IMPROVED SIMILARITY-BASED ONLINE FEATURE SELECTION IN REGION-BASED IMAGE RETRIEVAL

Fei Li, Qionghai Dai, Wenli Xu

Department of Automation, Tsinghua University, Beijing 100084, China

ABSTRACT

To bridge the gap between high level semantic concepts and low level visual features in content-based image retrieval (CBIR), online feature selection is really required. An effective similarity-based online feature selection algorithm in region-based image retrieval (RBIR) systems was proposed by W. Jiang etc., but some parts of the algorithm need to be improved. In this paper, the above algorithm is modified in two aspects: (1) Adaptive mixture models based on mutual information theory are adopted to determine the codebook size. (2) A new method is proposed, which can select not only feature axes parallel to the original ones, but also combined feature axes. Experimental results on 10000 images show that the proposed method can improve the retrieval performance, and save the computational time.

1. INTRODUCTION

Content-based image retrieval (CBIR) has been a very active research area in the last decade. In CBIR systems, each image is represented by a set of low level visual features, and the gap between high level semantic concepts and low level visual features hinders further performance improvement [1]. Relevance feedback [2,3] and region-based image retrieval (RBIR) [4,5,6] have been introduced as two effective mechanisms to bridge the gap. In general, the relevance feedback mechanism is considered as an iterative online supervised learning process. The systems can refine the learning algorithm through users' feedback and improve the retrieval performance. In RBIR, each image is segmented into several regions. Since region-based low level visual features accord with the human perception better than the global features, the systems can often achieve better retrieval results.

It is known that different low level visual features can be more representative of different query concepts. To bridge the gap between semantic concepts and visual features, online feature selection is really needed. However, CBIR online learning has three challenges [7]: small size of the training set, intrinsic asymmetry and fast response requirement. As a result, many traditional feature selection methods in pattern recognition can not satisfy the needs in CBIR. Additionally, in the RBIR context, it is difficult to represent the images of the database in a uniform feature space, because the number of regions in different images may not be the same. Therefore, not much work has been done to deal with the problem of online feature selection in RBIR systems.

In [7,8], a fuzzy codebook based on a generative model is extracted to represent the segmented images in a uniform feature space, and an effective similarity-based online feature selection algorithm in RBIR systems is proposed. The new algorithm can improve the retrieval performance consistently, and is fairly robust to the parameter setting. However, two aspects of the algorithm need to be modified. The first one is the method to determine the codebook size. In [7,8], since the adaptive method based on minimum message length criterion tends to get a small number, the authors determine the codebook size according to the experimental results, which requires heavy computational load and is lack of universality. Another one is that the algorithm can only select feature axes parallel to the original ones, while some combined feature axes may help separate the positive and negative samples better.

To solve the above two problems, an improved similarity-based online feature selection algorithm in RBIR systems is proposed. In this paper, adaptive mixture models based on mutual information theory [9] is adopted to determine the codebook size, and a new method, which can select combined feature axes, is proposed. The rest of the paper is organized as follows: Section 2 gives a brief description of the algorithm in [7,8]. And then how to determine the codebook size by adaptive mixture model is discussed in Section 3. Section 4 describes the new algorithm to select the feature axes online. Our experimental results are presented in Section 5, which is followed by some conclusions in Section 6.

2. SIMILARITY-BASED ONLINE FEATURE SELECTION ALGORITHM

1-4244-0367-7/06/$20.00 ©2006 IEEE

In [7,8], a novel similarity-based online feature selection algorithm in RBIR systems is proposed, and the whole process can be summarized as follows.

The JSEG algorithm is adopted to segment the images, then the region-based visual features are extracted and the feature dimensionality is reduced by Principle Component Analysis (PCA). Based on the assumption that all the features are generated by a Gaussian mixture model, a fuzzy codebook can be extracted by EM algorithm. After the saliency of each region is calculated, the segmented images can be represented in a new uniform feature space.

In order to determine an effective similarity-based feature selection criterion, Fuzzy Feature Contrast Model (FFCM) and Unified Feature Matching (UFM) are adopted to compute the similarity between images and that between image sets, respectively. Then the feature axis, along which the "relevant" and "irrelevant" sets have the smallest similarity, is selected to be the optimal one.

In each feedback round, the Real Adaboost framework is adopted. First, all the positive training samples are set with the same weight, and so are the negative ones. An optimal axis can be obtained according to the above similarity-based feature selection criterion, and a Fuzzy K-Nearest Neighbor (FKNN) classifier is designed along it. Based on the weight update mechanism, all the training samples are re-weighted in the next iteration, a new optimal axis can be selected, and a new FKNN classifier is designed. When the number of the selected axes is enough, a combined classifier is constructed by the ensemble mechanism. Then the retrieval results and the images to be labeled can be obtained based on the classification results of the combined classifier.

The whole algorithm introduces relevance feedback and feature selection into RBIR systems, and improves the retrieval performance. More details can be found in [7,8].

3. DETERMINING THE CODEBOOK SIZE

As discussed in Section 1, the method to determine the codebook size in [7,8] is based on the experimental results, so the computational load is quite heavy. Although the adaptive method based on minimum message length criterion is not practical in this situation, other methods to determine the number of components in the mixture model can be adopted.

A method based on mutual information theory is proposed in [9]. In the new method, the mutual information between one component and all the other components of the Gaussian mixture model can be calculated based on the estimated parameters. To determine the number of the components, a large enough set of components are chosen at first, then the component which has the largest positive mutual information would be removed. The above process repeats until the mutual information of each component with respect to all the others is non-positive.

Since the codebook size is usually quite large, and the computational complexity of EM algorithm is high, we should not choose too large set of components at first. The modified algorithm to determine the codebook size, namely the number of the components in the mixture model, is described in Figure 1:

a) Choose a small number of components $N = N_0$ (such as 100).

b) Estimate the model parameters and calculate the mutual information of each component with respect to all the others $I(\mathbf{u}_i \mid \Theta^{-i})$, $i = 1, 2, \cdots N$.

c) If all the values of mutual information are non-positive, go to the next step; otherwise, stop.

d) Increase the number of components by ΔN (such as 100), go to step b).

Figure 1. The algorithm to determine the codebook size

As the Gaussian mixture model in RBIR is often composed of hundreds of components, and the parameter estimation results of EM algorithm may be influenced by the original values. Therefore, in the experiment, even if the number of the components is smaller than the appropriate number, it can not be guaranteed that each value of the mutual information is non-positive. Therefore, a small threshold ε (such as 0.05) can be set beforehand. In the iteration, if the percentage of the components, each of which has non-positive mutual information, is greater than $1-\varepsilon$, then the iteration must continue until the percentage is smaller than $1-\varepsilon$.

4. FEATURE SELECTION ALGORITHM

The algorithm in [7,8] can only select feature axes parallel to the original ones, and a new algorithm is needed which can select combined feature axes effectively. Since there are operations of "max" and "min" in the calculation of the similarity between the "relevant" and "irrelevant" sets, no analytical optimal expression can be found. Therefore, we may turn to some optimization algorithms.

In [10], sequential 1D optimization algorithm is proposed to find optimal feature axes. However, in the similarity-based feature selection framework, the codebook size is usually quite large, and in each iteration step of the boosting process, a new optimal feature axis would be selected, so the computational complexity of sequential 1D optimization is too high to satisfy the requirement of fast response in CBIR.

In the light of the above optimization algorithm, a combined feature axis, which is the linear combination of the original axes, is supposed to be a better feature axis. At least, it should not be worse than the worst original axis, along which the "relevant" and "irrelevant" sets have the greatest similarity. The combined weight of each original axis can be determined by the similarity of the "relevant" and "irrelevant" sets along it. Obviously, the smaller the similarity of the two sets is, the greater the combined weight of the corresponding axis will be.

Let N be the suitable codebook size determined before, the similarity of the "relevant" and "irrelevant" sets along the i-th original axis be S_i, and the i-th column of the $N \times N$ identity matrix be $\mathbf{e}_i = \left([0 \cdots 0 1 0 \cdots 0]_{1\times N}\right)^T$, $i = 1, 2, \cdots N$, then the axes parallel to the original feature can be denoted as vectors:

$$\mathbf{a}_i = \mathbf{e}_i, \ i = 1, 2, \cdots, N \qquad (1)$$

The combined axis $\mathbf{a}_{N+1}$ can be calculated as follow:

$$\mathbf{a}_{N+1} = \frac{1}{Z}\sum_{i=1}^{N} \mathbf{a}_i \exp(-S_i) \qquad (2)$$

where Z is a normalization factor to make $\|\mathbf{a}_{N+1}\| = 1$. Then the first optimal feature axis can be selected from all the $N+1$ axes. When the sample weights and the obtained similarity results are updated, a new combined axis is constructed by combining the available axes linearly, and a new optimal feature axis can be selected.

As the number of the axes increases, the computational load will become heavier gradually. In order to lighten the load, the worst axis would be removed in each iteration step, and the number of total axes, from which the optimal axis is selected, would be kept unchangeable. Suppose that in each feedback round in CBIR, K optimal feature axes are needed, the whole feature selection algorithm can be summarized in Figure 2:

```
a) Calculate the similarity of the "relevant" and "irrelevant" sets along each original axis, and construct the first combined axis according to Eq.(1) and Eq.(2).
b) Repeat for $k = 1, 2, \cdots, K$
        Select the optimal axis among the $N+1$ axes.
        Remove the worst axis among the $N+1$ axes.
        Construct a new combined axis according to Eq.(2). Here, $\mathbf{a}_i$, $i = 1, 2, \cdots, N$ in the equation stand for the $N$ feature axes left.
        Update the sample weights and the obtained similarity results.
```

Figure 2. The feature selection algorithm

It is supposed that the combined axes should be left in the candidate pool, and some of them should be selected in the iteration process. Therefore, not all of the K optimal feature axes obtained in the process of boosting feature selection are parallel to the original ones and some effective combined axes can be found. Figure 3 gives an example to show the performance of the combined feature axes constructed according to Eq.(2). After reducing the dimensionality of the visual features to 2-D by PCA, images in the "Antelope" category (blue "+"), "Antique" category (green "*") and "Bench" category (red "o") are shown in the figure. Assume the images in "Antelope" category are positive samples, the images in the other two categories are negative samples. Based on the algorithm in [7,8], the similarities between the positive and negative samples along the horizontal and vertical axis are -0.2421 and -0.1203, respectively. According to the above algorithm shown in Figure 2, we can get the combined feature axes. The similarities between the positive and negative samples along the first two combined axes are -0.2573 and -0.2699, respectively, both of which are better than the original ones.

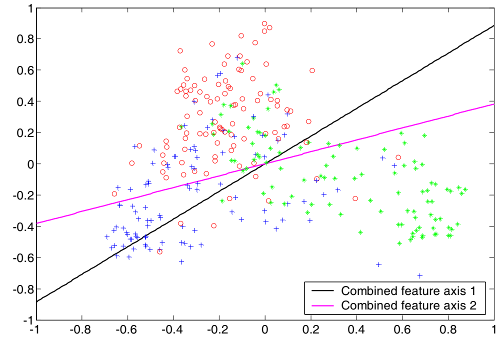


Figure 3. An example of combined feature axes

5. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated on the database of 10000 real-world images from Corel gallery. All the images belong to 100 semantic categories and 100 images in each category. The region-based low-level features adopted in the experiments are the 9-dimensional color moments in LUV color space, the 64-dimentional color histogram in HSV color space, the 10-dimensional coarseness vector and the 8-dimensional directionality, which compose a 91-dimentional feature space in total. In the experiments, the performance measurement used is the top-k precision P_k, which is the percentage of the "relevant" images in the top-k returned images. In order to make a reasonable and fair comparison, P_k is averaged by 1000 query sessions, in which the 1000 query images are selected randomly from the whole database and kept the same in different algorithms. 4 rounds of relevance feedback are conducted in each query session. During each round, the top 10 images among the returned images, which have not been labeled in previous feedback rounds, are labeled by users.

5.1. Determining the codebook size

When the codebook size N varies from 400 to 800, the numbers of the components having non-positive mutual information $N|_{MI\le 0}$ are shown in Table 1, and the average P_{20} and P_{50} in the 4-th feedback round are shown in Table 2. From the experimental results, it can be seen that both the two methods determine the codebook size to be 600, because when the codebook size exceeds 600, the percentage of the components, each of which has non-positive mutual information, is much smaller than 1; at the same time, the improvement of the precision becomes not very significant. Since the method based on adaptive mixture model only needs to calculate the mutual information, rather than to perform retrieval experiments, it cost much less time, obviously.

Table 1. The numbers of the components having non-positive mutual information with different codebook size

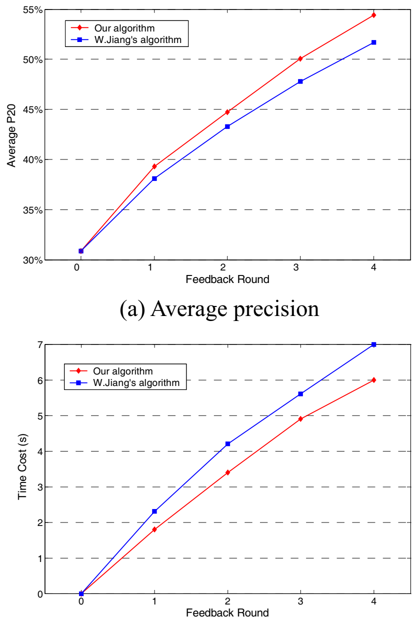
N	400	500	600	700	800
$N\|_{MI\le 0}$	396	494	593	621	634

Table 2. The retrieval results with different codebook size

N	400	500	600	700	800
P_{20}	0.423	0.471	0.544	0.560	0.567
P_{50}	0.212	0.246	0.282	0.293	0.301

5.2. Retrieval performance comparison

Let the codebook size be 600, the total numbers of axes to be selected in the feature selection algorithm in this paper and that in [7,8] be 60 and 150, respectively. The average P_{20} of each algorithm is shown in Figure 4(a), and the corresponding average time cost is shown in Figure 4(b). Although the number of selected axes is smaller, our algorithm can achieve better retrieval performance, because the classifiers constructed along the combined axes may have better classification ability. Since less feature axes are needed to be selected, our algorithm cost less computational time.



(a) Average precision

(b) Average time cost

Figure 4. Comparison between our proposal and W. Jiang's algorithm

6. CONCLUSIONS

In this paper, the similarity-based online feature selection algorithm, proposed by W. Jiang etc, is modified in two aspects. Adaptive mixture models based on mutual information theory are adopted to determine the codebook size, and a new method is proposed to select optimal feature axes, involving not only axes parallel to the original ones, but also combined axes. Experimental results show the effectiveness of our proposal.

7. ACKNOWLEDGEMENTS

This work is supported in part by the key project of NSFC (No.60432030) and in part by the Distinguished Young Scholars of NSFC (No.60525111).

8. REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* 22 (12), pp. 1349-1380, 2000.

[2] Y. Rui, T.S. Huang, M. Ortega and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Tran. Circuits and Systems for Video Technology*, 8 (5), pp. 644-655, 1998.

[3] X.S. Zhou and T.S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems,* 8, pp. 536-544, 2003.

[4] C.Carson, S. Belongie, H. Greenspan and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Analysis and Machine Intelligence,* 24 (8), pp. 1026-1038, 2002.

[5] J.Z. Wang, J. Li and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," *IEEE Trans. Pattern Analysis and Machine Intelligence,* 23 (9), pp. 947-963, 2001.

[6] F. Jing, M. Li, H. Zhang and B. Zhang, "An efficient and effective region-based image retrieval framework," *IEEE Trans. Image Processing,* 13 (5), pp. 699-709, 2004.

[7] W. Jiang, G. Er, Q. Dai and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE Trans. Image Processing,* 15 (3), pp. 702-712, 2006.

[8] W. Jiang, G. Er, Q. Dai, L. Zhong and Y. Hou, "Relevance feedback learning with feature selection in region-based image retrieval," *Proc. of IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, pp. 509-512, 2005.

[9] Z.R.Yang and M. Zwolinski, "Mutual Information Theory for Adaptive Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23 (4), pp. 396-403, 2001.

[10] C. Liu and H.Y. Shum, "Kullback-Leibler Boosting," *Proc. of IEEE Int. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp 587-594, 2003.