Image Retrieval Using Scale-Space Matching*

S. Ravela[1] R. Manmatha[2] E. M. Riseman[1]

[1] Computer Vision Research Laboratory
[2] Center for Intelligent Information Retrieval
University of Massachusetts at Amherst
{ravela, manmatha}@cs.umass.edu

Abstract. The retrieval of images from a large database of images is an important and emerging area of research. Here, a technique to retrieve images based on appearance that works effectively across large changes of scale is proposed. The database is initially filtered with derivatives of a Gaussian at several scales. A user defined template is then created from an image of an object similar to those being sought. The template is also filtered using Gaussian derivatives. The template is then matched with the filter outputs of the database images and the matches ranked according to the match score. Experiments demonstrate the technique on a number of images in a database. No prior segmentation of the images is required and the technique works with viewpoint changes up to 20 degrees and illumination changes.

1 Introduction

The advent of multi-media and large image collections in several different domains brings forth a necessity for image retrieval systems. These systems will respond to visual queries by retrieving images in a fast and effective manner. The application potential is enormous; ranging from database management in museums and medicine, architectural and interior design, image archiving, to constructing multi-media documents or presentations[3].

Simple image retrieval solutions have been proposed, one of which is to annotate images with text and then use a traditional text-based retrieval engine. While this solution is fast, it cannot however be effective over large collections of complex images. The variability and richness of interpretation is quite enormous as is the human effort required for annotation. To be effective an image retrieval system should exploit image attributes such as color distribution, motion, shape [1], structure, texture or perhaps user drawn sketches or even abstract token sets (such as points, lines etc.). Image retrieval can be viewed as an ordering of match scores that are obtained by searching through the database. The key challenges in building a retrieval system are the choice of attributes,

* This work was supported in part by the Center for Intelligent Information Retrieval, NSF Grants IRI-9208920, CDA-8922572 and ARPA N66001-94-D-6054

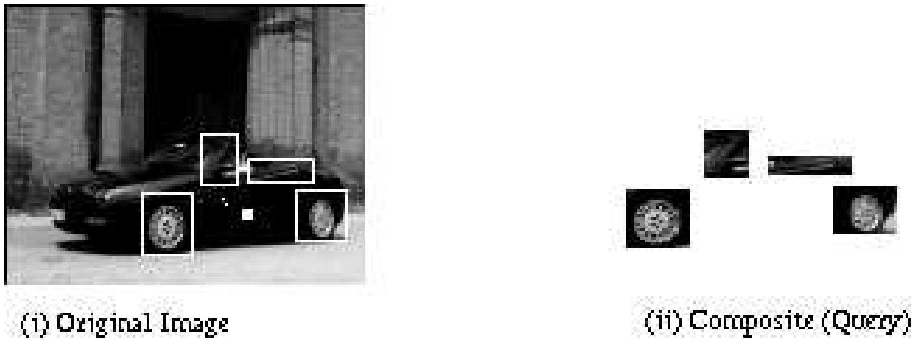

Fig. 1. *Construction of a query begins with a user marking regions of interest in an image, shown by the rectangles in (i). The regions of interest and their spatial relationships define a query, shown in (ii).*

their representations, query specification methods, match metrics and indexing strategies.

In this paper a method for retrieving images based on appearance is presented. Without resorting to token feature extraction or segmentation, images are retrieved in the order of their *similarity in appearance* to a *query*.

Queries are constructed from raw images, as illustrated in Figure 1. The regions in Figure 1(ii) along with their spatial relationship are conjunctively called as the query[1]. Images are then retrieved from the database in the order of their similarity of appearance to the query. Similarity of appearance is defined as the similarity of shape under small view variations. The proposed definition constrains view variations, but does not constrain scale variations.

A measure of similarity of appearance is obtained by correlating filtered representations of images. In particular a *vector representation*(VR) of an image is obtained by associating each pixel with a vector of responses to Gaussian derivative filters of several different orders. To retrieve similar looking images under varying scale a representation over the scale parameter is required and scale-space representations [6] are a natural choice. Lists of VRs generated using banks of Gaussian derivative filters at several different scales form a scale-space representation [6] of the object. A match score for any pair of images is obtained by correlating their scale-space vector representations.

Thus, the entire process of retrieval can be viewed as the following three-step process. The first is an off-line computation step that generates VRs of database images for matching (described in Section 3). The second is construction of queries and their VRs (described in Section 5). The third is an ordering of images ranked by the correlation of their VRs with that of the query (described in Section 4). In Section 6 experiments with this procedure demonstrate retrieval of similar looking objects under varying scale.

While one is tempted to argue that retrieval and recognition problems have

[1] The retrieved images for this case are shown in Figure 3.

a lot in common, one should also note the sharp contrasts between the two paradigms. First, putting a user in the "loop" , shifts the burden of the determination of feature saliency to the user. For example, only regions of the car in Figure 1(i) (namely, the wheels, side-view mirror and mid-section) considered salient by the user are highlighted. Second, user interaction can be used in a retrieval system of sufficient speed to evaluate the ordering of retrieved images and reformulate queries if necessary. Thus, in the approach presented in this paper, alternate regions could be marked if the retrieval is not satisfactory. Third, a hundred percent accuracy of retrieval is desirable but not at all critical (for comparison the best text-based retrieval engines have retrieval rates less than 50%). The user ultimately views and evaluates the results, allowing for tolerance to the few incorrect retrieval instances.

2 Related Work

A number of researchers have investigated the use of shape for retrieval [1, 9, 10]. However, unlike the technique presented in this paper, these methods all require prior segmentation of the object using knowledge of the contour or binary shape of the object.

It has been argued by Koenderink and van Doorn [5] and others that the structure of an image may be represented using Gaussian derivatives. Hancock et al [4] have shown that the principal components of a set of images containing natural structures may be modeled as the outputs of a Gaussian and its derivatives at several scales. That is, there is a natural decomposition of an image into Gaussian derivatives at several scales. Gaussians and their derivatives have, therefore, been successfully used for matching images of the same object under different viewpoints (see [12] for references). This paper is an extension to matching "similar" objects using Gaussian derivatives.

3 Matching Vector Representations

The key processing involves obtaining and matching vector-representations of a sample gray level image patch S and a candidate image C. The steps involved in doing this will now be described:

Consider a Gaussian described by it's coordinate $\mathbf{r}$ and scale σ

$$G(\mathbf{r},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mathbf{r}^2}{2\sigma^2}} \tag{1}$$

A vector-representation $\mathbf{I}$ of an image I is obtained by associating each pixel with a vector of responses to partial derivatives of the Gaussian at that location. Derivatives up to the second order are considered. More formally, $\mathbf{I}$ takes the form $\langle I_x, I_y, I_{xx}, I_{xy}, I_{yy} \rangle$ where I_x, I_y denote the the filter response of I to the first partial derivative of a Gaussian in direction x and y respectively. I_{xx},I_{xy} and I_{yy} are the appropriate second derivative responses. The choice of first and second Gaussian derivatives is discussed in [12].

The correlation coefficient η between images $\mathbf{C}$ and $\mathbf{S}$ at location (m,n) in $\mathbf{C}$ is given by:

$$\eta(m,n) = \sum_{i,j} \hat{C_M}(i,j) \cdot \hat{S_M}(m-i, n-j) \tag{2}$$

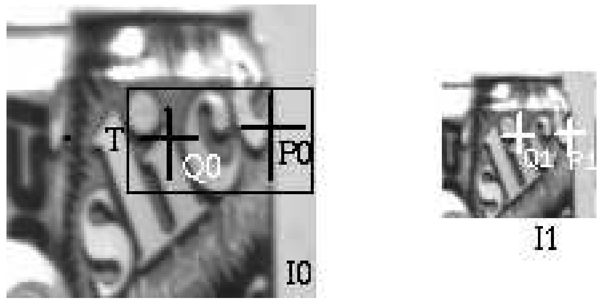

Fig. 2. *I1 is half the size of I0. To match points p_0 with p_1, Image I_0 should be filtered at point p_0 by a Gaussian of a scale twice that of the Gaussian used to filter image I_1 (at p_1). To match a template from I_0 containing p_0 and q_0, an additional warping step is required. See text in Section 4.*

where

$$\hat{S_M}(i,j) = \frac{\mathbf{S}(i,j) - S_M}{||\mathbf{S}(i,j) - S_M||}$$

and S_M is the mean of $\mathbf{S}(i,j)$ computed over S. $\hat{C_M}$ is computed similarly from $\mathbf{C}(i,j)$. The mean C_M is in this case computed at (m,n) over a neighborhood in C (the neighborhood is the same size as S).

Vector correlation performs well under small view variations. It is observed in [12] that typically for the experiments carried out with this method, in-plane rotations of up to 20^o, out-of plane rotation of up to 30^0 and scale changes of less than 1.2 can be tolerated. Similar results in terms of out-of-plane rotations were reported by [11].

4 Matching Across Scales

The database contains many objects imaged at several different scales. For example, the database used in our experiments has several diesel locomotives. The actual image size of these locomotives depends on the distance from which they are imaged and shows considerable variability in the database. The vector correlation technique described in Section 3 cannot handle large scale changes, and the matching technique, therefore, needs to be extended to handle large scale changes.

In Figure 2 image I_1 is half the size of image I_0 (otherwise the two images are identical). Thus,

$$I_0(\mathbf{r}) = I_1(s\mathbf{r}) \tag{3}$$

where $\mathbf{r}$ is any point in image I_0 and $s\mathbf{r}$ the corresponding point in I_1 and the scale change s = 0.5. In particular consider two corresponding points p_0 and p_1 and assume the image is Gaussian filtered at p_0. Then it can be shown that [7],

$$\int I_0(\mathbf{r})G(\mathbf{r} - \mathbf{p_0}, \sigma)d\mathbf{r} = \int I_1(s\mathbf{r})G(s\mathbf{r} - \mathbf{p_1}), s\sigma)d\,(s\mathbf{r}) \tag{4}$$

In other words, the output of I_0 filtered with a Gaussian of scale σ at p_0 is equal to the output of I_1 filtered with a Gaussian of scale $s\sigma$ i.e. the Gaussian has to be stretched in the same manner as the image if the filter outputs are to be equal. This is not a surprising result if the output of a Gaussian filter is viewed as a Gaussian weighted average of the intensity. A more detailed derivation of this result is provided in [7].

The derivation above does not use an explicit value of the scale change s. Thus, equation 4 is valid for any scale change s. The form of equation 4 resembles a convolution and in fact it may be rewritten as a convolution

$$I_0(\mathbf{r}) \star G(., \sigma) = I_1(s\mathbf{r}) \star G(., s\sigma) \tag{5}$$

Similarly, filtering with the first and second derivatives of a Gaussian gives [8]

$$I_0 \star \mathbf{G}^{'}(., \sigma) = I_1 \star \mathbf{G}^{'}(., s\sigma) \tag{6}$$

and,

$$I_0 \star \mathbf{G}^{"}(., \sigma) = I_1 \star \mathbf{G}^{"}(., s\sigma) \tag{7}$$

where the normalized first derivative of a Gaussian is given by

$$\mathbf{G}'(\mathbf{r}, s\sigma) = s\sigma \;\; dG(\mathbf{r}, s\sigma)/d\mathbf{r} \tag{8}$$

and the normalized second derivative of a Gaussian is given by

$$\mathbf{G}''(\mathbf{r}, s\sigma) = (s\sigma)^2 \;\; d^2G(\mathbf{r}, s\sigma)/d(\mathbf{r}\mathbf{r}^{\mathbf{T}}) \tag{9}$$

Note that the first derivative of a Gaussian is a vector and the second derivative of a Gaussian a 2 by 2 matrix.

The above equations are sufficient to match the filter outputs (in what follows assume only Gaussian filtering for simplicity) at corresponding points (for example at $\mathbf{p_0}$ and $\mathbf{p_1}$). A further complication is introduced if more than one point is to be matched while preserving the relative distances (structure) between the points. Consider for example the pair of corresponding points $\mathbf{p_0}, \mathbf{q_0}$ and $\mathbf{p_1}, \mathbf{q_1}$. The filter outputs at points $\mathbf{p_0}, \mathbf{q_0}$ may be visualized as a template and the task is to match this template with the filter outputs at points $\mathbf{p_1}, \mathbf{q_1}$. That is, the template is correlated with the filtered version of the image I_1 and a best match sought. However, since the distances between the points $\mathbf{p_1}, \mathbf{q_1}$ are different from those between $\mathbf{p_0}, \mathbf{q_0}$ the template cannot be matched correctly unless either the template is rescaled by a factor of $1/2$ or the image I_1 is rescaled by a factor of 2. The matching is, therefore, done by warping either the template or the image I_1 appropriately.

Thus, to find a match for a template from I_0, in I_1, the Gaussians must be filtered at the appropriate scale and then the image I_1 or the template should be warped appropriately. Now consider the problem of localizing a template T, extracted from I_0, in I_1(see Figure 2). For the purpose of subsequent analysis, assume two corresponding points ($\mathbf{p_0}$,$\mathbf{q_0}$) of interest in T and I_1 ($\mathbf{p_1}, \mathbf{q_1}$) respectively. To localize the template the following three steps are performed.

1. *Use appropriate Relative Scale:* Filter the template and I_1 with Gaussians whose scale ratio is 2. That is, filter T with a Gaussian of scale 2σ and I_1 with σ.

2. *Account for size change:* Sub-sample T by half. At this point the spatial and intensity relationship between the warped version (filtered and sub-sampled) of template points p_0 and q_0 should be exactly same as the relationships between filtered versions of p_1 and q_1.
3. *Translational Search:* Perform a translational search over I_1 to localize the template.

This three step procedure can be easily extended to match VRs of T and I_1 using Equations 6 and 7. In step(1) generate VRs of T and I_1 using the mentioned filter scale ratios. In step(2) warp the VR of T instead of just the intensity. In step(3) use vector-correlation(Equation 2 at every step of the translational search.

Without loss of generality any arbitrary template T can be localized in any I_1 that contains T scaled by a factor s.

4.1 Matching Queries over Unknown Scale

The aforementioned steps for matching use the assumption that the relative scale between a template and an image is known. However, the relative scale between structures in the database that are similar to a query cannot be determined *a priori*. That is, the query could occur in a database image at some unknown scale. A natural approach would be to search over a range of possible relative scales, the extent and step size being user controlled parameters.

One way of accomplishing this is as follows. First, VRs are generated for each image in the database over a range of scales, say $\frac{1}{4}\sigma, \frac{1}{2\sqrt{2}}\sigma, \ldots, 4\sigma$. Then, a VR for the query is generated using Gaussian derivatives of scale σ. The query VR is matched with each of the image VRs, thus traversing a relative scale change of $\frac{1}{4} \ldots 4$, in steps of $\sqrt{2}$. For each scale pairing the three step procedure for matching VRs is applied. In the warping step of this procedure either the query or the image is warped depending on the relative scale. If the relative scale between the query and a candidate image is less than 1 the candidate VR is warped and if it is greater than 1 the query VR is warped. After the query is matched with each of the image VRs, the location in the image which has the best correlation score is returned.

It is instructive to note that VR lists over scale are scale-space representations in the sense described by Lindeberg [6]. By smoothing an image with Gaussians at several different scales Lindeberg generates a scale-space representation. While VR lists are scale-space representations, however, they differ from Lindeberg's approach in two fundamental ways. First VRs are generated from derivatives of Gaussians and second, an assumption is made that smoothing is accompanied by changes in size (i.e. the images are scaled versions rather than just smoothed versions of each other). This is the reason warping is required during VR matching across scales. VR lists are proper scale-space representations unlike pyramidal representations [6, 12]

5 Constructing Query Images

The query construction process begins with the user marking salient regions on an object. VRs generated at several scales within these regions are matched with the database in accordance with the description in Section 4. Unselected regions are not used in matching. One way to think about this is to consider a composite template, such as one shown in Figure 1(ii). The unselected regions have been masked out. The composite template preserves inter-region spatial relationships and hence, the structure of the object is preserved. Warping the composite will warp all the components appropriately, preserving relative spatial relationships. That is, both the regions as well as distances between regions are scaled appropriately. Further, there are no constraints imposed on the selection of regions and the regions need not overlap.

Careful design of a query is important. It is interesting to note that marking the entire object does not work very well (see [12] for examples). Marking extremely small regions has also not worked with this database. There are too many coincidental structures that can lead to poor retrieval.

Many of these problems are, however, simplified by having the user interact extensively with the system. Letting the user design queries eliminates the need for detecting the saliency of features on an object. Instead, saliency is specified by the user. In addition, based on the feedback provided by the results of a query, the user can quickly adapt and modify the query to improve performance.

6 Experiments

The choice of images used in the experiments was based on a number of considerations. It is expected that when very dissimilar images are used the system should have little difficulty in ranking the images. For example, if a car query is used with a database containing cars and apes, then it is expected that cars would be ranked ahead of apes. This is borne out by the experiments done to date. Much poorer discrimination is expected if the images are much more 'similar'. For example, man-made vehicles like cars, diesel and steam locomotives should be harder to discriminate. It was, therefore, decided to test the system by primarily using images of cars, diesel and steam locomotives as part of the database.

The database used in this paper has digitized images of cars, steam locomotives, diesel locomotives, apes and a small number of other miscellaneous objects such as houses. Over 300 images were obtained from the internet to construct this database. About 215 of these are of cars, diesel locomotives and steam locomotives. There are about 80 apes and about 12 houses in the database. These photographs, were taken with several different cameras of unknown parameters, and, under varying but uncontrolled lighting and viewing geometry. The objects of interest are embedded in natural scenes such as car shows, railroad stations, country-sides and so on.

Prior to describing the experiments, it is important to clarify what a correct retrieval means. A retrieval system is expected to answer questions such as 'find all cars similar in view and shape to this car' or 'find all steam engines similar in appearance to this steam engine'. To that end one needs to evaluate if a query can be designed such that it captures the appearance of a generic steam engine or perhaps that of a generic car. Also, one needs to evaluate the performance of VR matching under a specified query. In the examples presented here the

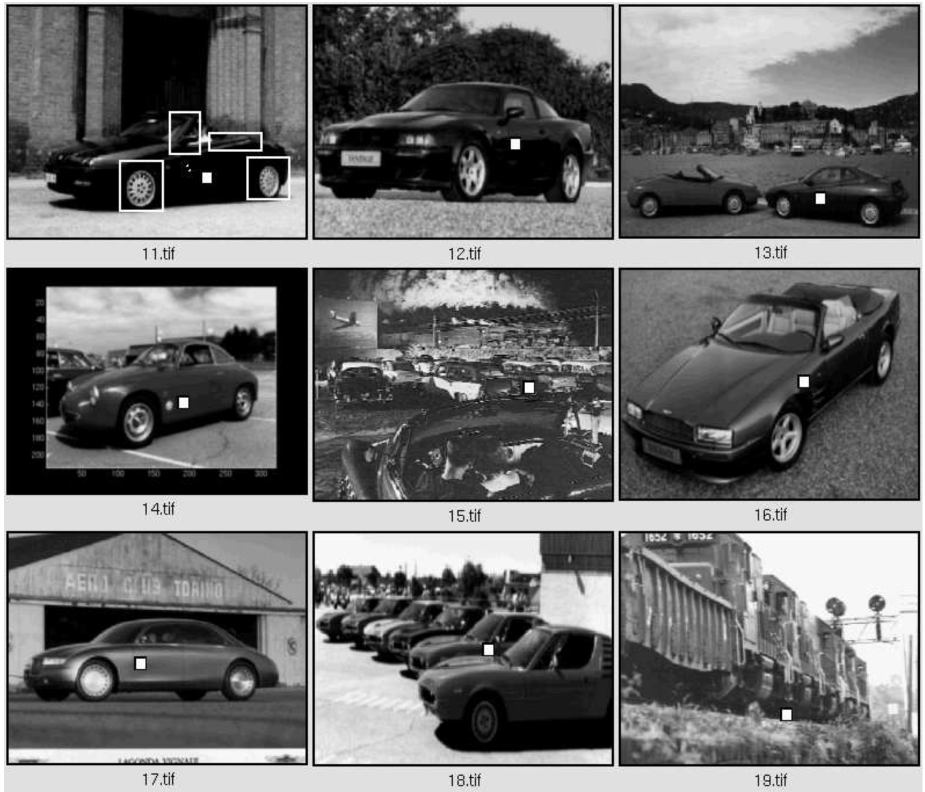

Fig. 3. *Retrieval results for Car.*

following method of evaluation is applied. First, the objective of the query is stated and then retrieval instances are gauged against the stated objective. In general, objectives of the form 'extract images similar in appearance to the query' will be posed to the retrieval algorithm.

Several different queries were constructed to retrieve objects of a particular type. It is observed that under reasonable queries at least 60% of m objects underlying the query are retrieved in the top m ranks. Best results indicate retrieval results of up to 85%. This performance compares very well with typical text retrieval systems[2]. To the best of our knowledge other image retrieval systems either need prior segmentation or work on restricted domains. Therefore, accurate comparisons cannot be made.

Several experiments were carried out with the database [12]. The results of the experiments carried out with a car query, a diesel query and a steam query are presented in table 6. The number of retrieved images in intervals of ten is charted in Table 6. The table shows, for example, that there are 16 car images "similar" in view to the car in the query and 14 of these are ranked in the top 20. For the steam query there are 12 "similar" images (as determined by a

[2] The average retrieval rate for text-based systems is 50%

person), 9 of which are ranked in the top 20. Finally, for the diesel query there are 30 “similar” images, 12 of which are found in the top 20 retrievals. Due to space limitations only the results of the *Car retrieval* are displayed (Figure 3) and analyzed in detail (for the others see [12]).

	No. Retrieved Images				
Query	1-10	11-20	21-30	31-40	41-50
Car	8	6	1	0	1
Steam	7	2	1	0	2
Diesel	7	5	5	6	4

Table 1. *Correct retrieval instances for the Car, Steam and Diesel queries in intervals of ten. The number of “similar” images in the database as determined by a human are 16 for the Car query, 12 for the Steam query and 30 for the Diesel query.*

The car image used for retrieval is shown in the top left picture of Figure 3. The objective is to 'obtain all similar cars to this picture'. Towards this end a query was marked by the user, highlighting the wheels, side view-mirror and mid section. The results to be read in text book fashion in Figure 3 are the ranks of the retrieved images. The white patches indicate the centroid of the composite template at best match. In the database, there are exactly 16 cars within a close variation in view to the original picture. Fourteen of these cars were retrieved in the top 16, resulting in a 87.5% retrieval. All 16 car pictures were picked up in the top 50. The results also show variability in the shape of the retrieved instances. The mismatches observed in pictures labeled 'car05.tif' and 'car09.tif' occur in VR matching when the relative scale between the query VR and the images is $\frac{1}{4}$.

Wrong instances of retrieval are of two types. The first is where the VR matching performs well but the objective of the query is not satisfied. In this case the query will have to be redesigned. The second reason for incorrect retrieval is mismatches due to the search over scale space. Most of the VR mismatches result from matching at the extreme relative scales.

Overall the queries designed were also able to distinguish steam engines and diesel engines from cars precisely because the regions selected are most similarly found in similar classes of objects. As was pointed out in Section 5 query selection must faithfully represent the intended retrieval, the burden of which is on the user. The retrieval system presented here performs well under it's stated purpose: that is to extract objects of similar shape and view to that of a query.

7 Conclusions and Limitations

This paper demonstrates retrieval of similar objects using vector representations over scale-space. There are several factors that affect retrieval results, including query selection, and the range of scale-space search. The results indicate that this method has sufficient accuracy for image retrieval applications.

One of the limitations of our current approach is the inability to handle large deformations. The filter theorems described in this paper hold under affine deformations and a current step is to incorporate it in to the vector-correlation routine.

While these results execute in a reasonable time they are still far from the high speed performance desired of image retrieval systems. Work is on-going towards building indices of images based on local shape properties and using the indices to reduce the amount of translational search.

Acknowledgments

The authors thank Prof. Bruce Croft and the Center for Intelligent Information Retrieval (CIIR) for continued support of this work. We also thank Jonathan Lim and Robert Heller for systems support. The pictures of trains were obtained from http://www.cs.monash.edu.au/image_lib/trains/. The pictures of cars were obtained from ftp.team.net/ktud/pictures/.

References

1. Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovix, David Steele, and Peter Yanker: Query By Image and Video Content: The QBIC System. IEEE Computer Magazine, September 1995, pp.23-30.
2. Gösta H. Granlund, and Hans Knutsson: Signal Processing in Computer Vision. Kluwer Academic Publishers, 1995, ISBN 0-7923-9530-1, Dordrecht, The Netherlands.
3. Venkat N. Gudivada, and Vijay V. Raghavan: Content-Based Image Retrieval Systems. IEEE Computer Magazine, September 1995, pp.18-21.
4. P. J. B. Hancock, R. J. Bradley and L. S. Smith: The Principal Components of Natural Images. Network, 1992, 3:61-70.
5. J. J. Koenderink, and A. J. van Doorn: Representation of Local Geometry in the Visual System. Biological Cybernetics, 1987, vol. 55, pp. 367-375.
6. Tony Lindeberg: Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, 1994, ISBN 0-7923-9418-6 , Dordrecht, The Netherlands.
7. R. Manmatha: Measuring Affine Transformations Using Gaussian Filters. Proc. European Conference on Computer Vision, 1994, vol II, pp. 159-164.
8. R. Manmatha and J. Oliensis: Measuring Affine Transform - I, Scale and Rotation. Proc. DARPA IUW, 1993, pp. 449-458, Washington D.C.
9. Rajiv Mehrotra and James E. Gary: Similar-Shape Retrieval In Shape Data Management.IEEE Computer Magazine, September 1995, pp. 57-62.
10. A. Pentland, R. W. Picard, and S. Sclaroff: Photobook: Tools for Content-Based Manipulation of Databases. Proc. Storage and Retrieval for Image and Video Databases II, 1994, Vol.2, 185, SPIE, pp. 34-47, Bellingham, Wash.
11. R. Rao, and D. Ballard: Object Indexing Using an Iconic Sparse Distributed Memory. Proc. International Conference on Computer Vision, 1995, pp. 24-31.
12. S. Ravela, R. Manmatha and E. M. Riseman: Retrieval from Image Databases Using Scale-Space Matching. Technical Report UM-CS-95-104, 1995, Dept. of Computer Science, Amherst, MA 01003.