International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:12, No:12, 2018

A Context-Sensitive Algorithm for Media Similarity Search

Guang-Ho Cha

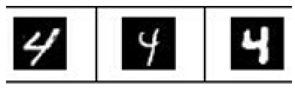
***Abstract*—**This paper presents a context-sensitive media similarity search algorithm. One of the central problems regarding media search is the semantic gap between the low-level features computed automatically from media data and the human interpretation of them. This is because the notion of similarity is usually based on high-level abstraction but the low-level features do not sometimes reflect the human perception. Many media search algorithms have used the Minkowski metric to measure similarity between image pairs. However those functions cannot adequately capture the aspects of the characteristics of the human visual system as well as the nonlinear relationships in contextual information given by images in a collection. Our search algorithm tackles this problem by employing a similarity measure and a ranking strategy that reflect the nonlinearity of human perception and contextual information in a dataset. Similarity search in an image database based on this contextual information shows encouraging experimental results.

***Keywords*—**Context-sensitive search, image search, media search, similarity ranking, similarity search.

I. Introduction

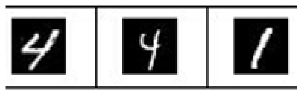
SIMILARITY search is essential to media information retrieval systems. In content-based media retrieval (CBMR), the media object is usually characterized by feature vector in a very high dimensional space, say, over 100. The degree of similarity between two media objects in CBMR is usually measured by Euclidean distance. However, users sometimes have experienced the mismatch between their requests and the results returned from the CBIR system. While the notion of similarity is usually based on high-level abstraction, the system-based low-level features used in the similarity comparison do not sometimes reflect the human perception. In this paper, we attempt to find a solution for negotiating this semantic gap by devising a new similarity measure and a new similarity ranking algorithm.

Let us consider a real example that shows the standard definitions of similarity fail to produce reasonable search results. We would like to select two images similar to a given query image in a handwritten digit image database. Fig. 1 (a) is a query image and Figs. 1 (b) and (c) are the results when a human selects two images similar to the query image from the database. On the other hand, Figs. 2 (b) and (c) are the actual results from the similarity search experiment when the Euclidean distance is used as a measure for similarity comparison. The digit '1' in Fig. 2 (c) is not matched to the query digit '4'. This result clearly shows that there exists a discrepancy between human perception and the distance metrics such as Euclidean distance. Therefore, in CBMR, it is necessary to establish the link that bridges the gap between human perception and distance calculation. Based on this concept, we define the semantics of an image dataset as the contextual relationship in the dataset, in other words, the information about the distribution and cluster structure of a given dataset. This concept also agrees with the cluster assumption [6]: (1) Points in the same local high density region are more similar to each other than to points outside this region; (2) Points in the same global structure are more similar to each other than to points outside this structure. This concept is the basis of our similarity metric.



(a) (b) (c)

Fig. 1 Human perception based retrieval: (a) is a query image; (b) and (c) are two images retrieved by similarity search



(a) (b) (c)

Fig. 2 Euclidean distance based retrieval: (a) is a query image; (b) and (c) are the two images retrieved by similarity search

In this paper, we attempt to tackle the semantic gap problem in image search by capturing the nonlinear relationships in contextual information given by an image collection. It has been widely acknowledged in media search that a query concept is typically a nonlinear combination of perceptual features [7]. In this paper, we first conduct the nonlinear transformation on the original dataset not only to capture nonlinear relationships but also to simulate human perception. For nonlinear transformation, we adopt a Gaussian function because it possesses an excellent nonlinear approximation capability [8], [9]. Compared to the conventional Minkowski metric (or L_p-norm), the similarity search based on the contextual information offers a more effective search results from a human viewpoint.

II. Related Work

A collection of research prototype and commercial CBMR systems have exploited the visual features of images, such as colors, shapes, and textures, to represent and index image

Guang-Ho Cha is with Seoul National Univesity of Science and Technology, Seoul 01811, South Korea (phone: +82-2-970-6710; fax: +82-2-977-9441; e-mail: ghcha@seoultech.ac.kr).

contents. However, it is widely noted that there is a semantic gap between the visual features and the semantic meanings of images, and it has been a major problem in CBMR systems [23].

The representation of semantic meaning perceived by users in a certain domain is a complex problem. Many researchers have proposed the use of relevance feedback to bridge the semantic gap between the automatically captured visual features and the human interpretation of image content. To take account of relevance feedback in image search, several researchers have explored *supervised learning* [10]. Tong and Chang [4] proposed $\text{SVM}_{\text{active}}$ for learning a decision boundary by iteratively adding the most informative samples as training data. Hoi and Lyu [11] developed a *soft-label* support vector machine (SVM) by taking the feedback confidence into consideration in learning the decision boundary. Despite these efforts, we note that though SVM [10] is effective for classification, the decision boundaries derived by the two schemes would be unstable when the feedback contains only a few image examples for on-line training.

The automatic image annotation methods [12]-[15] usually employ segmentation techniques to generate keyword-based annotations for the images being indexed to facilitate semantic searching. However, the segmentation techniques deployed on the images are often not robust enough to produce meaningful semantics. In addition, clustering of images based on the keyword output may include noise, and is usually error-prone [23].

The latest trend in the image search has shifted somewhat towards recovering the intrinsic structure for a proper image space of reduced dimensionality. Instead of working with the conventional Euclidean space, the main theme is to assume that the images are spread as a *manifold*, and the task is to learn the underlying structure of a dataset. Consequently, a similarity measure can be computed on the learned manifold. He et al. [16] used geodesic distances to approximate the distances between image pairs along the manifold. However, the main drawback is that the mapping is defined only on the set of *training data*, and thus needs additional mechanisms, such as radial basis function networks, to handle test data. Similarly, Wu et al. [7] also proposed a method for formulating a context-based distance function for measuring similarity. It uses the *kernel function* [10] to nonlinearly transform traditional distances into similarity in a transformed feature space. However, it requires human intervention to collect the contextual information and also needs the contextual information in the form of *training data*.

In [17], a non-metric distance function called dynamic partial function (DPF) was proposed to measure perceptual similarity. Although DPF proved that it works better than Minkowski metric for measuring perceptual similarity, it is actually difficult to dynamically select features to be used for distance computation.

In this paper, similar to the semantic manifold learning methods, we first conduct the nonlinear transformation on the original dataset not only to capture nonlinear relationships but also to simulate human perception. However, unlike [7], [16], we do not require human intervention to collect the contextual information and the training data.

III. Nonlinear Similarity Model

In this section, we discuss the difference between the assumptions under our work and the similarity models proposed in the literature and introduce our nonlinear similarity model, which captures the contextual information as well as simulates human perception for similarity evaluation.

A. Nonlinear Similarity Function

Constructing an effective image search system requires accurate characterization of visual information. Conventional models based on Minkowski distance metrics do not adequately capture all the aspects of the characteristics of the human visual system. The visual section of the human brain is known to use a *nonlinear* processing system for tasks such as pattern recognition and classification [8]. We refer to the systems that evaluate the degree of similarity between two images linearly proportional to the magnitude of their distances as the *linear* model-based retrieval methods.

In order to simulate human perception for similarity evaluation between images, we first establish a *nonlinear* model. The assumption for the nonlinear approach is that the same lengths of the distances do not always give the same degrees of similarity when judged by humans [18]. In other words, the linear model is not competent for the nonlinear nature of human perception and cannot cope with the complex decision boundary. We therefore propose to use a nonlinear criterion in performing similarity comparison.

The nonlinear model is constructed by an input-output mapping function $f(x)$ that uses feature values of input image x to evaluate the degree of similarity to a given query [19]. In its most common form, the input-output mapping function should be *smooth* in the sense that similar inputs correspond to similar outputs. We adopt the following Gaussian function as our basic similarity model for the input-output mapping function:

$$G(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right) \tag{1}$$

The Gaussian function possesses an excellent nonlinear approximation capability [8], [9]. The *exponential* similarity function is more sensitive to local changes and gives rise to better performance improvement with respect to human perception. Psychophysical studies find that the similarity, or the confusion frequency, between stimuli decays exponentially with some power of the perceptual measure of distance [20], [21].

B. Nonlinear Transformation

In our similarity model, it is needed to extract the global information to classify the dataset into clusters. While it is unclear how to extract this global information, it is easy to extract the local property of the points, i.e. the pairwise similarity of the points: Two points that are close by are 'similar' and two points that are far apart are 'different'. We can discover the global information of a dataset from the pairwise similarity of individual points. This idea is inspired by the work by Shi and Malik [22].

If a dataset contains m images we can construct an $m \times m$ similarity matrix S that describes the similarity properties between two images x_i and x_j according to (1) as:

$$S_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma^2),$$

where σ is a parameter chosen by user. When two images are very similar then S_{ij} has a high value and they are not similar at all then $S_{ij} \approx 0$. This similarity function has the following properties:

$$\forall\ i, j{:}\ S_{ij} \in [0, 1],\ S_{ii} = 0,\ S_{ij} = S_{ji}.$$

The first property imposes a normalization on S, the second property sets the self-similarity to zero, and the third property is a symmetry requirement.

C. Similarity Ranking

In order to consider the contextual relationship revealed by the dataset we introduce the concept of *similarity distribution*. Assume a set of points $X = \{x_1, x_2, \ldots, x_m\} \in R^n$ that we would like to rank based on the similarity to the query point. Let x_q be the query point. We define a vector $s_i = [s_{iq}, s_{i1}, s_{i2}, \ldots, s_{im}]^T$, where s_{ij} is the similarity value between two objects x_i and x_j. The similarity value s_{ij} is computed by (1), i.e., $s_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma^2)$. We consider the vector s_i as the *distribution of similarities* between the point x_i and all other points in a dataset including the query point. The vector $s_q = [s_{qq}, s_{q1}, s_{q2}, \ldots, s_{qm}]^T$ represents the distribution of similarities between the query point x_q and all other points including the query point itself. We define s_{qq} as 1.0 but s_{ii}, $i \neq q$, as 0.0 to avoid reinforcement of self-similarity value and the reason will be explained later in detail.

We define the *similarity value* of a point x_i to the query point x_q by the *dot product* of the similarity distribution for x_i and that for x_q in Gaussian feature space. The similarity value s_{iq} of point x_i to the query point x_q is computed by

$$s_{iq} = s_i^{\,T} \cdot s_q = s_{iq} s_{qq} + \sum_{j=1}^{m} s_{ij} \cdot s_{qj} = s_{iq} + \sum_{j=1}^{m} s_{ij} \cdot s_{qj} \qquad (2)$$

In (2), s_i and s_q are the similarity distribution vectors for points x_i and x_q, respectively, and the similarity values s_{ij} and s_{qj} are computed by (1). The similarity measure given by (2) denotes the actual similarity value between the query point and point x_i plus the linear combination of the similarity values between point x_i and its neighbors, weighted by its neighbors' similarity values to the query point. Therefore, the similarity value of a point affects its neighbors' similarity values, and if two points are close, they are more influenced by each other because their respective similarity values to query point are weighted by the similarity value between two points. With this similarity metric based on the similarity distribution, the points clustered near the query point are favored in similarity ranking.

Based on the similarity distribution we introduced, similarity ranking can be performed as:

[Input] A set of points $X = \{x_1, \ldots, x_q, x_{q+1}, \ldots, x_m\} \in R^n$, where $x_1, \ldots, x_q$ are query points and the rest $x_{q+1}, \ldots, x_m$ are the data points we would like to rank
[Output] The ranked list of data points

1. Construct a similarity matrix $S \in R^{m \times m}$ defined by
   $S_{ij} = \exp(-\|x_i - x_j\|^2 / \sigma^2)$ if $i \neq j$, and $S_{ii} = 0$
2. Construct a diagonal matrix D whose (i, i)-element is the sum of S's i-th row.
3. Form a normalized similarity matrix $S' = D^{-1/2} S D^{-1/2}$.
4. Create the initial similarity values s_{ij} between a point x_i and the query point x_j, $1 \leq j \leq q$, $q+1 \leq i \leq m$.
   $$s_{ij} = \begin{cases} 1 & \text{for } 1 \leq i \leq q \text{, i.e., both } x_i \text{ and } x_j \text{ are query points.} \\ S'_{ij} & \text{for } q+1 \leq i \leq m. \end{cases}$$
5. **for** $i = q+1$ to m **do** // for each data point
6. **for** $j = 1$ to q **do** // for each query point
7. $$s_{ij} = s_{ij} + \sum_{k=q+1}^{m} S'_{ik} S'_{kj}$$
8. **end for**
9. **end for**
10. Compute the similarity score s_i of x_i to q query points by $s_i = \max_{1 \leq j \leq q} \{s_{ij}\}$.
11. Sort the set $SS = \{s_{q+1}, s_{q+2}, \ldots, s_m\}$ in non-increasing order and return the top k points as the result.

In step 1, S_{ii} is set to 0 and the reason is explained as follows. Let us expand (2) to compute the similarity value S_{iq} between a point x_i and the query point x_q. If we set S_{ii} to 1, then S_{iq} is expanded as: $S_{iq} = S_{iq} \cdot S_{qq} + (S_{i1} \cdot S_{q1} + S_{i2} \cdot S_{q2} + \ldots + S_{ii} \cdot S_{qi} + \ldots + S_{im} \cdot S_{qm}) = S_{iq} + (S_{i1} \cdot S_{q1} + S_{i2} \cdot S_{q2} + \ldots + S_{qi} + \ldots + S_{im} \cdot S_{qm})$. This means that the component S_{iq} $(= S_{qi})$ is involved twice in computing the similarity value S_{iq} when we set $S_{ii} = 1$. In order to avoid this overestimating, we set the self-similarity value as $S_{ii} = 0$ (but $S_{qq} = 1$). Note that the steps 1-3 are performed in a *batch* mode before the search. The cost to compute $\Sigma_k S'_{ik} \cdot S'_{kj}$ in step 7 increases with the number of objects in a dataset, but it is not a heavy burden because the similarity values S'_{ik} and S'_{kj} are already computed and saved in the similarity matrix S'.

IV. Experiments

To illustrate the effect of our ranking algorithm, let us consider a toy dataset containing 100 data points shown in Fig. 3. The dataset has 3 cluster structures. Every point should be similar to the points in its neighborhood, and furthermore, points in one cluster should be more similar to each other than to points in the other clusters. The query point is denoted by + and the number beside each point denotes its similarity rank. As shown in Fig. 3, the ranking algorithm amazingly exploits the contextual information of the dataset.

In order to carry out experimental evaluation of our approach to the context-sensitive image search, we use the MNIST database [24] that contains 120,000 handwritten digit images with 28 × 28 pixels. The MNIST database is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting. In our experiments, we use only the first 6,000 images from the MNIST database and perform a similarity search to return k most similar images for given query images.

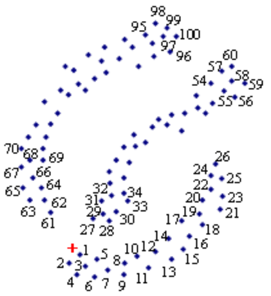


Fig. 3 Similarity ranking on a query point + over 100 data points

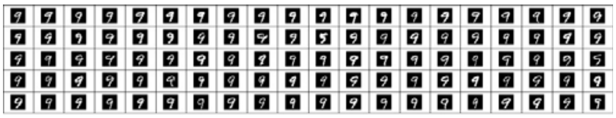


Fig. 4 Top 100 images by our similarity ranking, where the top-left image is the query image (precision: 98%)

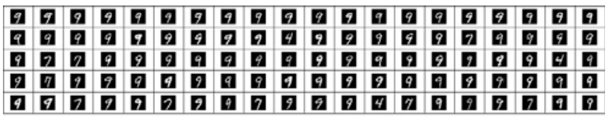


Fig. 5 Top 100 images by Euclidean distance based ranking (precision: 88%)

To obtain an objective measure of performance, we assume that a query concept is a category to which the query image is belonged, i.e., one of the labels '0', '1', …, '9' given to each digit category.

We evaluate *precision* performance for k nearest neighbor (k-NN) queries, where k is 10-100, and precision is computed by the fraction of the returned k images that belong to the query image category.

We perform k-NN queries 100 times and average their performances. The query images are randomly selected from the MNIST database. In order to provide the intuition for our method, we show the k-NN search results in Figs. 4 and 5.

Fig. 6 compares the precision performance for k-NN queries among our method, Euclidean distance based method, the SVM$_{active}$ method [4], and the general SVM-based method without the relevance feedback. In [4], it is stated that SVM$_{active}$ outperforms three query refinement methods: (1) query reweighting methods such as MARS [3], (2) the query point movement methods such as MARS [2] and MindReader [1], and (3) the query expansion methods such as Falcon [5]. Therefore, we compare our method with SVM$_{active}$. SVM$_{active}$ is a relevance feedback method based on active learning with SVM. It retrieves top-k images after a few relevance feedback rounds. In each round of relevance feedback, SVM$_{active}$ determines the images as "relevant" if they have the same label as the query image. In the experiment of SVM$_{active}$, we conduct four relevance feedback rounds and use 100 training samples per round. SVM$_{active}$ shows the worst performance. The poor performance of SVM$_{active}$ can be explained as follows. SVM$_{active}$ learner has no prior knowledge about the handwritten digit image categories and it learns a query concept only through a relevance feedback process. However, in many cases, classifiers perform poorly in a high-dimensional space given a small number of training samples. In our experiments, SVM$_{active}$ fails to achieve an average of 30% accuracy.

We also perform experiments with SVM to verify the classification effect of SVM in our domain. After trained 6,000 images through SVM, we achieve the averages of 86.1% and 82.9% accuracies on the top-10 results for single- and double-point queries, respectively. However, this prior training is not the approach employed by SVM$_{active}$ but learning the query concept dynamically is the motivation of SVM$_{active}$. Therefore, it seems that SVM$_{active}$ approach without prior training steps might fail in high-dimensional spaces.

As shown in Fig. 6, our method achieves at least 90% precision on the top-k results, whereas other methods cannot achieve our performance.

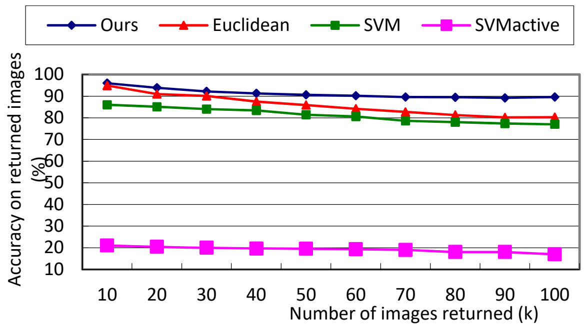


Fig. 6 Single-point queries: average top-k precision

Fig. 7 shows the precision result for multi-point k-NN searches. Our method achieves over 80% precision in any case, whereas SVM, SVM$_{active}$ and Falcon cannot achieve this performance.

Fig. 8 shows the average positions of irrelevant ones in top-k images returned. This position indicates where the irrelevant images appear in top-k results. It is desirable that the irrelevant images are found in rear positions. In our method, the positions of irrelevant images found are compared far later with other methods when the number of images returned is small, i.e., small k. This is a desirable result because users usually do not want to have a large number of images returned.

International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:12, No:12, 2018

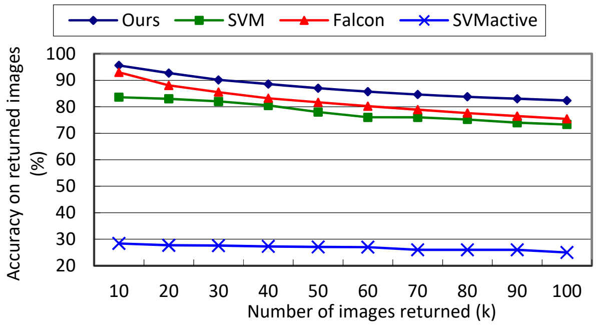


Fig. 7 Multi-point queries: average top-*k* precision

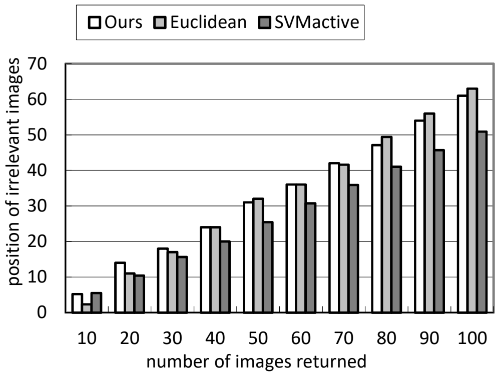


Fig. 8 Positions of irrelevant ones in top-*k* images

V. Conclusion

For effective image retrieval centered on a human viewpoint, we have captured the contextual relationship in a dataset and presented a similarity measure and a context-aware ranking algorithm based on a nonlinear similarity model. The similarity measure and the ranking algorithms proposed in this paper take into account the intrinsic structure and the data distribution revealed by the dataset to estimate image similarity. Our approach has demonstrated its effectiveness and outperformed the existing image retrieval methods such as SVMactive, Falcon, Euclidean distance-based method, and the SVM classification-based method. Our method takes advantage of the intuition that the same portions of the distances given by Minkowski metric do not always give the same degrees of similarity when judged by humans.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2059306).

References

[1] Y. Ishikawa, R. Subramanya, and C. Faloutsos, “MindReader: Querying databases through multiple examples,” *Proc. VLDB Conf.*, pp. 218-227, 1998

[2] Y. Rui, T.S. Huang, M. Ortega, and Mehrotra, S. “Relevance feedback: A Power tool for interactive content-based image retrieval,” *IEEE Trans. Circuits and Video Technology*, 8(5), 644-644, 1998

[3] Y. Rui, T.S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feedback in MARS,” *Proc. of Int’l Conf. on Image Processing*, 815-818, 1997

[4] S. Tong and E. Chang, “Support Vector Machine Active Learning for Image Retrieval,” *Proc. ACM Multimedia Conf.*, pp. 107-118, 2001

[5] L. Wu, C. Faloutsos, K. Sycara, and T.R. Payne, “FALCON: Feedback Adaptive Loop for Content-Based Retrieval,” *Proc. of VLDB Conf.*, pp. 297-306, 2000

[6] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, “Learning with Local and Global Consistency,” *Advances in Neural Information Processing Systems*, 16, 2004, MIT Press, Cambridge, MA.

[7] G. Wu, E.Y. Chang, and N. Panda, “Formulating Context-dependent Similarity Functions,” *Proc. ACM Multimedia*, pp. 725-734, 2005

[8] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1994, Maxmillan, NY.

[9] B. Schölkopf, S. Kung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, “Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers,” *IEEE Trans. on Signal Processing*, 45, 2758-2765, 1997

[10] V.N. Vapnik, *Statistical Learning Theory*, 1998, Wiley, NY.

[11] C.-H. Hoi and M. Lyu, “A novel log-based relevance feedback technique in content-based image retrieval,” *Proc. ACM Multimedia Conf.*, pp. 24-31, 2004

[12] K. Barnard and D. Forsyth, “Learning the semantics of words and pictures,” *Journal of Machine Learning Research*, Vol. 3, pp. 1107-1135, 2003

[13] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models,” *Proc. of ACM SIGIR Conf.*, 119-126, 2003

[14] J.Y. Pan, H.J, Yang, P. Duygulu, and C. Faloutsos, “Automatic Image Captioning,” *Proc. IEEE Int’l Conf. on Multimedia and Expo*, 1987-1990, 2004

[15] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan, “Exploiting Ontologies for Automatic Image Annotation,” *Proc. ACM SIGIR Conf.*, 552-558, 2005

[16] X. He, W.-Y. Ma, and H.-J. Zhang, “Learning an image manifold for retrieval,” *Proc. ACM Multimedia Conf.*, pp. 17-23, 2004

[17] K.-S. Goh, B. Li, and E. Chang, “DynDex: A Dynamic and Non-metric Space Indexer,” *Proc. ACM Multimedia*, pp. 466-475, 2002

[18] R.L.De. Valois and K.K.De. Valois, *Spatial Vision*, 1988, Oxford Science Publications, Oxford

[19] P. Muneesawang and L. Guan, “An Interactive Approach for CBIR Using a Network of Radial Basis Functions,” *IEEE Trans. on Multimedia*, 6(5), 703-716, 2004

[20] D. Ennis, “Modeling similarity and identification when there are momentary fluctuations in psychological amplitudes,” *Multidimensional Models of Perception and Cognition*, 279-298, 1998, Lawrence Erlbaum Assoc. Pub., Philadelphia, PA.

[21] R. Shepard, “Toward a universal law of generalization for psychological science,” *Science*, 237, 1317-1323, 1987

[22] J. Shi and J. Malik, “Normalized Cuts and Image Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-905, 2000

[23] G. Chen et al., “HISA: A Query System Bridging the Semantic Gap for Large Image Databases,” Proc. of VLDB Conf. 2006, pp. 1187-1190.

[24] Y. LeCun, C. Cortes, and C.J.C. Burges, *The MNIST Database of Handwritten Digits*, http://yann.lecun.com/exdb/mnist/