Multi-scale Document Description Using Rectangular Granulometries

Andrew D. Bagdanov and Marcel Worring

Intelligent Sensory Information Systems
University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam
The Netherlands
{andrew, worring}@science.uva.nl

Abstract. When comparing documents images based on visual similarity it is difficult to determine the correct scale and features for document representation. We report on new form of multivariate granulometries based on rectangles of varying size and aspect ratio. These rectangular granulometries are used to probe the layout structure of document images, and the rectangular size distributions derived from them are used as descriptors for document images. Feature selection is used to reduce the dimensionality and redundancy of the size distributions, while preserving the essence of the visual appearance of a document. Experimental results indicate that rectangular size distributions are an effective way to characterize visual similarity of document images and provide insightful interpretation of classification and retrieval results in the original image space rather than the abstract feature space.

1 Introduction

There are many applications in document image understanding where it is necessary to compare documents according to visual appearance before attempting high–level understanding of document content. Example applications include document genre classification, duplicate document detection, and document image retrieval.

Genre classification is useful for grouping documents for routing through office workflows, as well as identifying the type of document before applying class–specific strategies for document understanding [1]. Document image retrieval systems are of particular interest in some application areas [2]. Given an example image as a query, a document image retrieval system should return a ranked list of visually similar documents from an indexed collection. In document collections automatic conversion of documents is often expensive or impossible. In such cases image retrieval may be the only feasible means of providing access to a document database.

Whether document images are to be classified into a number of known document genres, or ranked by similarity to documents in a document database, it is necessary to establish meaningful measures of visual similarity between documents. To that end we must first define an appropriate document representation.

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 445–456, 2002.

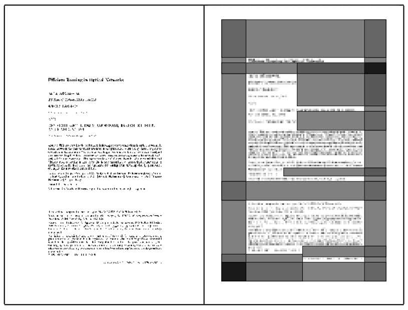


Fig. 1. Characterizing document images as a union of rectangles.

Consider the document shown in figure 1. The visual appearance of a document is determined by the foreground and background pixels in the document image. Document segmentation techniques using structural decompositions of the *background* are common in the literature on document understanding [3]. The background of a document image can be represented by rectangular regions of various sizes. Analysis of the structure of such rectangular decompositions can be used to derive useful descriptors of the appearance of document images. While most of the visual content of a document image can be described by analyzing the background in this way, for some documents it is necessary to perform the same type of decompositional analysis on the foreground. The most obvious example of this are documents containing reverse "video" regions.

The proper scale to use for document representation depends on the application, and hence a generic representation of visual content must be multi–scale. Some researchers, in fact, advocate exploration of an entire scale–space of potential document segmentations before committing to a single one [4]. Most techniques based on a single layout segmentation fail to take the multi–scale nature of visual perception into account. For documents this multi–scale nature is implicit in the scales distinguishing characters, words, textlines, paragraphs, columns, etc.

Our approach for representing visual content is based on morphological granulometric analysis of document images. A granulometry can be thought of as a morphological sieve, where objects not conforming to a particular size and shape are removed at each level of the sieving process. They were first introduced by Matheron for characterizing the probabilistic nature of random sets [5]. Granulometries, and the corresponding measurements taken on them, have been applied to problems of texture classification [6], image segmentation [7], and filtering [8]. Recent work by Vincent has shown how granulometries can be effectively and efficiently applied, particularly in the binary image domain [9].

The rest of this paper is organized as follows. We give in the next section a brief introduction to granulometries and the multivariate extensions used in our approach. Next, a description of our representation of document images derived from measurements on these granulometric filters is described. We also show how these measurements may be used to interpret the important features distinguishing between visually distinct document classes. To illustrate the effectiveness of our representation, we have applied our technique to the problems of document genre classification and document image retrieval. The results of these experiments are given in section 4.

2 Rectangular Granulometries

In this section we describe the properties of granulometries. Formally, a *granulometry* on $\mathcal{P}(R \times R)$, where $\mathcal{P}(X)$ is the power set of X, is a family of operators:

$$\Psi_t : \mathcal{P}(R \times R) \longrightarrow \mathcal{P}(R \times R)$$

satisfying for any $S \in \mathcal{P}(R \times R)$

A1: $\Psi_t(S) \subset S$ for all $t > 0$ (Ψ_t is anti–extensive)
A2: For $S \subset S'$, $\Psi_t(S) \subset \Psi_t(S')$ (Ψ_t is increasing).
A3: $\Psi_t \circ \Psi_{t'} = \Psi_{t'} \circ \Psi_t = \Psi_{\max(t,t')}$ for all $t, t' > 0$.

Of particular interest are granulometries generated by openings by scaled versions of a single convex structuring element B, i.e.

$$\Psi_t(S) = S \circ tB \ .$$

To capture the vertically and horizontally aligned regions of varying aspect ratios we use multivariate, rectangular granulometries to characterize document images. Let H and V be horizontal and vertical line segments of unit length centered at the origin. We define each opening in the rectangular granulometry as:

$$\Psi_{x,y}(S) = S \circ (yV \oplus xH).$$

The above definition makes use of the fact that any rectangle may be written as a dilation of its orthogonal horizontal and vertical components. Note that any increasing function $f(x)$ induces a univariate granulometry $\{\Psi_{x,f(x)}\}$ satisfying A1–A3. The extension to rectangular openings allows us to capture the information from all rectangular granulometries in a single parameterized family of operators. Figure 2 gives some example openings of this type for a document image.

3 Document Representation

In this section we describe our method for representing document images using measurements taken on rectangular granulometries. Note that it is not the openings constituting the rectangular granulometries described above, nor the filtered versions of the image S that are of most interest in describing the visual appearance of document images, but rather the *measurements* taken on the filtered images $\Psi_{x,y}(S)$.

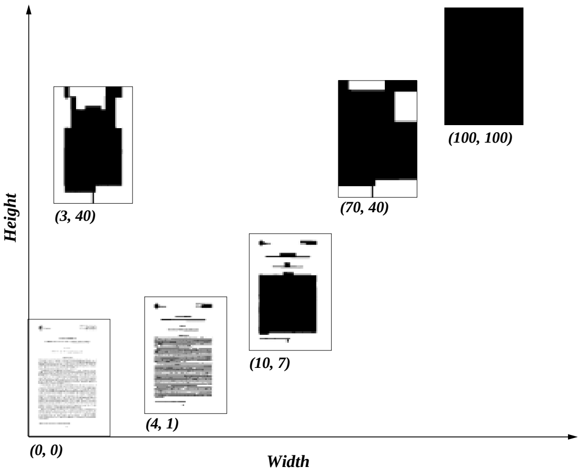


Fig. 2. Some examples of $\Psi_{x,y}(S)$ for a document S for various values of x and y. The multi–scale nature of documents is evident in the different structural relationships emerging at different levels in the granulometry: characters are merged into words, words into lines, and lines into textblocks. Eventually the margins are breached and the entire document is opened.

3.1 Rectangular Size Distributions

Maragos [6] has described two useful measurements for granulometries, the size distribution and pattern spectrum, which have subsequently been extended to multivariate granulometries [10]. We define the rectangular size distribution induced by the granulometry $G = \{\Psi_{x,y}\}$ on image S as:

$$\Phi_G(x, y, S) = \frac{A(S) - A(\Psi_{x,y}(S)))}{A(S)},$$

$A(X)$ denoting the area of set X. $\Phi_G(x, y, S)$ is a cumulative probability distribution, i.e. $\Phi_G(x, y, S)$ is the probability that an arbitrary pixel in S is opened by a rectangle of size $x \times y$ or smaller.

As mentioned in the introduction, documents with regions containing reverse video text, i.e. white text against a black background, are not thoroughly captured by the openings $\Psi_{x,y}$. To account for this, we extend the rectangular size distributions downward to include openings of the foreground. The definition becomes:

$$\Phi_G(x, y, S) = \begin{cases} \frac{A(S)-A(\Psi_{x,y}(S)))}{A(S)} & \text{if} \quad x, y \geq 0 \\ \frac{A(S^c)-A(\Psi_{x,y}(S^c)))}{A(S^c)} & \text{if} \quad x, y < 0 \end{cases}$$

The pattern spectrum is defined as the derivative of $\Phi_G(x, y, S)$, for which we have two choices in the case of rectangular granulometries. For document images there is no *a priori* evidence for preferring either horizontal or vertical

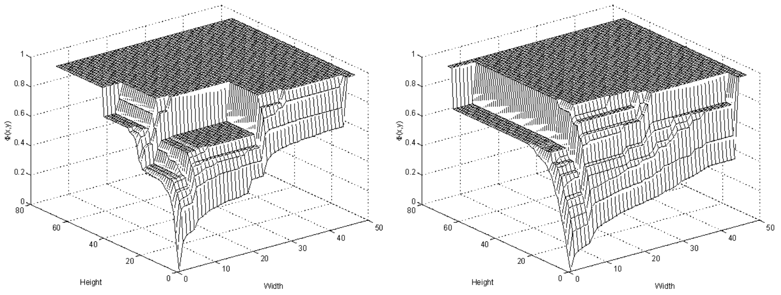


Fig. 3. Example rectangular size distributions for two documents from different genres in our test database. Note the prominent flat plateau regions indicating regions of stability in the granulometry. These most likely correspond to typographical parameters such as margin width, inter–line distance, etc. The size distribution on the left is constructed from the document shown in figure 1, and the one on the right from the document used to construct the example openings in figure 2.

directional derivatives, e.g. preferring emphasis on inter–column gap over inter–line spacing, and for now we concentrate on using the size distribution as our document representation.

Figure 3 gives two example size distributions. In these examples, we only plot the size distribution in the first quadrant, i.e. for $x, y > 0$. We see that the rectangular size distribution captures much information about the document image. Of specific interest are the plateau regions in the size distribution, which indicate islands of stability most likely corresponding to specific typographical features such as inter–line spacing, paragraph spacing, and inter–column gap.

3.2 Efficiency

It is not feasible to exhaust the entire parameter space for rectangular size distributions in a naïve way. This is especially true for document images, which tend to be large. We can take advantage of several properties of rectangular granulometries and size distributions in order to make their computation more tractable.

First, each rectangular opening may be decomposed into linear erosions and dilations as follows:

$$\begin{aligned}\varPsi_{x,y}(S) &= S \circ (yV \oplus xH) \\ &= (S \ominus (yV \oplus xH)) \oplus (yV \oplus xH) \\ &= (((S \ominus yV) \ominus xH) \oplus yV) \oplus xH. \end{aligned} \tag{1}$$

This eliminates the need to directly open a document image by rectangles of all sizes. Instead, the opening is incrementally constructed by the orthogonal components of each rectangle, which are increasing linearly in size rather than quadratically.

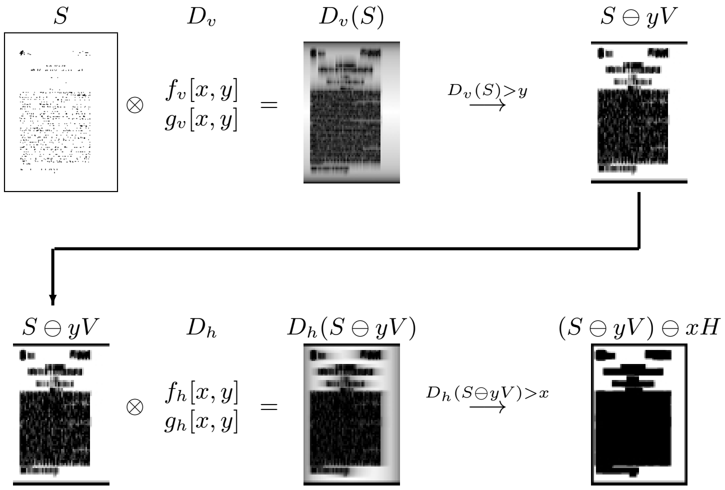


Fig. 4. Efficient computation of an arbitrary rectangular opening. Distance transforms are used to effectively encode all possible vertical and horizontal erosions. By thresholding these distance images we can obtain each desired erosion. The $\otimes$ operator is used above to indicate the application of the recursive filters described in equations 2 and 3. The first part of the opening, $(S \ominus yV) \ominus xH$, is illustrated above. The opening is completed by performing the same steps on $((S \ominus yV) \ominus xH)^c$.

Next, we can eliminate the need to erode and dilate the image by structuring elements increasing linearly in size. Using linear distance transforms for vertical and horizontal directions we can generate all needed erosions and dilations for each rectangular opening. The horizontal distance transform of an image S is defined as:

$$D_h(S,x,y) = \min\{\Delta x \mid (x \pm \Delta x, y) \in S\},$$

and the vertical distance transform as:

$$D_v(S,x,y) = \min\{\Delta y \mid (x, y \pm \Delta y) \in S\}.$$

These transforms can be efficiently performed using the following recursive forward/backward filter pairs defined on image S:

$$D_h \begin{cases} f_h[x,y] = \min\{f[x-1,y]+1,\ S(x,y)\} \\ g_h[x,y] = \min\{f[x,y],\ g[x+1,y]+1\} \end{cases} \tag{2}$$

$$D_v \begin{cases} f_v[x,y] = \min\{f[x,y-1]+1,\ S(x,y)\} \\ g_v[x,y] = \min\{f[x,y],\ g[x,y+1]+1\} \end{cases} \tag{3}$$

The use of these distance transforms to generate erosions of the original image represents a significant savings in computation time. To generate a vertical or horizontal erosion of arbitrary size we only have to apply two fixed–size recursive neighborhood operations, rather than eroding by structuring elements increasing in size. In this way each opening can be incrementally constructed as illustrated in figure 4.

Lastly, since rectangular size distributions are monotonically increasing in both parameters, i.e. if $x' \geq x$ and $y' \geq y$ then $\Phi_G(x', y', S) \geq \Phi_G(x, y, S)$, we can recursively search the parameter space, eliminating the need to explore large, flat regions.

3.3 Feature Space Reduction and Interpretation

The multi–scale representation developed in the previous two subsections captures much structural information about document images, but does little toward reducing the overall complexity of the problem. To that end we describe in this subsection our approach to dimensionality reduction, which also leads to interesting qualitative interpretations in the original document image space.

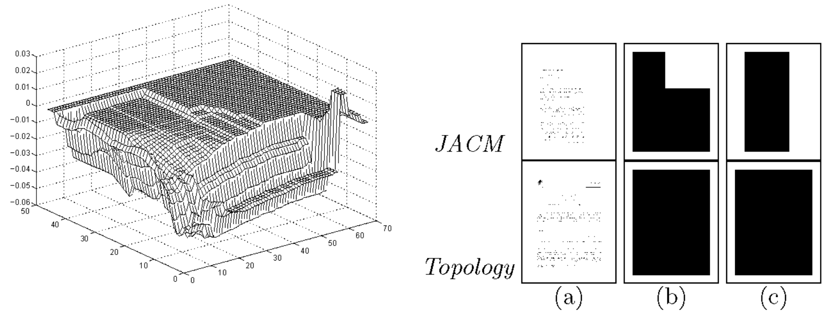


Fig. 5. Interpreting the PCA results. On the left are shown the coefficients in the principle eigenvector mapped back into the original feature space of the size distribution (i.e. the same feature space as shown in the examples given in figure 3). On the right, individual openings are interpreted: (a) shows the original images, (b) an opening emphasizing the presence of the *Topology* logotype, and (c) an opening emphasizing the differences in margins.

The dimensionality of the entire size distribution is too large to be applied effectively in a statistical pattern recognition setting. Some feature selection or reduction strategy must be applied. Principle Component Analysis (PCA) is a well–known approach to feature reduction, and can be applied to rectangular size distributions to reduce the dimensionality of our document representation, while preserving the maximum amount of variance in a document collection. The principle component mapping defines a rotation of the original feature space using the eigenvectors of the covariance matrix of the dataset. Since each eigenvector is of the same dimensionality as the original feature space, we can visualize them individually in the same way as size distributions. Figure 5 shows the coefficients of the first principle component computed for a two–class subset of our four document genres.

From inspection of the plot on the left in figure 5 it is evident that it is not necessary to sample much of the parameter space in order to account for most of the variance in the entire sample. In particular, most of the large openings do not contribute at all to the variance in the first principle component mapping. By selecting a coefficient of high magnitude in the first principle component, we can compute the corresponding opening $\Psi_{x,y}(S)$ on document images from our test sample. This allows us to interpret features important for distinguishing between documents in the original image space. The opening shown in figure 5b emphasizes the presence of the logotype appearing in the upper right corner of *Topology* articles, while in figure 5c the differences in margins are emphasized.

The principle component mapping is also useful for visualizing an entire genre of document images. Figure 6 shows a sample class of document images (from the *Journal of the ACM*) after mapping to the first two principle components. The clusters in the low–dimensional space represent the gross typographical differences between document images from this class. In this case, clusters indicating the paper size and gutter orientation are clearly defined. The outliers in this plot are page images not conforming to the standard layout style for articles, such as errata pages and editorials.

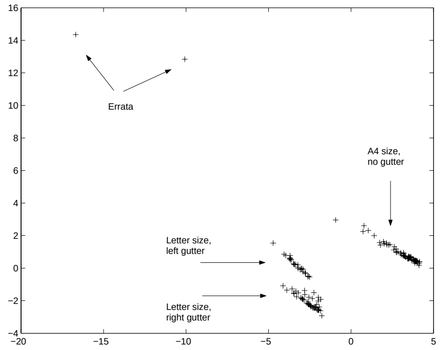


Fig. 6. The first two principle components for two document classes.

4 Experimental Results

To illustrate the effectiveness of rectangular granulometries, we have applied the technique to the problems of document genre classification and document image retrieval. A total of 537 PDF documents were collected from several digital libraries. The sample contains documents from four different journals, which determine the genres in our classification problem, and relevance for document retrieval. Note that these genres are not necessarily determined by visual similarity. Since we are using an inherently *logical* definition of document genre, i.e. coming from the same publication, there may be significantly different visual sub–genres within each genre (see figure 6). However, this does give us a non–subjective division of our document collection.

We consider only the first page of each document, as it contains most of the visually significant features for discriminating between document genres. The first page of each PDF document was converted to an image and subsampled to 1/4 of its original size. The rectangular size distribution described in section 2 was then computed for each image. Each quadrant of the size distribution is then sampled to form a rectangular size distribution of size 41 × 61. The resulting dimensionality of our feature space is 5002.

4.1 Genre Classification

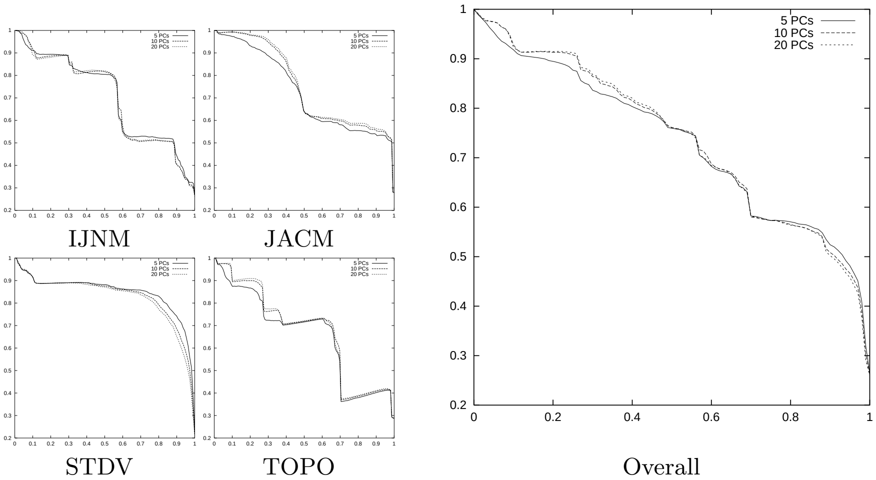
Table 1 gives the estimated classification accuracy for a training sample of 30 documents selected randomly from each document genre, with the remaining documents used as an independent test set. Estimated classification accuracy is shown for 5, 7, and 10 principle components computed from the training sample, and for a 1-nearest neighbor, quadratic discriminant, and linear discriminant classifier. These results indicate that, even with relatively few principle components, rectangular granulometries are capable of capturing the relevant differences between document genres.

Table 1. Genre classification results for 30 training samples per class and various numbers of principle components. Classification accuracy is estimated by averaging over 50 experimental trials. The PCA is performed independently for each trial.

	# PCs		
Classifier	5	7	10
1-Nearest Neighbor	94%	95%	98%
Quadratic Discriminant	93%	94%	98%
Linear Discriminant	76%	80%	93%

4.2 Document Image Retrieval

For the document image retrieval experiments, a single document image is given as a query, and a ranked list of relevant documents is returned. We use the rectangular size distributions described above as the representation for each document. Document ranking is computed using the Euclidean distance from the size distribution of the query document. For evaluation, a document is considered relevant if it belongs to the same genre as the query document (i.e. it is from the same publication).



IJNM JACM

STDV TOPO

Overall

Fig. 7. Average precision and recall plots for each genre in the test database. Results on the entire feature space and with 5, 10, and 20 principle components are shown. The graphs on the left show the precision and recall for each individual class, while the plot on the right gives the overall average precision and recall.

Precision and recall statistics can be used to measure the performance of retrieval systems. They are defined as:

$$\text{Precision} = \frac{\#\text{ relevant documents retrieved}}{\#\text{ documents retrieved}}$$
$$\text{Recall} = \frac{\#\text{ relevant documents retrieved}}{\#\text{ relevant documents}}$$

Rather than computing the overall precision, it is more useful to sample the precision and recall at several cutoff points. For a given recall rate, we can determine what the resulting precision is. That is, how many non–relevant documents must we inspect before finding that fraction of relevant documents.

Figure 7 gives the average precision/recall graphs for each document genre in our database. The graphs were constructed by using each document in a genre as a query, ranking all documents in the database against it, and computing the precision at each recall level. These individual precision/recall statistics are then averaged to form the final graph.

The graphs in figure 7 give a good indication of how well each individual genre is characterized by the rectangular size distribution representation, and also indicates the overall precision and recall for the entire dataset. The overall precision/recall graph is constructed by averaging the precision and recall rates over all classes. This graph indicates that, on average, 50% of all relevant documents can be retrieved with a precision of about 80%.

All of the precision/recall graphs have a characteristic plunging tail, indicating that there are some queries where relevant documents appear near the end of the ranked list. It is illustrative to examine some specific examples of this phenomenon. Figure 8 gives some example query images along with the highest ranked relevant document returned, excluding the query document itself, and the lowest ranked relevant document returned. In most cases these low ranking relevant documents represent pathologically different visual sub–classes of the document genre.

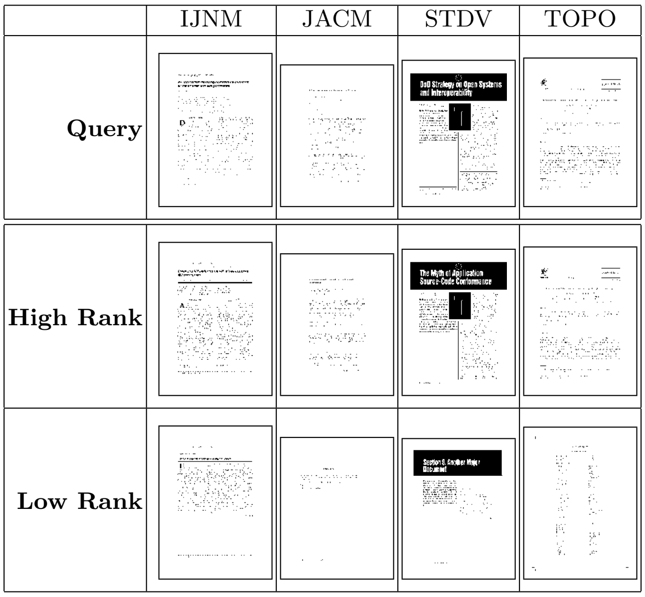


Fig. 8. Some illustrative query examples. A sample query image for each genre is shown, along with the highest and lowest ranked relevant images from the relevant genre. In most cases the least relevant document is pathologically different the query.

5 Conclusions

We have reported on an extension to multivariate granulometries that uses rectangles of varying scale and aspect ratio to characterize the visual content of document images. Rectangular size distributions are an effective way to describe the visual structure of document images, and by employing morphological decomposition techniques they can be efficiently computed. Experiments have shown that size distributions can be used to discriminate between specific document genres. Principle component analysis can be used to reduce the dimensionality of multivariate size distributions, while preserving their discriminating power. One of the attractive aspects of rectangular size distributions is the ability, even under dimensionality reduction, to interpret significant features back in the original image space.

Document retrieval experiments also indicate the effectiveness of rectangular size distributions for capturing visual similarity of documents. For our document database 50% of relevant documents can be retrieved with a precision of approximately 80%.

Principle component analysis has proved useful for accentuating the important features in size distributions. A non–linear PCA approach which maximizes inter–class variance while minimizing intra–class variance will certainly improve both the classification and retrieval results.

We plan to elaborate further on feature selection approaches in the near future. The entire parameter space for rectangular size distributions is expensive to sample for document images. Feature selection, as opposed to feature reduction such as PCA, is more desirable because of this. Feature subsets are also more natural to interpret in terms of the original document images. Research in currently focused on feature selection strategies which also (re–)introduce spatial information into the size distribution representation.

It should be noted that the techniques presented in this paper are not limited solely to visual similarity matching, but rather constitute a general approach to multi–scale analysis. As such, the granulometric approach may prove useful for applications such as table decomposition, text identification, and layout segmentation. A systematic study of the effects of noise on the representation is essential to establishing the widespread applicability of the granulometric technique to document understanding.

References

1. Shin, C.K., Doermann, D.S.: Classification of document page images based on visual similarity of layout structures. In: Proceedings of the SPIE Document Recognition and Retrieval VII. (2000)
2. Doermann, D.S.: The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding **70** (1998)
3. Antonacopoulos, A.: Page segmentation using the description of the background. Computer Vision and Image Understanding **70** (1998) 350–369
4. Breuel, T.: Thomas m. breuel. layout analysis by exploring the space of segmentation parameters. In: Proceedings of the Fourth International Workshop on Document Analysis Systems (DAS'2000). (2000)
5. Matheron, G.: Random Sets and Integral Geometry. John Wiley & Sons, New York (1975)
6. Maragos, P.: Pattern spectrum and multiscale shape representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989) 701–716
7. Dougherty, E.R., Pelz, J., Sand, F., Lent, A.: Morphological image segmentation by local granulometric size distributions. J. Electronic Imaging **1** (1992)
8. Haralick, R.M., Katz, P.L., Dougherty, E.R.: Model-based morphology: the opening spectrum. Graphical Models and Image Processing **57** (1995) 1–12
9. Vincent, L.: Granulometries and opening trees. Fundamenta Informatica **41** (2000) 57–90
10. Batman, S., Dougherty, E.R., Sand, F.: Heterogeneous morphological granulometries. Pattern Recognition **33** (2000) 1047–1057