ISSN:0975-9646
D.Mahidhar et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (5) , 2010, 329-331

A NOVEL APPROACH FOR RETRIEVING AN IMAGE USING CBIR

[1]D.Mahidhar, [2]M.V.S.N.Maheswar

Department of Computer Science Engineering PVP Siddhartha Institute of Technology,
Vijayawada, Andhra Pradesh, India.

[1]E-mail:mahidhar_dm@yahoo.co.in
[2]E-mail:marepalli.maheswar@gmail.com

Abstract— **In this paper we mainly focussed on how to retreive the images from large database. Generally in huge databases we will have large number of images with the same name. when we want to retrieve the image by giving a name,we may get some of the images with the same name,but in our method it retrieves the images based on the content in the image and it also show the comparision between present image in the database with the target image we are searching for.This is fast and efficient method for retriving the exact images from huge databases.**

Keywords: content based , image.

I. Introduction

In content-based image retrieval (CBIR), the search may be initiated using a query as an example. The top rank similar images are then presented to the user. Then, the interactive. Process allows the user to refine his request as much as necessary in a relevance feedback loop. Many kinds of interaction between the user and the system have been proposed [1], but most of the time, user information consists of binary labels indicating whether or not the image belongs to the desired concept. The positive labels indicate *relevant* images for the current concept, and the negative labels irrelevant images. To achieve the relevance feedback process, the first strategy focuses on the query concept updating. The aim of this strategy to refine the query according to the user labeling. A simple approach, called query modification, computes a new query by averaging the feature vectors of relevant images[2][3].Another approach, the query reweighting, consists in computing a new similarity function between the query and any picture in the database. A usual heuristic is to weigh the axes of the feature space [4]. In order to perform a better refinement of the similarity function, optimization-based techniques can be used. They are based on a mathematical criterion for computing the reweighting, for instance Bayes error [5], or average quadratic error [6], [7]. Although these techniques are efficient for target search and monomodal concept retrieval, they hardly track complex image concepts. Performing an estimation of the query concept can be seen as a statistical learning problem, and more precisely as binary classification task between the relevant and irrelevant classes [8]. In image retrieval, many techniques based on statistical learning have been proposed, as for instance Bayes classification [9], k-Nearest Neighbors [10], Gaussian Mixtures [11], Gaussian random fields [12], or Support Vector Machines. In order to deal with complex and multimodal concepts, we have adopted a statistical learning approach. Additionally, the possibility to work with kernel functions is decisive. However, a lot of these learning strategies consider the CBIR process as a classical classification problem, without any adaptations to the characteristics of this context. For instance, some discriminative classifiers exclusively return binary labels when real values are necessary for CBIR ranking purposes. Furthermore, the system has to handle classification with few training data, especially at the beginning of the search, where the query concept has to be estimated in a database of thousands of images with only a few examples. Active learning strategies have been proposed to handle this type of problem. Another point concerns the class sizes, since the query concept is often a small subset of the database. In contrast to more classical classification problems, relevant and irrelevant classes are highly imbalanced (up to factor 100). Depending on the application context, computational time has also to be carefully considered when an online retrieval algorithm is designed. To address this problem, we assume that any learning task must be at most $O(n)$, where n is the size of the database. In this paper, we focus on statistical learning techniques for interactive image retrieval. We propose a scheme to embed different active learning strategies into a general formulation.

II. RELATED WORK

We have experimented several classification methods for CBIR: Bayes classification with a Parzen density estimation-Nearest Neighbors, Support Vector Machines, Kernel Fisher Discriminant [13], and also a query-reweighting strategy [14]. Databases, scenario, evaluation protocol and quality measurement are detailed in the appendix. The results in terms of Mean Average Precision according to the training set size (we omit the KFD which gives results very close to inductive SVMs) for both ANN and Corel databases. One can see that the classificationbasedmethods give the best results, showing the power of statistical methods over geometrical approaches, like the one reported here (similarity refinement method). The

SVM technique performs slightly better than others in this context. As they have a strong mathematical framework and efficient algorithmic implementation, SVMs are used as the default method. We also made some comparisons between different kernels (linear, Polynomial, Triangle, Gaussian $L1$, Gaussian $L2$). In our experiments, the , Gaussian $L1$, Gaussian $L2$ gives the best performance, which is not a surprising result since histograms are used as image signatures, and the $\frown 2$ distance is dedicated for comparing distributions. In the following experiments, we always use a Gaussian kernel with a distance.As the whole data set is available during the training, it could be interesting to consider a semi-supervised or transductive framework. For instance, there are extended versions of Gaussian Mixtures, and transductive SVM . We have experimented with these methods. The computational time is very high and no improvement has been observed. Transduction does not seem to be an interesting approach for CBIR as already observed. Anyway, the global performances remain very low for any proposed methods for the Corel experiments. The MAP is under 20%, even when the number of training data is up to100 for Corel. The Corel database is much bigger than ANN and the simulated concepts are more difficult. The training set remains too small to allow classifiers to efficiently learn the query concept. According to this distinction, we have selected two popular active learning strategies for presentation and comparison.

1) Uncertainty-based sampling*:* In our context of binary classification, the learner of the relevance function has to classify data as relevant or irrelevant. Any data in the pool of unlabeled samples may be evaluated by the learner. Some are definitively relevant, others irrelevant, but some may be more difficult to classify. Uncertainty-based sampling strategy aims at selecting unlabeled samples that the learner is the most Uncertain about. To achieve this strategy, a first approach proposed by Cohn uses several classifiers with the same training set, and selects samples whose classifications are the most contradictory. Another solution consists in computing a probabilistic output for each sample, and selecting the unlabeled samples with the probabilities closest to $0\triangleright 5$ [14]. Similar strategies have also been proposed with SVM classifier, with a theoretical justification, and with nearest neighbor classifier [30]. In any case, a relevance function $f_{\mathcal{A}_y}$ is trained. This function may be adapted from a distribution, a membership to a class (distance to the hyper plane for SVM), or a utility function. The efficiency of these methods depends on the accuracy of the relevance function estimation close to the boundary between relevant and irrelevant classes.

2) Error Reduction-based strategy*:* active learning strategies based on error reduction aim at selecting the sample that, once added to the training set, minimizes the error of generalization of the new classifier.

III PRECISION ORIENTED SELECTION

As explained before, the straight estimation of the Average Precision is not efficient in our context. We opted for a different strategy but have kept in mind the objective of optimization of the ranking. We experimentally noticed that Selecting images close to the boundary between relevant and irrelevant classes are definitively efficient, but when considering subset of images close to the boundary, the active sampling strategy of selecting the closest image inside this subset is not necessarily the best strategy. We propose to consider the subset of images close to the boundary, and then to use a criterion related to the Average Precision in order to select the winner image of the selective sampling strategy. In order to compute a score related to Average Precision, we propose to consider the sub-database of the labeled pictures. We have the ground truth for this sub-database, since we have the labels of all its images. Thus, it becomes feasible to compute the Average Precision on this sub-database, without any estimation. We still need a ranking of this sub-database in order to compute the Average Precision. We compute the similarity $k(\mathbf{x}_i, \mathbf{x})$ of an unlabeled image $\mathbf{x}_i$ to any labeled images $\mathbf{x}$ (using the kernel function k as the similarity function), and then rank the labeled images according to these similarity values. The resulting factor $m_{\mathcal{A}_y}(\mathbf{x}_i)$ is the Average Precision on the sub-database with this ranking The aim of this factor $m_{\mathcal{A}_y}(\mathbf{x}_i)$ is to support the picture $\mathbf{x}_i$ that, once labeled, will have the most chance to increase the Average Precision.

IV SELECTION BASED ON PRECISION

As explained before, the straight estimation of the Average Precision is not efficient in our context. We opted for a Different strategy but have kept in mind the objective of Optimization of the ranking. We experimentally noticed that Selecting images close to the boundary between relevant and irrelevant classes is definitively efficient, but when considering a subset of images close to the boundary, the active sampling strategy of selecting the closest image inside this subset is not necessarily the best strategy. We propose to consider the subset of images close to the boundary, and then to use a criterion related to the Average Precision in order to select the winner image of the selective sampling strategy. In order to compute a score related to Average Precision, we propose to consider the sub-database of the labeled pictures.

We have the ground truth for this sub-database, since we have the labels of all its images. Thus, it becomes feasible to compute the Average Precision on this sub-database, without any estimation. We still need a ranking of this sub-database in order to compute the Average Precision. We compute the similarity $k(\mathbf{x}_i, \mathbf{x})$ of an unlabeled image $\mathbf{x}_i$ to any labeled images $\mathbf{x}$ (using the kernel function k as the similarity function), and then rank the labeled images according to these similarity values. The resulting factor $m_{\mathcal{A}_y}(\mathbf{x}_i)$ is the Average Precision on the sub-database with this ranking The aim of this factor $m_{\mathcal{A}_y}(\mathbf{x}_i)$ is to support the picture $\mathbf{x}_i$ that, once labeled, will have the most chance to increase the Average Precision.

V ACTIVE BOUNDARY CORRECTION

During the first steps of relevance feedback, classifiers are trained with very few data, about 0▹1% of the database size. At this stage, classifiers are not even able to perform a good estimation of the size of the concept. Their natural behavior in this case is to divide the database into two parts of almost the same size. Each new sample changes the estimated class of hundreds, sometimes thousands of images. Selection is then close to a random selection. A solution is to ensure that the retrieval session is initialized with a minimum of examples. For instance, Tong proposes to initialize with 20 examples. Beyond the question of the minimum number of examples required for a good starting boundary, initial examples are not always available without some third-party knowledge (for instance, keywords).Thus , we propose a method to correct the boundary in order to reduce this problem.

VI CONCLUSION

This is a fast and efficient method which is used to retrieve the images from huge database with the same name compared to other methods. This is content based image retrieving technique which also compares how much content is differ from other images

REFERENCES

[1] B. Sch¨olkopf and A. Smola, *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
[2] P. Yin, B. Bhanu, K. Chang, and A. Dong, "Integrating relevance Feedback techniques for image retrieval using reinforcement learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1536–1551, October 2005.
[3] T. Joachim's, "Transductive inference for text classification using Suppor vector machines," in *Proc. 16th International Conference on Machin Learning*. Morgan Kaufmann, San Francisco, CA, 1999, pp. 200–209.
[4] P. Gosselin, M. Najjar, M. Cord, and C. Ambroise, "Discriminative Classification *vs* modeling methods in CBIR," in *IEEE Advanced Concepts for Intelligent Vision Systems (ACIVS)*, Brussels, Belgium, September2004.
[5] P. Gosselin and M. Cord, "RETIN AL: An active learning strategy for Image category retrieval," in *IEEE International Conference on Image Processing*, vol. 4, Singapore, October 2004, pp. 2219–2222.
[6] N. Roy and A. McCallum, "Toward optimal active learning through Sampling estimation of error reduction," in *International Conference on Machine Learning*, 2001.
[7] D. Cohn, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.
[8] D. D. Lewis and J. Catlett, "Heterogeneous uncertainly sampling for Supervised learning," in *International Conference on Machine Learning*. W.W. Cohen and H. Hirsh, editors, July 1994, pp. 148–56.
[9] J. M. Park, "Online learning by active sampling using orthogonal Decision support vectors," in *IEEE Workshop on Neural Networks for Signal Processing*, vol. 77, no. 3, December 2000, pp. 263–283.
[10] M. Linden Baum, S. Marko itch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers," *Machine Learning*, pp. 54(2):125–152, February 2004.
[11] I. Cox, M. Miller, T. Minka, T. Papa Thomas, and P. Finials, "The Bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments," *IEEE Transactions on Image Processing*. vol. 9, no. 1, pp. 20–37, 2000.
[12] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *International Conference on Machine Learning*, 2003.
[13] A. Dong and B. Bhanu, "Active concept learning in image databases," *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*,vol. 35, pp. 450–466, 2005.
[14] N. Vasconcelos and M. Kunt, "Content-based retrieval from image Databases: current solutions and future directions," in *International Conference on Image Processing*, vol. 3, Thessalonica, Greece, October 2001, pp. 6–9.
[15] Y. Chen, X. Zhou, and T. Huang, "One-class svm for learning in image retrieval," in *International Conference in Image Processing (ICIP)*, vol. 1, Thessalonica, Greece, October 2001, pp. 34–37.
[16] S.-A. Berrani, L. Amsaleg, and P. Grogs, "Approximate searches: kneighbors