Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Image and Signal Processing (CIISP 2007)

Human-Controlled Vs. Semi-automatic Content-Based Image Retrieval

Kambiz Jarrah, *Student Member, IEEE*, and Ling Guan, *Senior Member, IEEE*
Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada
{kjarrah, lguan}@ee.ryerson.ca

Abstract— **The overall objective of this paper is to present a methodology for reducing the human workload through adapting an automatic scheme for Content-Based Image Retrieval (CBIR) engines. The proposed system utilizes an unsupervised hierarchical clustering algorithm, known as the Directed Self-Organizing Tree Map (DSOTM) that aims to closely mimic the process of information classification thought to be at work in the human brain [1, 2]. To further refine the search process and increase retrieval accuracy, a Semi-automatic relevance feedback approach is presented in this work. The Semi-automatic scheme refers to a relevance feedback CBIR engine, structured around the DSOTM algorithm. This system aims to learn from and adapt to different users' subjectivity under the guidance of an additional objective verdict provided by the DSOTM. Comprehensive comparisons with the Rank-based, relevance feedback, and automatic CBIR engines, demonstrate feasibility of adapting the Semi-automatic approach.**

I. Introduction

Rank-based CBIR engines simply neglect the semantic similarities among target images by searching and retrieving images according to degree of (statistical) similarities between the query and its neighboring images. Such systems assume direct association between statistical similarities and semantics of the query image and disregards inter-relationships among target images. Fig. 1 clearly illustrates the limitation of such an approach. In this figure, the query image is located on top-left corner of the figure and the top 16 images are ranked and retrieved – from left to right, and top to bottom – according to the decaying level of their likeness with respect to the query. It is evident that target images with high feature similarities may not be semantically similar to the query image due to the gap between low-level features used for image indexing and the high-level concepts used by human observers.

Past efforts to bridge this gap emphasized simulating human perception of visual contents *via* the human-computer interaction (HCI) – also known as Human-controlled or Relevance Feedback (RF) – scheme: a learning mechanism that allows retrieval systems to adapt to the users' needs by tuning the proximity matching process toward semantic levels using low-level visual features.

The Human-controlled approach is an extended application of the modern information retrieval (IR), proposed by Salton and McGill [3], in the image retrieval process. In IR systems, each document is represented by a set of key words and terms. These terms are then concatenated in a set of vectors (a.k.a. "Vector Models") and are then made available for search and retrieval. Some of the well-known implementations of the HCI approach in CBIR application are Multimedia Analysis and Retrieval System (MARS) [4], PicToSeek [5], DrawSearch [6], and Viper [7]. In all of the above systems, some kind of query refinement strategy (i.e., feature weighting) has been adapted to interactively create a new query with the goal of optimizing the search process.

Fig. 1: A sample query (top-left) to demonstrate behavior of the rank-based CBIR system.

There are few problems, however, associated with the above implementations of interactive learning approach: First, they suffer from limited degree of adaptivity due to incapability of distance measurement techniques used in above systems to adequately model perceptual differences as seen by the human user [8]. Secondly, these systems require a high degree of user involvement in providing feedback samples through many cycles of relevance feedback before convergence. Lastly, these systems suffer from potential human (subjective) errors due to their dependency on users' judgments on resemblance of retrieved images [16].

In view of the above problems, an unsupervised learning algorithm, namely, the Directed self-organizing tree map (DSOTM) was introduced [14]. The resulting search engine aims to *minimize* both human subjectivity and workload by replacing repetitive user interaction steps by the DSOTM module, which adaptively guides relevance feedback, to bridge the gap between low-level image descriptors and high level semantics. The proposed system also takes advantage of an adaptive technique based on non-linear radial-basis function (RBF) model [9] that aims to model human perceptual similarities among images. To further reduce this gap and achieve an enhanced performance for the CBIR system under study, a RF approach was proposed in conjunction with the DSOTM. The resulting framework,

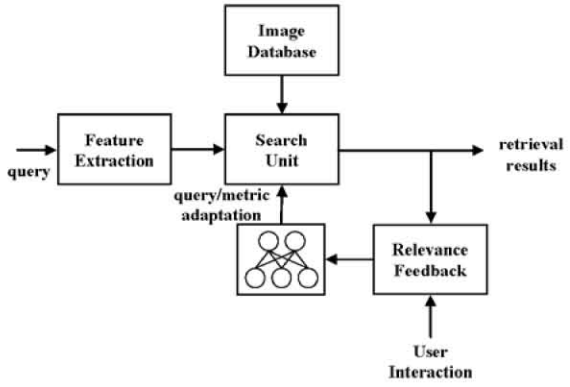


Fig. 2: Adaptive Human-controlled relevance feedback CBIR system [8].

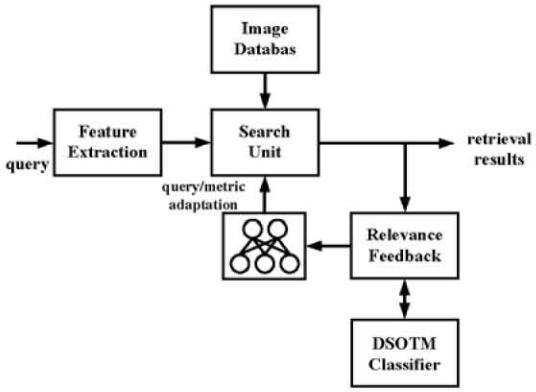


Fig. 3: Machine controlled CBIR system [11].

referred to as Semi-automatic CBIR, aims at learning different user subjectivity and increasing retrieval accuracy.

This paper provides some detailed descriptions on both RF and DSOTM algorithms in Sections 2 and 3; Section 4 discusses the Semi-automatic architecture in the CBIR application; a comprehensive comparison between Rank-based, fully interactive, Automatic, and Semi-automatic CBIR architectures is also presented in Section 5; Section 6 summarizes the paper with some remarks.

II. Interactive Approach in CBIR

As illustrated earlier in Fig. 1, image retrieval based on the statistical image representation and the linear similarity approximation is unable to completely articulate the users' requirements on semantic levels. This is due to the gap between low-level features and the high-level concepts, as pointed out previously.

Fig. 2 illustrates the adapted architecture for the RF-based CBIR system, proposed by Muneesawang et al. [8]. This system takes advantage of an adaptive technique based on non-linear RBF model for learning the users' notion of similarity between images. In this process, the user is provided with a set of retrieved images and is asked to select those with the highest (semantic) similarity with respect to the query image. Feature vectors extracted from selected images are then used as training seeds to determine centers and widths of different RBF units in the network. RBF is an attractive technique for simulating the human perception. Using the RBF-based learning model offers further adaptability to the retrieval system to refine the search to different users and various types of images rather than to enforce a fixed metric for comparisons.

RBF is a kernel function that has an outstanding approximation capability for *non-linear* proximity evaluation. One of the major properties of RBF is its localization capability: the trait that is determined by its exponentially decaying (or growing) behavior with respect to the distance from a mean point [9].

In this work, a one-dimensional Gaussian RBF is associated with each component of image feature vector and is used for the purpose of the nonlinear proximity evaluation between query, $\mathbf{z}$, and input image, $\mathbf{x}$, feature vectors. On the other hand, each RBF unit provides a nonlinear mapping of distance versus similarity where the highest similarity is achieved when $\mathbf{z} = \mathbf{x}$.

In addition to utilizing a non-linear metric for annotating perceptually resembled images, refining the query to a best representation of such likeness is also important. In the process of image retrieval, there are situations where the selected query can not entirely reflect users' preferences due to uncertainty of image interpretations as a result of ambiguous image contents (i.e., the presence of several objects of interest). Under such circumstances, the system often generates trivial or even irrelevant retrieval results due to its incapability of extracting all the required information that leads it to converge toward the query (relevant) class. In such situation, readjusting query location more toward a best representative class with the objective of retrieving more relevant images at subsequent RF iterations can significantly improve the retrieval accuracy. This is possible through the so-called Query Modification process [10]. These modifications are carried out based on information (or preferences) provided by the user from earlier iterations of RF.

Several schemes have been studied for the purpose of query modification in the literature. Tuning the query position to the center of mass of relevant samples *via* calculating the mean value of the training vectors associated with users' selected images can be a good indication of both the relent class itself and users' preferences. This method is effective in situations where there are significant numbers of relevant samples available for the user to select from (i.e., late stages of retrieval). In a situation where there is a small subset of the actual relevant class available (i.e., early stages of retrieval), this query modification scheme will not perform adequately since sparse data resolution can extensively impact the modified query by diverging it from the true position of the relevant cluster center. In such circumstances, the query can be modified by the information extracted from both relevant and irrelevant sub-samples. As a result, the query is adjusted to a new position by shifting it away from the irrelevant group and more toward the relevant image cluster.

The combination of user's interactions, query modification, and RBF learning unit enables the retrieval

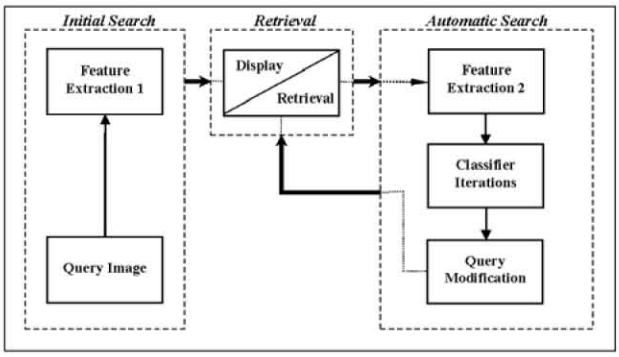


Fig. 4: Machine-controlled CBIR system - A closer look [8].

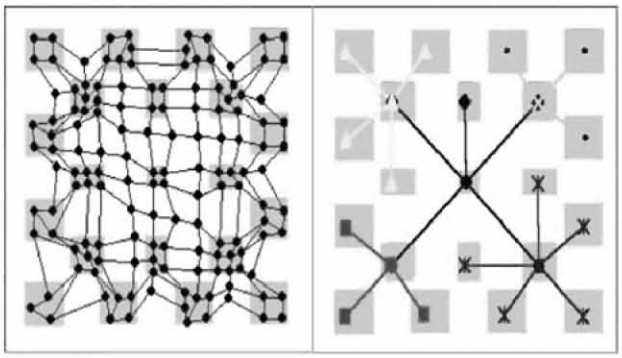
system to directly integrate users' preferences (semantics) into the retrieval process. Usually, depending on the nature of the query image, quality of feature descriptors, and learning curve of the system, duration of users' supervision is varied from a few to tens of interactions before the algorithm converges (i.e., until a satisfactory result is achieved). Such limitation makes the human-controlled CBIR engine essentially unsuitable in practice.

In view of the above limitation, an automatic image retrieval scheme is proposed here. Such system aims to minimize both human subjectivity and workload by replacing the required interactions with an unsupervised learning unit, namely, the Directed self-organizing tree map (DSOTM). This scheme is subject of the next section.

III. Automatic Approach in CBIR

Fig. 3 illustrates the proposed architecture for an automatic CBIR. The automatic image retrieval system in this figure differs from its interactive counterpart *via* integrating unsupervised data clustering principles into the retrieval process, thereby exploiting the DSOTM algorithm. Fig. 4 provides a detailed representation of Fig. 3. This figure can also be generalized to Fig. 2 by replacing the (*unsupervised*) *Classifier Iterations* module with the (*supervised*) *User Interactions*.

The DSOTM is an unsupervised machine learning algorithm and is inspired by principles found in Kohonen's self-organizing feature map (SOFM) [12] and Kong's Self Organizing Tree Map (SOTM) [13]. Similar to both SOFM and SOTM, DSOTM tends to follow the self-organization and competitive learning principles discussed in [1]; however unlike SOFM, DSOTM tends to grow a more dynamic topology (more plastic than SOFM) that not only extracts global intuition from an input pattern space but also injects some degree of localization into the discriminative process, such that maximal discrimination becomes a priority at any given resolution (or number of classes). Also, comparing with SOTM, DSOTM algorithm not only provides a partial supervision on cluster generation by forcing divisions away from the query class but also makes a gradual decision about the resemblance of the input patterns by constantly modifying each sample's memberships during the learning phase of the algorithm [14].



(*a*) (*b*)

Fig. 5: Self-organizing data clustering: (*a*) performed by SOFM, nodes converge to areas of zero data density; (*b*) performed by DSOTM, no nodes converge to join the areas of zero data density [15].

DSOTM is chosen in the current application, as the problem in image retrieval has different characteristics than other data classification applications: First, the training data set required by relevance feedback learning algorithms is very small, e.g., a few to tens of samples. Also, the feature space is of a very high dimension consisting of a combination of color, shape, and texture features. These tend to form sparsely distributed data. Secondly, a problem is caused by an unbalanced data distribution between relevant and irrelevant samples in the training set. It is expected that, after the first iteration of relevance feedback, the relevant items are retrieved more than irrelevant ones, and thus, the majority of relevant items will introduce an unbalanced space to the resulting clusters. To solve this problem, the DSOTM allows for a focus on maximization of discrimination within sub-regions of the (unbalanced) training data *via* competition among its hierarchically-discovered nodes. This efficient allocation and breakdown of class relationships minimizes classification errors compared to that achieved through SOFM, which unfold across data space, often leading to distortion within sparse interstices (see Fig. 5).

The algorithm for generating the DSOTM map is summarized in a simplified flowchart depicted in Fig. 6. Details associated with each of the main components are given in the following steps:

Initialization: A root node $\{\mathbf{w}_j\}_{j=1}^{J}$ is chosen from the available set of input vectors $\{\mathbf{x}_k\}_{k=1}^{K}$ in a random manner. J is the total number of centroids (initially set to 1) and K is the total number of input vectors (i.e., images);

Similarity Measurement: A new data point, $\mathbf{x}$, is randomly selected and the best-matching (winning) centroid, j^*, is found through the minimization of the predefined Euclidean distance criterion in (1):

$$\mathbf{w}_{j^*}(t) = \arg\min_j \left\| \mathbf{x}(t) - \mathbf{w}_j(t) \right\|, \qquad j = 1,2,\ldots,J. \tag{1}$$

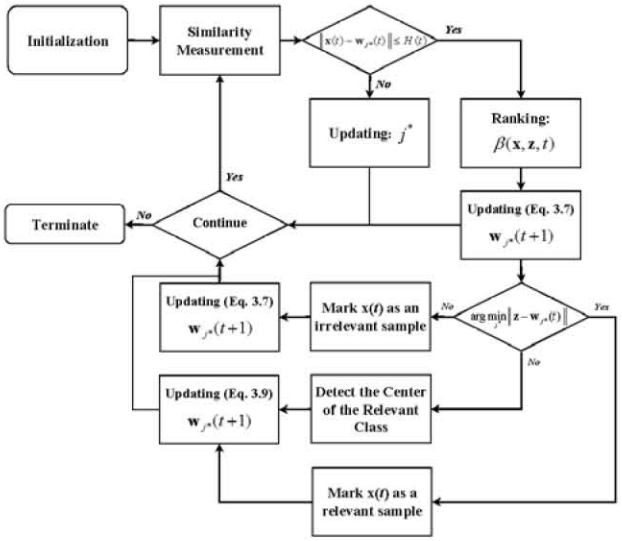


Fig. 6: The DSOTM Flowchart for the CBIR application.

Updating: ***If*** $\left\|\mathbf{x}(t)-\mathbf{w}_{j^*}(t)\right\| \le H(t)$, where $H(t)$ is the hierarchy function used to control the levels of the tree and decays exponentially over time from its initial value, $H(t_0) > \sigma_{\mathbf{x}}$, according to $H(t+1) = \lambda \cdot H(t) \cdot \exp(-t/\rho)$, where λ is the threshold constant, $0 < \lambda < 1$, and $\rho = \max(t)/\log_{10}[H(t)]$. The proposed threshold function is empirically established to decay faster than the one employed in the SOTM architecture [13]. As a result, the network is given a better opportunity to generate the required centers at its initial training phase and learn from them at the later stages of training. Alternatively, the preliminary training phase in the DSOTM algorithm is prioritized with the node generation process while the later stages are dominated with learning about the existing information. ***Then*** $\mathbf{x}(t)$ is assigned to the j^{th} centroid, and the synaptic vector is adjusted according to the reinforced learning rule:

$$\mathbf{w}_{j^*}(t+1) = \mathbf{w}_{j^*}(t) + \alpha(t) \cdot \beta(\mathbf{z},\mathbf{x},t) \cdot [\mathbf{x}(t) - \mathbf{w}_{j^*}(t)] \quad (2)$$

where $\alpha(t)$ is the learning rate, which decays exponentially over time as more neurons are allocated, $\alpha(t) = \alpha(t_0) \cdot \exp[-t/\max(t)]$, $0.01 \le \alpha(t) \le \alpha(t_0)$, and $\alpha(t_0) = 0.1$; and $\beta(\mathbf{z},\mathbf{x},t)$ is the exponential ranking function that measures the similarity between query feature vector, $\mathbf{z}$, and input feature vector, $\mathbf{x}$, at an automatic relevance feedback iteration from the previous search operation as is indicated in (3) [11]:

$$\beta(\mathbf{z},\mathbf{x},t) = \sum_{i=1}^{P} G_i(x_i - z_i) = \sum_{i=1}^{P} \exp\left(-\frac{(x_i - z_i)^2}{2\sigma_i^2}\right) \quad (3)$$

In this equation, P is the total number of features, $\sigma_i = \eta \max_i |x_i - z_i|$ is the tuning parameter, and η is an additional factor to ensure a large output for $\beta(\mathbf{z},\mathbf{x},t)$. A large value of $\beta(\mathbf{z},\mathbf{x},t)$ indicates a high relevance of the feature vector compared to the respective query feature at time t as is illustrated in Fig. 7. As a result, the synaptic vectors are adjusted so that they learn more from statistically similar inputs and less from statistically irrelevant ones. *Else* form a new centroid node starting with $\mathbf{x}$, reset the learning rate to its initial value (i.e., $\alpha(t_0) = 0.1$), and increment j by 1;

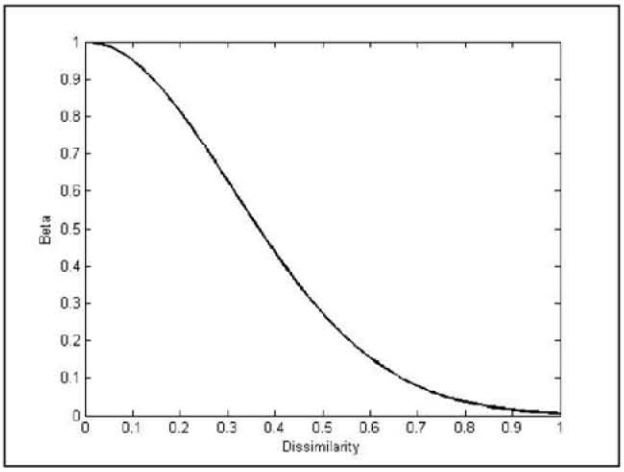


Fig. 7: The magnitude of $\beta(\mathbf{z},\mathbf{x},t)$ represents the similarity between the input vector $\mathbf{x}$ and the query $\mathbf{z}$, where the highest similarity is attained when $\mathbf{x} = \mathbf{z}$.

Cluster Adjustment and Relevance Identification: ***If*** $\left\|\mathbf{x}(t)-\mathbf{w}_{j^*}(t)\right\| \le H(t)$ and $\arg\min_j \left\|\mathbf{z}-\mathbf{w}_{j^*}(t)\right\|$, that is, if the closest center to the current input data is also the closest center to the query, then mark $x(t)$ as a relevant sample and update its centroid (winning neuron) toward the query position according to the degree of resemblance of the sample using:

$$\mathbf{w}_{j^*}(t+1) = \mathbf{w}_{j^*}(t) + \alpha(t) \cdot \beta(\mathbf{z},\mathbf{x},t) \cdot [\mathbf{z} - \mathbf{w}_{j^*}(t)], \quad (4)$$

else mark $x(t)$ as an irrelevant sample ***and*** update its centroid using (2). Subsequently, find and move center of relevant class further toward the query center using (4);

Continuation: The *Similarity Matching* step is repeated until the maximum number of iterations is reached, the maximum number of clusters is generated, and/or no noticeable changes in the feature map are observed.

The *Cluster Adjustment and Relevance Identification* step in the DSOTM algorithm imposes some constraints on cluster generation near the query position and, thus, avoids unnecessary boundaries to be formed around it. As a result, a better sense of relevance measurements can be achieved as the tree structure develops. Moreover, the growth of the DSOTM is biased *via* the ranking function to learn more from input vectors deemed to be similar to the query itself,

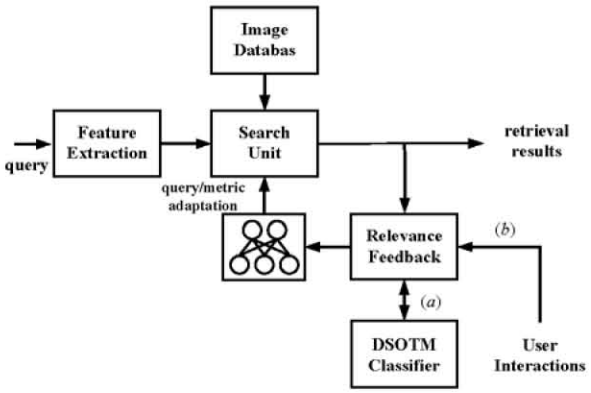


Fig. 9: Semi-automatic CBIR system [10].

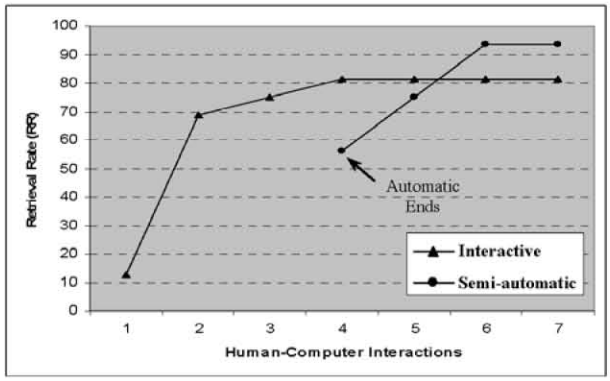


Fig. 10: A comparison of retrieval performance at convergence, between the Interactive and Semi-automatic methods.

and less from images far from the query. This promotes the generation of multiple irrelevant classes, while maintaining necessary plasticity in the relevant class [11, 14].

IV. Semi-automatic Approach in CBIR

The architecture of the proposed Semi-automatic CBIR engine is depicted in Fig. 9. The retrieval process in this scheme starts by an automatic search through the database (path (*a*)). The retrieval result is next displayed back to the user, when the user continues the search interactively through path (*b*). The interactive process continues until a satisfactory result is achieved [10]. The description of each path is given in previous sections. During this process, the DSOTM objectively decides on relevance of individual images and guides adaptations of an RBF-based relevance feedback network while the user coordinates the search subjectively by continuously highlighting and feeding relevant images to the retrieval system. As a result, the system requires minimum user interactions to achieve a more accurate performance.

V. Experimental Results

A number of experiments were conducted to compare the behaviors of Rank-Based, Interactive, Automatic, and Semi-automatic CBIR engines.

The simulations were carried out using a subset of the Corel image database consisting of nearly 12000 JPEG color images, covering a wide range of real-life photos, from 120 different categories [15]. Each category consisted of 100 visually associated objects to simplify the measurements of the retrieval accuracy during the experiments. 120 query images were randomly drawn from the database such that no two images were from the same class. Retrieval results were statistically calculated from top 16 most relevant images with respect to each query.

In this work, Color Histograms and Color Moments accompanied with Hu's seven moment invariants (HSMI) and Gabor Descriptors were used to construct feature vector for each image in the database [14].

Experimental results are illustrated in Table 1. Interactive approach clearly outperforms both Rank-Based and Automatic search, while Semi-automatic approach surpasses

TABLE I
Experimental Results in terms of *RR*

CBIR Engine	Rank-Based	Interactive	Automatic	Semi-automatic
Ave.	***43.4%***	***61.6%***	***56.4%***	***67.3%***

the Interactive search. These are expected results since the human-computer interaction method is directly steered by the users, while the Automatic search utilizes machine learning algorithm (DSOTM) to decide on relevance of input samples. Fig. 10 provides a vivid illustration of system's performance with respect to a sample query for each input of a user. Even though the Semi-automatic process still requires significant human supervision, the amount of required human-computer interactions for the system to converge is less than would be required for typical relevance feedback type CBIR systems. On the other hand, according to our experiments, the Semi-automatic CBIR approach requires, on average, 2.1 iterations while the Interactive approach requires 4.3 iterations out of 7 designated interactions to converge. By investigating the above results, it is evident that utilizing a Semi-automatic CBIR approach can successfully reduce required human interactions - therefore, human errors - and increase system's performance in terms of retrieval rate.

VI. Conclusion

This paper presents various architectures used in the current CBIR technology and compares their performance with regards to one another. It was mentioned that the simple architecture of the Rank-based CBIR engine is unable to incorporate semantic meaning to the retrieval process due to inadequateness of high-level concepts representation through statistically descriptive features. To incorporate semantics, Interactive CBIR architecture was introduced. Although such architecture integrates semantics in the CBIR, it suffers from high degree of human involvements. To tackle this problem, an automatic RF CBIR engine was introduced in this paper. In such a framework, DSOTM is incorporated with CBIR technology in order to perform the required decision making about the relevance of individual images and classify input patterns while preserving the integrity of the image clusters, mainly

the query class. Thus, a great degree of reduction in users' interaction can be achieved. A Semi-automatic CBIR engine was also discussed in this paper. Experimental results illustrate feasibility of adapting such architecture to reduce human interactions and increase systems' performance.

REFERENCES

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd Ed. Prentice Hall, New Jersey, 1999.

[2] S. Coren, L. M. Ward, J. T. Enns, *Sensation and Perception,* 6th Ed., John Wiley & Sons Inc., USA, 2004.

[3] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, New York, McGraw Hill, 1983.

[4] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based image retrieval with relevance feedback in MARS," *in Proc. of IEEE Int. Conf. Image Processing,* pp. 815-818, Santa Barbara, CA, 1997.

[5] T. Gevers and A. W. M. Smeulders, "PicToSeek: Combining color and shape invariant features for image retrieval," *IEEE Trans. on Image Processing,* Vol. 9, pp. 102-119, 2000.

[6] E. Di Sciascio and M. Mongiello, "DrawSearch: A tool for interactive content-based image retrieval over the net," in Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, Vol. 3656, pp. 561-572, 1999.

[7] H. Mller, W. Mller, S. Marchand-Maillet, and D. M. Squire, "Strategies for positive and negative relevance feedback in image retrieval," in Proc. IEEE Int. Conf. on Pattern Recognition, Vol. 1, pp. 1043-1046, Barcelona, Spain, Sep. 2000.

[8] P. Muneesawang, L. Guan, "A Non-Linear RBF Model For Adaptive Content-Based Image Retrieval," *Int. Symp. on Multimedia Information Processing*, pp: 188-191, University of Sydney, Australia, Dec. 2000.

[9] T. Sigitani, Y. Liguni, and H. Maeda, "Image interpolation for progressive transmission by using radial basis function networks," *IEEE Trans. on Neural Networks,* Vol. 10, No. 2, pp. 381-390, 1999.

[10] P. Muneesawang and L. Guan, "Minimizing user interaction by automatic and Semi-automatic relevance feedback for image retrieval," *in Proc. of IEEE Int. Conf. on Image Processing,* Vol.2, pp. 601-604, Rochester, USA, Sep. 2002.

[11] K. Jarrah, M. Kyan, I. Lee, and L. Guan, "Application of image visual characterization and soft feature selection in Content-Based Image Retrieval," *in Proc. of SPIE Multimedia Content Analysis, Management, and Retrieval (SPIE'06),* Vol. 6073, pp. 101-109, San Jose, California, USA, Jan. 2006.

[12] T. Kohonen, "The self-organizing map," *Proc. of the IEEE*, Vol. 78, Issue 9, pp. 1464 – 1480, Sept. 1990.

[13] H. S. Kong, "The Self-Organizing Tree Map, and its Applications in Digital Image Processing," PhD Thesis. University of Sydney, Australia, 1998.

[14] K. Jarrah, S. Krishnan, and L. Guan, "Automatic Content-Based Image Retrieval Using Hierarchical Clustering Algorithms," *International World Congress on Computational Intelligence (WCCI'06),* pp. 6564-6569, Vancouver, Canada, Jul. 2006.

[15] Corel Gallery Magic 65000, *http://www.corel.com,* 1999.

[16] K.-H. Yap and K. Wu, "A soft relevance framework in content-based image retrieval systems," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, pp. 1557- 1568, Dec. 2005.