LA-UR 94-3086

Conf-9410211--1

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: **EXPERIENCE WITH CANDID: COMPARISON ALGORITHM FOR NAVIGATING DIGITAL IMAGE DATABASES**

AUTHOR(S): Patrick Kelly and Michael Cannon

SUBMITTED TO: 23rd AIPR Workshop on Image and Information Systems: Applications and Opportunities
Washington, DC
October 12–14, 1994

DISCLAIMER

Los Alamos

Los Alamos National Laboratory
Los Alamos New Mexico 87545

Experience with CANDID: Comparison algorithm for navigating digital image databases

Patrick Kelly, Michael Cannon

Computer Research and Applications Group, MS B-265
Los Alamos National Laboratory, Los Alamos, New Mexico 87545

ABSTRACT

This paper presents results from our experience with *CANDID* (Comparison Algorithm for Navigating Digital Image Databases), which was designed to facilitate image retrieval by content using a query-by-example methodology. A global signature describing the texture, shape, or color content is first computed for every image stored in a database, and a normalized similarity measure between probability density functions of feature vectors is used to match signatures. This method can be used to retrieve images from a database that are similar to a user-provided example image. Results for three test applications are included.

1 Introduction

Future data management systems will be required to handle not only textual data, but also massive amounts of non-textual data such as raw system measurements, digital imagery, sound samples, and video clips. These systems will be extremely valuable if they can provide easy access to this diversity of data. Unfortunately, many systems will be simple archives where diverse types of data can only be retrieved by searching for desired dates, titles, subject keywords, and associated textual descriptions. The value of these systems can be greatly enhanced by adding the ability to search directly on the non-textual data, instead of searching only on the associated textual metadata.

Content-based retrieval of digital imagery is currently an active area of research. Several methods have been proposed for the comparison of pictorial or iconic images.[1,2] Other methods compare the relative geometries and positions of different objects in each image.[3,4] In the QBIC Project,[5,6] color, texture, and shape features are computed for each "object" in an image, as well as for each image overall. A Euclidean distance measure is then used to determine similarity between objects or images. Wavelet packet analysis has also been used as a basis for image comparison.[7,8] The structure of the quad-tree containing "significant" subbands is used as one basis for comparison, and specific features computed from these subbands are used as another. In this paper, we propose a method for comparing digital images that is based on ideas being explored for searching databases containing free-text documents.

Modern databases typically use keywords to search through large amounts of textual data. Although these techniques work well, a user is required to fully understand what is being sought by providing specific keywords on which to search. Some newer methods for searching textual databases use "global signatures" to represent the content (or topic) of an entire document instead of using a keyword indexing scheme. An example is the N-gram approach to document fingerprinting.[9,10]

When using the N-gram method for document comparison, a global signature is computed for each document in the database. This signature represents the content, or topic, of a document in an abstract sense. A signature is typically represented by a histogram of the number of times that each substring of length N occurs in the document, where N is a predetermined value. As an example, for a case-insensitive alphabet of 26 letters, there

are 26^3, or 17,576, different tri-grams ("aaa", "aab", "aac", $\cdots$, "zzz"). The signature for each document in this example is therefore a normalized vector of dimension 17,576. A dot-product between N-gram signatures determines the similarity between any two documents. Using this approach for retrieving documents from a database, a user can pose queries such as, "Show me all of the documents that are similar to this example". A user does not need to identify which specific keywords or phrases are to be searched on.

We are finding that this technique of using a global signature to characterize an entire set of data is also very useful in retrieving non-textual data such as digital imagery. The *CANDID* algorithm (Comparison Algorithm for Navigating Digital Image Databases) presented in this paper is analogous to the N-gram approach described above in the sense that we attempt to describe an entire image with a global signature, and then match signatures with some distance measure to determine image similarity. Each image stored in the database is characterized by a global signature that can represent features such as textures, shapes, and colors. When a user queries the database to retrieve images that are similar to a given example image, a global signature for that example image is first computed, and this signature is compared to the signatures of all images in the database. A handful of images having similar content, i.e. database images having a similar signature to the target image, is returned to the user.

2 Signature Computation

We must first recognize that similarity between images is an abstract concept; making judgements is very subjective. As an example, consider three different color pictures in an automobile magazine. One reader may think that image A is more similar to image B than to image C because both A and B contain red automobiles, and C contains a blue automobile. A second reader, however, might claim that image C resembles image A more than image B does, because the cars in both A and C are convertibles, whereas the car in B is not.

With this in mind, it is important to approach the problem of image comparison differently for every application. Shape descriptors, color features, and texture measures are all able to represent some of the information contained in an image, but the way in which they are used determine what we mean when we say that two images are "similar". The feature selection process for any application is one of the most important aspects in solving the problem.

In contrast to the QBIC method[5,6] where color, shape, and texture measurements are calculated for the entire image or for each user-specified object, we take an approach that more closely resembles the N-gram work for textual data. The general idea is that we first compute several features (local color, texture, and/or shape) at every pixel in the image, and then make a "histogram" of feature vector (pixel vector) occurrences for that image. Unlike textual data, we will most likely not have a finite number of unique feature vectors that can occur in our data, and we therefore calculate a continuous probability density function over the multidimensional feature space instead of an actual histogram. This probability density function is our content signature for the given image.

Probability density function estimation is a large problem in itself; we attempt to estimate the probability density function as a gaussian mixture. Each gaussian distribution function is defined by a mean vector $\underline{\mu}_i$ and a covariance matrix Σ_i. A general data clustering routine can provide clusters for which for $\underline{\mu}_i$ and Σ_i can be obtained. We use the k-means clustering algorithm[11,12] followed by an optional cluster merging process.[13] A mean vector and covariance matrix are computed for each of the resultant clusters, and the associated gaussian distribution function is weighted by the number of elements in the corresponding cluster. Any cluster having a singular covariance matrix is deleted and ignored in subsequent processing.

3 Signature Comparison

If each signature is represented by a continuous probability density function, many different distance measures can be used to compare them.[14] We have previously shown that the following distance measure can be used effectively by *CANDID* to retrieve images[15]:

$$dist\,(I_1, I_2) \;=\; \left[\int_{\Re} \left(\, P_{I_1}(\underline{x}) - P_{I_2}(\underline{x}) \,\right)^2 d\underline{x}\right]^{\frac{1}{2}} \tag{1}$$

In this equation, $P_{I_1}(\underline{x})$ and $P_{I_2}(\underline{x})$ are probability density functions over the feature space for images 1 and 2, respectively. In practice, we normalize this distance measure to yield a value between 0 (for a perfect match) and 1 (for completely non-overlapping signatures). Another measure that we have used is the normalized correlation coefficient between two functions:

$$sim\,(I_1, I_2) \;=\; \frac{\int_{\Re} P_{I_1}(\underline{x}) P_{I_2}(\underline{x}) d\underline{x}}{\left[\int_{\Re} P_{I_1}^2(\underline{x}) d\underline{x} \int_{\Re} P_{I_2}^2(\underline{x}) d\underline{x}\right]^{\frac{1}{2}}} \tag{2}$$

This measure will yield a maximum of 1 when both probability density functions $P_{I_1}(\underline{x})$ and $P_{I_2}(\underline{x})$ are identical. If these functions do not overlap, then $sim\,(I_1, I_2)$ will be 0. As discussed in the previous section, we can estimate these probability density functions as gaussian mixtures:

$$P_I(\underline{x}) \;\approx\; \sum_{i=1}^{K} w_i G_i(\underline{x}) \tag{3}$$

where $G_i(\underline{x})$ takes the form:

$$G_i(\underline{x}) \;=\; (2\pi)^{-\frac{N}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{\mu_i})^T \Sigma_i^{-1} (\underline{x} - \underline{\mu_i})\right] \tag{4}$$

We will differentiate between P_{I_1} and P_{I_2} with the following notation:

$$P_{I_1} \;=\; \sum_{i=1}^{K_1} w_i G_i(\underline{x}) \qquad P_{I_2} \;=\; \sum_{i=1}^{K_2} v_i F_i(\underline{x}) \tag{5}$$

Using this representation for the signatures, our similarity measure can be written as follows:

$$\begin{aligned} sim\,(I_1, I_2) \;&=\; \frac{\int_{\Re} \left(\sum_{i=1}^{K_1} w_i G_i(\underline{x})\right) \left(\sum_{i=1}^{K_2} v_i F_i(\underline{x})\right) d\underline{x}}{\left[\int_{\Re} \left(\sum_{i=1}^{K_1} w_i G_i(\underline{x})\right)^2 d\underline{x}\right]^{\frac{1}{2}} \left[\int_{\Re} \left(\sum_{i=1}^{K_2} v_i F_i(\underline{x})\right)^2 d\underline{x}\right]^{\frac{1}{2}}} \\ &=\; \left(\sum_{i=1}^{K_1}\sum_{j=1}^{K_2} w_i v_j \int_{\Re} G_i(\underline{x}) F_j(\underline{x}) d\underline{x}\right) \cdot \\ &\quad \left(\sum_{i=1}^{K_1} w_i^2 \int_{\Re} G_i^2(\underline{x}) d\underline{x} \;+\; 2\sum_{i=1}^{K_1}\sum_{j=i+1}^{K_1} w_i w_j \int_{\Re} G_i(\underline{x}) G_j(\underline{x}) d\underline{x}\right)^{-\frac{1}{2}} \cdot \\ &\quad \left(\sum_{i=1}^{K_2} v_i^2 \int_{\Re} F_i^2(\underline{x}) d\underline{x} \;+\; 2\sum_{i=1}^{K_2}\sum_{j=i+1}^{K_2} v_i v_j \int_{\Re} F_i(\underline{x}) F_j(\underline{x}) d\underline{x}\right)^{-\frac{1}{2}} \end{aligned} \tag{6}$$

This similarity measure is now in a form that can be computed. It contains $O(K_1^2 + K_2^2)$ terms, where each term contains an infinite integral over the product of two gaussians. These integrals can be computed as follows:

$$\int_{\Re} G_i(\underline{x})\, G_j(\underline{x})\, d\underline{x} \;=\; (2\pi)^{-\frac{N}{2}} \left|\Sigma_i + \Sigma_j\right|^{-\frac{1}{2}} \cdot \exp\left[-\frac{1}{2}(c_1 + c_2)\right] \tag{7}$$

where c_1 and c_2 are given by:

$$c_1 \;=\; \underline{\mu_i}^T \Sigma_i^{-1} \underline{\mu_i} \;+\; \underline{\mu_j}^T \Sigma_j^{-1} \underline{\mu_j} \tag{8}$$

$$c_2 \;=\; -(\Sigma_i^{-1}\underline{\mu_i} + \Sigma_j^{-1}\underline{\mu_j})^T \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1} (\Sigma_i^{-1}\underline{\mu_i} + \Sigma_j^{-1}\underline{\mu_j}) \tag{9}$$

Furthermore, for the special case where $\underline{\mu_i} = \underline{\mu_j}$ and $\Sigma_i = \Sigma_j$, we can simplify this even further:

$$\int_{\Re} G_i^2(\underline{x}) d\underline{x} \;=\; 2^{-N} \pi^{-\frac{N}{2}} \left|\Sigma_i\right|^{-\frac{1}{2}} \tag{10}$$

All results presented in this paper were obtained using this similarity measure $sim(I_1, I_2)$.

4 Experimental Results: Smithsonian Image Database

Several proposed content-based retrieval systems use color histograms as a basis for image comparison.[5,6,16] *CANDID* is similar to these approaches with the exception that a continuous probability density function, as opposed to a discrete histogram, is used to characterize the distribution of colors in an image. Although continuous probability density functions may be more expensive and more difficult to compute than histograms, they allow us to consider how "close" two specific colors (or any features) are to each other. Histograms, on the other hand, assign every color to a specific bin. Every color assigned to bin i is considered to be equivalent. Every color not assigned to bin i is considered to be completely dissimilar from colors that are assigned to bin i.

We downloaded 705 digital images from the Smithsonian Image Database [1] to be used as test data for *CANDID*. This database primarily consists of photographic images of items on display at the Smithsonian Institute in Washington, DC. We used this data set to determine *CANDID*'s utility in finding images with similar color content. Signatures for each image were computed over the 3-dimensional RGB color space. *CANDID* was then used to rank all of the database images according to their similarity to a user-specified query image. Our results were predictable; given an example image, images with similar color content were identified. We did notice, however, that some of the most similar images were selected because they had similar background colors. This is not always desirable since a user does not tend to notice the background of an image. The scheme used by QBIC[5,6] to select foreground objects in every image could be used to combat this problem.

Our results were predictable even when we used a single gaussian to represent the color content of each image. Images consisting of only one or two dominant colors were the best candidates for this test. The query image displayed on the left side of Figure 1 consists of a dark (black and deep blue) background and a reddish-brown foreground object. Using single-gaussian RGB color signatures, *CANDID* successfully retrieved images that also contain primarily reddish-brown and dark colors. The two best matches in the database to the query image are also displayed in Figure 1.

[1] The Smithsonian Image Database is available via anonymous ftp from photo1.si.edu, courtesy of Smithsonian Photographic Services.

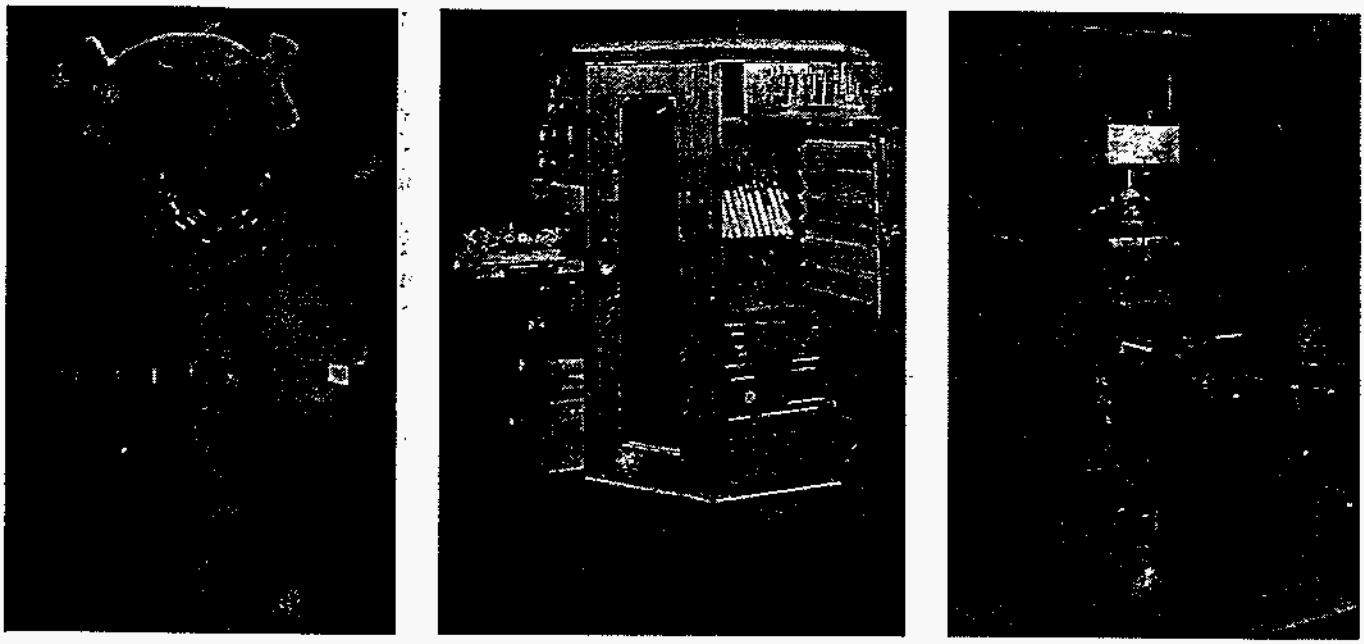

Figure 1: The image on the left, *nmex1.gif*, was used to query a database containing single-gaussian color signatures for all 705 images in the Smithsonian image database. The two best matches were *dental.gif* (with a match score of 0.87) and *press.gif* (with a match score of 0.82). All three of these images can be found in the *tech-history* subdirectory of the Smithsonian Image Database.

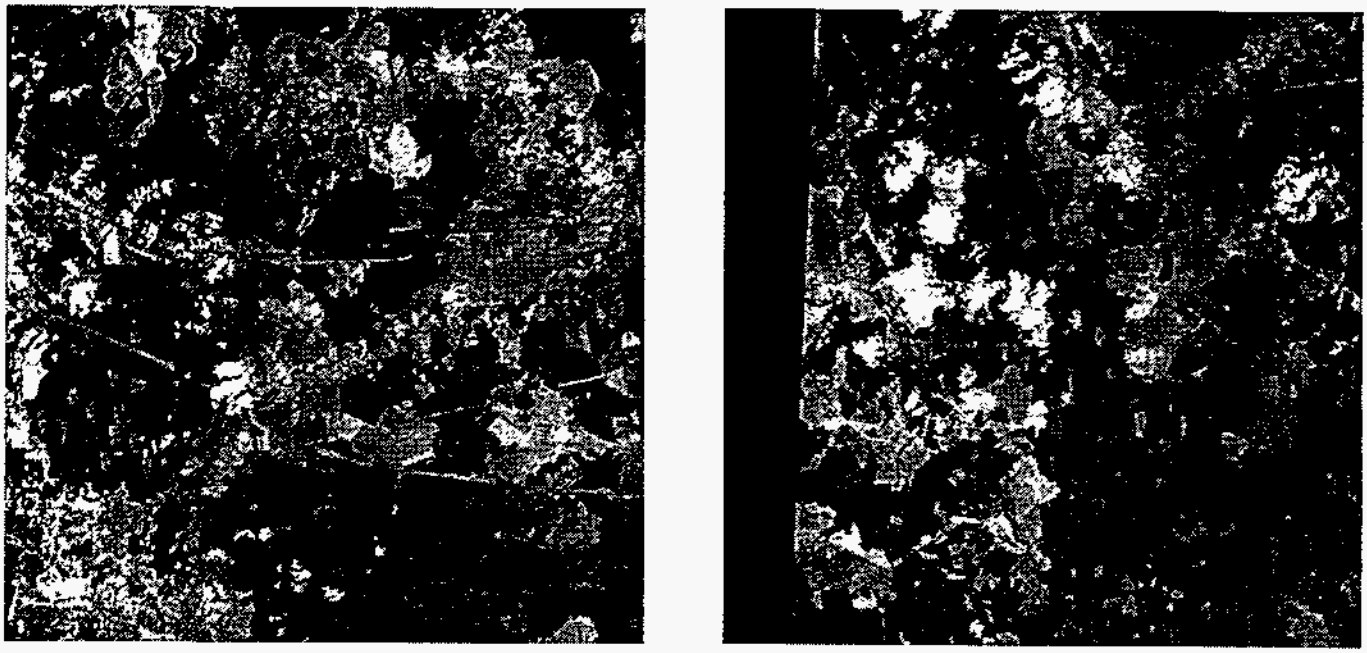

Figure 2: Retrieval of Landsat TM data. The query image on the left was selected from the Moscow scene. Using signatures represented by 20 gaussians in the 6-dimensional feature space, *CANDID* identified the image on the right as being the best match of the other database images (with a match score of 0.89). Both images are presented here as false-color composites made by mapping three of the TM bands to the red, green, and blue components of this picture. The black border on the retrieved image is due to the image border in the original data set; we ignored it for purposes of our experiment.

5 Experimental Results: Landsat TM Data

Remotely-sensed data can be used to locate underground oil reserves, monitor pollution from large factories, and track the disappearance of our world's rain forests. A database containing imagery collected by airborne sensors will prove much more valuable if scientists can access the data by searching on different attributes of image content instead of only being able to retrieve data by searching on associated textual metadata information. The ability to automatically locate areas having similar ground cover will enable scientists to search through terabyte-sized image databases in order to study environmental problems. As an example, if a coniferous forest in Oregon is rapidly disappearing for no apparent reason, then other areas around the world having similar vegetation can be retrieved to see if they are experiencing the same problem. Scientists would then know if this was a global phenomenon or if local conditions were to blame.

We have applied *CANDID* to the problem of retrieving multispectral satellite data (i.e. Landsat TM data) from a database. This enables queries such as, "Show me all images of areas with landcover similar to this example." Landsat Thematic Mapper (TM) data consists of seven different bands of spectral information as listed in Table 1. Each pixel has 28.5 meter resolution and is represented by 7 spectral values ranging from 0 to 255.

CHANNEL		WAVELENGTH (in microns)
Band 1	Visible Blue	0.45-0.52
Band 2	Visible Green	0.52-0.60
Band 3	Visible Red	0.63-0.69
Band 4	Near Infrared	0.76-0.90
Band 5	Mid Infrared	1.55-1.75
Band 6	Thermal Infrared	10.4-12.5
Band 7	Mid Infrared	2.08-2.35

Table 1: Spectral Bands for TM Data

As an experiment, we created a database containing 100 512x512, 6-banded images (the thermal infrared band in each image was ignored). The sample images used to populate our database were acquired from four different geographic locations, each having its own characteristic landscape (see Table 2). The Moscow area, for example, contains many diverse landcover types in every 512×512 subimage that was extracted. These landcover types include coniferous forest, deciduous forest, and agriculture. The Moscow images look nothing like the images around the other three geographic locations. Similarly, the Cairo landscape is unique and dissimilar to the Moscow, Albuquerque, and Los Alamos areas.

LOCATION	DOMINANT LANDSCAPE COVER
Moscow (Russia)	Coniferous Forest, Deciduous Forest, Agriculture, ...
Cairo (Egypt)	Agriculture, Dense Urban, ...
Albuquerque (USA)	Desert, Coniferous Forest, ...
Los Alamos (USA)	Desert, Coniferous Forest, ...

Table 2: Selected Geographic Locations

We calculated global spectral signatures for each database image by clustering the 6-dimensional pixel vectors into 20 clusters. We then used *CANDID* to query our database using an example image from the Moscow scene. Figure 2 shows the query image along with the best match from the database. We sorted the similarity scores between the query image and all 100 test images in the database, which we then plotted (see Figure 3). All

subimages from the Moscow area were retrieved before subimages around Cairo, Albuquerque, and Los Alamos. Furthermore, all Moscow images in the database yielded similarity scores between 0.5 and 1.0, whereas the other images produced similarity scores below 0.01. The point is that using a query image from the Moscow scene, all other images from the Moscow scene (which are all of similar landscape) are retrieved from the database first, after which the other images (from the Cairo, Albuquerque, and Los Alamos scenes) are retrieved with negligible match scores.

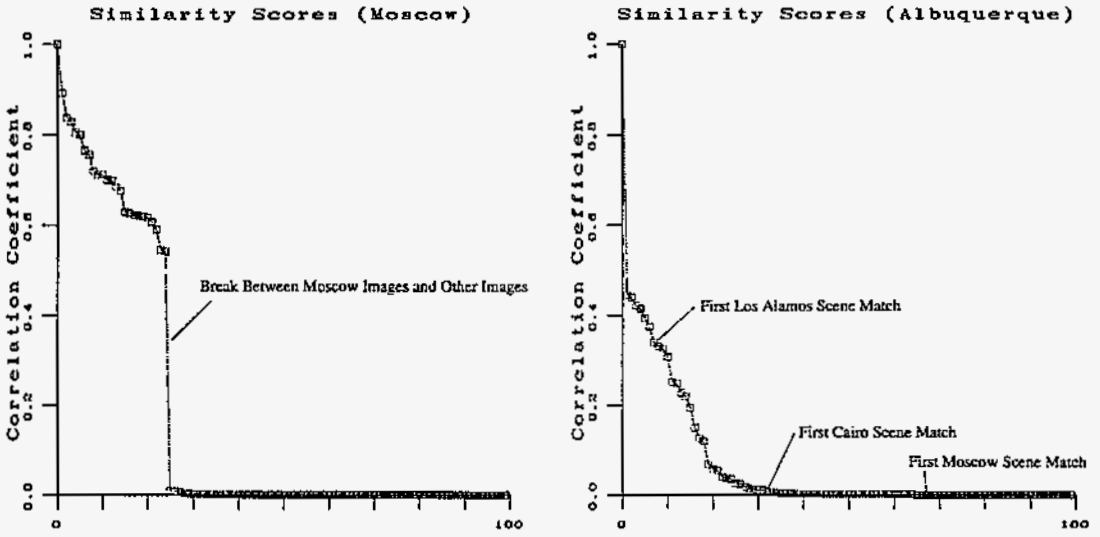

Figure 3: Sorted similarity scores using example images from the Moscow scene and from the Albuquerque scene. The Moscow example produced match scores greater than 0.5 when compared to all other Moscow images, while producing match scores under 0.01 when compared to all other images. The Albuquerque example, on the other hand, did not produce any match scores greater than 0.5 (except when compared with itself).

Unlike Moscow and Cairo, the Albuquerque and Los Alamos images vary quite a bit; 512×512 subimages that were adjacent to one another in the original data set do not necessarily have a lot in common. This is because the New Mexico landscape contains a combination of desert areas, mountainous areas (dominated by coniferous forest), and transitional areas. Again, we used *CANDID* to search the database for images similar to a query image from the Albuquerque scene. The sorted similarity scores are plotted in Figure 3. This plot shows that images from both the Albuquerque and Los Alamos scenes are retrieved first, but the corresponding similarity scores are typically low (less than 0.5). This reflects the fact that images in both the Albuquerque and Los Alamos scenes contain some geographically similar data, but they are not as homogeneous as, say, the Moscow scene. Images from the Moscow and Cairo scenes, which are not at all similar to the query image, are retrieved with negligible match scores. Results from these experiments were quite promising, indicating that further study into the application of *CANDID* to image retrieval for remote-sensing problems is warranted.

6 Experimental Results: Pulmonary CT Data

We have used the concept of global signature matching to retrieve medical imagery based on image content. Pulmonary CT scans reveal the gross pathology indicative of diseased lung tissue resulting from a variety of disorders such as lymphangioleiomyomatosis (LAM), idiopathic pulmonary fibrosis (IPF), scleroderma, emphysema, asthma, and vasculitis. Since CT data is acquired digitally, it can be easily stored in a computer database. It would be a natural extension of this process to search a database to retrieve images that exhibit the same pathology as the current study. These images would provide the radiologist with immediate access to past cases

where similar problems were encountered, thereby aiding with the current patient's diagnosis and treatment.

We applied *CANDID* to this problem of retrieving pulmonary CT imagery from a database containing a total of 220 lung images taken from pulmonary CT studies of 34 different patients (see Table 3). Each image was 512 x 512 pixels in size, consisting of 12-bit grayscale data. For this application, we are primarily interested in retrieving images containing similar textures. We have previously demonstrated that four Laws texture energy measures[17,18] were sufficient to discriminate between lungs affected by different diseases.[15] We recently used these same features on a larger, more diverse set of data in order to determine if *CANDID* could discriminate between diseases when the differences were not necessarily obvious to the untrained eye.

Diagnosis	Number of Patients	Total Number of Images
LAM	11	46
Scleroderma	1	20
IPF	6	12
Emphysema	2	10
Normal	3	6
Vasculitis	1	46
Asthma	10	80
TOTALS	34	220

Table 3: Contents of CT Image Database

To generate a global texture signature describing an image, we first calculated texture features for each pixel. For a given database image, we first convolved it with a number of Laws' convolution kernels. We then replaced each pixel value by the sum of the absolute values of the pixel values in a square neighborhood surrounding it:

$$I_{new}(x,y) = \sum_{i=x-N}^{x+N} \sum_{j=y-N}^{y+N} | I_{old}(i,j) | \tag{11}$$

We finished the process by normalizing our features for contrast.

L5	=	[	1	4	6	4	1]
E5	=	[	-1	-2	0	2	1]
S5	=	[	-1	0	2	0	-1]
W5	=	[	-1	2	0	-2	1]
R5	=	[	1	-4	6	-4	1]

Table 4: Center-Weighted Vectors

Table 4 lists the 5 one-dimensional center-weighted convolution kernels which are used to create the two-dimensional 5-by-5 convolution kernels. The names of these one-dimensional kernels are mnemonics for Level, Edge, Spot, Wave, and Ripple. Each two-dimensional kernel is created by convolving a horizontal kernel with a vertical kernel. For instance, an E5E5 kernel is formed by convolving a horizontal E5 kernel with a vertical E5 kernel. We built a total of 5 two-dimensional kernels – E5E5, R5R5, S5S5, W5W5, and L5L5 – which we used to process each database image.

After convolving an original image with one of the 5-by-5 convolution kernels, the associated Texture Energy Measure (TEM) for each pixel is calculated by summing the absolute pixel values of the convolved image within a 15x15 pixel window. A total of 5 TEM images were calculated during this stage of image processing. The

resultant E5E5, R5R5, S5S5, and W5W5 images were all divided by the L5L5 image to normalize features for contrast, as suggested by Laws,[17] after which the L5L5 image was discarded. The result was a set of 4 images, each representing some texture feature for the image. Each pixel in the image is now represented by a vector of 4 features.

For each database image, we computed a signature approximated by 20 gaussian distributions over the four-dimensional texture space that we defined. We then selected an image from the scleroderma data set and one from the vasculitis data set to use as our query images. To the layman, little difference is apparent between the two images. *CANDID* was able to successfully group together scleroderma database images with the scleroderma query image, and vasculitis database images with the vasculitis query image (see Figures 4, 5, and 6).

It is important to note that for these experiments, the signatures contain texture information about each *entire* CT image. It considers textures around the ribs, spinal cord, and heart as well as the textures inside of each lung. The external textures should generally be consistent between different CT studies.

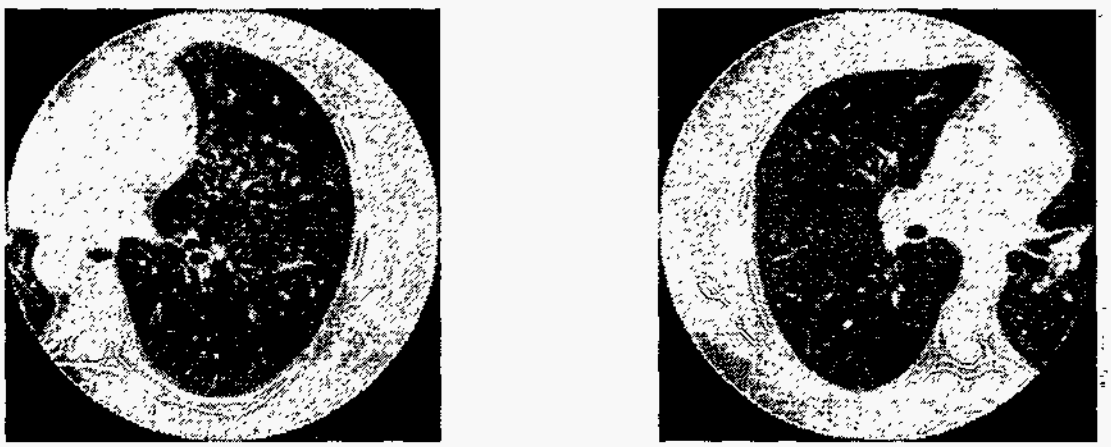

Figure 4: Example (query) images: The image on the left is from a patient with scleroderma. The image on the right is from a patient with vasculitis. The differences in these images are evident to a radiologist, but they are not necessarily obvious to the layman.

7 Conclusions

The problem of retrieving digital images from a database based on image content can be a difficult one. Not only must the meaning of "similarity" between images be determined for each application, but an algorithm must be developed to retrieve images in a manner consistent with the way that a human operator would. *CANDID* performed extremely well on our three test applications.

The general approach described in this paper is not limited to image retrieval problems. Since it attempts to characterize the distribution of features vectors in an abstract feature space, this approach can be used to work with almost any type of data and features. As an example, *CANDID* might be applied to the problem of 1-D signal matching. Many features (such as local frequency) can be computed at different positions along each signal. A signature for each signal could then be calculated and manipulated in a manner consistent with the approach we have presented.

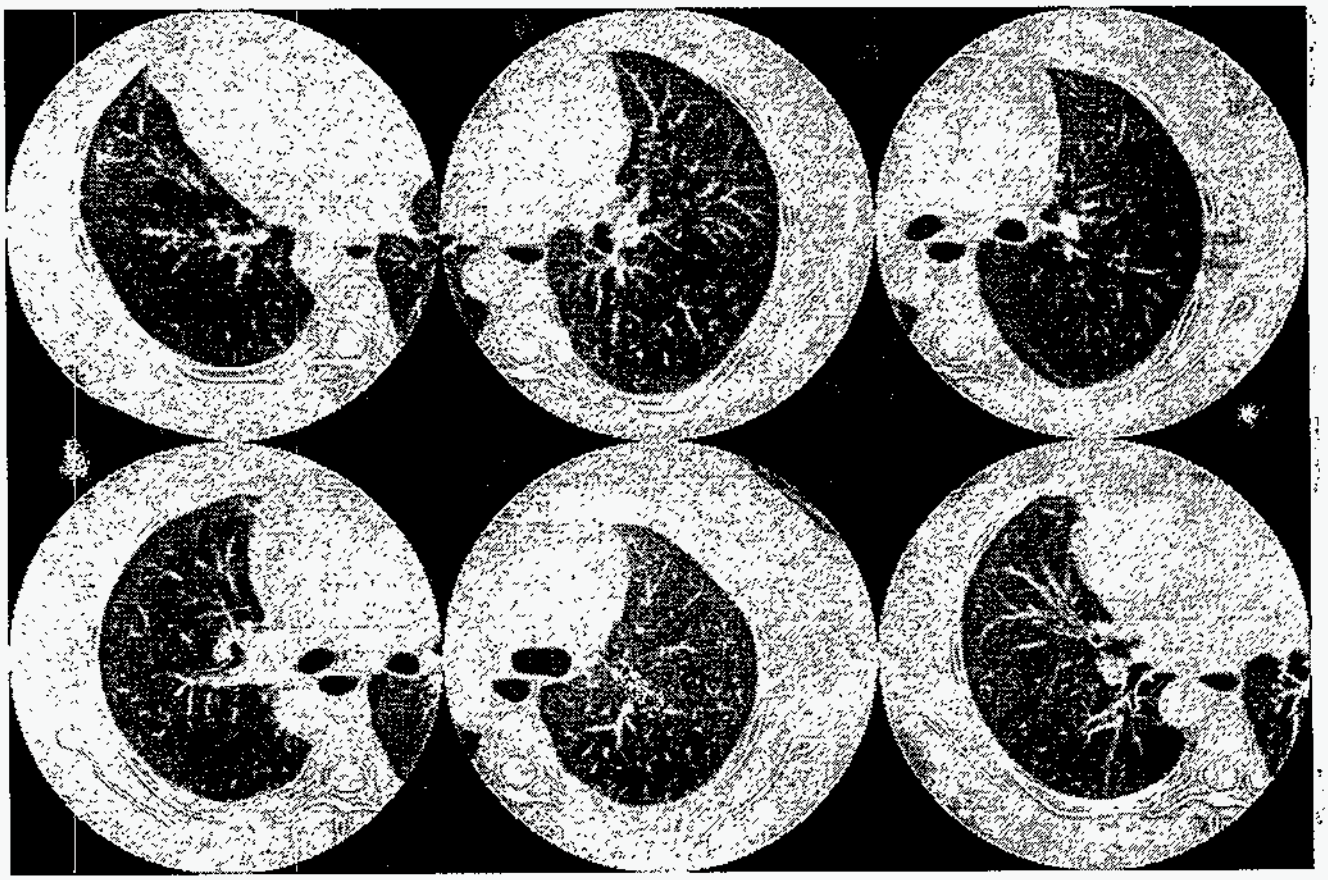

Figure 5: The best 6 matches to the first query image were also images of lungs affected by scleroderma.

8 Acknowledgements

We are grateful to those organizations that have provided us with data for our experiments. The Smithsonian Image Database is available via anonymous ftp from *photo1.si.edu*, courtesy of Smithsonian Photographic Services. Dr. John Newell and Dr. David Lynch at National Jewish Center for Immunology and Respiratory Medicine provided the CT data used to perform the medical image comparison work. This work was performed under a U.S. Government contract (W-7405-ENG-36) by Los Alamos National Laboratory, which is operated by the University of California for the U.S. Department of Energy.

9 REFERENCES

[1] C.C. Chang and T.C. Wu. Retrieving the most similar symbolic pictures from pictorial databases. *Information Processing and Management*, 28(5):581–588, 1992.

[2] F. Rabitti and P. Savino. Automatic image indexation to support content-based retrieval. *Information Processing and Management*, 28(5):547–565, 1992.

[3] T.Y. Hou, A. Hsu, P. Liu, and M.Y. Chiu. A content-based indexing technique using relative gometry features. In *SPIE Vol. 1662 Image Storage and Retrieval*, pages 607–720, 1992.

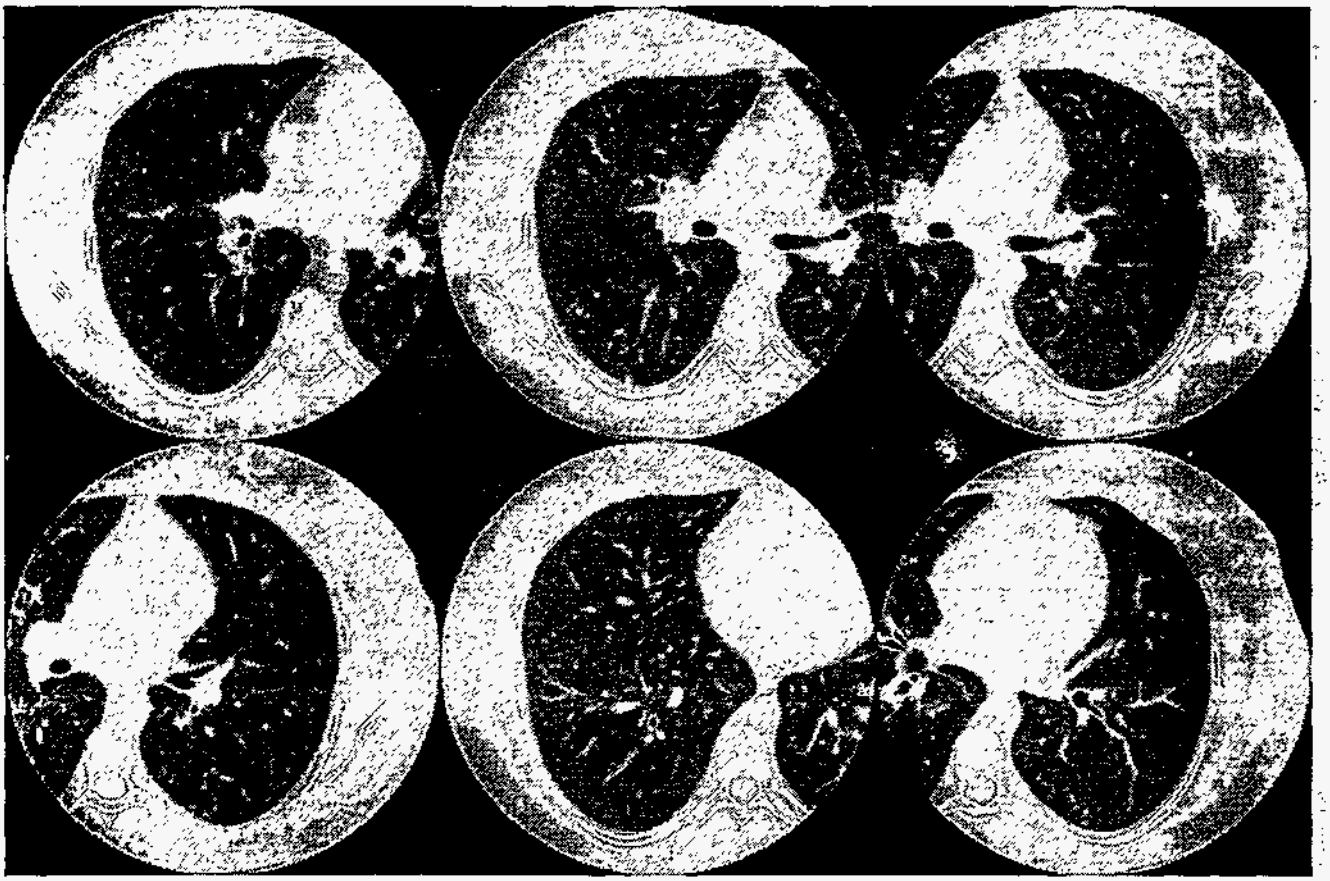

Figure 6: The best 6 matches to the second query image were also images of lungs affected by vasculitis.

[4] S.Y. Lee and F.J. Hsu. Spatial reasoning and similarity retrieval of images using 2d c-string knowledge representation. *Pattern Recognition*, 25(3):305–318, March 1992.

[5] W. Niblack, R. Barder, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Yaubin. The qbic project: Querying images by content using color, texture, and shape. In *SPIE Vol. 1908 Storage and Retrieval for Image and Video Databases*, pages 173–181, 1993.

[6] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Equitz, and R. Barber. Efficient and effective querying by image content. Technical Report RJ 9453 (83074), IBM Almaden Research Center, San Jose, CA, 1993.

[7] K. Perez-Lopez and A. Sood. Comparison of subband features for automatic indexing of scientific image databases. In *Proceedings of the SPIE 1994 International Symposium on Electronic Imaging: Science and Technology*, 1994. To appear.

[8] K. Perez-Lopez, A. Sood, and M. Manohar. Selecting image subbands for browsing scientific databases. In *Proceedings of the SPIE 1994 International Symposium on OE/Aerospace Sensing*, 1994. To appear.

[9] R.E. Kimbrell. Searching for text? send an n-gram! *BYTE*, pages 297–312, May 1988.

[10] T.R. Thomas. Document retrieval from a large dataset of free-text descriptions of physician-patient encounters via n-gram analysis. Technical Report LA-UR-93-0020, Los Alamos National Laboratory, Los Alamos, NM, 1993.

[11] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, NY, 1973.

[12] J.T. Tou and R.C. Gonzalez. *Pattern Recognition Principles.* Addison-Wesley, Reading, MA, 1974.

[13] P.M. Kelly, D.R. Hush, and J.M. White. An adaptive algorithm for modifying hyperellipsoidal decision surfaces. *Journal of Artificial Neural Networks.* In Press.

[14] T.Y. Young and K. Fu. *Handbook of Pattern Recognition and Image Processing.* Academic Press, Inc., New York, NY, 1986.

[15] P.M. Kelly and T.M. Cannon. Candid: Comparison algorithm for navigating digital image databases. In *Proceedings of the Seventh International Working Conference on Scientific Database Management*, 1994. To appear.

[16] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[17] K. Laws. *Textured Image Segmentation.* Ph.D. dissertation, Univ. of Southern Calif., January 1980.

[18] K. Laws. Rapid texture identification. In *SPIE Vol. 238 Image Processing for Missile Guidance*, pages 376–380, 1980.