

# Multiscale document description using rectangular granulometries

Andrew D. Bagdanov, Marcel Worring

Intelligent Sensory Information Systems, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Received: 16 November 2002 / Accepted: 20 June 2003

Published online: September 12, 2003 – © Springer-Verlag 2003

**Abstract.** When comparing document images based on visual similarity it is difficult to determine the correct scale and features for document representation. We report on a 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.

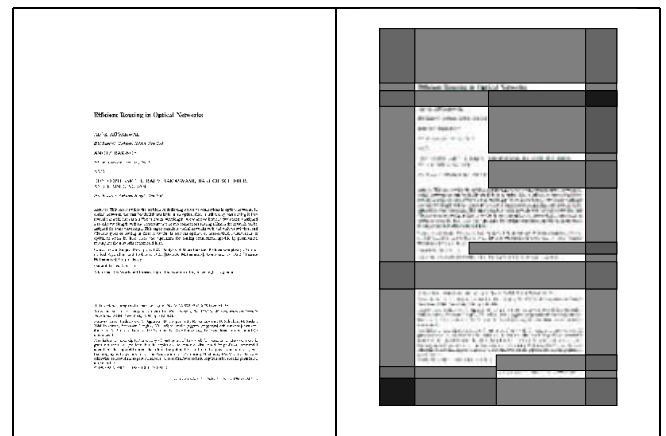
**Key words:** Mathematical morphology – Granulometries – Document image understanding – Document genre classification

## 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 for identifying the type of document before applying class-specific strategies for document understanding [11]. Document image retrieval systems are of particular interest in some application areas [5]. Examples of such application areas include digital libraries, ancient document collections,

Correspondence to: A.D. Bagdanov (andrew@science.uva.nl)



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

and technical drawing databases. Given an example image as a query, a document image retrieval system returns a ranked list of visually similar documents from an indexed collection. In some collections, automatic conversion of documents to electronic formats 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. Consider the document shown in Fig. 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 image processing [1]. The background of a document image can be represented by rectangular regions of various sizes. Analysis of the structure of such rectangular decompo-

sitions can be used to derive useful descriptors of the appearance of document images.

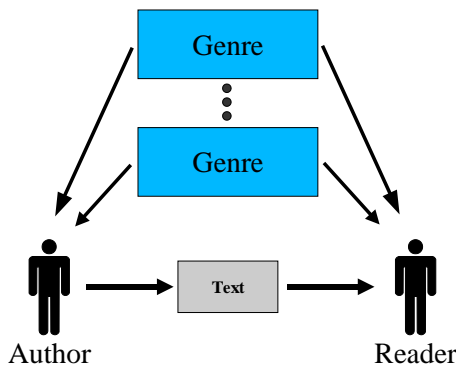
Note that most of the visual content of a document image can be described by analyzing the background in this way; for some documents it is necessary to perform the same type of compositional analysis on the foreground. The most obvious example of this are documents containing reverse “video” regions.

Documents have an intrinsic multiscale nature. This multiscale character is implicit in the scales distinguishing characters, words, textlines, paragraphs, columns, etc. The proper scale to use for document representation depends on the application, and hence a generic representation of visual content must be multiscale. Some researchers, in fact, advocate exploration of an entire scale-space of potential document segmentations before committing to a single one [3]. Most techniques based on a single layout segmentation fail to take the multiscale nature of visual perception into account.

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 [9]. Granulometries, and the corresponding measurements taken on them, have been applied to problems of texture classification [8], image segmentation [6], and filtering [7]. Recent work by Vincent has shown how granulometries can be effectively and efficiently applied, particularly in the binary image domain [12].

Traditional granulometries employ openings by homothetics, i.e., scaled versions, of a single structuring element to generate the filtered image at each level. Granulometries characterize the granular composition of images nicely when, as in the case of boolean random sets, they are constituted of homothetic versions of a single primary grain. While many natural textures fall into this category, the structural composition of document images is not so nicely captured by homothetic granulometries. The background of a document image typically contains rectangular regions of many different aspect ratios, and any homothetic filtering process will fail to capture both independent dimensions. For this reason, we propose a new multivariate rectangular granulometry that can be used to explore the entire space of rectangular image decompositions.

The rest of this paper is organized as follows. In the next section, we introduce the concept of document genre, which provides a context for understanding document similarity. We discuss the theory of granulometric analysis and the specifics of our multivariate extension to rectangular granulometries in Sect. 3. 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 that distinguish visually distinct document classes. To illustrate the effectiveness



**Fig. 2.** Genre as mediator. Knowledge of a genre, by both author and reader, create communication pathways for specific message components

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 Sect. 5.

## 2 Document genre

Humans rarely read documents outside of a specific context influencing their interpretation. Such contextual influences can be partially characterized by the concept of genre. Abstractly, document genre acts as a mediating factor between author and reader. Figure 2 illustrates this concept. Document genre consists of medium-specific rules or conventions that allow elements of an author’s message to be effectively encoded within a medium. Knowledge of a specific genre allows an author to effectively encode a message and a reader to decode it. For example, when presented with an unknown business letter, most people can easily decode the visual and typographical cues in it to identify the sender and recipient without the need to actually *read* any of the content. It is knowledge of the genre of business letters that makes this possible.

We can narrow this rather abstract conception of genre by adopting the following definition:

**Definition 1.** A document genre is a category of documents characterized by similarity of expression, style, form, or content.

Through these four elements document genre creates an implicit contract between author and reader. This contract manifests itself as expectation in the reader, expectation that specific components of the message are presented through known rules of expression, style, form, and content. In the semiotic research community, analysis of such structures is sometimes described as the study of *how* symbols mean as opposed to *what* they mean [4].

There are three major components of document genre for machine-printed texts:

- **visual genre**, which dictates the overall appearance of a document

- **typographical genre**, which determines the characteristics of the various font styles, sizes, and forms of emphasis applied to different information components
- **textual genre**, which consists of the rules of expression, terminology, and rhetorical devices used to express linguistic content

Written communication is highly structured, and document typesetting systems exploit this in organizing logical content into a physical realization of that content in geometric layout structures. Document understanding systems likewise exploit this structure in decomposing document images into typographically homogeneous regions before attempting high-level understanding. Just as genre plays a key role in mediating author/reader communication, it can play a similar role in document understanding systems. Figure 3 illustrates the conceptual pipeline of processing steps in a document understanding system and indicates the genre characterization stages inserted along the processing flow. Immediately after scanning, the visual components of a document’s genre can be analyzed. After layout analysis, the typographical constructs can then be characterized. Lastly, the textual components of genre can be extracted from the text extracted by an OCR system. All of these components of genre are finally indexed in a document retrieval system.

Note the bidirectional communication between the logical analysis stage and the retrieval system. Logical analysis algorithms may utilize genre characterization by exploiting genre-level similarity with known documents already indexed in the system. As shown in Fig. 3, this work focuses on the characterization of visual components of document genre. In the following sections, we detail our approach to characterizing the visual appearance of documents.

### 3 Granulometries

The visual appearance of a document is wholly determined by the foreground and background pixels in the document image. While complete, this representation is cumbersome due to the enormous semantic gap between the sensor space, i.e., the field of pixels acquired by the document scanner, and the conceptual space in which documents are interpreted. Documents are intrinsically multiscale, and their multiscale nature is evident in the scales distinguishing characters from words, words from textlines, textlines from paragraphs, etc. A multiscale approach is consequently an obvious choice for a genre characterization technique. Our approach is based on multiscale decompositions of the background of document images. We can imagine a document image being decomposed into a collection of maximal rectangles that “fit” into the background of the image, as shown in Fig. 1. Such decompositions supply information about the information-bearing portions of a document or class of documents. In this and the following section, we show how such decompositions may be constructed and ana-

lyzed in order to characterize visual similarity between document genres.

As mentioned above, the intrinsic multiscale nature of documents lends itself well to multiscale analysis techniques. Morphological scale-spaces possess conceptual and practical advantages that make them particularly suitable in the document image domain. In this section, we first introduce some necessary concepts and terminology from mathematical morphology and then describe the specific extensions we use to measure visual structure in document images.

We are primarily concerned with scanned, binary document images, and all of our morphological operations will be defined on subsets of the Euclidean plane, or constant Euclidean images in morphological parlance. All of our notation and conventions follow those of Serra [10].

The basic operations in mathematical morphology are the *erosion* and *dilation*. An erosion of image  $S$  by structuring element  $B$  is defined in terms of Minkowski subtraction:

$$e_B(S) = \bigcap_{y \in B} S_{-y} = S \ominus \check{B}$$

where  $S_{-y}$  denotes the translate of  $S$  by  $-y$ , and  $\check{B}$  denotes the reflection of  $B$  about the origin. All of the erosions we will discuss use symmetric structuring elements, and we will therefore denote erosion simply as  $S \ominus B$ . Dilation is similarly defined in terms of Minkowski addition:

$$d_B(S) = \bigcup_{y \in B} S_y = S \oplus B$$

Note that erosion and dilation are dual with respect to complementation, i.e.,  $S^c \oplus B = (S \ominus B)^c$ .

Combinations of erosions and dilations can be constructed that perform more elaborate transformations of images. The most basic of these are the opening and closing operations. The opening of an image by structuring element  $B$  is:

$$S \circ B = (S \ominus B) \oplus B$$

and the closing

$$S \bullet B = (S \oplus B) \ominus B$$

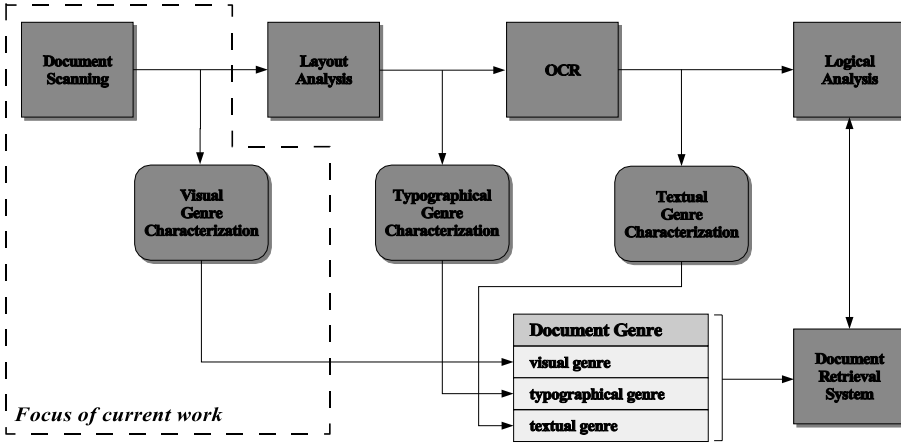
One of the most useful tools in mathematical morphology is the granulometry, which is constructed through sequential opening of an image.

Formally, a *granulometry* on  $\mathcal{P}(R \times R)$ , where  $\mathcal{P}(X)$  is the power set of  $X$  and  $R$  is the set of real numbers, 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  $t > 0$  ( $\Psi_t$  is antiextensive)
- A2:** For  $S \subset S'$ ,  $\Psi_t(S) \subset \Psi_t(S')$  ( $\Psi_t$  is increasing)
- A3:**  $\Psi_t \circ \Psi_{t'} = \Psi_{t'} \circ \Psi_t = \Psi_{\max(t,t')}$  for all  $t, t' > 0$



**Fig. 3.** Characterizing document genre along the way. At each stage in the document analysis process some type of genre characterization may be performed. Genre information is then indexed in a document retrieval system along with the extracted logical information

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

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

Maragos [8] has described two useful measurements on granulometries, the size distribution and the pattern spectrum. The size distribution induced by the granulometry  $G = \{\Psi_t\}$  on image  $S$  is:

$$\Phi_G(t, S) = \frac{A(S) - A(\Psi_t(S))}{A(S)} \quad (1)$$

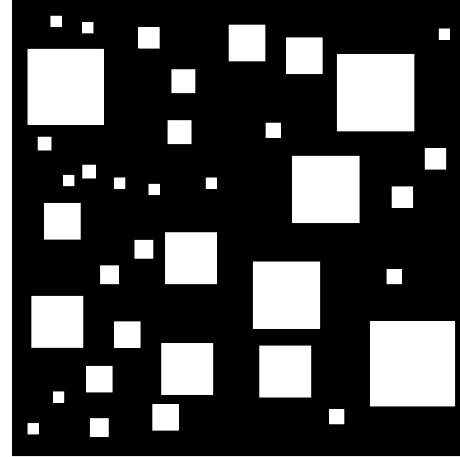
$A(X)$  denoting the area of set  $X$ .  $\Phi_G(t, S)$  is a cumulative probability distribution. The pattern spectrum is defined as the derivative of the size distribution and is a probability density function. Intuitively, the pattern spectrum represents the frequency of grain sizes occurring in the image, where the grains are defined as scaled versions of the convex structuring element used to construct the granulometry. Figure 4 provides an example size distribution and pattern spectrum for a synthetic image.

Univariate size distributions, i.e., size distributions constructed from granulometries with a single scale parameter as in Eq. 1, are generally incapable of capturing all of the free variables controlling grain placement and orientation. All of the example abstract images shown in Fig. 5 generate identical size distributions. For document images such discriminations are vital, as the arrangement of both vertically and horizontally aligned rectangles is key to background decomposition. In such cases, multivariate granulometries are more appropriate [2].

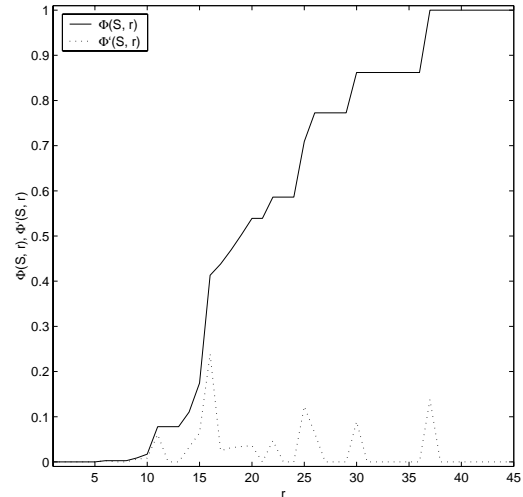
To capture the vertically and horizontally aligned regions of varying aspect ratios, we use multivariate, rectangular granulometries to characterize document images. Let  $H$  and  $V$  be horizontal and vertical line segments of unit length centered at the origin. We define each opening in the rectangular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

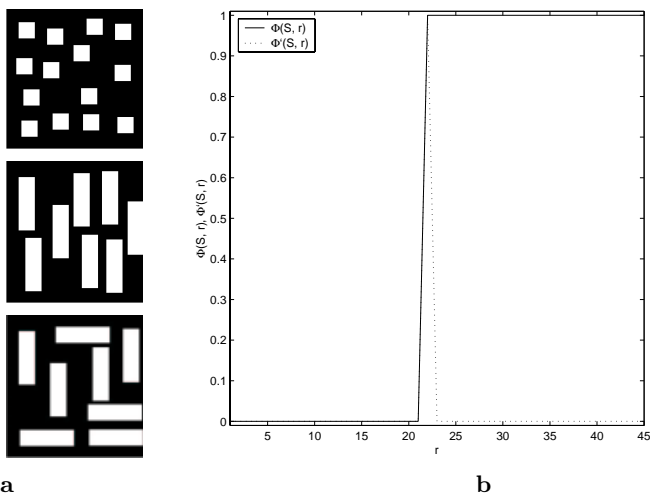


**a**

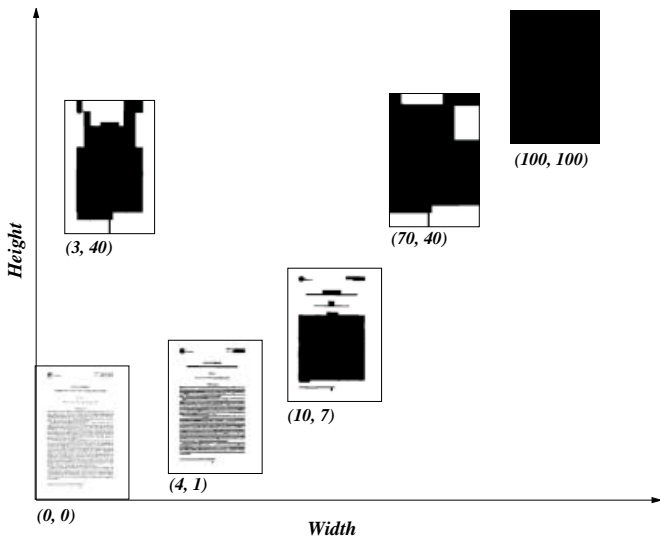


**b**

**Fig. 4a,b.** The size distribution and pattern spectrum of an image consisting entirely of square grains at various scales. The *solid line* is the size distribution, and the *dotted line* is the pattern spectrum



**Fig. 5a,b.** Ambiguity in size distributions. All of the images in **a** generate the size distribution shown in **b**. In many cases, univariate size distributions are incapable of capturing all of the degrees of freedom associated with grain sizing and orientation



**Fig. 6.** Some examples of  $\Psi_{x,y}(S)$  for a document  $S$  for various values of  $x$  and  $y$ . The multiscale 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 text blocks. Eventually the margins are breached and the entire document is opened

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 6 gives some example openings of this type for a document image.

## 4 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 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  $\Psi_{x,y}(S)$ .

### 4.1 Rectangular size distributions

The size distributions and pattern spectra introduced by Maragos [8] have been subsequently extended to multivariate granulometries [2]. The rectangular size distribution induced by the granulometry  $G = \{\Psi_{x,y}\}$  on image  $S$  is:

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

$A(X)$  denoting the area of set  $X$ .  $\Phi_G(x, y, S)$  is also 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 on 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:

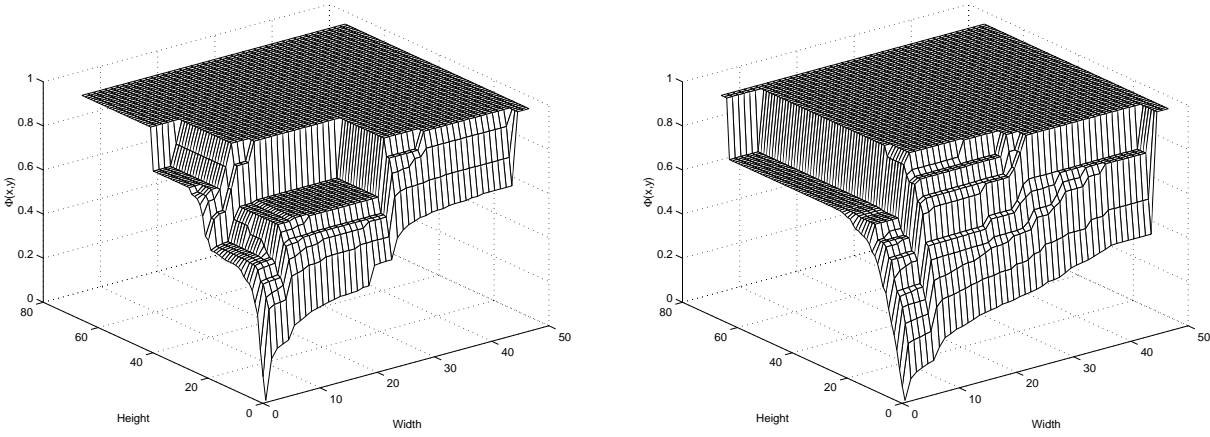
$$\Phi_G(x, y, S) = \begin{cases} \frac{A(S) - A(\Psi_{x,y}(S))}{A(S)} & \text{if } x, y \geq 0 \\ \frac{A(S^c) - A(\Psi_{x,y}(S^c))}{A(S^c)} & \text{if } x, y < 0 \end{cases} \quad (2)$$

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

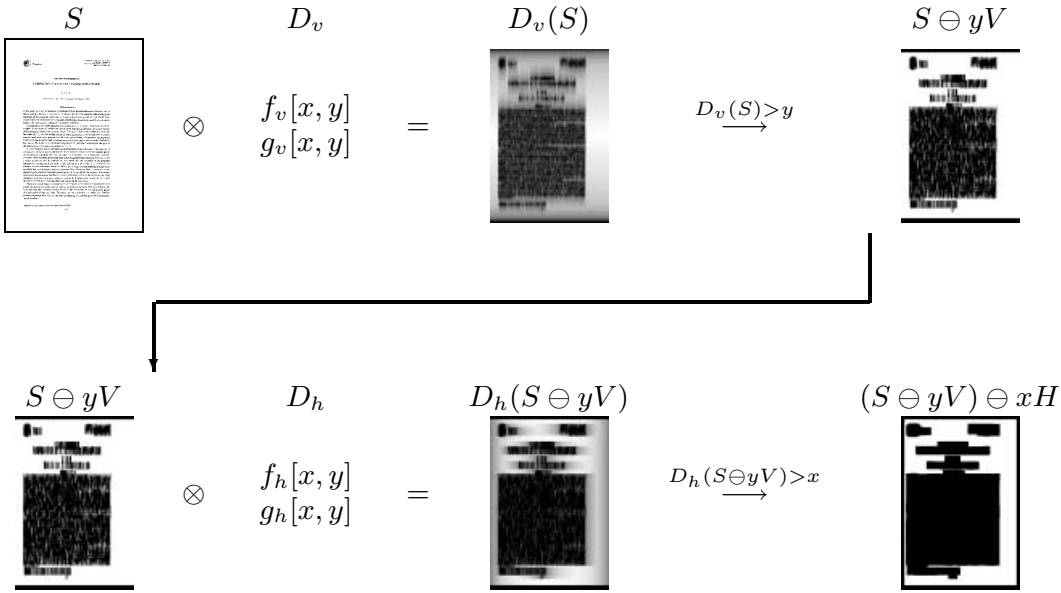
Figure 7 gives two example size distributions. In these examples, we only plot the size distribution in the first quadrant, i.e., for  $x, y > 0$ . We see that the rectangular size distribution captures much information about the document image. Of specific interest are the plateau regions in the size distribution, which indicate islands of stability most likely corresponding to specific typographical features such as interline spacing, paragraph spacing, and intercolumn gap.

### 4.2 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



**Fig. 7.** 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, interline distance, etc. The size distribution on the left is constructed from the document shown in Fig. 1 and the one on the right from the document used to construct the example openings in Fig. 6



**Fig. 8.** 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 Eqs. 4 and 5. The first part of the opening,  $(S \ominus yV) \ominus xH$ , is illustrated above. The opening is completed by performing the same steps on  $((S \ominus yV) \ominus xH)^c$

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:

$$\begin{aligned}
 \Psi_{x,y}(S) &= S \circ (yV \oplus xH) \\
 &= (S \ominus (yV \oplus xH)) \oplus (yV \oplus xH) \\
 &= (((S \ominus yV) \ominus xH) \oplus yV) \oplus xH
 \end{aligned} \tag{3}$$

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.

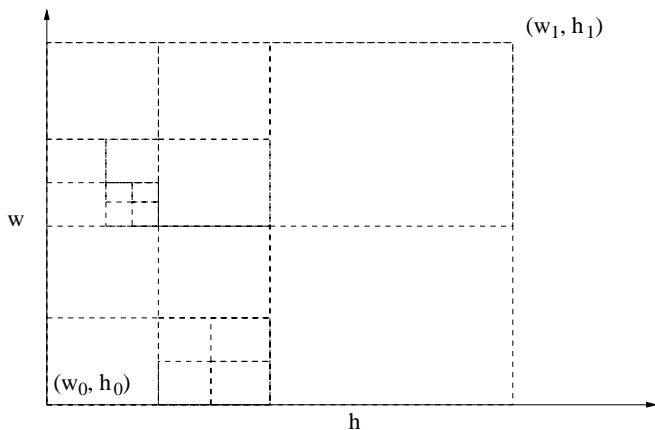
Next, we can eliminate the need to erode and dilate the image by structuring elements increasing linearly in size. Using linear distance transforms for vertical and horizontal directions we can generate all needed erosions and dilations for each rectangular opening. The horizontal distance transform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



**Fig. 9.** Recursive exploration of the rectangular parameter space. If  $\Phi_G(w_0, h_0, S) = \Phi_G(w_1, h_1, S)$ , then *every* opening in the rectangle defined by these points will open the same area. If not, the same strategy is applied recursively on the four subrectangles

on image  $S$ :

$$D_h \begin{cases} f_h[x, y] = \min\{f[x-1, y] + 1, S(x, y)\} \\ g_h[x, y] = \min\{f[x, y], g[x+1, y] + 1\} \end{cases} \quad (4)$$

$$D_v \begin{cases} f_v[x, y] = \min\{f[x, y-1] + 1, S(x, y)\} \\ g_v[x, y] = \min\{f[x, y], g[x, y+1] + 1\} \end{cases} \quad (5)$$

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 Fig. 8. The computational complexity of this algorithm, for the computation of a single opening, is linear in the number of pixels in the image. Since the total number of openings computed for a rectangular size distribution is typically a linear function of the size of the image (e.g., a linear subsampling of all possible widths and heights), the total running time of the algorithm is  $O(n^2)$ , where  $n$  is the number of pixels in the image.

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. The recursive decomposition process is illustrated in Fig. 9.

#### 4.3 Feature space reduction and interpretation

The multiscale representation developed in the previous two subsections captures much structural information about document images, and we have also shown how the computational complexity of computing rectangular size distributions can be reduced. However, the complexity of the representation itself remains unchanged. 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.

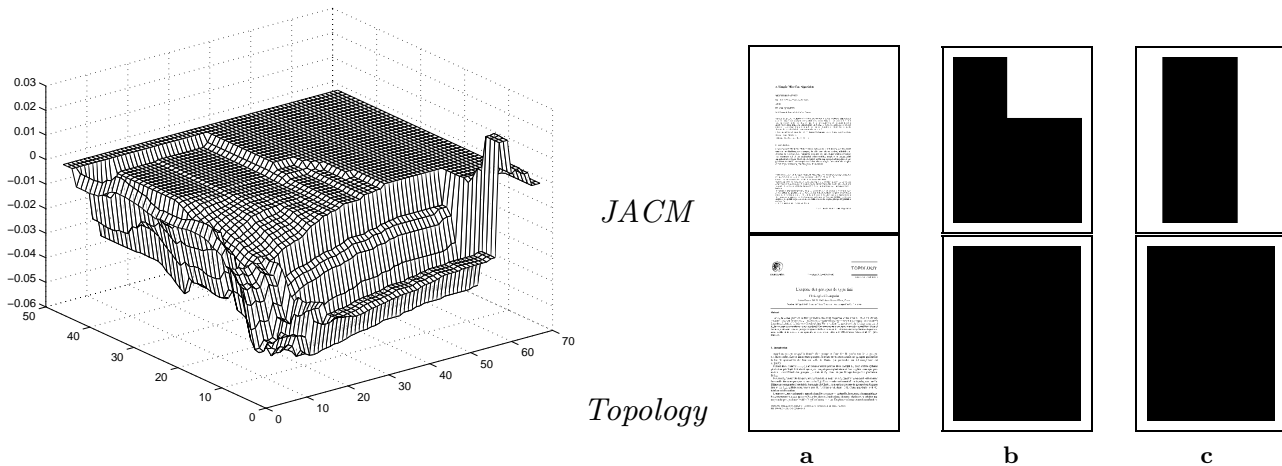
The dimensionality of the entire size distribution is too large to be applied effectively in a statistical pattern recognition setting. Some feature selection or reduction strategy must be applied. Principal Component Analysis (PCA) is a well-known approach to feature reduction and can be applied to rectangular size distributions to reduce the dimensionality of our document representation while preserving the maximum amount of variance in a document collection. The principal 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 10 shows the coefficients of the first principal component computed for a two-class subset of our four document genres (see Sect. 5 for a description of the document collection used in our experiments).

From inspection of the plot on the left in Fig. 10 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 principal component mapping. By selecting a coefficient of high magnitude in the first principal 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 Fig. 10b emphasizes the presence of the logotype appearing in the upper right corner of *Topology* articles, while in Fig. 10c the differences in margins are emphasized.

The principal component mapping is also useful for visualizing an entire genre of document images. Figure 11 shows a sample class of document images (from the *Journal of the ACM*) after mapping to the first two principal 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.

## 5 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 that determine the genres in our classification problem and the relevance for document retrieval. Table 1 gives an overview of the journals comprising our dataset. Note that these genres are not



**Fig. 10a–c.** Coefficients of the first principal component for a document collection. On the left are shown the coefficients in the principal 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 Fig. 7). 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

**Table 1.** Journals sampled for our dataset along with the abbreviations used throughout the experimental results section and the number of articles in each class

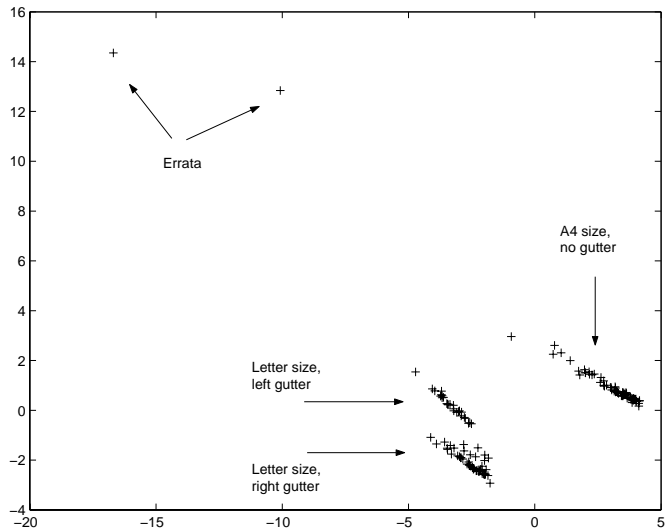
Journal	Abbrev	# in class
<i>Int J Netw Management</i>	IJNM	132
<i>Journal of the ACM</i>	JACM	147
<i>Standard View</i>	STDV	109
<i>Topology</i>	TOPO	149

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 subgenres within each genre (see Fig. 11). However, this does give us a nonsubjective 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 Sect. 3, Eq. 2 was then computed for each image. Each quadrant of the size distribution is then sampled to form a rectangular size distribution of size  $41 \times 61$ . The resulting dimensionality of our feature space is 5002.

### 5.1 Genre classification

Table 2 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 five, seven, and ten principal components computed from the training sample and



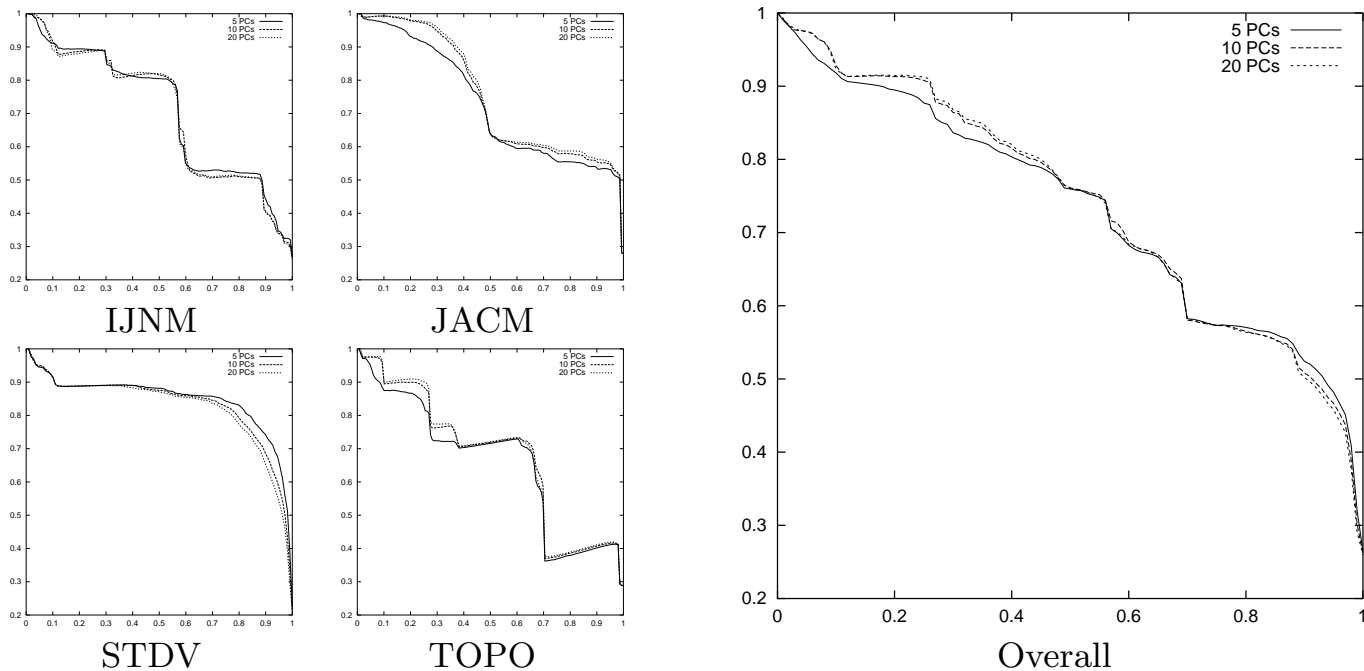
**Fig. 11.** The first two principal components for two document classes

for a 1-nearest neighbor, quadratic discriminant, and linear discriminant classifier. These results indicate that, even with relatively few principal components, rectangular granulometries are capable of capturing the relevant differences between document genres.

### 5.2 Document image retrieval

For our 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 to each document in the database.





**Fig. 12.** Average precision and recall plots for each genre in the test database. Results on the entire feature space and with 5, 10, and 20 principal 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

**Table 2.** Genre classification results for 30 training samples per class and various numbers of principal components. Classification accuracy is estimated by averaging over 50 experimental trials. The PCA is performed independently for each trial

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

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). Note that this definition of relevance does not take into account the existence of visually distinct subclasses within a single publication.

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, the number nonrelevant documents that must be inspected before finding that fraction of relevant documents.

Figure 12 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 Fig. 12 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 13 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 subclasses of the document genre.

Another interesting phenomenon in graphs of Fig. 12 is the presence of the nonmonotonic regions in the TOPO class. This occurs only when there is a dramatic change in the recall rate. In this case, the TOPO class contains articles with styles similar to those in the other classes. It is common in the ranked retrieval lists to encounter ranges of nonrelevant documents from another class fol-

	IJNM	JACM	STDV	TOPO
<b>Query</b>				
<b>High Rank</b>				
<b>Low Rank</b>				

**Fig. 13.** Some illustrative query examples. A sample query image for each genre is shown along with the highest-ranked and lowest-ranked relevant images from the relevant genre. In most cases, the least relevant document is pathologically different from the query

lowed by a range of TOPO documents. Essentially, the TOPO class does not have as much internal layout consistency as the other classes.

## 6 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 with morphological decomposition techniques they can be efficiently computed. Experiments have also shown that rectangular size distributions can be used to discriminate between specific document genres. Furthermore, principal 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%.

Principal component analysis has proved useful for accentuating the important features in size distributions. A nonlinear PCA approach that maximizes interclass variance while minimizing intraclass 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 too 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. Re-

search is currently focused on feature selection strategies that also (re-)introduce spatial information into the size distribution representation.

The effects of noise on rectangular granulometries remains an important open question. All of the document images considered in our experiments are clean images, generated directly from a PDF source. While the feature reduction techniques discussed in Sect. 4.3 will compensate for noise to a certain degree, it is unknown what precise effect noise will have on the resulting size distributions. A systematic theoretical and experimental investigation is still needed.

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 multiscale 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. Antonacopoulos A (1998) Page segmentation using the description of the background. *Comput Vision Image Understand* 70(3):350–369
2. Batman S, Dougherty ER, Sand F (2000) Heterogeneous morphological granulometries. *Patt Recog* 33:1047–1057
3. Breuel T (2000) Layout analysis by exploring the space of segmentation parameters. In: *Proceedings of the 4th international workshop on document analysis systems (DAS'2000)*, Rio de Janeiro, 10–13 December 2000
4. Chandler D (2001) *Semiotics: the basics*. Routledge, London
5. Doermann DS (1998) The indexing and retrieval of document images: a survey. *Comput Vision Image Understand* 70(3):287–298

6. Dougherty ER, Pelz J, Sand F, Lent A (1992) Morphological image segmentation by local granulometric size distributions. *J Electron Imag* 1:46–60
7. Haralick RM, Katz PL, Dougherty ER (1995) Model-based morphology: the opening spectrum. *Graph Models Image Process* 57(1):1–12
8. Maragos P (1989) Pattern spectrum and multiscale shape representation. *IEEE Trans Patt Analysis Mach Intell* 11:701–716
9. Matheron G (1975) *Random sets and integral geometry*. Wiley, New York
10. Serra J (1982) *Image analysis and mathematical morphology*. Academic, New York
11. Shin CK, Doermann DS (2000) Classification of document page images based on visual similarity of layout structures. In: *Proceedings of SPIE Document Recognition and Retrieval VII*, San Jose, 26–27 January 2000, pp 182–190
12. Vincent L (2000) Granulometries and opening trees. *Fundamenta Informatica* 41(1–2):57–90



and functional programming languages.

**Andrew Bagdanov** received his B.S. and M.S. in mathematics and computer science from the University of Nevada, Las Vegas, where he was a member of the Information Science Research Institute. He is currently a Ph.D. student in computer science at the University of Amsterdam, working in the field of multimedia information analysis. His research interests include document understanding, pattern recognition, image processing,



and information space interaction, conducted in close cooperation with industry. In 1998, he was a visiting research fellow at the University of California, San Diego. He has published over 50 scientific papers and serves on the program committee of several international conferences.

**Marcel Worrington** received his master's degree (honors) and doctoral degree, both in computer science, from, respectively, the Free University Amsterdam ('88) and the University of Amsterdam ('93), The Netherlands. He is currently an associate professor at the University of Amsterdam. His interests are in multimedia information analysis and systems. He leads several multidisciplinary projects covering knowledge engineering, pattern recognition, image and video analysis,