Statistical structuring of pictorial databases for content-based image retrieval systems

Thierry Pun and David Squire[1]

Computer Science Department, University of Geneva
24, rue du Général Dufour, CH - 1211 Geneva 4, Switzerland
E-mail: pun@cui.unige.ch

Abstract

This letter presents a two-stage statistical approach for "exploring and explaining" a pictorial database, for content-based image retrieval systems. First, we describe how correspondence analysis provides image classes, as well as facilitates the understanding of the role of image primitives and attributes used to index pictures. Such understanding allows an intelligent choice of features, and thus computational savings, to be made. Second, ascendent hierarchical classification permits the structuring of the database, in order to ease picture indexing and retrieval.

Keywords

Image databases, content-based image retrieval systems, exploratory statistics, correspondence analysis, ascendant hierarchical classification.

Running head

Statistical structuring of pictorial databases.

Corresponding author:
Thierry Pun, address above (phone: +41 (22) 705 7627, fax: +41 (22) 320 29 27).

1. This research is supported by the Swiss National Fund for Scientific Research, grant 20.40239.94. Part of this work was realized while the first author was on sabbatical leave at the Computer Science Department, University of Melbourne, Australia, which is gratefully acknowledged. Preliminary work was carried out by Catherine De Garrini. We thank all the contributors to our image database, amongst others INRIA-France, ETHZ-Switzerland.

1 Introduction

The rapid development of multimedia information technologies has markedly fostered research on Content-Based Image Retrieval Systems (CBIRS). In what follows, “Content” refers to pictorial attributes that can be extracted by image analysis methods. CBIRS allow exploratory “data mining” in large pictorial databases: search and retrieval of images whose attributes satisfy specific criteria (e.g. Gudivada and Raghavan, 1995). Typical image attributes are the dominant colors of the images (e.g. Gong and Sakauchi, 1995), their textural components (e.g. Pentland *et al*, 1994), or the shape of the objects contained in the pictures (e.g. Kato, 1992, Flickner *et al*, 1995).

CBIRS should offer tools for the construction and precompilation of an *index* to the pictorial database, for the *indexing* (search) and *retrieval* operations to answer specific queries, for the *browsing* and selection amongst answers, and for the *refinement* of the search. Human interaction in such operations is fundamental: for classifying the images in order to structure the database, for selecting the most descriptive image attributes, for formulating and refining queries, and for browsing amongst responses to these queries. Simple approaches to the above problems are reaching their limits in terms of descriptive power, retrieval accuracy, and ability to cope with large datasets. There is need for more sophisticated methods based, for example, on computer vision techniques (e.g. Gudivada and Raghavan, 1995; Pun and Milanese, 1995). In this context, exploratory statistics offer alternative approaches, well suited to the handling of large datasets as well as to human interaction.

We concentrate here on the index creation phase. Given a training set composed of a very large number of images, the problem is to obtain a hierarchical index whose nodes point towards the various image classes, whose links express inclusion relationships, and whose leaves are the individual images. This index will also be used as a decision tree for retrieving images from the database. This is basically a learning problem for which families of solutions exist in computer vision and other domains (e.g. Diday and Lechevallier, 1991; Chen *et al*, 1993; Bhanu and Poggio, 1994). In the present case, the training set will be composed of a large collection of images, of various types. It seems therefore unrealistic to base the index creation on any sophisticated structural object recognition procedure, likely to fail in the absence of more precise assumptions. Also, it is known to be difficult to extract hierarchically-structured classes using connectionist methods, and that explanations for classifications in terms of the original feature set are similarly difficult to obtain.

We argue here that exploratory statistics (e.g. Jambu 1991) provide paradigms and methods that can greatly ease the major operations involved in CBIRS. Methods from exploratory statistics allow the dimensionality of the classification problem to be reduced, permitting hierarchical structuring of large datasets. In addition, these methods provide a list of pertinent image features and attributes that led to the classification obtained. Due to the nature of the data (images of unconstrained content) and to the constraint of designing CBIRS with acceptable response times, features and attributes used here are global (e.g. luminance and chrominance, edge statistics, statistics on regions, etc.) rather than local or structural.

2 Exploratory statistics

2.1 Overview

Exploratory statistics (e.g. Jambu 1991) offers a collection of methods aimed at better understanding of large datasets, i.e. at explaining the underlying structure of the dataset. We describe below *correspondence analysis* (Benzécri 1973) for helping “explore and explain” the database, and *ascendant hierarchical classification* (e.g. Lebart *et al*, 1979) for providing a classification.

Correspondence analysis (CA) belongs to the family of factor analysis methods; as such, it provides a synthetic representation in a low dimension *factor space* of large sets of numerical data. These data constitute a cloud of points in a feature space. Factor analysis methods rely on finding a new, ordered, orthogonal set of axes, the *factor axes*, so that the sum of the norms of the projections of the data onto the axes are maximized, for each axis in turn.

Perhaps the best-known factor analysis method is Principal Component Analysis (PCA). The term PCA is frequently used without fully-defining the method by which the coordinates of the points in the original cloud were determined. In fact, the specification of these coordinates, and thus the metric used for calculating the distance between data points, is crucial in determining just what sort of factor analysis is performed. PCA is often performed using the covariance of the mean-centered data as the metric. The analysis thus provides factors that characterize the *variance* of the original data. If the raw data for each feature is also normalized by its empirical standard deviation as well as being mean-centered, then the *correlation* between data points becomes the metric for the factor analysis.

In CA, the coordinates of the data points are defined so that the usual Euclidean metric in the original feature space corresponds to the χ^2 distance between the points, and thus the analysis is in terms of the *independence* of the data. It will be seen below that CA offers other advantages also. It is often useful to adopt the terminology of PCA, and to describe factor axes as explaining a certain percentage of the variance of the original cloud of points. Here this should be understood to refer to the distribution of the points according to the χ^2 metric, rather than the covariance of the raw data as used by PCA.

CA has as starting point a numerical data table $[g_{ij}]$; in our case, rows correspond to images $i \in [1\ldots N]$, columns to attributes $j \in [1\ldots M]$ with usually $M \ll N$, and cells to the measure g_{ij} of a given attribute for a particular image. A specificity of CA with respect to other factor analysis methods is that rows and columns play a similar role. An "object" can be either one of the N images (rows) described each by M values, or one of the M attributes (columns) described each by N values. As with any factor analysis method, CA allows one to project in the factor space of dimension $L \leq min\{M, N\}$ the N image "objects"; in addition, CA permits the simultaneous representation in the same factor space of the M attribute "objects". This projection into a common space allows one to determine which particular attribute is near a cluster of images, and thus to know which are the important parameters for a given class of images. In other words, this method not only provides classes, but also explains why these classes are obtained. CA is commonly used in data analysis, but not as frequently in image analysis (for an example of application of CA to biomedical image analysis, see Pun *et al*, 1988). CA can be performed on a representative subset of the image dataset, which plays the role of a *training set*. Other images in the dataset, or new images, can then be classified on the basis of the factors derived from the training set. As with any such technique, there is a trade-off between the computational cost of analyzing a large training set, and the reliability of the subsequent classification of new data.

Ascendant hierarchical classification (AHC) is a computationally simple method amongst the large family of clustering algorithms (e.g. Diday *et al*, 1981). After projecting the images into factor space, AHC allows iterative clustering of images that are the closest in this space. This process yields a binary decision tree reflecting the hierarchy of similarities between images from the training set. It is presented here because of its hierarchical nature, and its simplicity. Any other unsupervised clustering technique could be applied to the data in factor space.

2.2 Correspondence analysis

Correspondence analysis involves the following sequence of steps: (1) computation of a normalized observation table $[p]$ from the original data table $[g]$; (2) determination of its χ^2 matrix $[c]$; (3) computation of eigenvalues and eigenvectors of $[c]$, the latter defining the factor space; (4) representation of images and attributes in this space. The corresponding equations follow (e.g. Lebart *et al*, 1979).

The sum G of all elements in $[g_{ij}]$ is denoted $G = \sum_{i=1}^{N} \sum_{j=1}^{M} g_{ij}$. The marginal sums (vectors) are obtained by $g_{\_j} = \sum_{i=1}^{N} g_{ij}$ and $g_{i\_} = \sum_{j=1}^{M} g_{ij}$. Each element $[p_{ij}]$ of the normalized table is determined by:

$$p_{ij} = g_{ij} / \sqrt{g_{i\_} g_{\_j}} \ . \tag{1}$$

The factor space in which images and attributes are projected is spawned by the eigenvectors of the χ^2 matrix $[c] = [c_{kl}] = [p]^T [p]$, of rank $min\{M, N\}$. Each element of $[c]$ is given by:[1]

$$c_{kl} = \frac{1}{\sqrt{g_{\_k} g_{\_l}}} \cdot \sum_{i=1}^{N} \frac{g_{ik} g_{il}}{g_{i\_}} \tag{2}$$

The number of eigenvalues and eigenvectors of $[c]$ is $min(M, N)$, i.e. M in what follows since we assume $M \ll N$. Eigenvalues are denoted λ_a, with $a \in [1 \ldots M]$; the corresponding eigenvectors of dimension M are $e_{a,j}$, j denoting the coordinates. The largest eigenvalue is trivial ($\lambda_1 \equiv 1.0$). The following ratio indicates the percentage of the total variance of the system that is explained by each factor $a \in [2 \ldots M]$:

$$\lambda_a / \sum_{a=2}^{M} \lambda_a . \tag{3}$$

In the factor space, the scalar coordinate of image $i \in [1 \ldots N]$ along factor axis a is:

$$I_{a,i} = \frac{\sqrt{G}}{g_{i\_}} \cdot \sum_{j=1}^{M} \left[\frac{g_{ij}}{\sqrt{g_{\_j}}} \cdot e_{a,j} \right] . \tag{4}$$

1. This is actually a simplified form, rather than the full χ^2 matrix. This form is less computationally costly, and takes advantage of the fact that its eigenvalues and eigenvectors are identical to those of the original matrix, except that the first becomes trivial. See Lebart et al, 1979, for a derivation.

Similarly, the scalar coordinate of attribute $j \in [1 \ldots M]$ along factor axis a is:

$$A_{a,j} = \frac{1}{g_{\_j}\sqrt{\lambda_a}} \cdot \sum_{i=1}^{N} [g_{ij} \cdot I_{a,i}] . \tag{5}$$

Equation (5) indicates that each attribute can be interpreted as the center of mass of all images, with proper weighting of their coordinates. Conversely, each image can be interpreted as the center of mass of the properly weighted attributes:

$$I_{a,i} = \frac{1}{g_{i\_}\sqrt{\lambda_a}} \cdot \sum_{j=1}^{M} [g_{ij} \cdot A_{a,j}] . \tag{6}$$

Equations (5) and (6) establish the correspondence between the projections in factor space of the images and their attributes. The respective position of images and attributes indicates which attributes characterize the best a given group of images or, conversely, which images are the most typical of given attributes. Another useful element of information is provided by the absolute contribution of image i to one given factor axis a (scalar):

$$AI_{a,i} = \frac{g_{i\_} \cdot I_{a,i}^2}{G \cdot \lambda_a} \tag{7}$$

This value specifies the amount by which each image i contributes to axis a. Images i with the highest $AI_{a,i}$ are the most significant ones to be taken into account for axis a. Similarly, the absolute contribution of an attribute j to one given factor axis a is:

$$AA_{a,j} = \frac{g_{\_j} \cdot A_{a,j}^2}{G \cdot \lambda_a} . \tag{8}$$

2.3 Ascendant hierarchical classification

The purpose of the ascendant hierarchical classification is to cluster the objects (here in factor space) into meaningful groups. If the objects are the images, this provides classes of similar images. In the factor space spawned by the L most significant factors (typically $L = 2 \ldots 4$), the Euclidean distance between two images is defined by:

$$d^2(i, i') = \sum_{a=1}^{L} [I_{a,i} - I_{a,i'}]^2 . \tag{9}$$

When using the distance defined by Equation (9), and due to Equation (1), each image i can be implicitly given a “mass” $g_{i\_}/G$. This ability to define such a simple metric in the parameter space allows the clustering to be based on the geometrical proximity of elements in this space. Ideally, the aggregation criterion should build a binary tree by minimizing the within-class centered second moment while maximizing the between class moment. After the grouping of two images i and i' or more generally, of two groups of images x and y, the added within-class moment is:

$$\nu(n) = \frac{F_x F_y}{F_x + F_y} \cdot \sum_{a=1}^{L} [I_{a,x} - I_{a,y}]^2 \ , \tag{10}$$

with

$$F_{x\ or\ y} = \frac{1}{G} \cdot \left[\sum_{i \in\ cluster\ x\ or\ y} g_{i\_} \right] , \tag{11}$$

and where the cluster index $n \in [1 \ldots 2N-1]$ numbers the nodes of the binary tree. $\nu(n) = 0$ for $n \in [1 \ldots N]$, since the leaves of the tree are the individual images; otherwise $\nu(n) > 0$ for $n \in [N+1, \ldots, 2N-1]$.

A suboptimal algorithm for building the classification tree, which guarantees minimum within-class moments, consists of: (1) initializing the symmetric array $[\nu_{x,y}]$ of size $N \times N$ with the intra-class variances of all possible pairs of images $\{x = i, y = i'\}$ (Equation (9)); (2) finding the entries $\{x, y\}$ for which $[\nu_{x,y}]$ is minimum, and aggregating the corresponding two gels or group of gels; (3) computing $\nu(n)$; (4) recomputing the array $[\nu_{x,y}]$, whose rank has diminished by 1; (5) going to (2). The resulting binary tree indicates which images are to be clustered in order to obtain a meaningful hierarchy.

3 Results and discussion

3.1 Images, features and attributes

Our image database currently contains several thousand B/W and color pictures (news photographs, objects, textures, biological specimen, etc.). Color is obviously a fundamental cue in discriminating images. However, for brevity, we concentrate below on the analysis of 54 grey-scale images from our database. The images are unconstrained as to their content (see Figure 5). The choice of representative features and attributes is conditioned by various factors, including their ease of computation as well as their statistical independence. The most basic feature used here is the *image intensity.* We have also retained simple geometric features, namely *line segments* and *circular arcs* approximating the image contours, as well as *regions* obtained through a classical region growing process. Details regarding the algorithms we used for features extraction can be found in e.g. Milanese *et al*, 1994; these details do not matter much in the following analysis. The above features lead to the following $M = 12$ attributes:

- average μ_G (attribute **1**) and standard deviation σ_G (attribute **2**) of the intensity;
- average μ_{SL} (attribute **3**) and standard deviation σ_{SL} (**4**) of the segment lengths;
- average μ_{SO} (attribute **5**) and standard deviation σ_{SO} (**6**) of the segment orientations;
- average μ_{AL} (attribute **7**) and standard deviation σ_{AL} (**8**) of the circular arc lengths;
- average μ_{AR} (attribute **9**) and standard deviation σ_{AR} (**10**) of the circular arc radii;
- average μ_{Reg} (attribute **11**) of the standard deviation of the region intensities (global measure of the homogeneity of the regions);
- standard deviation σ_{Reg} (attribute **12**) of the averages of the region intensities (global measure of the image homogeneity).

This choice of attributes provide translational, rotational and scale invariance in terms of the factor coordinates. Note that for color images, we replace attributes 1 and 2 by six attributes, namely the mean and standard-deviations of the R, G, B channels. It is difficult to a-priori select good representative features for a large image database. Our selection of attributes certainly involves a degree of arbitrariness, and could be straightforwardly modified. In fact, attribute 5 was found to be unsuitable, and was excluded from the following analysis. The purpose here, however, is not to propose the best possible choice. Our goal is rather to show and stress the role of exploratory methods to assess the pertinence and role of each feature and attribute from a given image set, and therefore to ease their selection.

3.2 Correspondence analysis

The first step when applying CA to a dataset is to determine the number of factors to be taken into account. Eigenvalues are used to this effect: each of them quantifies the amount of the total variance of the system explained by each corresponding factor (Equation (3)). In our case, the percentage explained by the 10 non-trivial eigenvalues is for λ_2 : 42.3%, λ_3 : 24.7%, λ_4 : 14.7%, λ_5 : 8.9%, λ_6 : 6.4%, λ_7 : 1.4%, λ_8 to λ_{11} : 1.5%. The first four factors therefore explain 90.6% of the total variance, and should be sufficient to analyze the influence of the various factors.

The data (images, attributes) can be projected into the factor space spawned by any pair of eigenvectors. The most significant are eigenvectors e_2 and e_3, respectively called first and second factor axis. The result is given in Figure 1. A first inspection readily shows the existence of distinct image clusters (e.g. cluster of images nr. 13, 23, 35, 38), of isolated elements (e.g. image 5), and of large clouds. Similarly, some attributes appear significant (e.g. attributes 1, 2, 4) while others seem correlated (e.g. attributes 6, 7, 8, 10). The correlation of these features is unsurprising, as σ_{SO}, μ_{AL}, σ_{AL} and σ_{AR} are are related to the distribution of curved features (arcs) in the images. The segment orientation will have a larger variance in the presence of curved features, and it is not surprising that the variances in arc length and arc radius are correlated.

INSERT Figure 1.

Equation (7) allows one to rank images according to their contribution to each factor axis. For the first factor axis e_2 (λ_2 = 42.3%), the most important images by decreasing order of influence are nr. 32(absolute contribution 9.39%), 47 (8.22%), 46 (7.60%), 45 (7.12%), 12 (4.64%), 42 (4.32%). These images are shown in Figure 2. For the second factor axis e_3 (λ_3 = 24.7%), the most important images are nr. 5 (11.35%), 13 (6.88%), 23 (6.38%), 48 (6.18%), 10 (5.86%), 32 (4.29%) (Figure 3).

INSERT Figure 2.

INSERT Figure 3.

In order to better understand the causes behind this distribution of images in factor space, it is necessary to know which are the attributes that contribute the most to each factor axis. Similarly to the case of images, Equation (8) allows attributes to be ranked according to their absolute contribution to each axis. For the first factor axis, the most important attributes by decreasing order of influence are nr. 1 = μ_G (absolute contribution 56.24%), 12 = σ_{Reg} (12.19%), 4 = σ_{SL} (8.21%) and 2 = σ_G (7.33%); these four attributes explain 83.97% of the axis. The dominant attribute in explaining the first factor axis is the average grey value of the image. This is apparent when looking at the images shown in Figure 2: of the images with the largest contributions, those with negative factor coordinates are very dark, while those with positive coordinates are light. The second and third most important attributes, σ_{Reg} and σ_{SL}, both characterize image homogeneity. This is again verified by examining the images: e.g. images 12 and 42 show greater variation in region intensities and segment lengths compared to 32, 47, 46 or 45. Finally, looking at the factor plane shown in Figure 1, one observes that images 45, 46, 47 and 32 are clustered together on one side of the factor plane, while images 12 and 42 are on the other. The distances in factor space between these images correspond well to the subjective differences between them.

Regarding the second factor axis, the most important attributes are nr. 2 = σ_G (41.21%), 9 = μ_{AR} (19.57%), 10 = σ_{AR} (12.02%) and 3 = μ_{SL} (8.03%); these four attributes explain 80.84% of the axis. The dominant attribute is the standard-deviation of the image intensity. This is confirmed by looking at Figure 3: images 10 and 48, with positive factor coordinates, are highly contrasted, with a clearly bimodal intensity histogram. Their variances are much larger than those of images 5, 13, 23 and 32, whose contrasts are low. The second and third most important attributes are the average and standard-deviation of the radii of the circular arcs approximating the image contours. In addition to this difference in terms of variance, images 10 and 48 show an evident difference in structural complexity compared to 5, 13, 23 and 32. These differing characteristics account for the clearly separated positions of these images in the factor plane.

Examination of Figure 1 provides additional insights regarding the choice of primitives and attributes. For example, attributes 6, 7, 8 and 10 (respectively σ_{SO}, μ_{AL}, σ_{Al} and σ_{AR}) are very close in the space spawned by the first two factors. If one decides to use only these factors for all subsequent analysis, several of these attributes could be dropped. In this way it is possible to select the features that really need evaluation, hence diminishing computational cost. As an additional example, it is interesting to observe that attributes 2 and 12 (σ_G and σ_{Reg}), that one could have expected to yield very similar results, are actually quite separated in factor space.

An advantage of correspondence analysis with respect to other factor analysis methods, is that the geometrical proximity of both images and attributes in the *same* factor plane can be exploited to aid in the interpretation (Equations (5) and (6)). This is exemplified by looking in Figure 1 at the proximity of attribute 2, and images 10 and 48 (top of figure). Correspondence analysis allows one to infer from this observation that images 10 and 48 are very representative of attribute 2, or, conversely, that attribute 2 is a key factor in positioning images 10 and 48 in factor plane. With other factor methods, the analysis must be accomplished separately for the rows or columns of $[g_{ij}]$.

3.3 Ascendant hierarchical classification

AHC, as defined in §2.3, is a sub-optimal algorithm for aggregating images in factor space. Therefore, the hierarchy of classes obtained solely depends on the choice of primitives and attributes. Despite however the simplicity of the attributes used here, the following experiment shows that meaningful classes can be obtained. Prior to applying the AHC algorithm, it is necessary to select the number L of factors that will be used in Equations (9), (10) and (11). Unlike the factor space analysis where at most three factors can be *graphically* depicted, no such limit applies here. It would therefore be interesting to compare results obtained with several values of L. For the sake of brevity, we concentrate here on results obtained with the four most important factors ($L = 4$), explaining 90.6% of the total variance. The binary tree obtained by AHC is shown in Figure 4. A brief analysis of the results obtained for other values of L will also be given.

INSERT Figure 4.

The results of the clustering are presented in Figure 5, where the 8 classes (32, 74, 5, 97, 99, 96, 54, and 103) corresponding to the level 3 (where the root node is level 0) of the tree are shown. These classes are ordered vertically by their coordinate on the first factor axis. Since these classes have widely varying intra-class variance, they contain differing numbers of images. For the two large classes, the images have been grouped to indicate how these classes are split further down the tree. Class 99 is split into classes 81 and 91 at level 4, and class 103 is split into classes 93, 95, 80 and 89 at level 5.

INSERT Figure 5.

The analysis of the contributions presented above indicated that average grey level and variance were the most significant attributes for the first and second factor respectively. This is confirmed by Figure 5, were the influence of the grey scale distribution plays a role. As one moves across the tree (down Figure 5) from class 32 to class 103, it is apparent that the images become lighter. The classes in the "middle" (99, 96, 54 and 93) contain images with high contrast, suggesting that the second factor axis was significant in determining these classes. The classes at the extremes of the first factor axis (32, 74, 80 and 89) are composed of pictures with narrower intensity distribution. Geometrical features also play an important role in the classification (e.g. σ_{Reg},

σ_{SL}, μ_{AR} and σ_{AR}). Classes 32, 74, 5, and 97 contain simple objects with large scale features, whereas classes 80 and 89 contain objects with fine detail and small features. In fact there is a trend from large scale to small scale features as one moves across the tree.

Some classes contain a mixture of objects without a clear *subjective* relationship (e.g. classes 93, 95); this is to be expected, not only since the attributes used are global, but also due to the fact that only the first four factors are used. Other classes, however, do establish subjectively meaningful relationships between pictures: class 74 contains only paperclips; class 99 interior scenes; class 91 contains all the stamps in the dataset; class 96 contains all the playing cards; classes 80 and 89 (children of class 98) contain mainly biological images.

3.3.1 Stability

The analysis described above was repeated using only the first two factor axes ($L = 2$), and also considering all the factors ($L = 10$). There is insufficient space in this letter to present a detailed analysis. When two factors were used classes at level three identical to classes 32, 74, 5 and 97 in Figure 5 were obtained. Images 49 and 50 were classified with those in class 96, and there were minor differences in the splitting of class 103. When all the factors were considered, the results were almost identical to those in Figure 5: class 74 remained unchanged, as did classes 99 and 103. Image 52 was moved to class 96, and image 5 to class 97. there were again minor differences in the splitting of class 103. These results indicate that CA provides factors that are robust to the clustering system. Computational savings can thus be made by using a reduced number of factors, with confidence that the classes obtained will not be radically altered.

The analysis was also carried out with features 6, 7 and 8 suppressed, as suggested by their proximity in factor space, discussed above. In this case, the classes obtained were identical to those in Figure 5, except that images 27, 28 and 41 were moved from class 96 to class 91. This is a convincing demonstration of the usefulness of the joint representation of images and features in the same graphical factor space representation provided by CA: these features were selected for exclusion after a quick perusal of Figure 1, rather than analysis of large tables of numbers. It also indicates the stability of the technique when number of features is reduced on the basis of dependence.

3.3.2 Subjectivity

It is important to note that any judgement on the “goodness” of this clustering is inherently subjective. Whether or not images are considered to be similar depends on the perceptions of the user of the system, and also on the task for which the system is to be used. A given user may be disappointed with a classification, such as the one presented here, in which intensity plays an important role, since they may be more interested in morphological features. Another user, perhaps choosing images for an advertising campaign, might be more interested in the colors, or “mood”, of an image than in the objects it contains.

Correspondence analysis can not solve this problem. It can, however, indicate to the user clearly which features are most responsible for the factor axes, allowing the user to exclude them if they are not desired. Moreover, the facility to project both images and features into the same plane allows the user to observe, in a simple, graphical manner, which images are correlated with which features. This might indicate that some images should be excluded from the correspondence analysis stage of the procedure, so that they do not unduly influence the factors, and then included only at the classification stage.

The subjective nature of the image clustering problem can not be solved independently of the tastes and needs of the user. This suggests that a fruitful area for future research would be a system that interacts with the user, allowing the user to review automatically-obtained classes, and to move images between classes. Such a system could implement an on-line learning scheme, so that it adapted to the preferences of its user. This could take the form of adapting the features used and the number of factors considered, but it could also involve a further transformation, from the factor space obtained from CA to a “user-space”, in which the metric reflects the past preferences of the specific user.

4 Conclusion

The purpose of this note is to suggest the use of well-established exploratory statistical methods for “exploring and explaining” a pictorial database, in the framework of content-based image retrieval systems. Correspondence analysis provides a drastic data reduction, which permits better understanding and explanation of the underlying relationships between the elements composing a given dataset, and thus an intelligent structuring of the database. In particular, the role of the diverse image features and attributes can be analyzed; this permits the insignificant ones to be eliminated. Together with an appropriate clustering method, such as ascendent hierarchical classification, we show how a method for building hierarchical classifications of large pictorial databases. These techniques have been shown to produce subjectively good clusters, and to be robust under changes in the number of factors considered. These statistical methods allow the structuring of such databases in a manner more appropriate for indexing and retrieval. Finally, the nature of exploratory statistics lends them well to an interactive usage; our view is that such methods should be utilized as tools for organizing image databases as a complement to other more classical pattern recognition approaches.

References

Benzécri, J.P. (1973). *L'analyse des Données; Tome II: l'Analyse des Correspondances.* Dunod, Paris.

Bhanu, B. and T. Poggio, Eds. (1994). Special Section: Learning in Computer Vision. *IEEE Trans. PAMI*, 16, 9, September.

Chen, C.H., L.F. Pau and P.S.P. Wang, Eds. (1993). *Handbook of Pattern Recognition and Computer Vision*. World Scientific, Singapore.

Diday, E., G. Govaert, Y. Lechevallier and J. Sidi (1981). Clustering in pattern recognition. In: J.-C. Simon, R. Haralick, Eds., Digital Image Processing, D. Reidel Publ. / Kluwer Inc., 19-58.

Diday, E., and Y. Lechevallier, Eds. (1991). *Symbolic-Numeric Data Analysis and Learning*. Nova Science Publish, New-York.

Flickner, M., H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker (1995). Query by image and video content: the QBIC system. *IEEE Computer*, September, 23-32.

Gong, Y. and M. Sakauchi (1995). Detection of regions matching specified chromatic features. *Comp. Vision and Image Understanding*, 61, 2, March, 263-269.

Gudivada, V.N. and V.V. Raghavan, Eds (1995). Special Issue: Finding the right image: Content-based image retrieval systems, *IEEE Computer*, September.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. Statistical Modeling and Decision Science Series, Academic Press.

Kato, T. (1992). Data-base architecture for context-based image retrieval. *Proc. SPIE,* 1662: Image storage and retrieval systems, 112-123.

Lebart, L., A. Morineau, and J.-P. Fénélon (1979). *Traitement des données statistiques.* Dunod

Méthodes et Programmes, Paris.

Milanese, R. H. Wechsler, S. Gil, J.-M. Bost and T. Pun (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In: *Proc. IEEE - Computer Vision and Pattern Recognition '94*, Seattle, Washington, June 20-23, 781-785.

Pentland, A., R.W. Picard, and S. Sclaroff (1994). Photobook: tools for content based manipulation of image databases. *MIT Media Lab. Report*, July 25, 1994.

Pun, T., D. Hochstrasser and C. Pellegrini (1988). Correspondence analysis and hierarchical classification of complex images: application to two-dimensional gel electrophoretograms. In: J.L. Lacoume, A. Chehikian, N. Martin and J. Malbos, Eds., *Signal Processing IV, Theories and Applications*, North-Holland.

Pun, T. and R. Milanese (1995). Computer vision and multimedia information systems. In: M. Sakauchi and R. Jain, Eds., *Proc. International Workshop on Multimedia Information Systems and Hypermedia*, Tokyo, Japan, March 23-24, 29-37.

Figure 1: Factor plane spawned by the first and second factor axes. Images are indicated by a dot (.) and numbered from 1 to 58; circles correspond to images which contribute the most to the factor axes. Attributes are indicated by a cross (+) surrounded by a grey square; they are numbered from 1 to 12 (with 5 excluded), corresponding respectively to μ_G, σ_G, μ_{SL}, σ_{SL}, σ_{SO}, μ_{AL}, σ_{AL}, μ_{AR}, σ_{AR}, μ_R, σ_R.

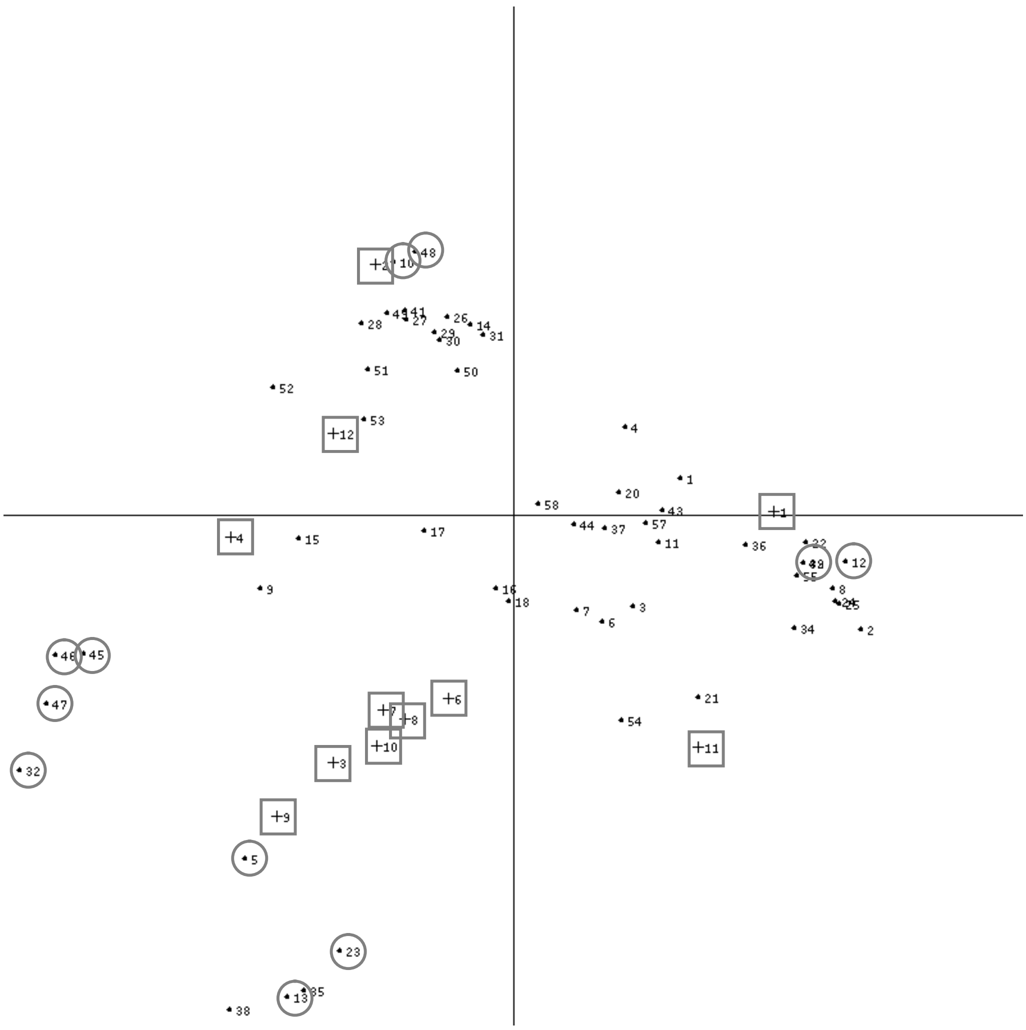
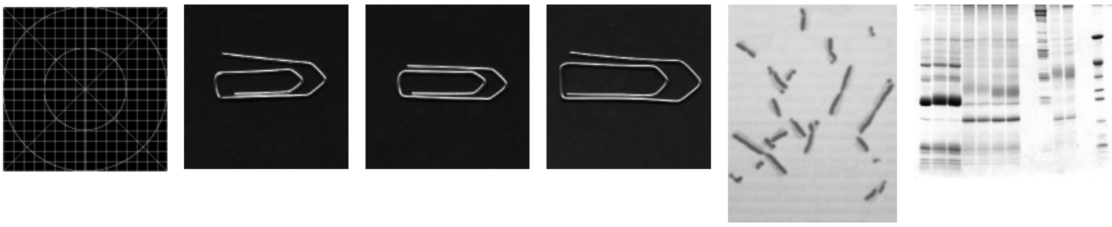


Figure 2: The 6 images that contribute the most to the first factor axis, ranked left to right by decreasing order of their absolute contribution. The plus (+) or minus (-) signs indicate whether the images had positive or negative coordinates on this axis. Image nr. 32: absolute contribution 9.39%; 47: 8.22%; 46: 7.60%; 45: 7.12%; 12: 4.64%; 42: 4.32%. Image numbers refer to Figure 1.



32 (-) 47 (-) 46 (-) 45 (-) 12 (+) 42 (+)

Figure 3: The 6 images that contribute the most to the second factor axis, ranked left to right by decreasing order of their absolute contribution. The plus (+) or minus (-) signs indicate whether the images had positive or negative coordinates on this axis. Image nr. 5: absolute contribution 11.35%; 13: 6.88%; 23: 6.38%; 48: 6.18%; 10: 5.87%; 32: 4.29%. Image numbers refer to Figure 1.

5 (-) 13 (-) 23 (-) 48 (+) 10 (+) 32 (-)

Figure 4: The binary classification tree obtained by ascendant hierarchical classification. The horizontal coordinate corresponds to the image coordinate on the first factor axis (Equation (4) with $a = 4$); the vertical coordinate is the logarithm of the added intra-class variance obtained when creating a new class (Equation (10)).

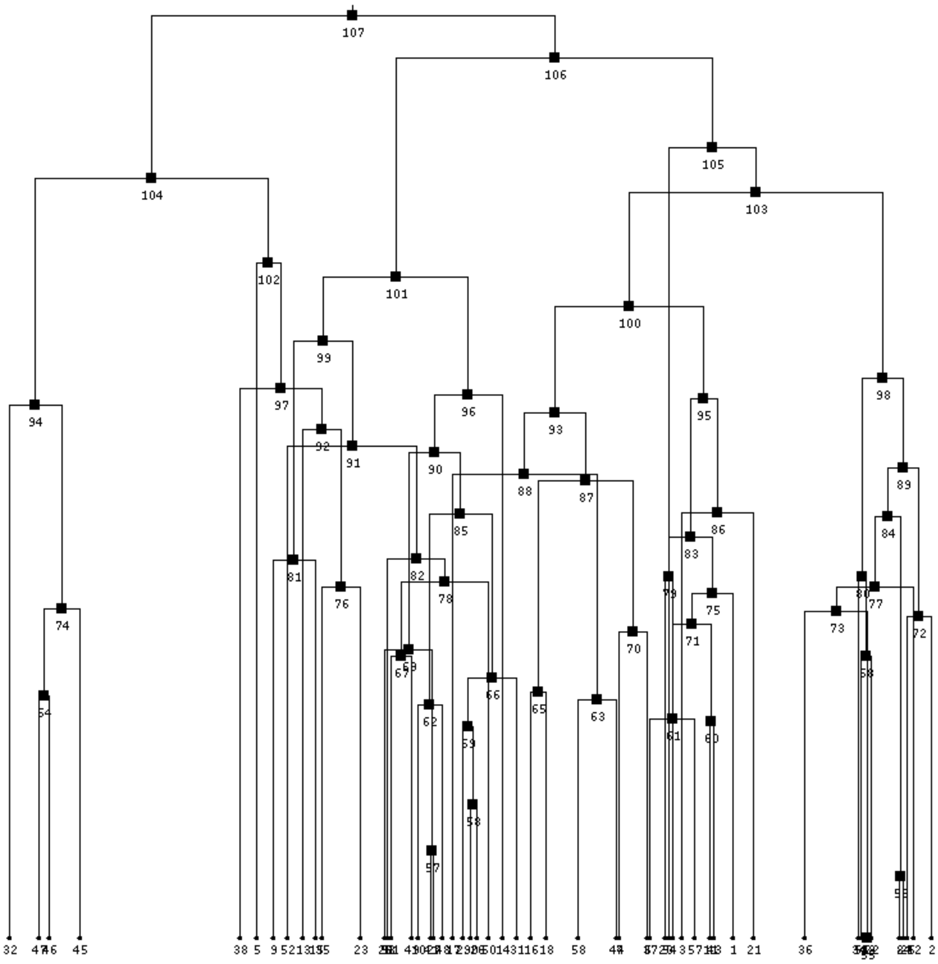


Figure 5: Classes at level 3. The class numbering refers to Figure 4, the image numbering (in parenthesis) refers to Figure 1. The vertical ordering of the classes corresponds to their coordinates on the first factor axis.