M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 205
Appeared: IEEE ICASSP, Minneapolis, MN, Vol. V., pp.161–164, Apr. 1993

Finding Similar Patterns in Large Image Databases

R. W. Picard and T. Kabir *

Abstract

We address a new and rapidly growing application, automated searching through large sets of images to find a pattern "similar to this one." Classical matched filtering fails at this problem since patterns, particularly textures, can differ in every pixel and still be perceptually similar. Most potential recognition methods have not been tested on large sets of imagery. This paper evaluates a key recognition method on a library of almost 1000 images, based on the entire Brodatz texture album. The features used for searching rely on a significant improvement to the traditional Karhunen-Loéve (KL) transform which makes it shift-invariant. Results are shown for a variety of false alarm rates and for different subsets of KL features.

1 Introduction

As vastly increasing amounts of image and video are stored in computers it becomes harder for humans to locate a particular scene or video clip. It is currently impossible, in the general case, to semantically describe an image to the computer and have it retrieve it. A simpler and more immediate solution might be to have the user show the computer example image data, or to speak to it keywords with which it has previously associated image features, and then have the computer search for similar patterns or features. However, no existing image processing tools are known to solve this problem for a large general set of images.

Most people know of the ease with which a computer can perform text-string matching. Similar techniques based on matched filters have been successful for signal detection in noise. These types of solutions, however, fail for the problem described above. Two uncorrupted patterns, especially two textures, can be visually similar and still differ in every pixel. Consider for example two video clips of a waterfall taken several minutes apart. Although the pixel values in these two "temporal texture" images will differ, their semantic appearances are likely equivalent. When a human asks a computer to find a particular picture, one would like the computer to understand the human's similarity criteria.

Ideally, we could define a measure of perceptual or semantic similarity and use it instead of the ubiquitous mean-squared error measure of similarity. A step toward this ideal is to transform the data so that perceptually similar things become measurably close to one another in some new space. The mean-squared error, or a weighted version of it is then used to measure "closeness" in this new space.

Most studies of potential transformation algorithms have been run on small sets of test data, typically four to sixteen images at once from the standard Brodatz library of natural textures [1]. Moreover, selected test images have typically exhibited a lot of visual and semantic dissimilarity, as well as a lot of within-class homogeneity. The selected subsets usually do not include the less homogeneous samples from the Brodatz library or those samples which are different shots of the same material, possessing "semantic" similarity. In this study, the inclusion of non-homogeneous patterns gives a scenario closer to what one would expect searching through an image database.

The Brodatz library is limited in that most patterns do not include perspective distortions, most are uniformly illuminated, and there is not much diversity in orientation. Nevertheless, it still challenges the existing texture discrimination tools. To our knowledge, this study is the first which uses the entire available Brodatz library[1]. From the interior of each of 111 original 8 bit 512×512 Brodatz images, nine 128×128 subimages were cropped. This yields a set of $d = 999$ 128×128 images which are called the "Brodatz database."

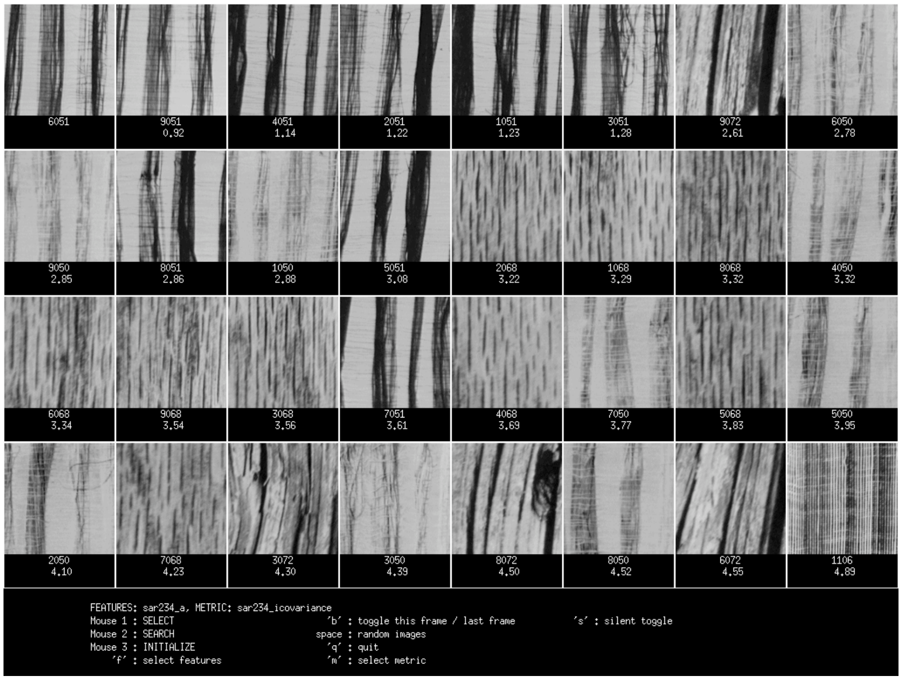
2 Feature selection

Optimal feature selection remains an open research problem. For texture, it has been shown that local second order statistics are important, and that statistics such as co-occurrences incorporate perceptually significant changes [2]. However, features based purely on co-occurrences have been out-performed by features based on outputs of various local filters, and emphasis has shifted to the choice of these filters [3, 4, 5]. This study begins with the "eigenfilters," or principle components of the texture covariance. Eigenfilters have been shown to provide good texture discrimination on small sets of data [3, 4].

The eigenfilter method is more commonly known in the image coding community as the Karhunen-Loéve (KL) transform or principal components analysis, and is optimal for decorrelating the features. Although there is no direct evidence that the human visual system uses an eigenvector-based method, there are some who believe that humans

*This work was supported in part by BT, U.K.

[1] Note the borders were not used since in several cases their data was corrupted by the imaging process. Also, the original Brodatz library has 112 images, one of which appears to have been omitted from the digital library.



```
FEATURES: sar234_a, METRIC: sar234_icovariance
Mouse 1 : SELECT                         'b' : toggle this frame / last frame          's' : silent toggle
Mouse 2 : SEARCH                       space : random images
Mouse 3 : INITIALIZE                     'q' : quit
    'f' : select features                'm' : select metric
```

Figure 1: Photobook displays images in raster scan order, by their similarity to the upper left image.

construct basic "templates" of commonly occurring configurations and then use combinations of these for recognition. However, based on currently available understanding of the human visual system, it is highly unlikely anyone can prove that a given algorithm imitates the human notion of "visual similarity" on more than a trivial set of data.

3 Principal components analysis

The principal components analysis is conducted as follows. Let $\mathbf{x}_i \in \mathbf{R}^{n \times 1}$, $i = 1, \ldots, d$, $n = 128^2$ be a vector representing the DFT magnitude of one of the images in the database. The vectors $\mathbf{x}_i$ are formed by raster scan ordering the rows of the image into one long vector. Ten percent of the images, $p = 100$, are picked at random for training, $\mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_p}$, and used to form an estimate of the "pooled covariance", $\mathbf{C} = \frac{1}{p}\sum_{j=1}^{p}(\mathbf{x}_{t_j} - \mathbf{m})(\mathbf{x}_{t_j} - \mathbf{m})' = \mathbf{X}\mathbf{X}'$, where $\mathbf{m}$ is the training set sample mean, $\mathbf{m} = \frac{1}{p}\sum_{j=1}^{p}\mathbf{x}_{t_j}$. Each principal component, $\mathbf{q}_j$, is an eigenvector of this covariance matrix having associated eigenvalue λ_j:

$$\mathbf{C}\mathbf{q}_j = \mathbf{X}\mathbf{X}'\mathbf{q}_j = \lambda_j \mathbf{q}_j \tag{1}$$

Taking into account that $\mathbf{C}$ will have at most $p << n$ eigenvectors we can save computation by first solving for the eigenvectors of the problem:

$$\mathbf{X}'\mathbf{X}\mathbf{u}_j = \lambda_j \mathbf{u}_j. \tag{2}$$

Left multiplying both sides of (2) by $\mathbf{X}$ gives

$$\mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{u}_j) = \lambda_j(\mathbf{X}\mathbf{u}_j)$$

so that the desired vectors for projection can be obtained by $\mathbf{q}_j = \mathbf{X}\mathbf{u}_j$. With this analysis the calculations are greatly reduced. The associated eigenvalues are used to order the eigenvectors. Note that in recognition, unlike in coding, $\lambda_j > \lambda_i$ does not imply that $\mathbf{u}_j$ will be more useful in reducing error than $\mathbf{u}_i$ [6].

For a given database, these values and the projection of each $\mathbf{x}_i$, $i = 1, \ldots, d$, onto the p eigenvectors are precomputed and stored. These projections (or KL transform coefficients) are the features used for comparing patterns.

Earlier texture applications of eigenfilters computed the coefficients directly from the covariance estimate of the spatial data. However, using the DFT magnitudes makes the KL features invariant to spatial translation. This one difference made an immediately noticeable improvement in the recognition ability of the algorithm. This improvement is similar to one noticed by Akamatsu, *et. al.* [7] in face recognition. Incorporating translation-invariance makes the recognition algorithm perform a little more closely like a human. Other invariants not investigated here [8] may also be similar to those used by humans. Although phase is important for structural image reconstruction, the linear part of the phase appears to be unimportant for much of pattern recognition. We suggest that selective use of phase components should be better than discarding all the phase.

It should be noted that horizontal/vertical boundary effects were mitigated by first windowing each image with a 2D isotropic Gaussian filter, $\sigma = 24$, before computing its DFT. The windowing eliminated the corner regions, in most cases reducing the information available for comparisons.

4 Photobook test environment

The Photobook test environment is illustrated in Fig. 1 running with features based on the phase invariant KL coefficients. For the example shown, the first twenty coefficients were used as features. The Euclidean distance was used on these features to measure similarity. Using twenty coefficients gave surprisingly good performance while maintaining the interactive speed of the system. (Actually, the re-display time was found to be the limiting factor speed wise.)

Initially, Photobook displays forty randomly selected images from the database. The user selects an image of interest, and after about 2 seconds on a DECstation 5000 the display is updated with the "nearest" patterns. In the case shown, the user selected D51-raffia and the algorithm found the eight other subimages from the original D51 image. It also found several other textures, such as D72-tree stump which has subimages visually similar to D51.

5 Recognition results

First using a zero "false alarm" constraint, the very strictest, we determined the classification performance for every sample in the Brodatz database. This test required that Photobook return the eight other samples from the source image as the first eight nearest samples. This was repeated for all nine samples of each image, averaging the number recognized in the first eight. The patterns are ordered by these recognition rates in Table 1.

These results can be used in many ways. If a researcher picks only a subset of the Brodatz library, and happens to pick patterns at the high end of this table, then they can claim better than 90-100% performance. If they pick patterns located near the left, then this method will result in reduced performance. Hence, these results can be used to estimate a measure of "difficulty" for a particular subset of patterns. It is interesting to observe where the patterns people have traditionally used appear in this table. One can also use these results to typify the types of patterns the eigenvector method seems best/worst suited for.

Note that the results reported here correspond to a "worst case" scenario, or a lower bound on performance. Typically researchers evaluate performance on a smaller subset of the Brodatz patterns where the discrimination has less potential to choose a wrong pattern. These results are also conservative for another reason. As mentioned, the Brodatz database contains different images (e.g. D3 and D36, both of reptile skin) that humans may judge to look similar, but which our results declare to be a "miss" because the source images differ. This case appears to occur more frequently than the opposing case where two visually different samples from the same source image are declared to match. When a source image was very inhomogeneous, the performance was poor as one would expect for both the human and the computer algorithms.

The data presented above was for the strictest case, zero false-alarms. Fig. 2 shows the results of a study that allows up to 32 false alarms. This limit is appropriate as the screen only displays 40 images at once. It is acceptable in database search applications to declare success if all the patterns it should have found get displayed. Each point on this plot was formed from an average of the performance for all the 999 database images. Similar to the Neyman-Pearson operating characteristic [6], these curves show that the performance of the shift-invariant principal components will increase monotonically with the permitted number of false detections. In the limit of course, as

Avg. Recogn. Rate	Brodatz Texture	Avg. Recogn. Rate	Brodatz Texture
12.35	D88	60.49	D64
14.81	D89	60.49	D75
17.28	D31	60.49	D92
18.52	D43	60.49	D68
18.52	D58	62.96	D10
19.75	D42	64.20	D47
20.99	D107	65.43	D106
22.22	D40	65.43	D2
23.46	D72	65.43	D79
23.46	D30	66.67	D93
24.69	D23	66.67	D18
24.69	D69	66.67	D3
25.93	D60	66.67	D46
25.93	D41	66.67	D71
25.93	D91	67.90	D52
25.93	D98	69.14	D24
25.93	D13	69.14	D87
27.16	D90	70.37	D15
27.16	D73	71.60	D80
32.10	D27	74.07	D76
34.57	D100	74.07	D12
34.57	D7	75.31	D82
34.57	D108	76.54	D81
35.80	D44	76.54	D38
35.80	D97	80.25	D35
35.80	D63	81.48	D4
35.80	D112	82.72	D51
37.04	D28	82.72	D1
37.04	D54	82.72	D50
38.27	D39	85.19	D25
39.51	D62	86.42	D16
39.51	D99	86.42	D85
40.74	D59	86.42	D105
42.00	D36	87.65	D17
42.00	D111	88.89	D8
42.00	D66	90.12	D84
43.21	D61	90.12	D95
43.21	D70	91.36	D32
45.68	D48	95.06	D57
45.68	D109	96.30	D21
46.91	D5	97.53	D83
48.15	D86	97.53	D37
49.38	D74	97.53	D56
50.62	D104	98.77	D65
51.85	D29	98.77	D49
51.85	D94	98.77	D110
51.85	D33	98.77	D6
51.85	D67	98.77	D77
54.32	D19	100.00	D34
55.56	D11	100.00	D55
55.56	D78	100.00	D101
56.79	D26	100.00	D102
56.79	D22	100.00	D20
58.02	D9	100.00	D14
58.02	D96	100.00	D53
59.26	D103		

Table 1: Brodatz textures ordered by recognition rates.

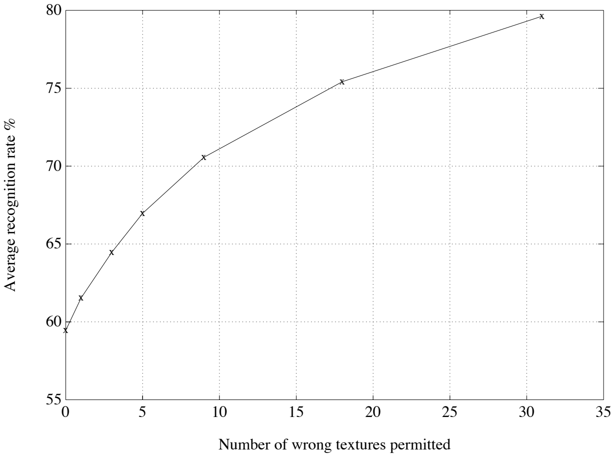


Figure 2: An approximate "operating characteristic" for the Brodatz database.

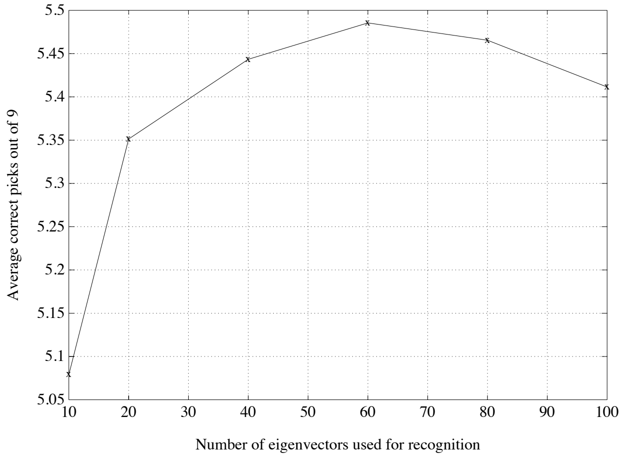


Figure 3: How average recognition rate, with zero false detects, varies while increasing the number of features.

the number of false detects goes to 990, all the curves will reach 100%. A method which has a curve lying above this one can be considered an improvement. Thus, this data can be used as a bench mark for subsequent methods.

As noted, the coefficients for the largest eigenvalues (except DC) may be optimal for representation, but this is not necessarily true for discrimination. To find the best optimal subset of size p requires a search over the power set of coefficients, a size 2^p problem. Clearly this is unreasonable for large p. A suboptimal alternative was taken here, picking subsets of size $s = 10$, 20, 40, 60, 80, and 100 from the coefficients corresponding to the s largest eigenvalues. The DC coefficient was only included in the last case.

The results of this study are shown in Fig. 3. Each point here is the average over a whole table of data like that in Table 1. This study verifies that using more coefficients, although better for representation, is not necessarily better for discrimination or classification. After the first sixty coefficients the performance degrades. The greatest increase can be seen to occur at twenty coefficients, where we concluded the best price-performance was for the interactive Photobook system.

6 Summary

The "similar pattern" recognition problem has been investigated in the context of a set of 999 images based on the entire Brodatz library of textures, the standard source in texture classification. The performance of a shift-invariant principal components algorithm was characterized for each pattern, for various false alarm rates, and for various subsets of features. These results provide a bench mark for comparing alternate algorithms. Additionally, they provide one ordering of the images from "difficult" to "easy" for their recognition from within the Brodatz database.

Acknowledgments

The authors would like to thank A. P. Pentland for the idea of the Photobook interface, T. E. Starner for its initial implementation, and C. H. Perry for his help implementing some of the experiments.

References

[1] P. Brodatz. *Textures: A Photographic Album for Artists and Designers.* Dover, New York, 1966.

[2] R. Haralick. Statistical and structural approaches to texture. *Proc. IEEE*, 67:786–804, May 1979.

[3] F. Ade. Characterization of textures by 'Eigenfilters'. *Signal Processing*, 5:451–457, 1983.

[4] M. Unser. Local linear transforms for texture measurements. *Signal Processing*, 11:61–78, 1986.

[5] R. Vistnes. Texture models and image measures for texture discrimination. *Int. J. of Comp. Vis.*, 3:313–336, 1989.

[6] C. W. Therrien. *Decision Estimation and Classification.* John Wiley and Sons, Inc., New York, 1989.

[7] S. Akamatsu, T. Sasaki, H. Fukamachi, and Y. Suenaga. A robust face identification scheme – KL expansion of an invariant feature space. In *Proc. SPIE Conf. On Intell. Robots and Comp. Vis.*, volume 1607, pages 71–84, Boston, Ma, Nov. 1991.

[8] L. Jacobson and H. Wechsler. Invariant analogical image representation and pattern recognition. *Patt. Rec. Lett.*, 2:289–299, 1984.