On computing global similarity in images*

S. Ravela and R. Manmatha
Computer Science Department
University of Massachusetts, Amherst, MA 01003
Email: {ravela,manmatha}@cs.umass.edu

Abstract

The retrieval of images based on their visual similarity to an example image is an important and fascinating area of research. Here, a method to characterize visual appearance for determining global similarity in images is described.

Images are filtered with Gaussian derivatives and geometric features are computed from the filtered images. The geometric features used here are curvature and phase. Two images may be said to be similar if they have similar distributions of such features. Global similarity may, therefore, be deduced by comparing histograms of these features. This allows for rapid retrieval and examples from collection of gray-level and trademark images are shown.

1 Introduction

The advent of large multi-media collections and digital libraries has led to a need for good search tools to index and retrieve information from them. For text available in machine readable form (ASCII) a number of good search engines are available. However, there are as yet no good tools to retrieve images. The traditional approach to searching and indexing images using manual annotations is slow, labor intensive and expensive. In addition, textual annotations cannot encode all the information available in an image. There is thus a need for retrieving images using their content. The indexing and retrieval of images using their content is a difficult problem. A person using an image retrieval system usually seeks to find semantically relevant information. This entails solutions to such hard problems as automatic segmentation, robust feature detection and recognition, all of which are as yet unsolved. However, many image attributes like color, texture, shape and "appearance" are often directly correlated with the semantics of the problem. For example, logos or product packages (e.g., a box of Tide) have the same color wherever they are found. The coat of a leopard has a unique texture while Abraham Lincoln's appearance is uniquely defined. These image attributes can often be used to index and retrieve images.

A common model for retrieval, and one that is adopted here, is that images in the database are processed and described by a set of feature vectors. A priori these vectors are indexed. During run-time, a query is provided in the form of an example image and its features are compared with those stored. Images are then retrieved in the order indicated by the comparison operator. In this paper, objects similar in visual appearance to a given query object are retrieved by comparing with a set of database images using a characterization of their image intensity surfaces. Arguably an object's visual appearance in an image is closely related to several factors including, among others, its three dimensional shape, albedo, surface texture and the imaged viewpoint. It is non-trivial to separate the different factors constituting an object's appearance. For example, the face of a person has a unique appearance that cannot just be characterized by the geometric shape of the 'component parts'. In this paper a characterization of the shape of the intensity surface of imaged objects is used for retrieval. The experiments conducted show that retrieved objects have similar visual appearance, and henceforth an association is made between 'appearance' and the shape of the intensity surface.

Specifically, this paper focuses on a representation for computing global similarity. That is, the task is to find images that, as a whole, appear visually similar. The utility of global similarity retrieval is evident, for example, in finding similar scenes or similar faces in a face database. In addition, practical applications such as finding similar trademarks in a trademark database significantly benefit from global similarity retrieval.

The image intensity surface is robustly characterized using features obtained from responses to multi-scale Gaussian derivative filters. Koenderink [8] and others [3] have argued that the local structure of an image can be represented by the outputs of a set of Gaussian derivative filters applied to an image. That is, images are filtered

*This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, in part by the United States Patent and Trademarks Office and the Defense Advanced Research Projects Agency/ITO under ARPA order number D468, issued by ESC/AXS contract number F19628-95-C-0235, in part by the National Science Foundation under grant IRI-9619117 and in part by NSF Multimedia CDA-9502639. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

with Gaussian derivatives at several scales and the resulting response vector locally describes the structure of the intensity surface. By computing features derived from the local response vector and accumulating them over the image, robust representations appropriate to querying images as a whole (global similarity) can be generated. One such representation uses histograms of features derived from the multi-scale Gaussian derivatives. Histograms form a global representation because they capture the distribution of local features (A histogram is one of the simplest ways of estimating a non parametric distribution). This global representation can be efficiently used for global similarity retrieval by appearance and retrieval is very fast.

The choice of features often determines how well the image retrieval system performs. Here the task is to robustly characterize the 3 dimensional intensity surface. A 3-dimensional surface is uniquely determined if the local curvatures everywhere are known. Thus, it is appropriate that one of the features be local curvature. The principal curvatures of the intensity surface are invariant to image plane rotations, monotonic intensity variations and further, their ratios are in principle insensitive to scale variations of the entire image. However, spatial orientation information is lost when constructing histograms of curvature (or ratios thereof) alone. Therefore we augment the local curvature with local phase, and the representation uses histograms of local curvature and phase.

Local principal curvatures and phase are computed at several scales from responses to multi-scale Gaussian derivative filters. Then histograms of the curvature ratios [7, 1] and phase are generated. Thus, the image is represented by a single vector (multi-scale histograms). During run-time the user presents an example image as a query and the query histograms are compared with the ones stored, and the images are then ranked and displayed in order to the user.

The rest of the paper is organized as follows. Section 2 surveys related work in the literature. In section 3, the notion of appearance is developed further and characterized using Gaussian derivative filters and the derived global representation is discussed. Comparisons are made in the context of trademark retrieval with the traditional moment invariants. A discussion and conclusion follows in Section 4.

2 RELATED WORK

Several authors have tried to characterize the appearance of an object via a description of the intensity surface. In the context of object recognition [14] represent the appearance of an object using a parametric eigen space description. This space is constructed by treating the image as a fixed length vector, and then computing the principal components across the entire database. The images therefore have to be size and intensity normalized, segmented and trained. Similarly, using principal component representations described in [5] face recognition is performed in [19]. In [17] the traditional eigen representation is augmented by using most discriminant features and is applied to image retrieval. The authors apply eigen representation to retrieval of several classes of objects. The issue, however , is that these classes are manually determined and training must be performed on each. The approach presented in this paper is different from all the above because eigen decompositions are not used at all to characterize appearance. Further, the method presented uses no learning and, does not require constant sized images. It should be noted that although learning significantly helps in such applications as face recognition, however, it may not be feasible in many instances where sufficient examples are not available. This system is designed to be applied to a wide class of images and there is no restriction per se.

In earlier work we showed that local features computed using Gaussian derivative filters can be used for local similarity, i.e. to retrieve parts of images [12]. Here we argue that global similarity can be determined by computing local features and comparing distributions of these features. This technique gives good results, and is reasonably tolerant to view variations. Schiele and Crowley [16] used such a technique for recognizing objects using grey-level images. Their technique used the outputs of Gaussian derivatives as local features. A multi-dimensional histogram of these local features is then computed. Two images are considered to be of the same object if they had similar histograms. The difference between this approach and the one presented by Schiele and Crowley is that here we use 1D histograms (as opposed to multi-dimensional) and further use the principal curvatures as the primary feature.

The use of Gaussian derivative filters to represent appearance is motivated by their use in describing the spatial structure [8] and its uniqueness in representing the scale space of a function [9, 6, 21, 18] The invariance properties of the principal curvatures are well documented in [3].

In the context of global similarity retrieval it should be noted that representations using moment invariants have been well studied [13]. In these methods global representation of appearance may involve computing a few numbers over the entire image. Two images are then considered similar if these numbers are close to each other (say using an L2 norm). We argue that such representations are not able to really capture the "appearance" of an image, particularly in the context of trademark retrieval where moment invariants are widely used. In other work [12] we compared moment invariants with the technique presented here and found that moment invariants work best for a single binary shape without holes in it, and, in general, fare worse than the method presented here.

Texture based image retrieval is also related to the appearance based work presented in this paper. Using Wold

modeling, in [10] the authors try to classify the entire Brodatz texture and in [4] attempt to classify scenes, such as city and country. Of particular interest is work by [11] who use Gabor filters to retrieve texture similar images.

The earliest general image retrieval systems were designed by [2, 15]. In [2] the shape queries require prior manual segmentation of the database which is undesirable and not practical for most applications.

3 Global representation of appearance

Three steps are involved in order to computing global similarity. First, local derivatives are computed at several scales. Second, derivative responses are combined to generate local features, namely, the principal curvatures and phase and, their histograms are generated. Third, the 1D curvature and phase histograms generated at several scales are matched. These steps are described next.

A. Computing local derivatives: Computing derivatives using finite differences does not guarantee stability of derivatives. In order to compute derivatives stably, the image must be regularized, or smoothed or band-limited. A Gaussian filtered image $I_\sigma = I * G$ obtained by convolving the image I with a normalized Gaussian $G(\mathbf{r}, \sigma)$ is a band-limited function. Its high frequency components are eliminated and derivatives will be stable. In fact, it has been argued by Koenderink and van Doorn [8] and others [3] that the local structure of an image I at a given scale can be represented by filtering it with Gaussian derivative filters (in the sense of a Taylor expansion), and they term it the N-jet.

However, the shape of the smoothed intensity surface depends on the scale at which it is observed. For example, at a small scale the texture of an ape's coat will be visible. At a large enough scale, the ape's coat will appear homogeneous. A description at just one scale is likely to give rise to many accidental mis-matches. Thus it is desirable to provide a description of the image over a number of scales, that is, a scale space description of the image. It has been shown by several authors [9, 6, 21, 18, 3], that under certain general constraints, the Gaussian filter forms a unique choice for generating scale-space. Thus local spatial derivatives are computed at several scales.

B. Feature Histograms: The normal and tangential curvatures of a 3-D surface (X,Y,Intensity) are defined as [3]:

$$N(\mathbf{p}, \sigma) = \left[\frac{I_x^2 I_{yy} + I_y^2 I_{xx} - 2 I_x I_y I_{xy}}{\left(I_x^2 + I_y^2\right)^{\frac{3}{2}}} \right] (\mathbf{p}, \sigma)$$

$$T(\mathbf{p}, \sigma) = \left[\frac{\left(I_x^2 - I_y^2\right) I_{xy} + \left(I_{xx} - I_{yy}\right) I_x I_y}{\left(I_x^2 + I_y^2\right)^{\frac{3}{2}}} \right] (\mathbf{p}, \sigma)$$

Where $I_x(\mathbf{p}, \sigma)$ and $I_y(\mathbf{p}, \sigma)$ are the local derivatives of Image I around point $\mathbf{p}$ using Gaussian derivative at scale σ. Similarly $I_{xx}(\cdot, \cdot)$, $I_{xy}(\cdot, \cdot)$, and $I_{yy}(\cdot, \cdot)$ are the corresponding second derivatives. The normal curvature N and tangential curvature T are then combined [7] to generate a shape index as follows:

$$C(\mathbf{p}, \sigma) = atan\left[\frac{N + T}{N - T} \right] (\mathbf{p}, \sigma)$$

The index value C is $\frac{\pi}{2}$ when $N = T$ and is undefined when either N and T are both zero, and is, therefore, not computed. This is interesting because very flat portions of an image (or ones with constant ramp) are eliminated. For example in Figure 2(middle-row), the background in most of these face images does not contribute to the curvature histogram. The curvature index or shape index is rescaled and shifted to the range $[0, 1]$ as is done in [1]. A histogram is then computed of the valid index values over an entire image.

The second feature used is phase. The phase is simply defined as $P(\mathbf{p}, \sigma) = atan2\left(I_y(\mathbf{p}, \sigma), I_x(\mathbf{p}, \sigma)\right)$. Note that P is defined only at those locations where C is and ignored elsewhere. As with the curvature index P is rescaled and shifted to lie between the interval $[0, 1]$.

Although the curvature and phase histograms are in principle insensitive to variations in scale, in early experiments we found that computing histograms at multiple scales dramatically improved the results. An explanation for this is that at different scales different local structures are observed and, therefore, multi-scale histograms are a more robust representation. Consequently, a feature vector is defined for an image I as the vector $V_i = \langle H_c(\sigma_1) \ldots H_c(\sigma_n), H_p(\sigma_1) \ldots H_p(\sigma_n) \rangle$ where H_p and H_c are the curvature and phase histograms respectively. We found that using 5 scales gives good results and the scales are $1 \cdots 4$ in steps of half an octave.

C. Matching feature histograms: Two feature vectors are compared using normalized cross-covariance defined as

$$d_{ij} = \frac{V_i^{(m)} \cdot V_j^{(m)}}{\left|\left|V_i^{(m)}\right|\right| \left|\left|V_j^{(m)}\right|\right|}$$

where $V_i^{(m)} = V_i - mean(V_i)$.

Retrieval is carried out as follows. A query image is selected and the query histogram vector V_q is correlated with the database histogram vectors V_i using the above formula. Then the images are ranked by their correlation score and displayed to the user. In this implementation, and for evaluation purposes, the ranks are computed in advance, since every query image is also a database image.

3.1 Experiments

The curvature-phase method is tested using two databases. The first is a trademark database of 2048 im-

ages obtained from the US Patent and Trademark Office (PTO). The images obtained from the PTO are large, binary and are converted to gray-level and reduced for the experiments. The second database is a collection of 1561 assorted gray-level images. This database has digitized images of cars, steam locomotives, diesel locomotives, apes, faces, people embedded in different background(s) and a small number of other miscellaneous objects such as houses. These images were obtained from the Internet and the Corel photo-cd collection and were taken with several different cameras of unknown parameters, and under varying uncontrolled lighting and viewing geometry.

In the following experiments an image is selected and submitted as a query. The objective of this query is stated and the relevant images are decided in advance. Then the 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. A measure of the performance of the retrieval engine can be obtained by examining the recall/precision table for several queries. Briefly, recall is the proportion of the relevant material actually retrieved and precision is the proportion of retrieved material that is relevant [20]. It is a standard widely used in the information retrieval community and is one that is adopted here.

Queries were submitted each to the trademark and assorted image collection for the purpose of computing recall/precision. The judgment of relevance is qualitative. For each query in both databases the relevant images were decided in advance. These were restricted to 48. The top 48 ranks were then examined to check the proportion of retrieved images that were relevant. All images not retrieved within 48 were assigned a rank equal to the size of the database. That is, they are not considered retrieved. These ranks were used to interpolate and extrapolate precision at all recall points.In the case of assorted images relevance is easier to determine and more similar for different people. However in the trademark case it can be quite difficult and therefore the recall-precision can be subject to some error. The recall/precision results are summarized in Table 1 and both databases are individually discussed below.

Figure 1 shows the performance of the algorithm on the trademark images. Each strip depicts the top 8 retrievals, given the leftmost as the query. Most of the shapes have roughly the same structure as the query. Note that, outline and solid figures are treated similarly (see rows one and two in Figure 1). Six queries were submitted for the purpose of computing recall-precision in Table 1.

Experiments are also carried out with assorted gray level images. Six queries submitted for recall-precision are shown in Figure 2. The left most image in each row is the query and is also the first retrieved. The rest from-left to right are seven retrievals depicted in rank order. Note that, flat portions of the background are never considered because the principal curvatures are very close to zero and therefore do not contribute to the final score. Thus, for example, the flat background in Figure 2(second row) is not used. Notice that visually similar images are retrieved even when there is some change in the background (row 1). This is because the dominant object contributes most to the histograms. In using a single scale poorer results are achieved and background affects the results more significantly.

The results of these examples are discussed below, with the precision over all recall points depicted in parentheses. For comparison the best text retrieval engines have an average precision of 50%:

1. Find similar cars(65%). Pictures of cars viewed from similar orientations appear in the top ranks because of the contribution of the phase histogram. This result also shows that some background variation can be tolerated. The eighth retrieval although a car is a mismatch and is not considered.

2. Find same face(87.4%) and find similar faces: In the face query the objective is to find the same face. In experiments with a University of Bern face database of 300 faces with a 10 relevant faces each, the average precision over all recall points for all 300 queries was 78%. It should be noted that the system presented here works well for faces with the same representation and parameters used for all the other databases. There is no specific "tuning" or learning involved to retrieve faces. The query "find similar faces" resulted in a 100% precision at 48 ranks because there are far more faces than 48. Therefore, it was not used in the final precision computation.

3. Find dark textured apes (64.2%). The ape query results in several other light textured apes and country scenes with similar texture. Although these are not mis-matches they are not consistent with the intent of the query which is to find dark textured apes.

4. Find other patas monkeys. (47.1%) Here there are 16 patas monkeys in all and 9 within a small view variation. However, here the whole image is being matched so the number of relevant patas monkeys is 16. The precision is low because the method cannot distinguish between light and dark textures, leading to irrelevant images. Note, that it finds other apes, dark textured ones, but those are deemed irrelevant with respect to the query.

5. Given a wall with a Coca Cola logo find other Coca Cola images (63.8%). This query clearly depicts the limitation of global matching. Although all three database images that had a certain texture of the wall

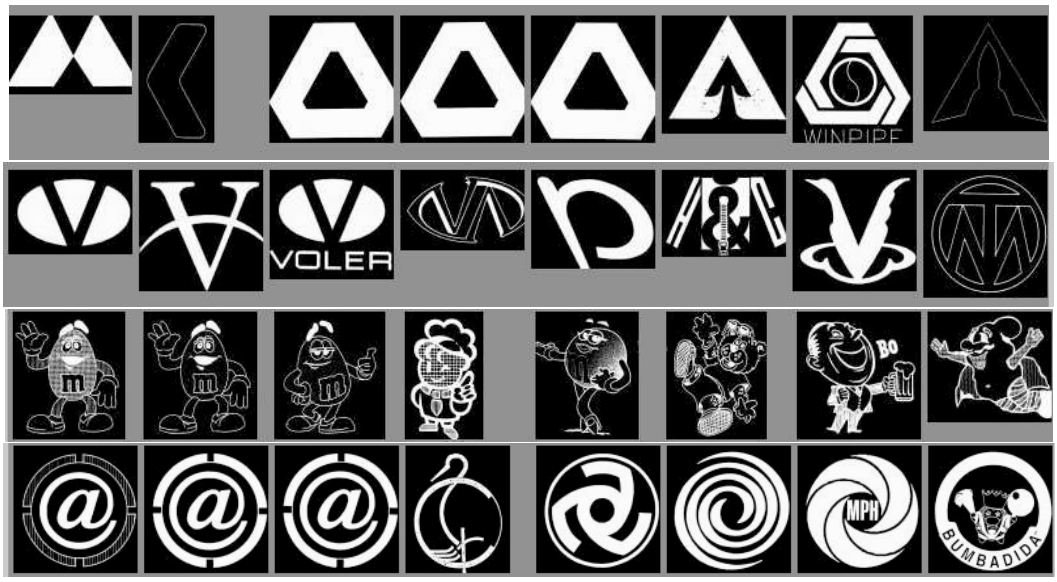

Figure 1: Trademark retrieval using Curvature and Phase

Figure 2: Image retrieval using Curvature and Phase

(also had Coca Cola logos) were retrieved (100% precision), two other very dissimilar images with coca-cola logos were not.

6. Scenes with Bill Clinton (72.8%). The retrieval in this case results in several mismatches. However, three of the four are retrieved in succession at the top and the scenes appear visually similar.

While the queries presented here are not "optimal" with respect to the design constraints of global similarity retrieval, they are however, realistic queries that can be posed to the system. Mismatches can and do occur. The first

Table 1: Precision at standard recall points for six Queries

Recall	0	10	20	30	40	50	60	70	80	90	100
Precision(trademark) %	100	93.2	93.2	85.2	76.3	74.5	59.5	45.5	27.2	9.0	9.0
Precision(assorted) %	100	92.6	90.0	88.3	87.0	86.8	83.8	65.9	21.3	12.0	1.4

average(trademark)	61.1%
average(assorted)	66.3%

is the case where the global appearance is very different. The Coca Cola retrieval is a good example of this. Second, mismatches can occur at the algorithmic level. Histograms coarsely represent spatial information and therefore will admit images with non-trivial deformations. The recall/precision presented here compares well with text retrieval. The time per retrieval is of the order of milliseconds. In on going work we are experimenting with a database of 63000 images and the amount of time taken to retrieve is still less than a second. The space required is also a small fraction of the database. These are the primary advantages of global similarity retrieval. That is, to provide a low storage, high speed retrieval with good recall/precision.

4 Conclusions and Limitations

This paper demonstrates retrieval of similar objects on the basis of their visual appearance. Visual appearance is characterized using filter responses to Gaussian derivatives over scale space. In addition, we claim that global representations are better constructed by representing the distribution of robustly computed local features. Currently we are investigating two issues. First is to scale the database up to about 100000 images and second is to provide a mechanism for combining global and local similarity matching in a single framework.

5 Acknowledgements

We would like to thank Yasmina Chitti for providing results on the faces. We would also like to thank Bruce Croft and CIIR for supporting this work.

References

[1] Chitra Dorai and Anil Jain. Cosmos - a representation scheme for free form surfaces. In *Proc. 5th Intl. Conf. on Computer Vision*, pages 1024–1029, 1995.

[2] Myron Flickner et al. Query by image and video content: The qbic system. *IEEE Computer Magazine*, pages 23–30, Sept. 1995.

[3] L. Florack. *The Syntactical Structure of Scalar Images*. PhD thesis, University of Utrecht, Utrecht, Holland, 1993.

[4] M. M. Gorkani and R. W. Picard. Texture orientation for sorting photos 'at a glance'. In *Proc. 12th Int. Conf. on Pattern Recognition*, pages A459–A464, October 1994.

[5] M Kirby and L Sirovich. Application of the kruhnen-loeve procedure for the characterization of human faces. *IEEE Trans. Patt. Anal. and Mach. Intel.*, 12(1):103–108, January 1990.

[6] J. J. Koenderink. The structure of images. *Biological Cybernetics*, 50:363–396, 1984.

[7] J. J. Koenderink and A. J. Van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8), 1992.

[8] J. J. Koenderink and A. J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[9] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[10] Fang Liu and Rosalind W Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. PAMI*, 18(7):722–733, July 1996.

[11] W. Y. Ma and B. S. Manjunath. Texture-based pattern retrieval from image databases. *Multimedia Tools and Applications*, 2(1):35–51, January 1996.

[12] R. Manmatha, S. Ravela, and Y. Chitti. On computing local and global similarity in images. In *Proc. SPIE conf. on Human and Electronic Imaging III,*, 1998.

[13] B.M. Methre, M.S. Kankanhalli, and W.F. Lee. Shape Measures for Content Based Image Retrieval: A Comparison. *Information Processing and Management*, 33, 1997.

[14] S. K. Nayar, H. Murase, and S. A. Nene. Parametric appearance representation. In *Early Visual Learning*. Oxford University Press, February 1996.

[15] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of databases. In *Proc. Storage and Retrieval for Image and Video Databases II,SPIE*, volume 185, pages 34–47, 1994.

[16] Bernt Schiele and James L. Crowley. Object recognition using multidimensional receptive field histograms. In *Proc. 4th European Conf. Computer Vision*, Cambridge, U.K., April 1996.

[17] D. L. Swets and J. Weng. Using discriminant eigen features for retrieval. *IEEE Trans. Patt. Anal. and Mach. Intel.*, 18:831–836, August 1996.

[18] Bart M. ter Har Romeny. *Geometry Driven Diffusion in Computer Vision*. Kluwer Academic Publishers, 1994.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive NeuroScience*, 3:71–86, 1991.

[20] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[21] A. P. Witkin. Scale-space filtering. In *Proc. Intl. Joint Conf. Art. Intell.*, pages 1019–1023, 1983.