2013 Sixth International Symposium on Computational Intelligence and Design

An Image Classification Method Based on Matching Similarity and TF-IDF Value of Region

Donghua Xu, Zhiyi Qu*
School of Information Science & Engineering
Lanzhou University
Lanzhou, China
E-mail: 807831808@qq.com, quzy@lzu.edu.cn*

***Abstract*—Traditional image classification methods are mainly based on the overall color statistics and content semantics of the image itself. However, due to the poor distinctiveness of overall color statistics and content semantics, traditional methods often cannot acquire very accurate classification results. In this study, an image classification method based on matching similarity and TF-IDF value of region is proposed. First, an image is divided into some fixed-size regions according to its size, and features of each region are extracted and stored. Then, the matching similarity and TF-IDF value of each region are obtained by comparing with and matching regions in the standard image region dataset. Finally, constructing corresponding feature vectors based on features of each region to get the final image classification result. Experimental results show that the proposed method can be used for fast classification of vast images and gains good classification results.**

Keywords-image classification; matching similarity; TF-IDF; feature vector

I. INTRODUCTION

Image classification is a key technology in the field of computer vision and pattern recognition, and has been widely applied in the process of automatically annotating images, searching massive images and recognizing and filtering the content of images. Traditional image classification methods are mainly implemented by analyzing the low-level vision features of image to gain high-level content semantics [1], [2]. However, poor distinctiveness and irregularity of the low-level vision features and requirement of vast prior knowledge and experience to gain high-level content semantics make image classification face a lot of difficulties.

Researchers have used diverse methods to implement image classification. Berens [3] and Yihong Gong [4] used color histogram descriptors to describe contents of an image. That is, using some quantifying methods to quantify the color components of an image in the color space, calculating the proportion of each quantitative channel in the whole image and describing the content of an image by comparing the proportion of each channel to implement the final image classification. However, this method only considers the overall color statistics and ignores the color distribution information in the image, which may result in that different images with same overall color statistics and different color distribution information are mistakenly classified into the same class. Sheikholeslami [5] proposed an algorithm called SemQuery, constructs a top-down clustering structure of image content by semantically clustering color, shape and other low-level visual features, and classifies an image by querying the clustering structure of image content. But this method requires vast high-quality images used for semantic cluster, and may get bad classification results because of obvious noises in the clustering structure of image content caused by the similarity of visual features in different images. Similarly, on the basis of the potential Latent Dirichlet Allocation Model (LDA), Blei [6] uses a three-layer-structure Bayes model and trains a group of images with semantic features to gain high-level semantics for image classification. This method also requires vast well-trained prior knowledge and experience.

To overcome the disadvantages of above methods, an image classification method based on matching similarity and TF-IDF value of region is proposed in this study. The proposed method constructs a standard image region dataset containing class, TF-IDF value and other information by training different class of standard images. That is, extract features of each region of images in the standard image dataset and update TF-IDF value of each region by the proposed method for calculating matching similarity. In the proposed image classification method, images are divided into some fixed-size regions by size, and features of each region are extracted and stored in the database. Then, calculating the TF-IDF value of each region based on matching similarity by comparing with and matching regions in the standard image region dataset. Finally, constructing corresponding feature vectors based on features of each region and analyzing them to gain the final image classification result. The proposed method for image region extraction is based on color distribution in the image, extracts value of each pixel in the region as the image region feature and effectively avoids wrong classifications caused by the similarity of overall color statistics. Meanwhile, the TF-IDF value of regions in the trained standard image region dataset can be regarded as the importance of current region to decide the image class, which successfully avoids the inaccurate classifications

978-0-7695-5079-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ISCID.2013.142

caused by the poor distinctiveness of content semantics and effectively improves the accuracy of image classification.

II. MATERIAL AND METHODS

A. Image region feature extractios

In this study, an image is divided into some fixed-size regions (containing M*N pixels) according to its size, and features of each region are extracted and stored. First, implement the preprocessing operations of removing the superfluous image regions by its width W and height H. That is, averagely removing W%N pixels in the left and right sides and H%M pixels in the up and down sides of the image. Then, the preprocessed image is averagely divided into some regions and each region contains M*N pixels. Finally, extract the R, G, B color component value for every pixel in each region of the image [7], calculate the average value of color components for every pixel and take all values of pixels in the region of the image as the features of current region. In the experiments, M=N=25.

B. Image region matching similarity calculation

A calculation method for region matching similarity based on features of image region is proposed, and method of gaining the best matching similarity threshold by analyzing the number distribution curve of different similar region is also introduced.

For different image region R_i and R_j, using the formula (1) to calculate the image region matching similarity based on the features of the region:

$$S(R_i,R_j)=\frac{1}{N}\sum_{n=1}^{N}\left(1-\frac{\left|R_i(n)-R_j(n)\right|}{Max[R_i(n),R_j(n)]}\right) \quad (1)$$

Where $R_i(n)$ and $R_j(n)$ are the feature value of pixel n in the image region R_i and R_j respectively, $Max[R_i(n),R_j(n)]$ is the larger of $R_i(n)$ and $R_j(n)$(If $R_i(n)=R_j(n)=0$, $Max[R_i(n),R_j(n)]=1$.), and N is the total number of pixels in the image region (in the experiments, the total number is 625.). If $S(R_i,R_j)=1$, it means that the image region R_i completely matches with R_j.

This study defines the following four concepts: (1) if the matching similarity of two regions is larger than the threshold, the two regions are similar; (2) the number of regions (including itself) similar to a region is called base number of similar region of current region; (3) if the base number of similar region of an image region is larger than 1, the region is a similar region; (4) the number of different similar regions in all image regions is called number of similar region. Randomly select 1000 regions from images of a same class, choose different matching similarity and count the number of similar region to construct the number distribution curve of different similar region (shown in Fig. 1). When the matching similarity is smaller, the number of similar region is less because of larger base number of similar region. While the matching similarity is larger, the number of similar region is less because of less similar regions. When the matching similarity value at the crest of the distribution curve is chosen as the best matching similarity threshold, the number of similar region is the largest and distinguishing different image regions gains the best effects. The best matching similarity threshold is 0.82 in the experiments.

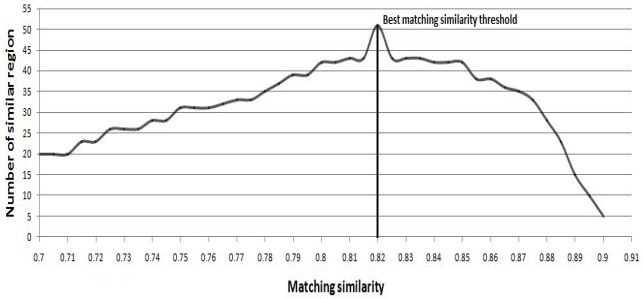

Figure 2. Number distribution curve of different similar region

C. Image region TF-IDF value calculation

TF-IDF method [8] is a classic algorithm in the field of text classification. The TF-IDF value indicates the importance of a word for a file from a file-set or corpus. This study proposes a calculation method for image region TF-IDF value based on the matching similarity. Formula (2) is used to train and calculate the TF-IDF value to the image class T of each image region in different images from the standard image dataset:

$$\begin{aligned} F_T(R) &= tf(R)*idf(R) \\ &= tf(R)*\log(Num/df(R)) \end{aligned} \quad (2)$$

Where Num is the total number of all images, tf(R) is the base number of similar region for image region R to all regions in images of the class T, and df(R) is the number of all image regions from the standard image region dataset which are similar to region R. On the basis of TF-IDF value of each image region, construct a feature vector for all the regions from the standard image region dataset to describe different image class:

$V_T=\{F_{T_1}(R_1),F_{T_2}(R_1),...,F_{T_N}(R_1),F_{T_1}(R_2),F_{T_2}(R_2),...,F_{T_N}(R_2),...,F_{T_1}(R_M),F_{T_2}(R_M),...,F_{T_N}(R_M)\}$, where R_i is a different image region of all regions from the standard image region dataset, M is the total number of all different regions, $F_{T_j}(R_i)$ is the TF-IDF value of region R_i to all images of class T_j, and the standard image dataset contains a total of N different classes of images.

D. Image classification method

To implement image classification, a method to gain the class of images based on the matching similarity and TF-IDF value of regions is proposed. The specific method is as follows:

Step 1: Divide the unclassified image I into some fixed-size (25*25) regions by its size, and extract the features of each image region r_i, where i is the serial number of region.

Step 2: Match each image region r_i from Step 1 with all different image regions in the standard image region dataset, calculate the matching similarity and find out the image region R_j which has the best matching similarity $S(r_i,R_j)$. Construct a feature vector of image I based on its image region matching similarity: $V_R = \{S_{R_1}, S_{R_2}, ..., S_{R_M}\}$, where M is the total number of different regions, and $S_{R_j} = \sum_{r_i \in I} S(r_i, R_j)$ is the sum of matching similarity between image region R_j and all regions in image I that are similar to region R_j. If there is no region in the image I similar to the region R_j, $S_{R_j} = 0$.

Step 3: Use the feature vector V_T of all image regions based on region TF-IDF value and the feature vector V_R of image I based on region matching similarity to construct a feature vector of image I based on the TF-IDF value and matching similarity of regions: $V' = \{S'_{T_1}(R_1), S'_{T_2}(R_1), ..., S'_{T_N}(R_1), S'_{T_1}(R_2), S'_{T_2}(R_2), ..., S'_{T_N}(R_2), ..., S'_{T_1}(R_M), S'_{T_2}(R_M), ..., S'_{T_N}(R_M)\}$, where N is the total number of different classes in the standard image dataset, $S'_{T_i}(R_j) = S_{R_j} * F_{T_i}(R_j)$, S_{R_j} is the sum of matching similarity between image region R_j and all regions in image I that are similar to region R_j and $F_{T_i}(R_j)$ is the TF-IDF value of region R_j to all images of class T_i.

Step 4: Use each dimension of feature vector V' to construct the final feature vector of image I to describe the confidence of different image class: $V = \{S(T_1), S(T_2), ..., S(T_N)\}$, where $S(T_i) = \sum_{j=1}^{M} S'_{T_i}(R_j)$ is the confidence of that image I is classified into the class T_i.

Step 5: On the basis of the confidence of different classes in the feature vector V, use the KNN classification algorithm [9] to gain the class with the largest confidence and take the gained class as the final classification result of image I.

III. Experimental Results

Programs in this study are written in Microsoft Visual C++ 6.0 and OpenCV2.3, and the standard image region dataset is stored in the MySQL database. In the experiments, 10000 images from the standard image dataset Corel are chosen to test the proposed method, which contain bus, flower, dinosaur, beach, building etc. 100 classes and the number of images of each class is 100. 80 images of each class are trained to construct the standard image region dataset and the other 20 are chosen as test samples.

Construct the standard image region dataset by the proposed method for calculating the TF-IDF value of different image regions. As shown in Fig. 2, the TF-IDF value of regions with a white dot is relatively large, such as bus body in images of bus, petals in images of flower and dinosaur body in images of dinosaur. By the proposed calculation method for image region TF-IDF value, the TF-IDF value can be regarded as the importance of current region to decide the image class, that is, the larger the value is, the more important it is for the region to decide the class of the current image.

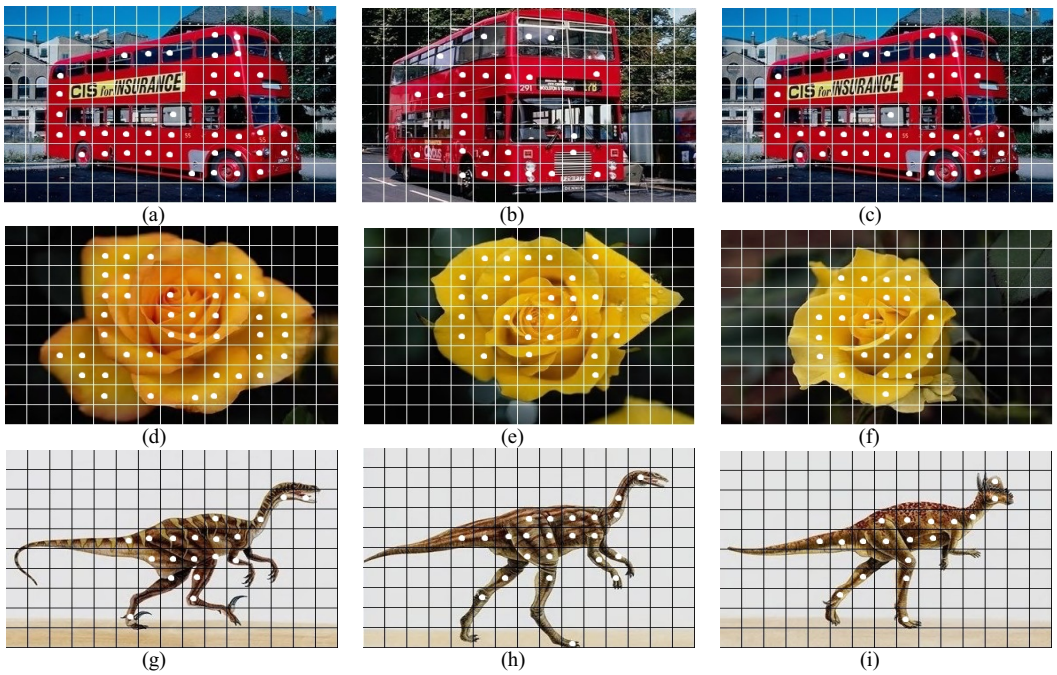

Figure 3. Dotted image regions with large TF-IDF value

A comparison of the DBT-SVM method [10], BT-SVM method [11], T-SVM method [12] and the proposed method to classify 2000 images from 100 different classes is shown in the Table I. The proposed classification method based on the matching similarity (based on color distribution in the image) and TF-IDF value (regarded as the importance of current region to decide the image class) of the region can effectively avoid the inaccurate classifications caused by the poor distinctiveness of overall color statistics and content semantics. The comparison indicates that the classification results of the proposed method are significantly more accurate than the above three methods [10-12], and it can be used for fast classification of vast images and gains accurate classification results.

TABLE I. ACCURACY OF DIFFERENT METHOD

Method	DBT-SVM	BT-SVM	T-SVM	Our method
Accuracy	78.8%	80.2%	80.4%	85.2%

IV. CONCLUSIONS

An image classification method based on matching similarity and TF-IDF value is proposed in this study, and experimental results tested in the standard image dataset Corel show that it is feasible and effective. Using the proposed calculation method for matching similarity and TF-IDF value of image region, we construct a standard image region dataset containing TF-IDF value of each region based on region matching similarity, which effectively avoids the inaccurate classifications caused by the poor distinctiveness of overall color statistics and content semantics and gains accurate classification results in the process of new images classification. However, with the growing size of the standard image region dataset, the speed of adding new regions into the standard image region dataset and classifying new images will decrease. Additionally, classifications of images captured in a different angle have poor robustness. Our future research work will focus on optimizing the storage of the standard image region dataset and increasing the speed of image classification.

ACKNOWLEDGEMENT

This work has been supported by the Key project in the National Science & Technology pillar program (2011BAK08B02).

REFERENCES

[1] Chang S F, Chen W, Sundaram H. Semantic visual templates: linking visual features to semantics [A]. Proceedings of 1998 International Conference on Image Processing [C]. Chicago, Illinois, 1998.4-7.

[2] Zhao R, Grosky W I. Narrowing the semantic gap-improved text-based web document retrieval using visual features [J]. IEEE Transaction on Multimedia, 2002, 4(2):189-200.

[3] Berens J, Finlayson G D, Qiu G. Image indexing using compressed color histograms [J]. IEEE Proceedings, Vision, Image and Signal Processing, 2000, 147(4):349-355.

[4] Yihong Gong, Chua Hock Chuan, Guo Xiaoyi. Image indexing and retrieval based on color histograms [J]. Multimedia Tools and Applications, 1996, 2(2):133-156.

[5] Sheilkholeslami G, Chang W, Zhang A. Semantic clustering and querying on heterogeneous features for visual data. In: Proc. of the ACM Multimedia. Bristol: ACM Press, 1998.3-12.

[6] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation [J]. Journal of Machine Leaning Research, 2003, 3(5):993-1022.

[7] Smith JR, Chang SF. Single color extraction and image query [C]. //Image Processing, 1995. Proceedings., International Conference on vol.3.1995:528-531.

[8] Salton G, Wong A, Yang CS. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11):613-620.

[9] Tan S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus [J]. Expert Systems with Applications, 2005, 28(4):667-671.

[10] Guixiong Liu, Xiongping Zhang. Multi-class classification of Support Vector Machines based on Double Binary Tree [C]. //Fourth International Conference on Natural Computation, 2008:103-105.

[11] Fei B, Liu J. Binary tree of SVM: a new fast multi-class training and classification algorithm [J]. IEEE Transactions on Neural Networks, 2006, 17(3):696-704.

[12] Yong Yin, Yichao lv. Tree structure SVM for image semantic classification [J]. Computer Engineering and Applications, 2012, 48(12):186-189.