PERGAMON

Information Processing and Management 35 (1999) 605–632

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

A relevance feedback mechanism for content-based image retrieval

G. Ciocca, R. Schettini*

Istituto Tecnologie Informatiche Multimediali, Consiglio Nazionale delle Ricerche, Via Ampere 56, 20131 Milan, Italy

Received 22 October 1998; accepted 2 April 1999

Abstract

Content-based image retrieval systems require the development of relevance feedback mechanisms that allow the user to progressively refine the system's response to a query. In this paper a new relevance feedback mechanism is described which evaluates the feature distributions of the images judged relevant, or not relevant, by the user and dynamically updates both the similarity measure and the query in order to accurately represent the user's particular information needs. Experimental results demonstrate the effectiveness of this mechanism.

1. Introduction

Images, next to text, are the media most widely used to convey information. The low cost of scanning and storage devices encourages the creation of large image databases for use in a wide range of application domains, such as medicine, journalism, trademarks, fashion design, museology, etc. (Apers, Blanken & Houtsma, 1997; Jain, Pentland & Petkovic, 1995). As content-based image retrieval is highly desirable in many of these applications, it is not surprising that it has been such a popular topic of investigation in recent years, producing many papers describing a broad range of techniques (Aigrain, Zhang & Petkovic, 1996; Gudivada & Rahavan, 1997). In particular, techniques exploiting low-level visual features have become a promising research issue (Barolo, Gagliardi & Schettini, 1997; Del Bimbo & Pala, 1997; Della Ventura, Gagliardi & Schettini, 1998; Gagliardi & Schettini, 1997a,b; Gimel'Farb

* Corresponding author. Tel.: +39-2-706-43288; fax: +39-2-706-43292.
E-mail addresses: ciocca@itim.mi.cnr.it (G. Ciocca), centaura@itim.mi.cnr.it (R. Schettini)

PII: S0306-4573(99)00021-7

& Jain, 1996; Hafner, Sawhney, Esquitz, Flickner & Niblack, 1995; Mehtre, Kankanhalli & Lee, 1997, 1998; Picard & Minka, 1995). General-purpose systems of this type are also now available (Faloutsos et al., 1994; Smith & Chang, 1996). These systems usually make it possible to extract image representations in terms of color, texture, shape and layout features from the images and define the relative search/matching functions that can be used to retrieve those of interest. However, notwithstanding the substantial progress made, the integrated management of the different features remains complex and application dependent (Mehtre et al., 1998; Minka & Picard, 1997). In fact, several factors may intervene when choosing the aggregation operator to integrate the results of a query based on single features (Binaghi, Della Ventura, Rampini & Schettini, 1993; Binaghi, Gagliardi & Schettini, 1994): different tasks in the same context deal with similarity at different levels of precision; similarity depends greatly on the nature of the objects to which it is applied and on the features selected for their description; different users from different backgrounds may interpret image content differently, and the objective of their queries may also differ. All these factors, which are interrelated and consequently influence each other, make it quite impossible to determine in advance the most suitable aggregation operator for the different similarity measures.

We have developed a relevance feedback mechanism for automatically updating the similarity measure and the query that represents the information needs, by exploiting feedback from the user about the relevance (or non relevance) of the retrieved images. This mechanism is a part of a visual information retrieval system currently under development that indexes the images in terms of color, texture, shape, and spatial relationships among meaningful regions (Ciocca, Gagliardi & Schettini, 1998), is actually description-independent, that is, the index can be modified or extended to include other features without requiring any change in the algorithm.

Our approach differs from other recently presented studies (Mitra, Huang & Kumar, 1997; Ortega, Rui, Chakrabarti, Mehrotra & Huang, 1997; Rui, Huang, Mehrotra & Ortega, 1997; Sclaroff, Taycher & La Cascia, 1997; Taycher, La Cascia & Sclaroff, 1997) in both the strategy for learning the query representing the information needs and the similarity function. These differences are outlined in Section 2 describing the relevance feedback.

2. The relevance feedback mechanism

Users do not find it difficult to provide interactively examples of similar and dissimilar images. However, if the image database queried is large and heterogeneous, or the retrieval task particularly complex, the user may find not enough examples of images that are actually very similar to the query in the first screens and, to avoid the time-consuming visual browsing of the database, may mark as relevant images that are only partially similar. The user's information needs may also be rather vague, such as: find all the images containing people. In both cases images judged relevant may differ widely. Treating all these images in the same way, for example, averaging the features of the relevant images to compute a new query vector or updating the similarity measure, may consequently produce very poor results, while processing all the relevant images as single queries and then combining the retrieval outputs may create an unacceptable computational burden when the database is large. Last but not least, relevant images may have some features, color for example, that are only casually similar. If the system is not able to identify these features and treat them differently, subsequent retrieval iterations will be biased.

The key concept of the relevance feedback we propose is the statistical analysis of the feature distributions of the images the user has judged relevant, or not relevant, in order to understand what features have been taken into account (and to what extent) by the user in formulating this judgment, so that we can then accentuate the influence of these features in the overall evaluation of image similarity as well as in the formulation of a new query.

We assume here that an appropriate image representation is already available. We also assume that the images may have been divided into individually indexed sub-images to provide a more detailed spatial information. Fig. 2 shows the simple splitting strategy implemented in our system. The relevance feedback mechanism proposed is actually independent of the strategy applied to split the images and of the order in which these are evaluated.

The sub-vectors of features (the color histogram, for example) are indicated by $\mathbf{X}_{hs}$, where h is the index of the feature and s the index of the sub-image to which the feature refers. D_{hs} indicates the distance associated with the feature hth of region s.

In content-based retrieval the global metric used to evaluate the similarity between two images of the database is, in general, a linear combination of the distances between the individual features:

$$\mathrm{Sim}(\mathbf{X},\ \mathbf{Y}) = \sum_{s=1}^{q}\sum_{h=1}^{p} w_{hs} D_{hs}(\mathbf{X}_{hs},\ \mathbf{Y}_{hs}) \tag{1}$$

in which the w_{hs} are weights. There are two problems in this formulation of image similarity. First, since the single distances may be defined on intervals of widely varying values, they must be normalized over a common interval to place equal emphasis on every feature score. Second, the weights must often be heuristically set and this is often rather difficult for the user to do as there may be no clear relationship between the features used to index the image database and those considered in the user's subjective evaluation of image similarity. The solutions with these problems are described here below.

2.1. Distance normalization

Let us consider a database containing n images and the corresponding feature vectors $\mathbf{X}^k = \{\mathbf{X}^k_{hs},\ h = 1, \ldots, p,\ s = 1, \ldots, q\}$, with $k = 1, \ldots, n$.

The average distance between the sub-vectors of the database items is computed as follows:

$$\mu_{hs} = \frac{2}{n(n-1)} \sum_{i=1}^{n}\sum_{j-i+1}^{n} D_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) \tag{2}$$

The vector of the normalized distance between two images having indices i and j, respectively, is

$$D(\mathbf{X}^i,\ \mathbf{X}^j) = \left[\frac{D_{11}(\mathbf{X}^i_{11},\ \mathbf{X}^j_{11})}{\mu_{11}},\ \frac{D_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs})}{\mu_{hs}},\ \ldots,\ \frac{D_{pq}(\mathbf{X}^i_{pq},\ \mathbf{X}^j_{pq})}{\mu_{pq}} \right]^{\mathrm{T}} \tag{3}$$

The advantage of this type of normalization is that the averages are computed once when the database is indexed and, if the database is large enough, it is not necessary to recompute them when new items are added. Moreover the computational cost is low (Sclaroff et al., 1997). This normalization, however, does not guarantee that all the features' distances are defined on a common interval, but simply that half of the values will be in the range of [0, 1] and the other half in the range of [1, x], where x is a function of the maximum value of the set.

Another possibility would be to normalize the distances according to the smallest ($\min_{hs}$) and biggest ($\max_{hs}$) distances values among the $n(n-1)/2$ possible image pairs, as follows:

$$D(\mathbf{X}^i,\ \mathbf{X}^j) = \left[\frac{D_{11}(\mathbf{X}^i_{11},\ \mathbf{X}^j_{11}) = \min_{11}}{\max_{11} - \min_{11}},\ \frac{D_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) - \min_{hs}}{\max_{hs} - \min_{hs}},\ \ldots,\right.$$

$$\left.\frac{D_{pq}(\mathbf{X}^i_{pq},\ \mathbf{X}^j_{pq}) - \min_{pq}}{\max_{pq} - \min_{pq}} \right]^{\mathrm{T}} \tag{4}$$

This approach, however, may compress the feature distance values into a very small range if even a single abnormally large distance is present.

A better approach (Ortega et al., 1997) makes use of Guassian normalization as follows:

$$D(\mathbf{X}^i,\ \mathbf{X}^j) = \left[\frac{D_{11}(\mathbf{X}^i_{11},\ \mathbf{X}^j_{11}) - \mu_{11}}{K\sigma_{11}},\ \frac{D_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) - \mu_{hs}}{K\sigma_{hs}},\ \ldots,\ \frac{D_{pq}(\mathbf{X}^i_{pq},\ \mathbf{X}^j_{pq}) - \mu_{pq}}{K\sigma_{pq}} \right]^{\mathrm{T}} \tag{5}$$

where the standard deviation is

$$\sigma_{hs} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} [D_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) - \mu_{hs}]^2 \tag{6}$$

Assuming that the features' distance distributions have a Gaussian distribution, it can be shown that there is a 68% probability that the feature values will lie in the range of [−1, 1] if $K = 1$ and a 99% probability if $K = 3$ (Mood, Graybill & Boes, 1988). As we can not assume a-priori that the distances have Gaussian distributions, the following general relationship holds:

$$P\left[-1 \leq \frac{D_{hs} - \mu_{hs}}{K\sigma_{hs}} \leq 1 \right] \geq 1 - \frac{1}{K^2} \tag{7}$$

According to this relationship the probability that the distance lie in the range [−1, 1] is of 89% if we set K at 3 and 94% setting K at 4 (Mood et al., 1988). The latter is the default value used in all our experiments.

A simple additional shift moves the distances into the [0, 1] range,

$$d_{hs} = \frac{D_{hs} + 1}{2} \tag{8}$$

Out-of-range values are mapped to the extreme values, so that they do not bias further processing.

2.2. Similarity evaluation

At this point our similarity function has the following form:

$$\mathrm{Sim}(\mathbf{X}^i,\ \mathbf{X}^j) = \sum_{s=1}^{q}\sum_{h=1}^{p} w_{hs} d_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) \tag{9}$$

The algorithm must now determine the weights for the individual distances by a statistical analysis of the distance feature values of the images of the relevance-set and these weights are then used to accentuate, or diminish, the influence of a given feature in the overall evaluation of similarity (Sclaroff et al., 1997).

Letting m be the cardinality of the relevance set $\mathbf{R}^+$ and $\mathbf{d}^+_{hs} = \{d^+_{hs}(\mathbf{X}^1_{hs},\ \mathbf{X}^2_{hs}),\ \ldots,\ d^+_{hs}(\mathbf{X}^{m-1}_{hs},\ \mathbf{X}^m_{hs})\}$, the set of normalized distances among the elements of $\mathbf{R}^+$, in formula we have:

$$w_{hs} = \frac{1}{\varepsilon + \mu^+_{hs}} \qquad \mu^+_{hs} = \frac{1}{|\mathbf{d}^+_{hs}|}\sum_{i=1}^{n}\sum_{j=i+1}^{n} d^+_{hs}(\mathbf{X}^i_{hs},\ \mathbf{X}^j_{hs}) \quad h = 1,\ \ldots, p;\ \ s = 1,\ \ldots,\ q \tag{10}$$

where ε is a positive constant (set at 0.01 in our experiments). We need at least three images in the relevance set to update the weights (Eq. (10)); otherwise the weights' values are all set at $1/\varepsilon$.

In our experience in content based retrieval, relevant images are sometimes selected because they resemble the query image in just some pictorial features. Consequently, after an initial query, one retrieved image may be considered relevant because it is the same color as the query and a second be selected for its similarity in shape, although the two are quite different from each other. One solution to this problem could be to disregard the largest and the smallest elements of $\mathbf{d}^+_{hs}$ when the weights are computed.

Assuming that the user has marked some images as not-relevant, the following weight updating algorithm can be applied:

Let $\mathbf{R}^-$ be the set of not relevant images, and $\mathbf{d}^-_{hs}$, the set of normalized distances among the elements of $\mathbf{R}^-$. Consider the union between $\mathbf{R}^+$ and $\mathbf{R}^-$ and compute the corresponding distance sets $\mathbf{d}^{+-}_{hs}$, let $\mathbf{d}^*_{hs}$ be $\mathbf{d}^{+-}_{hs} \backslash \mathbf{d}^-_{hs}$. The weight terms are defined as:

$$w_{hs} = \begin{cases} \dfrac{1}{\varepsilon + \mu_{hs}^{+}} - \dfrac{1}{\varepsilon + \mu_{hs}^{*}} & \text{if } |\mathbf{R}^{+}| \geq 2 \text{ and } |\mathbf{R}^{-}| \geq 1 \text{ and } \dfrac{1}{\varepsilon + \mu_{hs}^{+}} \geq \dfrac{1}{\varepsilon + \mu_{hs}^{*}} \\ 0 & \text{if } |\mathbf{R}^{+}| \geq 2 \text{ and } |\mathbf{R}^{-}| \geq 1 \text{ and } \dfrac{1}{\varepsilon + \mu_{hs}^{+}} < \dfrac{1}{\varepsilon + \mu_{hs}^{*}} \\ \dfrac{1}{\varepsilon + \mu_{hs}^{+}} & \text{if } |\mathbf{R}^{+}| \geq 3 \text{ and } |\mathbf{R}^{-}| = 0 \\ \dfrac{1}{\varepsilon} & \text{otherwise} \end{cases} \tag{11}$$

When at least three positive examples and a negative example are available, we can take into account negative examples in tuning the similarity weights. When this is not possible, Eq. (11) is reduced to Eq. (10). For any given feature the first term is high when there is some form of agreement among the feature values of the relevant (positive) set, while the second term is high when there is a similarity between positive and negative examples. In fact, when positive and negative examples share some similar features, these must considered as non-discriminant, and their corresponding weight decreased.

2.3. Query processing

Query processing consists in modifying the feature vector of the query by taking into account the feature vectors of the images judged relevant by the user. One way of doing so is to take a weighted average of the query feature vector and of the relevant images as follows (Mitra et al., 1997):

$$(1-\beta)\mathbf{X}^{Q} + \beta \frac{1}{|\mathbf{R}^{+}|} \sum_{\mathbf{X}^{i} \in \mathbf{R}^{+}} \mathbf{X}^{i} \tag{12}$$

This presents two drawbacks: the user must set the β values heuristically and, again, the algorithm does not provide for the fact that relevant images may differ from the original query with respect to some features. Our approach is to let $\mathbf{R}^{+}$ be the set of relevant images the user has selected (including the original query) and proceed as follows:

$$\bar{\mathbf{Q}} = \frac{1}{|\mathbf{R}^{+}|} \sum_{\mathbf{X}^{i} \in \mathbf{R}^{+}} \mathbf{X}^{i} \qquad \bar{\sigma} = \sqrt{\frac{1}{|\mathbf{R}^{+}|} \sum_{\mathbf{X}^{i} \in \mathbf{R}^{+}} (\mathbf{X}^{i} - \bar{\mathbf{Q}})^{2}} \tag{13}$$

$$\mathbf{Y}_{hs}(j) = \{X_{hs}(j) \mid \left| X_{hs}(j) - \bar{Q}_{hs}(j) \right| \leq 3\bar{\sigma}_{hs}(j)\} \quad \forall h,\ s \text{ and } j \tag{14}$$

$$\tilde{\mathbf{Q}}_{hs}(j) = \frac{1}{\left|\mathbf{Y}_{hs}(j)\right|} \sum_{X_{hs}(j) \in \mathbf{Y}_{hs}(j)} X_{hs}(j) \tag{15}$$

The query processing formulates a new query $\tilde{\mathbf{Q}}_{hs}$ that can better represent the images of interest to the user, taking into account the features of the relevant images without allowing a single different feature value to bias query computation.

The query process could be similarly applied to compute a query representing non-relevant examples. This seems of little practical interest as non-relevant examples are usually not similar to each other and are, consequently, scattered in the feature space.

2.4. Query classification

At the first iteration, when the user has selected just one image to be searched, all the weights in the similarity function (Eq. (16)) are set at the value of $1/\varepsilon$. For faster tuning of the similarity function we can exploit previous query sessions performed by the user on the same database. To this end the user is allowed to register satisfactory queries together with the corresponding weights in the similarity measure. When the user has already formulated a query 'similar' to the new one, the algorithm sets the initial weights of the similarity function at the value of the former query reducing the time and effort needed to adapt the similarity measure by means of the relevance feedback algorithm.

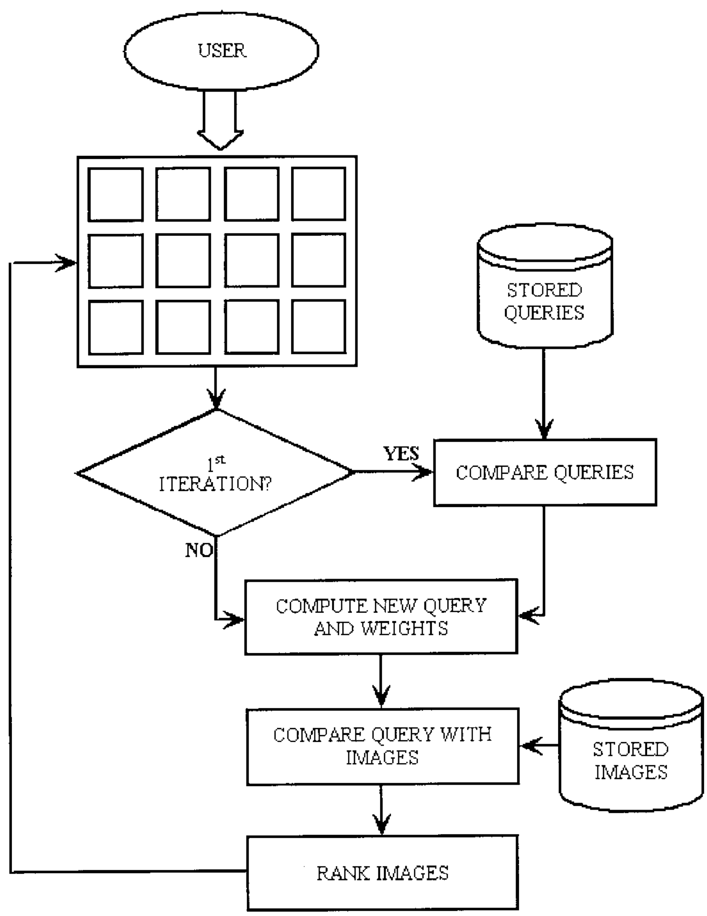

Fig. 1. The retrieval process.

We let $\hat{\mathbf{Q}} = \{\langle \tilde{\mathbf{Q}}^k,\ w_{hs}^k \rangle$ be the set of queries and the corresponding weights. When a new query $\mathbf{Q}^n$ is submitted, the system first evaluates its similarity with respect to all the 'old' ones, using the corresponding weights in the similarity function; it then selects the closest one as follows:

$$\langle \tilde{\mathbf{Q}}^k,\ w_{hs}^k \rangle = \arg\min_{\langle \tilde{Q}^k, w_{hs}^k \rangle \in \hat{Q}} \sum_{s=1}^{q} \sum_{h=1}^{p} w_{hs}^k d_{hs}(Q^n,\ Q^k) \tag{16}$$

The initial weights selected in Eq. (17) are now set as follows:

$$w_{hs}^{k(\text{initial})} = \begin{cases} w_{hs}^k & \text{if } \dfrac{\sum_{s=1}^{q} \sum_{h=1}^{p} w_{hs}^k d_{hs}(Q^n,\ \tilde{Q}^k}{(1/\varepsilon)pq} \leq T \quad 0 < T \leq 1 \\ \dfrac{1}{\varepsilon} & \text{otherwise} \end{cases}. \tag{17}$$

The T parameter allows the user to tune the sensitivity. In our implementation the default value is set at 0.1, that is, the new query is considered truly similar to an old one when the two differ less than the 10% of $(1/\varepsilon)pq$, the maximum distance value allowed.

2.5. Summary of the retrieval process

The retrieval process is summarized below and graphically depicted in Fig. 1.

1. The feature vector of the first query image is compared with the feature vectors of the stored queries and the initial weights are set (Eqs. (16) and (17)).
2. The query vector is compared with the stored images according to the similarity measure (Eq. (9)) and the top ranking ones are displayed.
3. The user may mark the retrieved documents as relevant or not relevant.
4. A new query vector is computed (Eqs.(13)–(15)).
5. The features weights in the similarity measure are updated (Eq. (11)).
6. The new query is submitted and the algorithm moves to step 2 and starts a new iteration of retrieval.

3. Test results and discussion

The relevance feedback mechanism described above has been implemented as part of a visual information retrieval system that is currently under development. A full description of all the functionalities of this system is beyond the scope of the paper. We should like to evaluate here only the improvements that the use of the proposed relevance feedback mechanism produces. However, as the absolute performances of an image retrieval system are strongly related to the nature and quality of the features used to represent the image content, for the sake of completeness the features used to index the images in the experiments reported are listed below:

- The color coherence vectors (CCV) in the CIELAB color space quantized in 64 colors (Pass, Zabih & Miller, 1996).
- A histogram of the transition in colors (CIELAB color space quantized in 11 colors) (Gagliardi & Schettini, 1997a,b).
- The moments of inertia of the distribution of colors in the unquantized CIELAB color space (Striker & Orengo, 1995).
- A histogram of contour directions opportunely filtered (only high gradient pixels are considered) and using box widths of 15 (Ciocca et al., 1998; Jian & Vailaya, 1996).
- The mean and variance of the absolute values of the coefficients of the sub-images of the first three levels of the multi-resolution wavelet transform of the luminance image (Scheunders, Livens, Van de Wouwer, Vautrot & Van Dyck, 1997).
- The neighborhood gray-tone difference matrix (NGTDM), i.e. coarseness, contrast, busyness, complexity and strength, as proposed by Amadasum and King (1989).
- The spatial composition of the color regions identified by the process of quantization in 11 colors (Ciocca et al., 1998).

In order to have more accurate spatial information the images are divided into sub-images and each of these is then indexed. There are many possible strategies for splitting images. Stricker and Dimai (1996), for example, split an image into a oval central region and four corners. Their system evaluates and combines color feature similarity for each of these sub-images, attributing more weight to the central region. However, this is a strictly domain-dependent solution which might not be acceptable in other applications. In our system the features are calculated on the global image and on 5 sub-images obtained by dividing the original image in the manner shown in Fig. 2. The image matching may be restricted to just some subimages if the user so desires.

Finally, all the features are compared with the L_1 distance measure, as it is statistically more robust than the L_2 distance measure (Rousseeuw & Leroy, 1987).

The archives on which the algorithms were tested contain: photographs of real scenes, such as landscapes, people and animals; reproductions of antique textiles of the Poldi Pezzoli Museum of Milan (Carrara, Della Ventura & Gagliardi, 1996); a collection of paintings of the Accademia Carrara of Bergamo; and a collection of ceramics belonging to the Museo Internazionale della Ceramica of Faenza, for a total of some 5000 images.

When a query is submitted the system reorders the database images by similarity with respect to the query and then returns to the user displays the most similar images. We consider

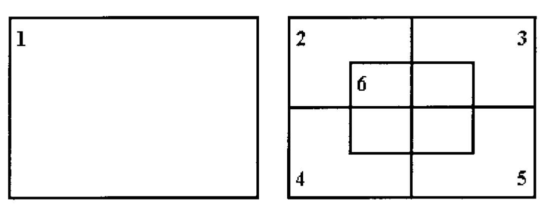

Fig. 2. Image regions.

an image that is truly similar to the query correctly retrieved if it appears within the first set of displayed images (short list). The user may iteratively refine the results by relevance feedback. Performance measures for text retrieval have been extensively studied (McGill & Salton, 1983) and some of these methods can be adapted to image content-based retrieval (Desai Narasimhalu & Kankunhalli, 1997). In order to quantify the improvement in performance obtained by applying the relevance feedback mechanism a measure called effectiveness (efficiency of retrieval or fill ratio), is applied here. This measure, proposed by Mehtre, Kankanhalli, Desai Narasimhalu and Man (1995), has also been applied recently in the comparison of shape similarity measures (Mehtre et al., 1998) and color similarity measures (Barolo et al., 1997) in content-based image retrieval.

We let S be the number of relevant items the user wanted to retrieve when posing a query (in our implementation S was set at 24, the number of images in the short list); $\mathbf{R}_q^{\mathrm{I}}$ is the set of relevant images and $\mathbf{R}_q^{\mathrm{E}}$, the set of images retrieved in the short list. The effectiveness measure is defined as:

$$\eta_S = \begin{cases} \dfrac{\left|\mathbf{R}_q^{\mathrm{I}} \cap \mathbf{R}_q^{\mathrm{E}}\right|}{\left|\mathbf{R}_q^{\mathrm{I}}\right|} & \text{if } \left|\mathbf{R}_q^{\mathrm{I}}\right| \leq S \\ \dfrac{\left|\mathbf{R}_q^{\mathrm{I}} \cap \mathbf{R}_q^{\mathrm{E}}\right|}{\left|\mathbf{R}_q^{\mathrm{E}}\right|} & \text{if } \left|\mathbf{R}_q^{\mathrm{I}}\right| > S \end{cases} \tag{18}$$

If $|\mathbf{R}_q^{\mathrm{I}}| \leq S$, the effectiveness is reduced to the traditional recall measure, while if $|\mathbf{R}_q^{\mathrm{I}}| > S$, the effectiveness corresponds to precision.

The effectiveness of the algorithms was tested on the single databases, and on a combination of them, to evaluate the system's capacity for adaptation.

In Table 1 we have summarized the experimental results for twenty queries for each of the different databases considered. No query classification has been performed: the first iteration corresponds to a similarity measure in which all the features have the same importance. Each row corresponds to a different database and reports the average effectiveness value at each of the first three retrieval iterations. The percentage of improvement in effectiveness obtained by relevance feedback is reported in parentheses below the means.

Relevance feedback improves the effectiveness of the retrieval considerably for all the databases and, in general, the second iteration (that is the first relevance feedback iteration) corresponds to the largest single improvement. We have observed, to the contrary, little benefit

Table 1
Retrieval effectiveness

Database	η_{T} (first iteration)	η_{T} (sec. iteration)	η_{T} (third iteration)
Photos (1745 images)	0.50	0.74 (+24%)	0.81 (+31%)
Paintings (1768 images)	0.52	0.64 (+12%)	0.82 (+30%)
Ceramics and ancient textiles (942 images)	0.56	0.69 (+13%)	0.77 (+21%)

(a)

Fig. 3. (a) Painting database: initial retrieval results. (b) Painting database: retrieval results after the first iteration of relevance feedback. (c) Painting database: retrieval results after the second iteration of relevance feedback.

(b)

Fig. 3 (*continued*)

(c)

Fig. 3 (*continued*)

(a)

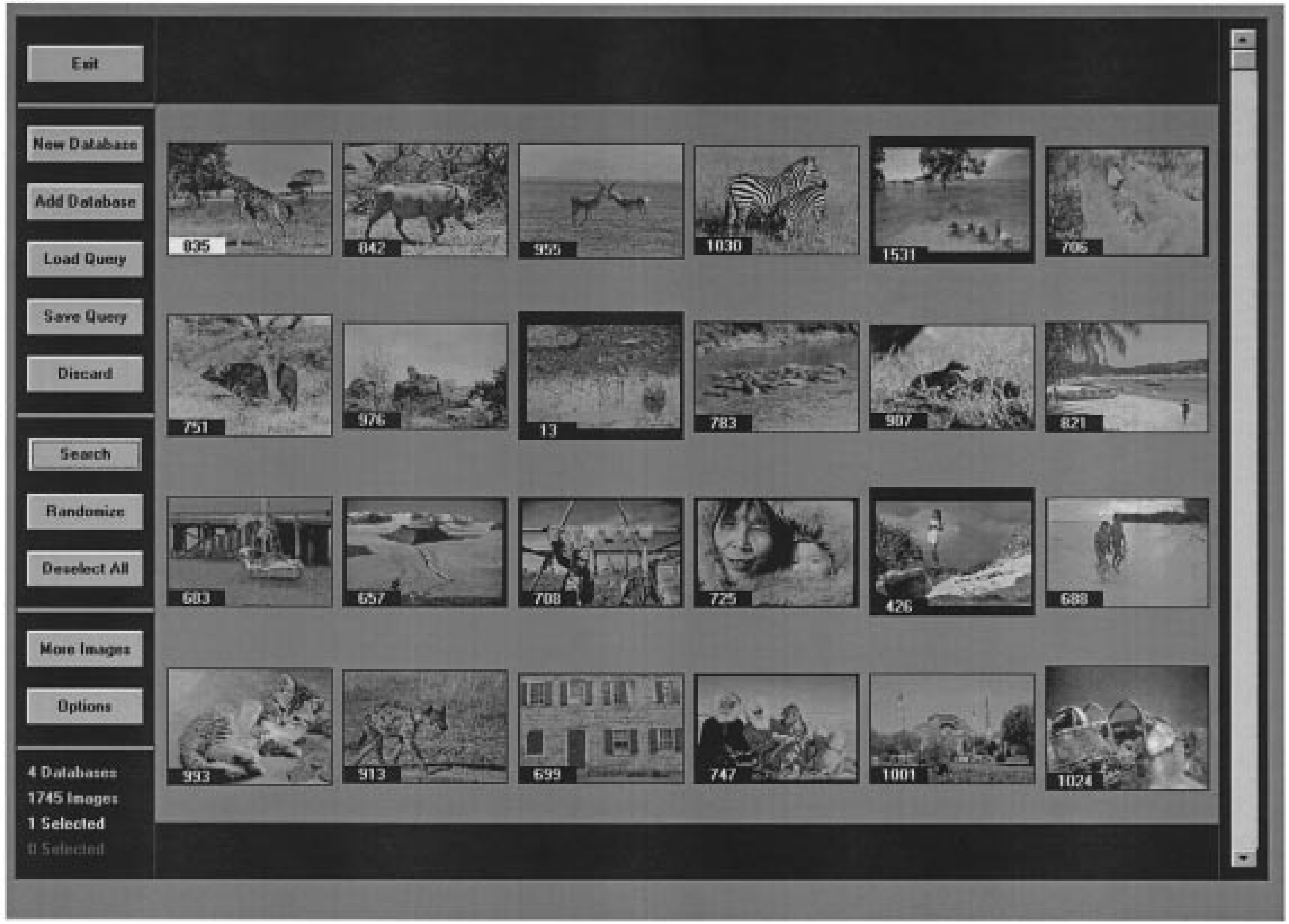

Fig. 4. (a) Photo database: initial retrieval results. (b) Photo database: retrieval results after the first iteration of relevance feedback. (c) Photo database: retrieval results after the second iteration of relevance feedback.

(b)

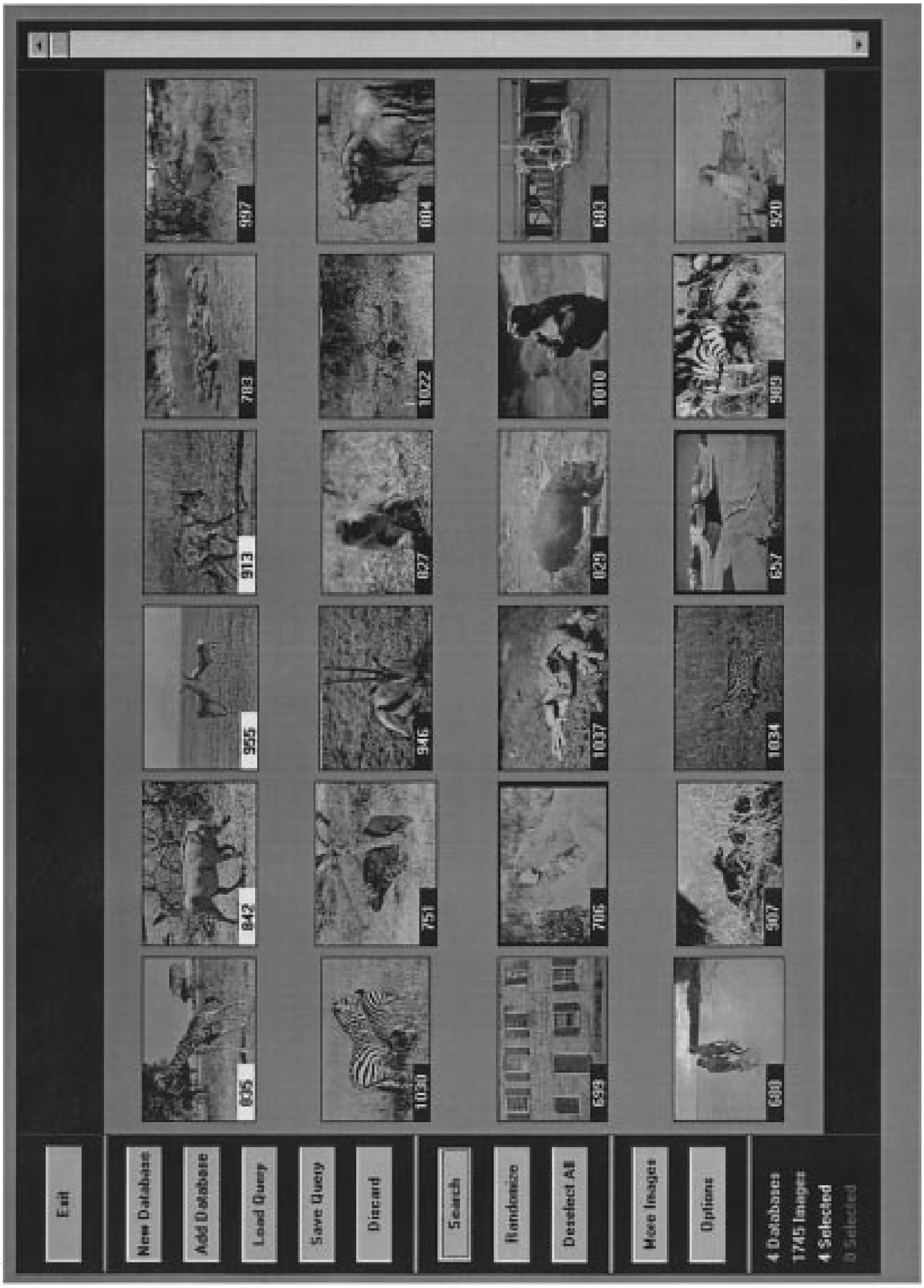

Fig. 4 (*continued*)

(c)

Fig. 4 (*continued*)

(a)

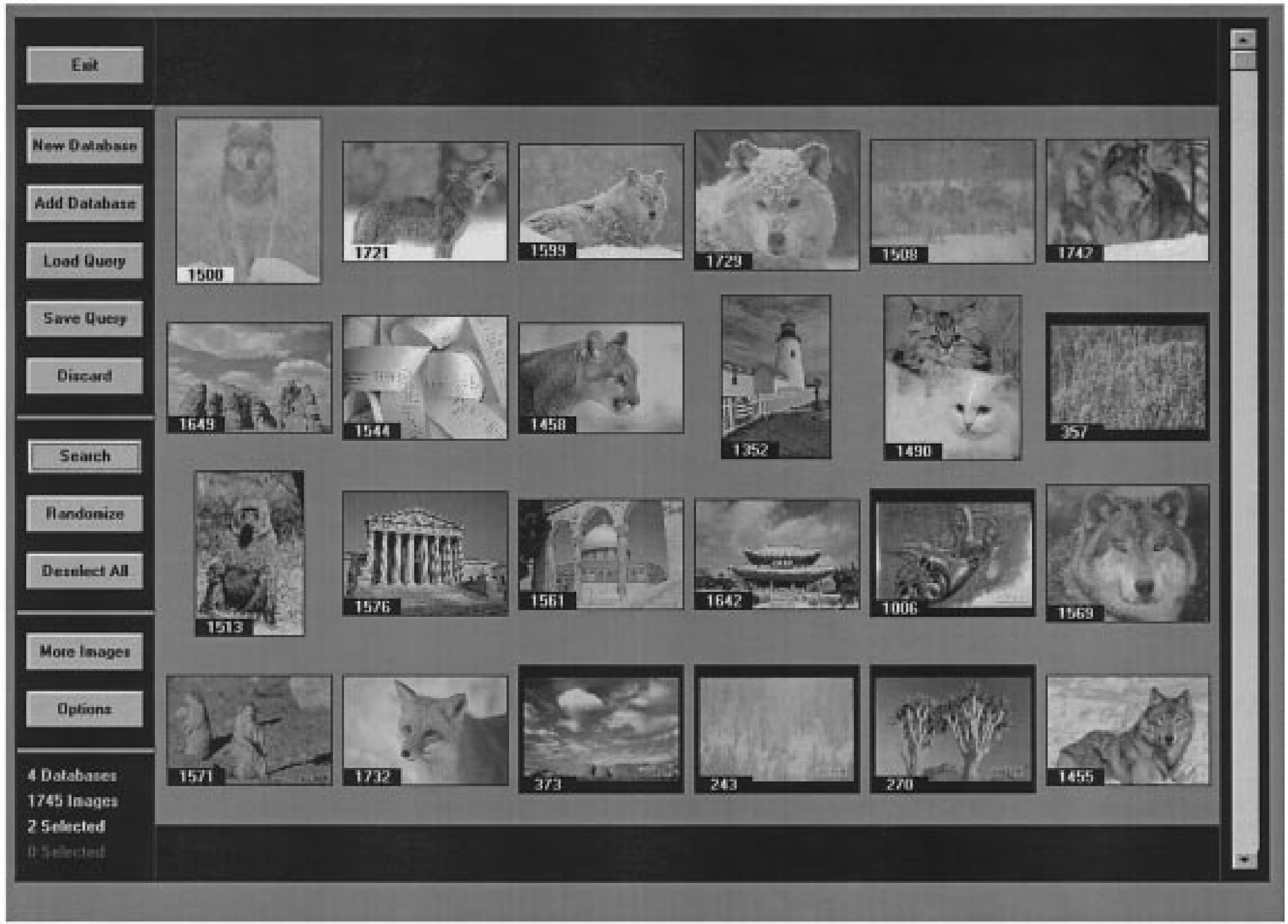

Fig. 5. (a) Photo database: initial retrieval results. (b) Photo database: retrieval results after the first iteration of relevance feedback. (c) Photo database: retrieval results after the second iteration of relevance feedback.

(b)

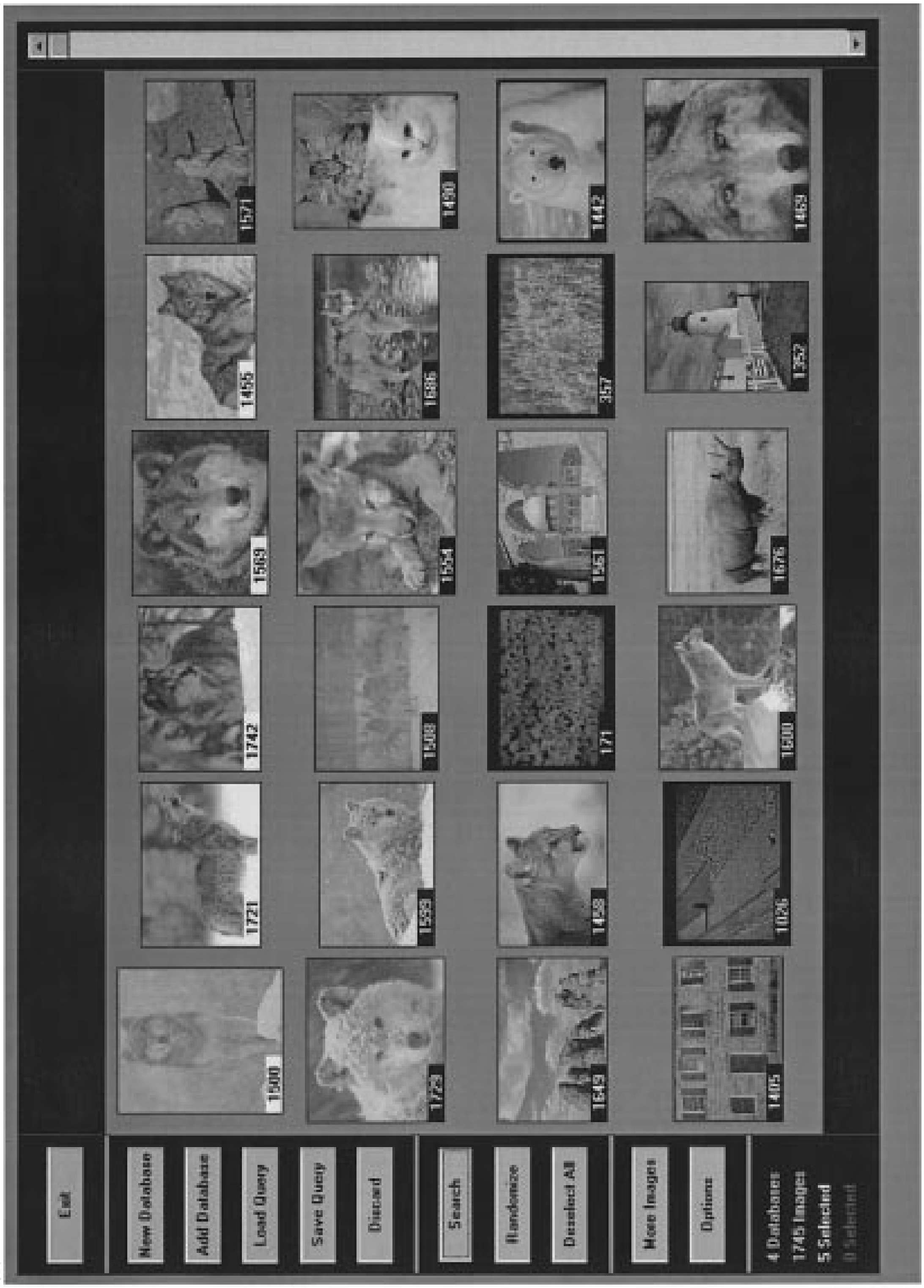

Fig. 5 (*continued*)

(c)

Fig. 5 (*continued*)

(a)

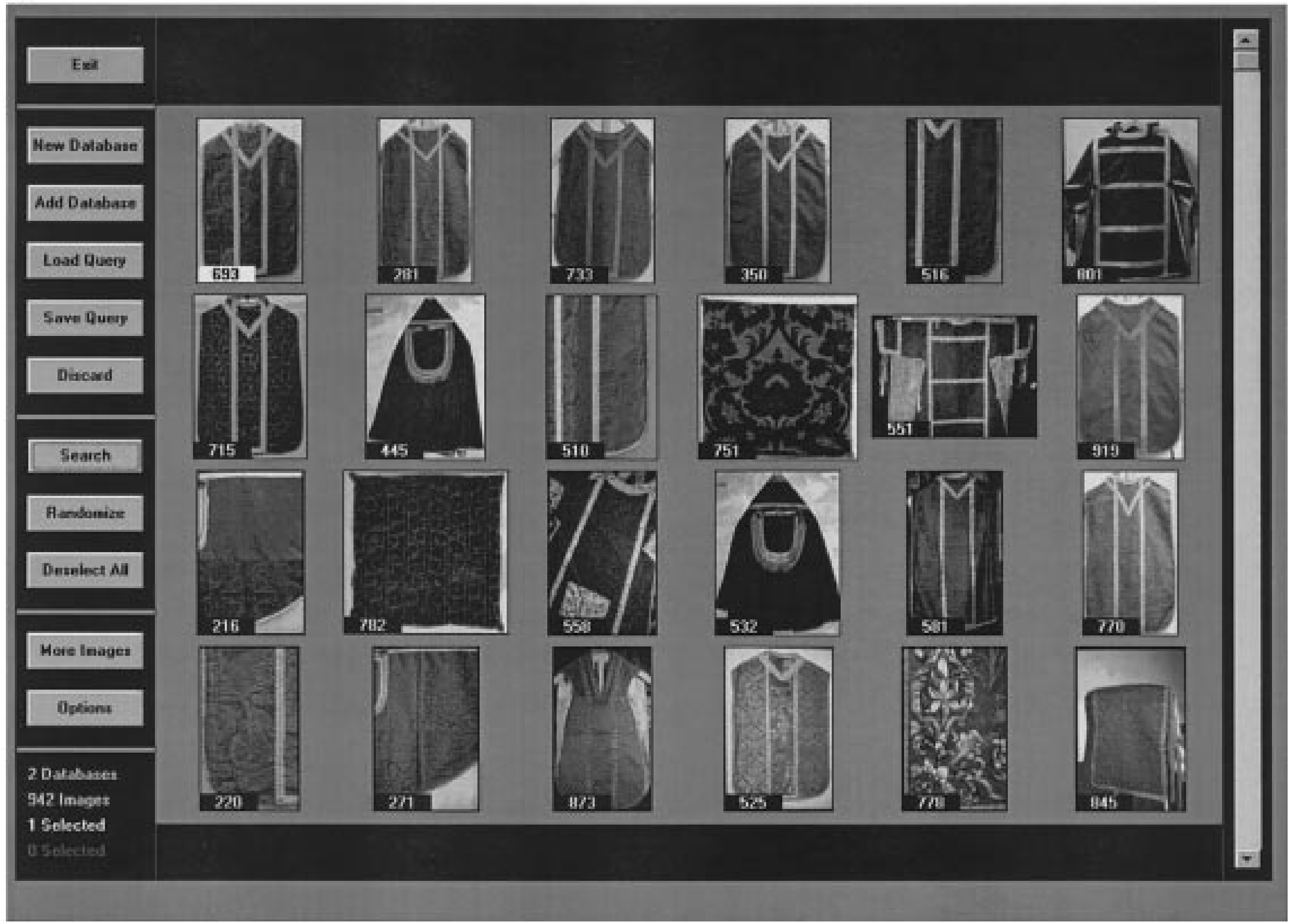

Fig. 6. (a) Ancient textiles database: initial retrieval results. (b) Ancient textiles: retrieval results after the first iteration of relevance feedback. (c) Ancient textiles: retrieval results after the second iteration of relevance feedback.

(b)

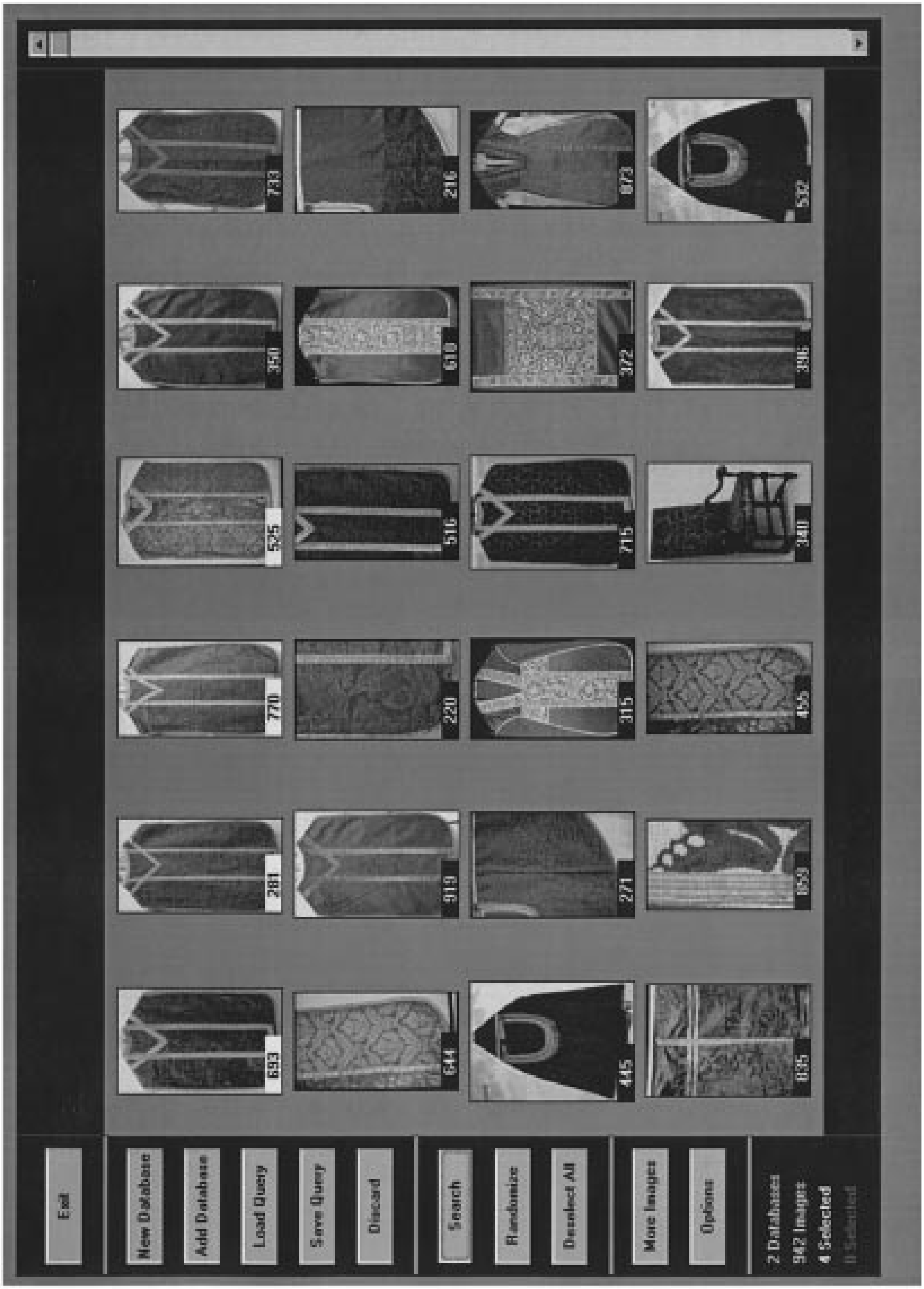

Fig. 6 (*continued*)

(c)

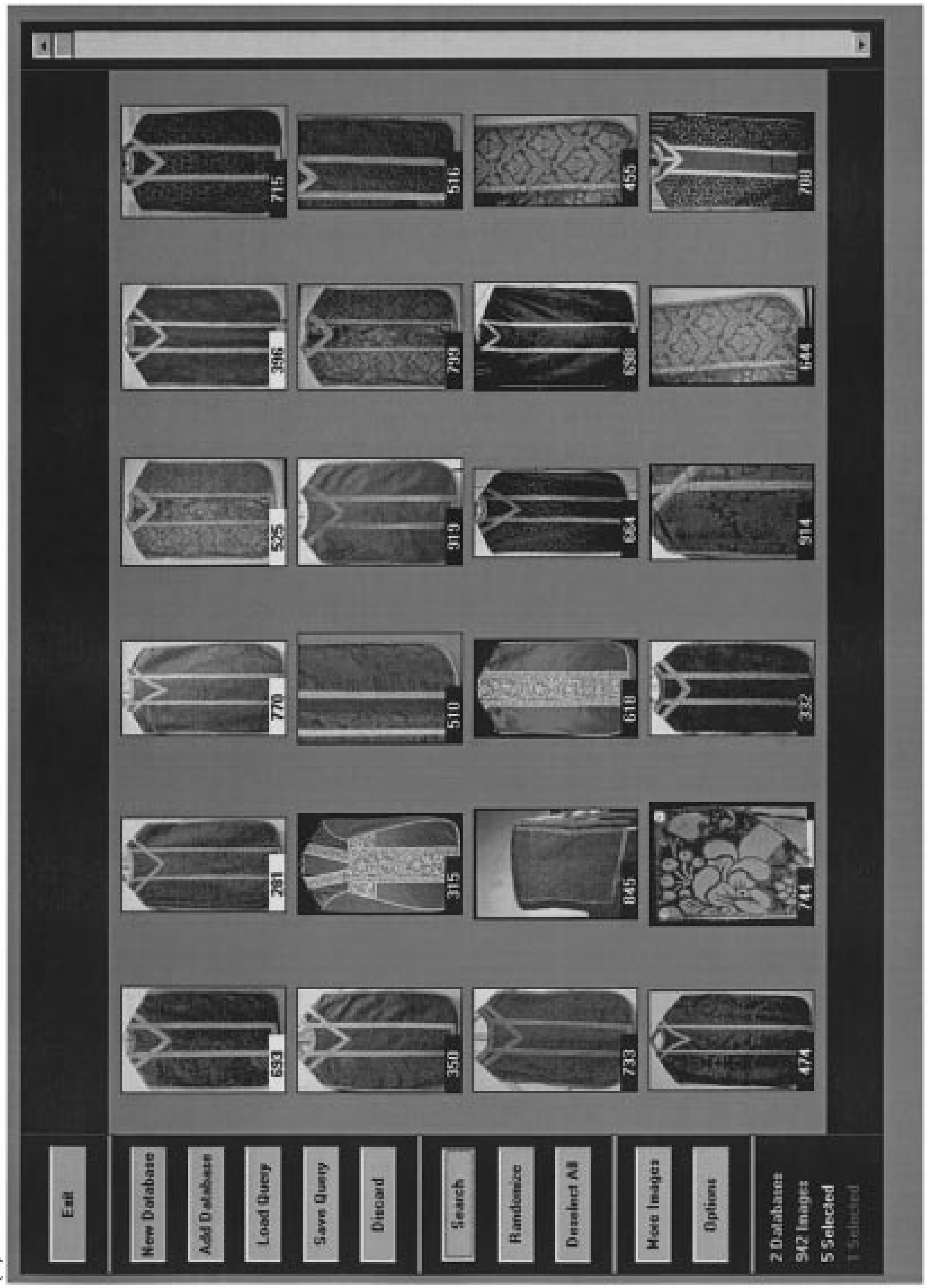

Fig. 6 (*continued*)

(a)

Fig. 7. (a) Ceramics: initial retrieval results. (b) Ceramics: retrieval results after the first iteration of relevance feedback. (c) Ceramics: retrieval results after the second iteration of relevance feedback.

(b)

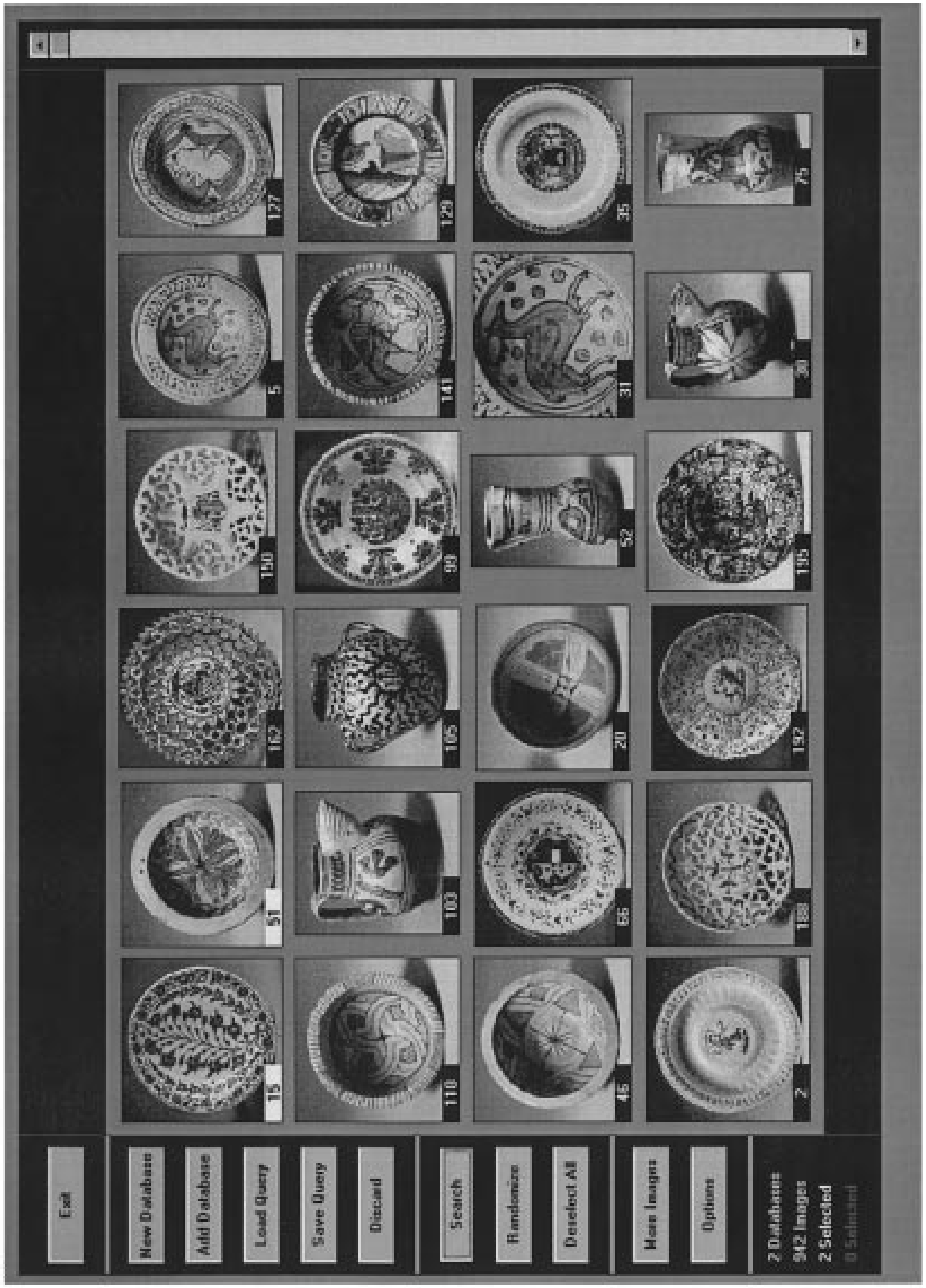

Fig. 7 (*continued*)

(c)

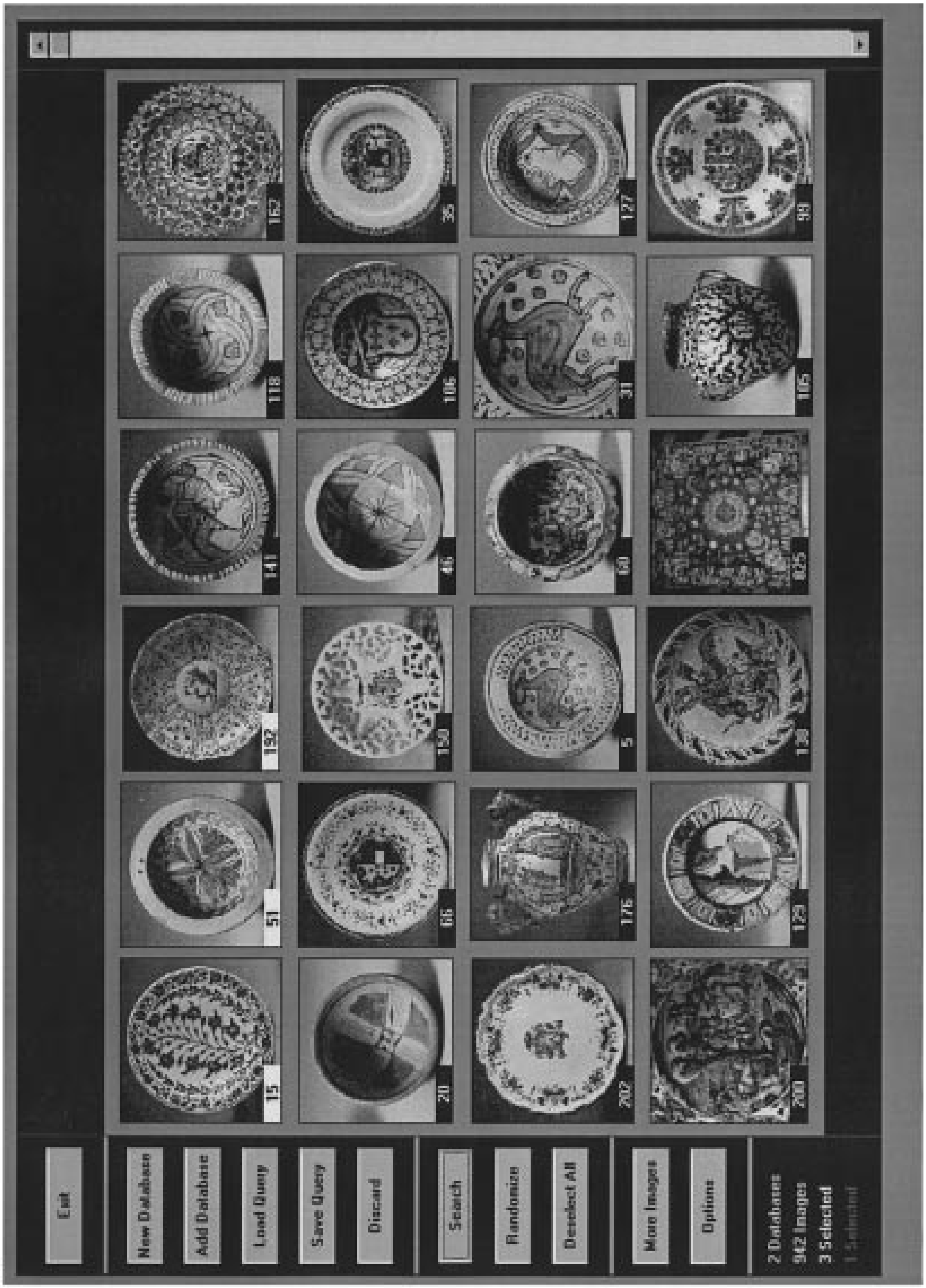

Fig. 7 (*continued*)

in repeating the procedure for more than five or six times. It can reasonably be argued that this is due to the limited capability of the low-level features used to exhaustively describe the image content, and not to the mechanism itself.

Figs. 3–7 present some examples of the system's application to explain its operation. Interested readers may find the corresponding color images at the following address: http://wwwtest.itim.mi.cnr.it/sitoitim/schettini/relfeme.htm.

4. Conclusions

The availability of high quality image compression algorithms has alleviated storage requirements and digital archives now often contain of thousands of images. Content-based retrieval systems operate on these collections to extract relevant images in response to a visual query. Unfortunately, the concept of 'relevance' is most likely to be associated with image semantics and encoding and exploiting semantic information in a general purpose retrieval system is still an unsolved issue. However, in many practical situations low-level image features are correlated with image semantic contents. The performance of an image retrieval system is, consequently, closely related to the nature and quality of the features used to represent image content, but it is not limited to this. Another important issue is that the measure adopted to quantify image similarity is user- and task-dependent (Minka & Picard, 1997) and this dependence is not in general understood well enough to permit careful, a priori selection of the optimal measure. In this paper we have described an mechanism that allows the user to query the database and progressively refine the system's response to the query by indicating the relevance, or irrelevance, of the retrieved items. Some experimental evidence has been provided to demonstrate that our mechanism makes it possible to approximate the user's information needs in a great variety of applications.

References

Aigrain, O., Zhang, H., & Petkovic, D. (1996). Content-based representation and retrieval of visual media: a state-of-the-art review. *Multimedia Tools and Applications*, *3*, 179 and 182.

Amadasun, M., & King, R. (1989). Textural features corresponding to textural properties. *IEEE Transaction on System, Man and Cybernetics*, *19*(5), 1264–1274.

Apers, P. M. G., Blanken, H. M., & Houtsma, M. A. W. (Eds.) (1997). *Multimedia database in perspective*. New York: Springer.

Barolo, B., Gagliardi, I., & Schettini, R. (1997). An effective strategy for querying image databases by color distribution. *Computer and the History of Art Journal (Special issue on Electronic Imaging and the Visual Arts)*, *7*(1), 3–14.

Binaghi, E., Della Ventura, A., Rampini, A., & Schettini, R. (1993). A fuzzy reasoning approach to similarity evaluation in image analysis. *International Journal of Intelligent Systems*, *8*(7), 749–769.

Binaghi, E., Gagliardi, I., & Schettini, R. (1994). Image retrieval using fuzzy evaluation of color similarity. *International Journal of Pattern Recognition and Artificial Intelligence*, *8*(4), 945–968.

Carrara, P., Della Ventura, A., & Gagliardi, I. (1996). Designing hypermedia information retrieval systems for multimedia art catalogues. *The New Review of Hypermedia and Multimedia*, *2*, 175–195.

Ciocca, G., Gagliardi, I., & Schettini, R. (1998). Retrieving color images by content. Presented at the *Image and Video Content-Based Retrieval Workshop*.

Del Bimbo, A., & Pala, P. (1997). Visual image retrieval by elastic matching of user sketches. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *19*(2), 121–132.

Della Ventura, A., Gagliardi, I., & Schettini, R. (1998). Indexing color-texture image patterns. Image description and retrieval. In E. Vicario, *Advance in computer vision and machine intelligence* (pp. 105–117). New York: Plenum Press.

Desai Narasimhalu, A., & Kankanhalli, M. S. (1997). Benchmarking multimedia databases. *Multimedia Tools and Applications*, *4*, 333–356.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., & Petrovic, D. (1994). Efficient and effective querying by image content. *Journal of Intelligent Systems*, *3*, 231–262.

Gagliardi, I., & Schettini, R. (1997a). A method for the automatic indexing of color images for effective image retrieval. *The New Review of Hypermedia and Multimedia*, *3*, 201–224.

Gagliardi, I., & Schettini, R. (1997b). Content-based image indexing using wavelet transform. Paper presented at the *Third International Workshop on Multimedia Information Systems*.

Gimel'Farb, G. L., & Jain, A. K. (1996). On retrieving textured images from an image database. *Pattern Recognition*, *29*, 1461–1483.

Gudivada, V. N., & Rahavan, V. V. (1997). Modeling and retrieving images by content. *Information Processing and Management*, *33*(4), 427–452.

Hafner, J., Sawhney, H. S., Esquitz, W., Flickner, M., & Niblack, W. (1995). Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *17*, 729–736.

Jain, A. K., & Vailaya, A. (1996). Image retrieval using color and shape. *Pattern Recognition*, *29*(8), 1233–1244.

Jain, R., Pentland, A., & Petkovic, D. (1995). Report on the *NFS-ARPA Workshop on Visual Information Management Systems Cambridge, MA, USA*.

McGill, M. J., & Salton, G. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Mehtre, B. M., Kankanhalli, M. S., Desai Narasimhalu, A., & Man, G. C. (1995). Color matching for image retrieval. *Pattern Recognition Letters*, *16*, 325–331.

Mehtre, B. M., Kankanhalli, M. S., & Lee, W. F. (1997). Shape measures for content based image retrieval: a comparison. *Information Processing and Management*, *33*(3), 319–337.

Mehtre, B. M., Kankanhalli, M. S., & Lee, W. F. (1998). Content-based image retrieval using a composite color-shape approach. *Information Processing and Management*, *34*(1), 109–120.

Minka, T., & Picard, R. W. (1997). Interactive learning with a society of models. *Pattern Recognition*, *30*(4), 565–581.

Mitra, M., Huang, J., & Kumar, S. R. (1997). Combining supervised learning with color correlograms for content-based image retrieval, Paper presented at the *Fifth ACM Multimedia 97 Conference*.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1998). *Introduzione alla statistica*. New York: McGraw-Hill.

Ortega, M., Rui, Y., Chakrabarti, K., Mehrotra, S., & Huang, T. S. (1997). Supporting similarity queries in MARS. Paper presented at the *Fifth ACM Multimedia 97 Conference*.

Pass, G., Zabih, R., & Miller, J. (1996). Comparing images using color coherence vectors. Paper presented at the *Fourth ACM Multimedia 96 Conference*.

Picard, R. W., & Minka, T. P. (1995). Vision texture for annotation. *Multimedia Systems*, *3*, 3–14.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rui, Y., Huang, T. S., Mehrotra, S., & Ortega, M. (1997). A relevance feedback architecture in content-based multimedia information retrieval systems. Paper presented at the *IEEE Workshop on Content-based Access of Image and Video Libraries*.

Scheunders, P., Livens, S., Van de Wouwer, G., Vautrot, P., & Van Dyck, D. (1997). Wavelet-based texture analysis. *International Journal Computer Science and Information management* (wcc.ruca.ua.ac.be/ ~ livens/WTA/).

Sclaroff, S., Taycher, L., & La Cascia, M. (1997). ImageRover: a content-based image browser for the world wide web. Paper presented at the *IEEE Workshop on Content-based Access Image and Video Libraries*.

Smith, J. R., & Chang, S.-F. (1996). VisualSEEK: a fully automated content-based image query. Paper presented at the *Fourth ACM Multimedia 96 Conference*.

Stricker, M. A., & Orengo, M. (1995). Similarity of color images. Paper presented at the *SPIE Storage and Retrieval for Image and Video Databases III Conference*.

Stricker, M. A., & Dimai, A. (1996). Color indexing with weak spatial constrains. Paper presented at the *SPIE Storage and Retrieval for Image and Video Databases IV Conference*.

Taycher, L., La Cascia, M., & Sclaroff, S. (1997). Image digestion and relevance feedback in the ImageRover WWW Search Engine. Paper presented at the *Visual 1997 Conference*.