Relevance Feedback in Content-based Image Retrieval Systems, an Overview and Analysis

Divna Djordjevic and Ebroul Izquierdo, *Member, IEEE*

Abstract — **In this paper an overview of relevant developments in visual relevance feedback based image retrieval is presented. Important problems of content-based image retrieval are analyzed and relevant findings from the evaluation of our framework are reported.**

***Keywords* — Image retrieval, relevance feedback, semantic meaning, taxonomy.**

I. Introduction

THE growth of visual digital media often fused with audio and written semantic content is dictating new approaches in information based retrieval not only limited to one kind of content but aiming at joining all in an ongoing effort to teach, modify and adopt a machine to human based reasoning. Only in the latter case would the end user in either professional or personal oriented scenarios be satisfied with the outcome of a retrieval system. However, we have to keep in mind that the subjectivity and fuzziness of human reasoning will always attribute a component of noise to the overall picture.

In modern and continuously growing media databases the inability of accessing accurate and desired content can be as limiting as the lack of content itself. Research in information retrieval whether based on textual semantic description or low-level content is aiming at overcoming the drawbacks of machine limited behaviour and incorporating human understanding into complex equation of machine responses.

Although the information retrieval problems have appeared many years ago and have been investigated in the past this was manly done for text based databases. Textual annotation is extremely time-consuming and not practical though it was initially developed to preserve knowledge. The availability of properly annotated content it real world consumer oriented scenarios is limited and is dependent on subjective interpretations of the professional annotator. Unlike textual information which is human define and precise in meaning, a picture, audio-video content has a hidden component of creative reasoning of the human brain giving it an overall shape and meaning far beyond capabilities of any language based representation.

In the last decades we have moved from text only retrieval systems to multimedia content databases fusing all information together. Throughout the last decade the idea of simulating human understanding is closely related to iterative human feedback and incorporated the obtained knowledge into a learning approach that could eventually be able to "think", "behave" as a human. As means of achieving a step closer to bridging the well known "semantic gap" between human and machine driven reasoning, iterative short term and low-effort relevance feedback has been presented as unavoidable step into training any leaning algorithm. In this paper we present analyses of the "semantic problem" concentrating on image retrieval with relevance feedback and the classification problem and presenting the results and some issues coming out of such a complex cognitive problem.

In section II the research problem is defined and in section III the related research areas dealing with content-based image search and retrieval are presented. Section IV gives a description of our evaluation framework based on retrieval and iterative user feedback and in section V some experimental results are presented. The final section presents the conclusion.

II. Definition of the problem

Several relevance feedback (RF) algorithms have been proposed over the last few years. The idea behind most RF-models is that the distance between images labelled as relevant and other similar images in the database should be minimal. The key factor here is that the human visual system does not follow any mathematic metric when looking for similarity in visual content and distances used in image retrieval systems are well-defined metrics in a feature space. Therefore numerous issues appear when modelling a content-based image retrieval system:

1. Appropriate definition and selection of a feature space in order to present not only image content but the interpretational characteristics of the human visual system.

2. How to define distances to simulate similarity matrices between representative content-based feature vectors, again to express high-level semantic similarity observed by humans.

3. Interoperations of high-level conceptual content are not only subjective to a particular user and the appropriate knowledge he might posses but also to the circumstances of the search and retrieval scenario.

Divna Djordjevic is with the Multimedia and Vision Lab, Electrical Engineering Department, Queen Mary University of London, UK (phone: +44 (0)20 7882 7880; fax:+44 (0)20 7882 7997; e-mail: divna.djordjevic@elec.qmul.ac.uk).

Prof. Ebroul Izquierdo is head of the Multimedia and Vision Research Lab, Electrical Engineering Department, Queen Mary University of London, UK (phone: +44 (0)20 7882 5354 fax:+44 (0)20 7882 7997; e-mail: ebrou.izquierdo@elec.qmul.ac.uk).

The issues mentioned above present difficulties of enabling an efficient content-based image retrieval system and are closely related to a number of similar pattern recognition problems.

III. Related Research

Based on the feature space, the matrices used and the underlying learning process inferring leaning preferences of the end users a number of different approaches have been developed.

The PicHunter system presented in [1] uses "stochastic-comparison search". Once a user selects relevant images, a comparison search is done over the whole database. In the simplest one-dimensional case, a binary search comparing any element from the meta-database and the target vector is conducted. Other models use feature re-weighting or query point movement strategies [2]-[4] in these cases feature re-weighting is based on "term frequency" and "inverse document frequency" techniques similar to those studied and used in text retrieval [5]. Early probabilistic RF systems assumed that positive images follow a single Gaussian distribution [5] with the more advanced systems being based on Gaussian mixture models. To estimate the parameters of the Gaussian mixture model it is assumed that positive examples can be grouped in feature space and that they are mutually separated by negative examples. There are also a number of techniques for image retrieval integrating neural network learning approaches and relevance feedback. Among different strategies, self-organizing maps are used to index images according to low-level features. Self-organizing maps are unsupervised topologically ordered neural networks, which project a high-dimensional input space into a low-dimensional lattice. The latter, usually being a two-dimensional grid with n-dimensional neighbours connected in appropriately weighted nodes. Using query-by-example the model combines self-organizing maps and RF in an iterative refinement process. Some schemas, e.g., the PicSOM system [6], reduce the complexity inherent to the training of large self-organizing maps using a hierarchical structure called Tree-structured Self-Organizing Maps. Another class of approaches combines neural network based learning with fuzzy user RF is described in [7]. In this model the user provides a fuzzy judgment about the relevance of an image. In response to user's perception images are weighted with different fuzzy factors. The learning process involves user's preferences and visual content interpretation combined in a single layer radial bases function neural network. In another set of approaches to relevance feedback discriminative analysis is often used to address the two class relevant/irrelevant problem. In this case separability of the two user defined classes is measured by how far apart are the projected means of each class and how large is the variance of the data along the projected direction. Discriminative Expectation Maximization [8] considers image retrieval as a learning problem with labelled and unlabeled data samples and combines expectation maximization and discriminative analysis. It is used to estimate both, the parameters of the probabilistic density model and the linear transformation that maps the original feature space into another features space. The overall goal of this approach is to find a linear transformation based on the labelled data set and generalize it to the unlabelled dataset so that the inter-class scattering is maximized and intra-class scattering minimized. In [9] authors use a combination of weighted retrieval system with Mahalanobis distance as a similarity measure and support vector machines (SVM) for estimating the weight of relevant images in the covariance matrix. This approach is a combination of already exploited techniques and new statistical learning algorithm SVM. The overall similarity for a particular image in the database is obtain by linearly combining similarity measures for each feature, as in many other approaches already mentioned. The approach proposed in [4] is a region-based method for extracting local region features. In this approach the authors combine regions and perform image-to-image similarity matching by using Earth Mover's Distance, which allows different dimensions of feature vectors. In SVM-based classification both positive and negative labelled images are used as training data to learn the classifier how to separate the unknown part of the database, the test set, into two or more classes. In our evaluation framework we also use a SVM as a binary classifier, which incorporates user provided knowledge into the learning process with the implementation based on sequential minimization optimization algorithm.

IV. Evaluation Framework

Generalized content-based image retrieval systems with relevance feedback (CBIR-RF) Fig. 1, should satisfy several conditions:

1. Display images to the user without repetition.
2. Take relevant and/or irrelevant, fuzzy information as an input provided by the user on iterative bases.
3. Learn user's preferences by adapting behaviour of the system based on provided knowledge accumulating in time.

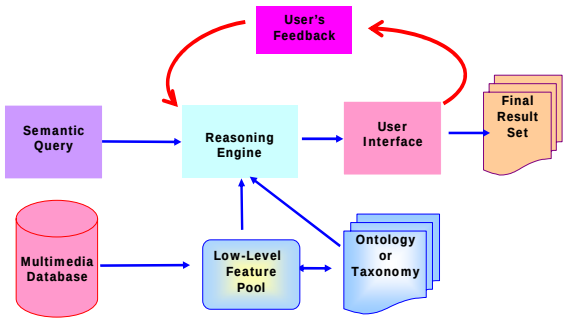

Fig. 1. Generalized architecture of a query based retrieval engine with relevance feedback.

The general system displayed in Fig. 1 would not only use low-level feature information from the available content itself but also prior knowledge incorporated into ontology or taxonomy. In our case we have a number of predefined categories and randomly choose images to iteratively refine the search. Our leaning method is a machine leaning

classifier, SVM which is a kernel based approach for defining a separating hyperplane between two different classes. We have used a radial basis function kernel and kept the kernel parameters fixed since in a real on-line process we can not search for optimal parameters. So rather than long term personalized search an online adaptive content retrieval is needed and relevance feedback approaches manage to satisfy this need.

V. Experimental setup and results

Our experimental database consists of 2100 images from the Corel collection (corel.com), for the purpose of experimenting we limit ourselves to random selection of relevant and irrelevant images from appropriate classes to be trained and classified by support vector machines (SVMs). What we are expecting is increased confusion with increased number of relevant images per concept, due to the overall background noise present in images and disagreement of low-level features and human envisaged concepts.

In Fig.1 some of the representative examples of the image categories taken from the Corel database are presented. It can be seen that all of the classes are conceptually very different and there are no overlaps in concepts Fig. 2.

Fig. 2. Some of the representative images of different categories forming our image database, each category has 100 images.

The problem of the semantic gap is reencountered in many aspect of any pattern recognition problem, here specified for a classification scenario. Images can conceptually satisfy many different categories all based on a form of predefined taxonomy and have appropriately high variance in low level features. In Fig. 3 some of the alternative category definitions could be: “a rose, close-up”, “cluster of roses on a green background”, “a rose with visible parts of the background”. Though conceptually the same all of the mentioned sub-categories can have numerous definitions and when a user is searching for particular content, the human brain is automatically processing and creating the picture and concepts.

For all the images belonging to a same conceptual category there is a considerable variance in low-level features understandable by a machine.

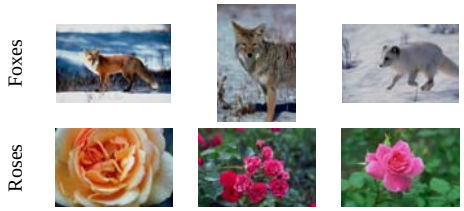

Fig. 3. Conceptual consistency does not imply descriptive low-level features consistency in colour and texture across images within a category, the semantic gap.

For each image low-level extracted features are Colour Structure and Edge Histogram descriptor as representatives of MPEG-7 Standard descriptors [10] for colour and texture. The colour structure descriptor represents colour distribution of a content (as in colour histogram) and local spatial structure of this content. Edge histogram describes the local edge distribution of an image. Five types of edges are defined (horizontal, vertical, diagonal 45 degrees, diagonal 135 degrees and non-directional).

Several scenarios are defined for assessing classification performance on database with whole images taking into account the objects of interest and background (noisy) information. We experiment trying to identify categories as tigers and elephants from our database, in Fig. 2, the representatives of these categories are in second row middle and on the far left, respectively.

Scenario I: First, a particular concept is considered with the assumption that the number of training samples is increasing for all classes and in every iteration. Starting from 10 samples for each class, in each step 10 samples are added for both positive and negative classes (e.g. elephants and not elephants). We also take into account that each category has 100 images, and reduce the testing set in every iteration accordingly to the increased training set.

Scenario II: A particular concept is considered with the assumption that the number of positive training samples is increasing in every iteration. Starting from 10 samples for each class and adding each time 10 samples for positive class, with keeping the number of negative samples fixed to 10; only information about samples that are “good” representatives of the relevant class is added.

Experiments were done for the mentioned scenarios, three cases of descriptors and for two categories: tigers and elephants. The training and testing sets are generated based on either Scenario I or II. Combining SVM classifiers and mentioned descriptors we judge the classification performance in various scenarios mentioned above. In our case relevant class is either category elephants or tigers and the irrelevant class is the rest of the database consisting of many non-relevant classes.

We calculate overall accuracy in our retrieval task as the fraction of the total number of correctly predicted cases for both classes; therefore it is a measure of how well the relevant class is predicted and at the same time how well are the rest of examples predicted as irrelevant. To achieve

good distinction properties it is not only required that the relevant class is predicted correctly but also that irrelevant samples are "far" from relevant ones, therefore high accuracy is an important measure Fig. 4.

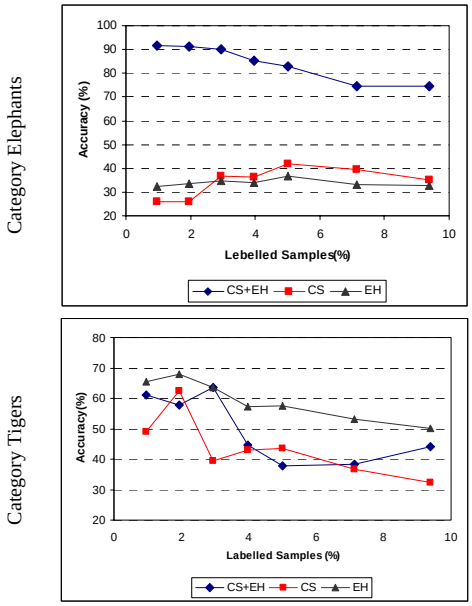

Fig. 4. Accuracy for the two categories elephants and tigers, depending on the percentage of labelled training set over the whole testing database for Scenario I. The accuracy is calculated for three different cases of descriptors: Colour Structure (CS), Edge Histogram (EH) and combination (CS+EH).

As it can be seen for Fig. 4 though for category elephants we have an increase in performance when using more descriptors this is not a case for the category tigers. The difference in performance comes from the fact that not all low-level descriptors are good enough representatives of a particular concept in mind and that the diversity in images can often lead to worse performance even though more information is available and the images belong to the same semantic concept. In the case of combined descriptor it can be observed that category elephants have better accuracy whereas in case of separate descriptors category tigers perform slightly better. The overall conclusion is that though we might be able to fit the available descriptors to best represent a category in question using the ones that give higher accuracy, finding a generalized case with good performances for various image categories is very difficult and when dealing with various images with diverse background it is highly unstable.

When combining Scenario I and Scenario II in Fig. 5 it can be seen that in most cases with increasing the number of training samples the combination taking into account both positive and negative examples performs better. Herein more information helps the classifier perform better but the problem of noisy data within images is persistent across both the relevant and irrelevant classes.

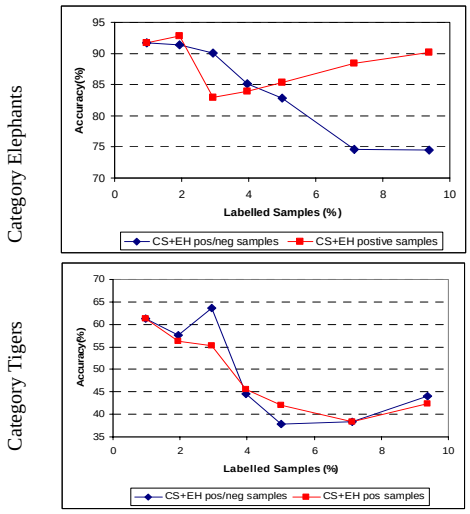

Fig. 5. Accuracy for category Elephants and category Tigers and for combined descriptor CS+EH, for the two scenarios: Scenario I taking into account both positive and negative samples and Scenario II taking into account only positive samples.

VI. Conclusion

Our experiments give a strong message in two general directions when dealing with CBIR systems; one is that these systems are highly subjective and therefore unstable. And second is that only by selecting representative features of particular objects and avoiding the nosy background we can correctly model image data and achieve better retrieval performances and stability.

References

[1] J. Cox, M. Miller, T. Minka, P. Yianilos,"An optimized interaction strategy for Bayesian relevance feedback", *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998, pp. 553-558.

[2] M. Koskela, J. Laaksonen, E. Oja, "Comparison of techniques for content-based image retrieval", Proceedings *of 12- th Scandinavian Conference on Image Analysis*, Norway, 2001.

[3] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra,"Relevance feedback: A power tool in interactive content-based image retrieval", *IEEE Tran. Circuits and Systems for Video Technology*, 1998, Vol. 8, No 5, pp. 644-655.

[4] F. Jing, M. Li, H.-J. Zhang, B. Zhang, "Relevance feedback in region-based image retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, 2004, Vol. 14, No. 5

[5] Y. Rui, T.S. Huang, S. Mehrotra,"Content-based image retrieval with relevance feedback in MARS", *Proceedings of IEEE Int. Conf. on Image Processing*, 1997, pp. 26-29.

[6] M. Koskela, J. Laaksonen, E. Oja, "Use of image subsets in image retrieval with self-organizing maps", *Proceedings for International Conference on Image and Video Retrieval*, 2004, pp. 508-516.

[7] K. Wu, K.H. Yap, "Fuzzy relevance feedback in content-based image retrieval", *Proc. Int. Conf. Information and Signal Processing and Pacific-Rim Conf. Multimedia*, Singpore, 2003.

[8] Q. Tian, Y. Wu, T.S. Huang, "Incorporate discriminate analysis with EM algorithm in image retrieval", *In Proc. IEEE International Conf. on Multimedia and Expo*, 2000.

[9] Q. Tian, P. Hong, T. S. Huang, "Update relevant image weights for content-based image retrieval using support vector machines", *IEEE International Conference on Multimedia and Expo*, Hilton New York & Towers, New York, 2000.

[10] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, A. Yamada, "Color and texture descriptors**",** *IEEE Transactions on Circuits and Systems for Video Technology*,Volume: 11, Issue: 6, Jun 2001, pp. 703-715.