computed:

$$q_m(c_k) = p(c_k|I_m, \Theta) = \frac{p(I_m|c_k, \Theta)p(c_k)}{\sum_{k=1}^{K} p(I_m|c_k, \Theta)p(c_k)}$$
$$= \frac{\left[\prod_{n=1}^{N} p(\boldsymbol{X}_{mn}|c_k, \Theta)\right]p(c_k)}{\sum_{k=1}^{K}\left[\prod_{n=1}^{N} p(\boldsymbol{X}_{mn}|c_k, \Theta)\right]p(c_k)} \quad (10)$$

In M-step, let the partial differential of $B(\Theta;c_k)$ for $\boldsymbol{\mu}_{kn}$, $\boldsymbol{\Sigma}_{kn}$ and c_k equals to zero respectively, and then we get the iterative solution for each parameter:

$$c_k^{\text{new}} = \frac{1}{M}\sum_{m=1}^{M} p(c_k|I_m, \Theta) \quad (11)$$

$$\boldsymbol{\mu}_{kn}^{\text{new}} = \frac{\sum_{m=1}^{M} \boldsymbol{X}_{mn} p(c_k|I_m, \Theta)}{\sum_{m=1}^{M} p(c_k|I_m, \Theta)} \quad (12)$$

$$\boldsymbol{\Sigma}_{kn}^{\text{new}} = \frac{\sum_{m=1}^{M} p(c_k|I_m, \Theta)(\boldsymbol{X}_{mn} - \boldsymbol{\mu}_{kn}^{\text{new}})(\boldsymbol{X}_{mn} - \boldsymbol{\mu}_{kn}^{\text{new}})^{\mathrm{T}}}{\sum_{m=1}^{M} p(c_k|I_m, \Theta)} \quad (13)$$

Because the number of salient patch is various in different images, the representative patch must be selected for each visual keyword to keep the probability multiplication of visual keywords in the same number. The selection of the representative patch can be implemented in two ways. In the first way, the patch nearest to the center of visual keyword is selected. In the second way, the average patch in the same visual keyword is selected.

This model can be explained at three levels. At the first level, the probabilities of images are multiplied based on the same category concept, because the images are assumed as i.i.d. Then at the second level, a finite mixture model is applied for each image category. At the third level, the probability multiplication of the visual keywords is computed for each image.

As the parameter estimation results, there are K components in one category, which are modeled as Gaussian functions. And for each component, there are N visual keywords, which are modeled as independent Gaussian functions with mean $\boldsymbol{\mu}_{kn}$ and covariance $\boldsymbol{\Sigma}_{kn}$. So the total number of parameters is $2KN + K$.

For a test image, its posterior probability is calculated in each category at first. Salient patches are detected and labeled as the nearest visual keywords. Then the 64-dimensional feature is extracted. The posterior probability of component in a category is computed as

$$p(c_k|I_m, \Theta) \propto p(I_m|c_k, \Theta)p(c_k)$$
$$= \left[\prod_{n=1}^{N} p(\boldsymbol{X}_{mn}|c_k, \boldsymbol{\mu}_{kn}, \boldsymbol{\Sigma}_{kn})\right]p(c_k) \quad (14)$$

The largest posterior probability is taken as the component-prediction of the image in this category. Then the component-prediction probabilities in all the image categories are compared, and the largest one is taken as the category-prediction for the image. Finally the image is labeled as the corresponding category. The posterior probability of the test image is also the multiplication of the salient patch probabilities. Therefore, we call this model as the Integrated Patch model.

For a certain image category, the IP model learns the image appearance from the selected salient parts, i.e. the visual keywords. Although the visual keywords are represented by the color and texture features, they can describe the image by the multiplication of the probabilities. Based on the visual keywords, this model is guaranteed for image categorization in two aspects: (1) The probabilities of visual keywords for each image are multiplied in the same order. The visual keywords are constructed from the selected salient patches. If a visual keyword is a necessary part of the concept, it will be presented in most of the images in the category and selected as the common features. Then all the selected parts of each image are considered in the same order so that the appearance of an object or a scene is assembled. (2) The diversity of the image appearance, especially the luminance and color, is described by the mixture components. In terms of the statistical theory, an image as a sample is contained in one of the components according to its luminance and color.

4. Experimental results

The image dataset employed in our empirical study consists of 5000 images taken from the Corel image database, in which each image category with 100 images represents one distinct topic of interest. Within each of 50 image categories, images are randomly divided into a training set and a test set. A keyword is assigned to describe each image category. The randomly selected image categories with names and sample images are shown in Fig. 4.

We will evaluate the proposed approach in three aspects: (1) performances of 35 image categories; (2) performances in different numbers of image categories; (3) performances in different image number ratios between training set and test set. We also provide comparisons: (1) comparison between the proposed approach and the SVM classifier; (2) comparisons between the proposed model and some other methods.

The benchmark metric for categorization evaluation is *precision* α, defined as

$$\alpha = \frac{\phi}{\phi + \varepsilon} \tag{15}$$

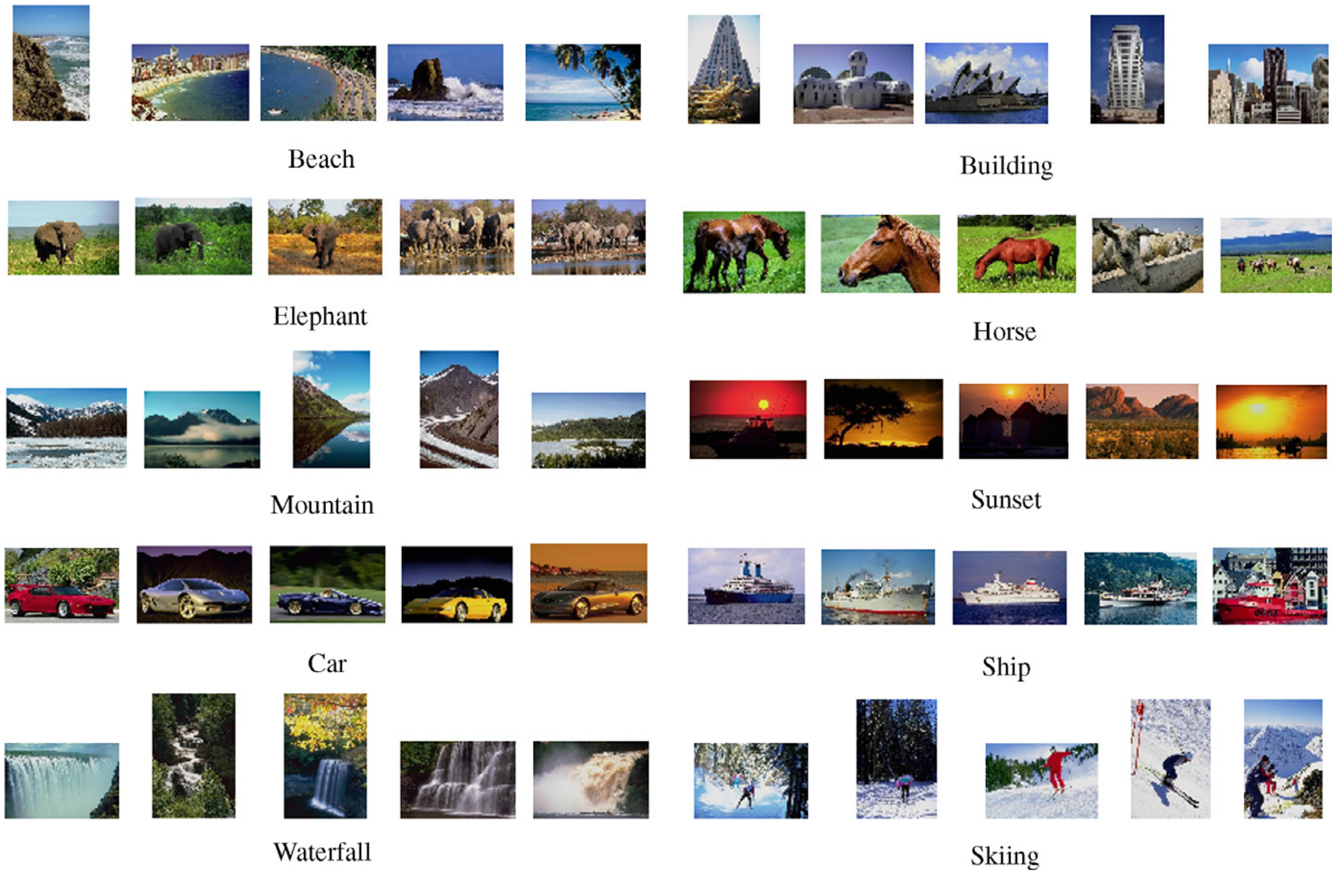

Fig. 4. Sample images taken from 10 image categories.

where ϕ is the number of true positive samples that are correctly categorized into the corresponding semantic category, ε is the number of true negative samples that are irrelevant to the corresponding semantic category and are categorized incorrectly. The categorization recall is 100%.

4.1. Parameters

In our approach, several parameters need to be tuned to obtain better performance.

Firstly, the total number of clusters is important in visual keywords construction. Too few visual keywords may lead to low discriminative results between category models while too many visual keywords may lead to low distribution entropy of salient patches. According to Eq. (2), we give the curve of the ratio in Fig. 5, in which the horizontal axis denotes the number of clusters in an image category and the vertical axis denotes the ratio.

Although only a few points are calculated and illustrated, it is clear to see the trend that the number of clusters has a value to minimize the ratio. In the experiments, we set this number as 50, around the minimum of the ratio curve. Thus, there are 50 visual keywords for each image category.

Secondly, the number of selected visual keywords is significant for the description of image category. In the experiments, we set the selected number of visual keywords by ROD and the salient entropy respectively. The intersection clusters between them are used as the final visual keywords. Thus the number of selected visual keywords is adaptively determined in different image categories.

Thirdly, the number of mixture components in the model should be determined according to the number of patterns in the image category. In our experiments, the component number is variable in the range from 2 to 6.

In the test, the salient patches are selected by the distance from the centers of visual keywords. For each test image, 100 salient patches nearest to the visual keywords are preserved to represent the image.

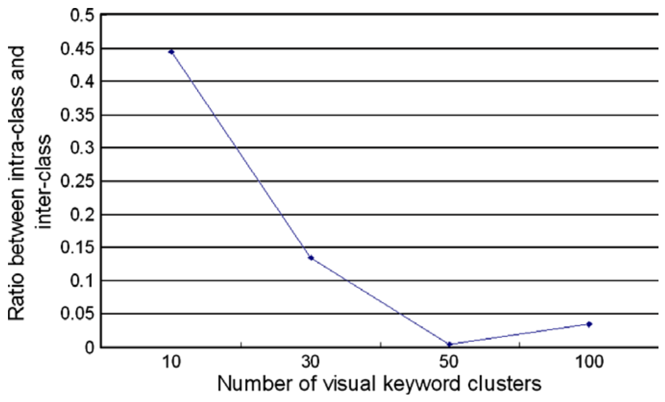

Fig. 5. Curve of the ratio between scatters of intra-class and inter-class.

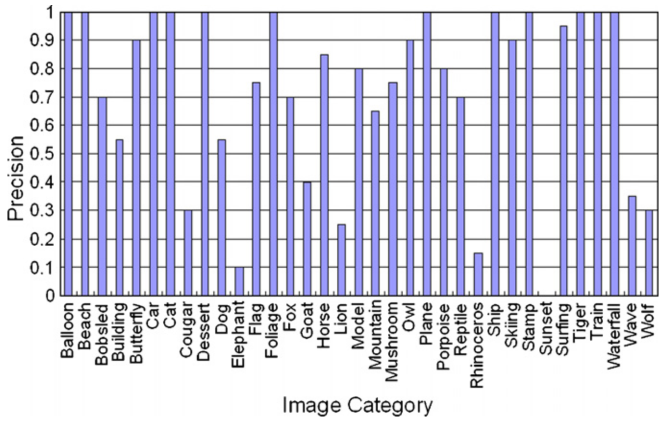

Fig. 6. Categorization precisions of 35 image categories.

4.2. Performance evaluation

4.2.1. Categorization precisions of 35 categories

When the image number ratio between training set and test set is 4:1, the categorization precisions of 35 image categories are shown in Fig. 6.

For most of the image categories, the categorization precisions are impressively high. This suggests that the proposed approach is effective in learning concepts of image categories. However, the performances of some categories are poor. These low precisions are mainly caused by misclassification between two similar categories. Fig. 7 presents some misclassified images (in at least one experiment) from categories "Elephant" and "Rhinoceros". Some images in these two categories contain the similar foreground and background, and even the animals have the similar poses. Hence, the misclassification is unavoidable and leads to low precisions. For the "Sunset" category, the precision is almost zero. Images in this category are in great diversities, which is the main reason for the large error.

4.2.2. Average categorization precisions in different numbers of image categories

In this experiment, we investigate the categorization performance varies with the number of image categories. The image number ratio between training set and test set is 4:1. When the number of image categories equals 10, 20, 30, 40 and 50 respectively, the average precisions are illustrated in Fig. 8.

We observe that the average precision decreases as the number of categories increases. When the number of categories increases from 10 to 50, the average precision drops from 82.5% to 45.3%. That is, when the number of categories increases to 5 times, the average precision decreases 37.2%. However, when the number of categories is less than 30, the difference of the average precision is much less. Compared with the binary classifier in discriminative models, the scalability is another advantage except for the better performance. The IP model is a generative model and can incrementally learn the added categories so that it is capable of being extended to larger scale image datasets.

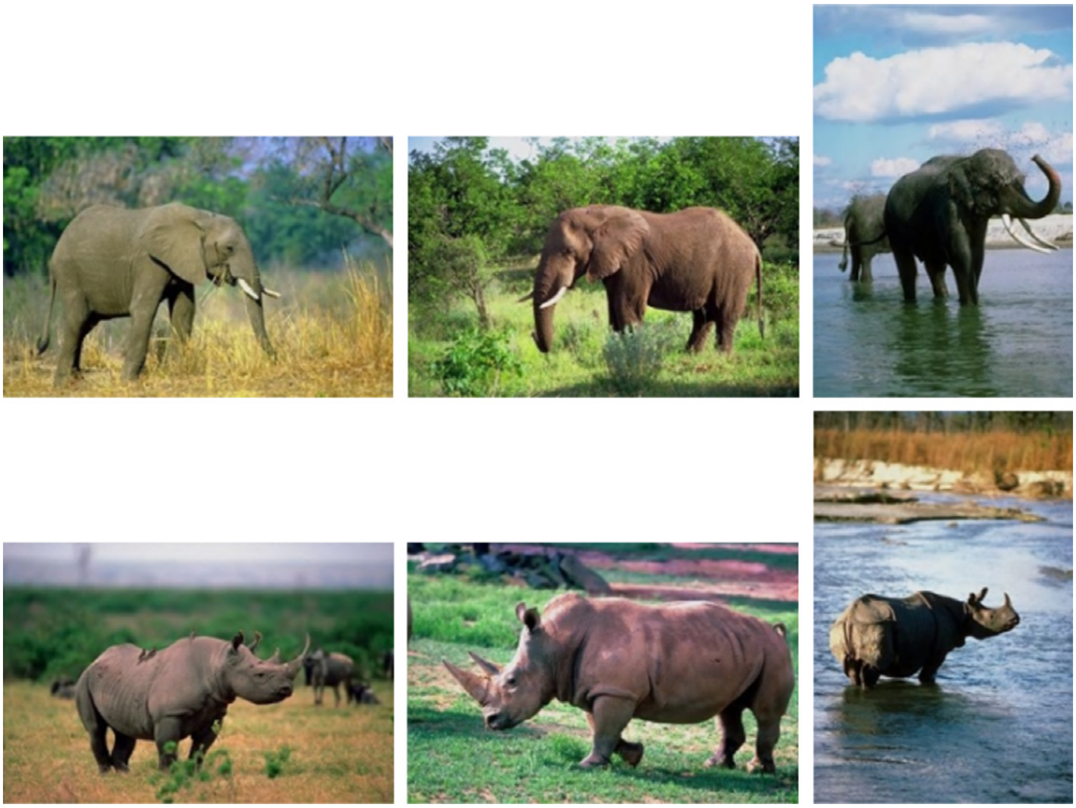

Fig. 7. Misclassified image examples for low categorization precisions (Elephant vs. Rhinoceros).

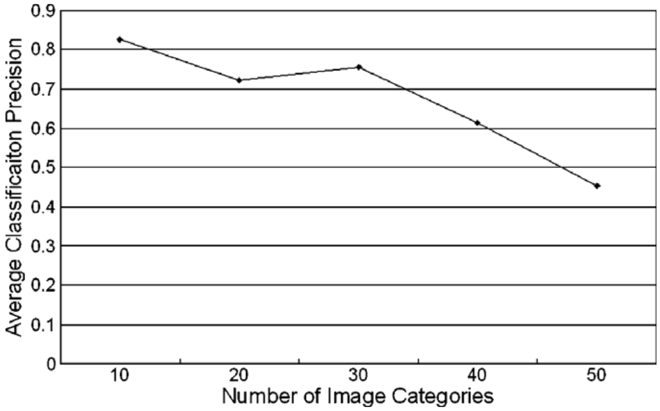

Fig. 8. Average categorization precisions in different numbers of image categories.

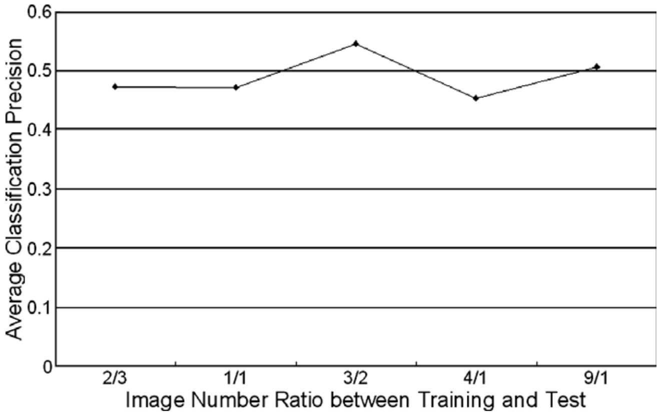

Fig. 9. Average categorization precisions in different image number ratios between training and test.

4.2.3. Average categorization precisions in different image number ratios between training and test

In this experiment, we investigate the categorization performance varies with the image number ratio between training set and test set. All the 50 image categories are used. In each category, the image number ratio between training and test is 2:3, 1:1, 3:2, 4:1 and 9:1 respectively. The average precisions are illustrated in Fig. 9.

As indicated in Fig. 9, when the number of training images increases, the average categorization precision fluctuates around 50%, instead of increasing monotonically. It can be conclude that the IP Model with a small quantity of training images can achieve the same performance as that with large numbers of training images. This is useful in some applications, in which users need not gather plenty of images with the same semantic concept to train the model.

4.3. Performance comparison

4.3.1. Comparison between the IP model and the SVM classifier

In this experiment, the 35 image categories same as those in Section 4.2.1 are used. For the proposed IP model, the features are extracted as the preceding process. For the SVM classifier with feature selection, the salient patches are arranged as the cluster labels. For the SVM classifier without feature selection, the salient patches are also detected. Because the numbers of salient patches in different images are different, the image is partitioned into grids and the average salient patch in each grid is arranged.

The average categorization precisions are listed in Table 1.

From the table, we can conclude that the proposed feature selection strategy apparently improves the performance of the categorization. For the same SVM classifier, the precision with feature selection is 11.3% higher than that without feature selection. For the semantic concept of an image category, the most important and common features are selected; hence the categorization precision could be improved. On the other hand, based on the same features, the precision by the proposed approach is slightly lower than that by the SVM classifier. We can believe that the proposed model is at least as effective as the SVM classifier, though as two different kinds of models.

4.3.2. Comparison between the IP model and other methods

To evaluate the proposed approach, we compare the IP model with the DD-SVM (Chen and Wang, 2004) and the S-C + C-T model (Datta et al., 2006) on the same dataset in the Corel image database. The average categorization precisions are listed in Table 2 and illustrated in Fig. 10, respectively.

From Table 2, it can be found that the precisions of IP model are higher than those of DD-SVM on both 10 classes and 20 classes. Moreover, the IP model can be used for more image categories, incrementally learning the added images. From these two precisions, it can be modestly concluded that the IP model outperforms DD-SVM for image categorization. In Fig. 10, the average categorization precision of 10 image categories varies with the number of mixture components. Although the meaning of the two component numbers is different, we can observe the influence of the parameters to the two models respectively.

Table 1
Comparison between the average precisions (%) of IP model and SVM classifier

	IP model	SVM
Without feature selection	N/A	63.7
With feature selection	72.3	75.0

Table 2
Comparison between the average precisions (%) of IP model and DD-SVM

	IP model	DD-SVM
10 Classes	82.5	81.5
20 Classes	72.1	67.5

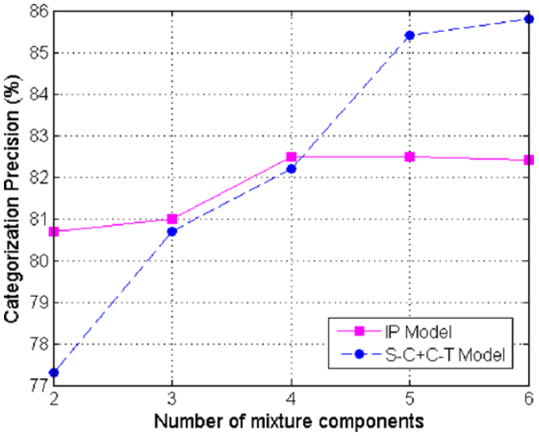

Fig. 10. Comparison of categorization precisions between IP model and S-C + C-T model with varying number of mixture components.

When the component number is from 2 to 4, the precision of IP model is higher than that of the S-C + C-T model. When the component number is 5 or 6, the precision of IP model is lower than that of the S-C + C-T model. Compared with the S-C + C-T model, the IP model is less sensitive to the parameter. This property is useful in the case that the component number cannot be tuned according to the dataset.

4.4. Discussions

From these results, we discuss some interesting points of note.

4.4.1. Robustness

We conclude that the proposed approach is robust from two facts. First, we use the robust local salient features. The salient point detector has been proposed for object recognition (Kadir and Brady, 2001) and has been proven to be robust to rotation and scaling. Second, we exclude features from different backgrounds and emphasize the category concept. What the strategy selects and the model learns is the common parts of images in the same category. If the most salient features have been caught by the model, the concept labels can be propagated to the new images regardless of rotation, scaling, and even degradation.

4.4.2. Examples of visual keywords

The visual keywords play an important role in the proposed approach. We examine an example to show how the visual keywords represent the image. In Fig. 11, an example image in “Ship” category with six visual keywords marked by color circles is illustrated. For each visual keyword, six patches extracted from six images are shown. We can observe that the patches of the same visual keyword are in similar appearance (intensity and texture). Moreover, each visual keyword corresponds to a sub-concept of the whole concept, that is, the patches of the same visual keyword are located on the corresponding parts of the

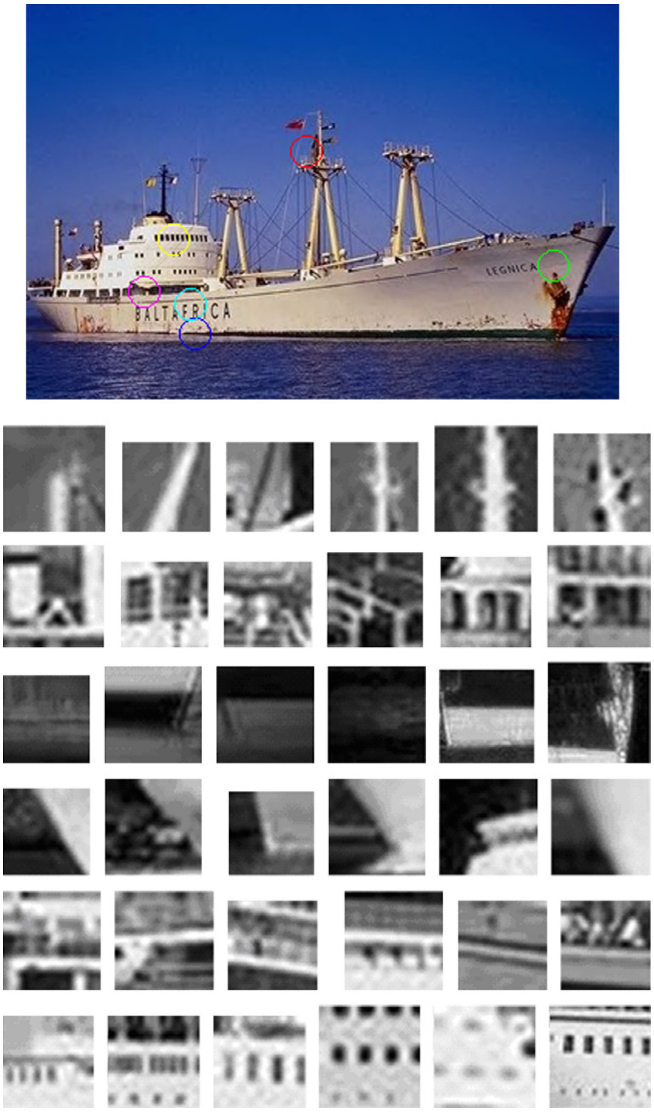

Fig. 11. Examples of visual keywords for "Ship" category. Visual keywords on *backstay* (red circle), *bottom in water* (blue circle), *windows* (yellow circle), *fore or stern* (green circle), *hull* (purple circle) and *baluster* (pink circle). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

object or the scene in different images. Then each image is represented as plenty of salient patches in different visual keywords. For each image category, there are dozens of visual keywords to describe the concept.

5. Conclusions and future work

We have presented a novel approach to categorize images based on the semantic concept of the image category. The feature selection strategy is effective in generating visual keywords to describe image categories. Through the feature selection, objects and scenes are emphasized while similar and non-common noises are reduced. For image categorization, the IP model is proposed to represent the image appearance. The experimental results on the Corel image dataset demonstrate that the proposed feature selection and image categorization approach are effective for image categories with cognitive semantic concepts. As the IP model is category-independent, it can be potentially extended to large-scale image databases.

However, there are also some limitations in our approach. First, the discriminative features between image categories are not well leveraged. The discriminative information can possibly improve the categorization performance. Second, the EM algorithm may lead to local extrema, and therefore it is worthy of investigating adaptive parameter estimation algorithms in the future.

Acknowledgement

This work has been supported by the Grant SRFDP-20050003013.

References

Chen, Y., Wang, J.Z., 2004. Images categorization by learning and reasoning with regions. J. Mach. Learning Res. 5, 913–939.

Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. In: Proc. 8th Eur. Conf. on Computer Vision, pp. 11–14.

Datta, R., Ge, W., Li, J., Wang J.Z., 2006. Toward bridging the annotation–retrieval gap in image search by a generative modeling approach. In: Proc. ACM Internat. Conf. on Multimedia, pp. 977–986.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38.

Deng, Y., Manjunath, B.S., Kenney, C., Moore, M.S., Shin, H., 2001. An efficient color representation for image retrieval. IEEE Trans. Image Process. 10 (1), 140–147.

Fergus, R., Perona, P., Zisserman, A., 2003. Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271.

Fergus, R., Perona, P., Zisserman, A., 2004. A visual category filter for Google images. In: Proc. 8th Eur. Conf. on Computer Vision, pp. 242–256.

Huang, J., Kumar, S.R., Mitra, M., Zhu, W.-J., Zabih, R., 1997. Image indexing using color correlograms. In: Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 762–768.

Kadir, T., Brady, M., 2001. Scale, saliency and image description. Internat. J. Comput. Vis. 45 (2), 83–105.

Li, F.-F., Fergus, R., Perona, P., 2003. A Bayesian approach to unsupervised one-shot learning of object categories. In: Proc. IEEE Internat. Conf. on Computer Vision, pp. 1134–1141.

Liu, X., Zhang, L., Li, M., Zhang, H.J., Wang, D., 2005. Boosting image classification with LDA-based feature combination for digital photograph management. Pattern Recognition 38, 887–901.

Liu, Y., Collins, R.T., 2000. A computational model for repeated pattern perception using frieze and wallpaper groups. In: Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 537–544.

Szummer, M., Picard, R., 1998. Indoor–outdoor image classification. In: IEEE Internat. Workshop on Content-based Access of Image and Video Databases, pp. 42–51.

Vailaya, A., Jain, A., Zhang, H.J., 1998. On image classification: City vs. landscape. Pattern Recognition 31 (12), 1921–1935.

Vailaya, A., Zhang, H.-J., Yang, C., Liu, F.-I., Jain, A.K., 2002. Automatic image orientation detection. IEEE Trans. Image Process. 11 (7), 746–754.

Vasconcelos, N., Vasconcelos, M., 2004. Scalable discriminant feature selection for image retrieval and recognition. In: Proc. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 770–775.

Yu, H., Li, M., Zhang, H.-J., Feng, J., 2002. Color texture moments for content-based image retrieval. In: Internat. Conf. on Image Processing, vol. 3, pp. 929–932.