# Evaluating the Effectiveness and the Efficiency of a Vector Image Search Tool

Patrizia Di Marco, Tania Di Mascio, Daniele Frigioni, and Massimo Gastaldi

Department of Electrical and Information Engineering,
University of L'Aquila, Poggio di Roio, I-67040 L'Aquila, Italy
{tania.dimascio,daniele.frigioni,massimo.gastaldi}@univaq.it

**Abstract.** In this paper we develop VISTO (Vector Images Search TOol), along two directions: (1) we present a new interface for VISTO which is more sophisticated than the original one, since it has been developed having in mind the users and their retrieval requests; (2) we provide a much deeper evaluation of the effectiveness and the efficiency of VISTO in the specific domain of the Blissymbolic images.

## 1 Introduction

The research in the field of Content Based Image Retrieval (CBIR) has been concentrated in the past mainly on raster images. It was maybe the wide variety of formats available for vector images, along with their strong dependence on application programs, which discouraged research in CBIR systems for vector images. It may also be noticed that raster images rule the roost on the World Wide Web, notwithstanding the convenience of vector images on the web, for their reduced size, as well as for the possibility of client-side scaling, which avoids new images to be sent. However, in the last years, the growing popularity of new vectorial-based web design programs, such as Macromedia Flash, along with the new format SVG (Scalable Vector Graphic) proposed by W3C, are changing this trend, promising to bring vector graphics to ordinary web pages soon.

Notwithstanding this increasing interest, the great majority of CBIR systems proposed in the literature still deal with raster images (for a complete survey we refer to [16]). To the best of our knowledge, the unique proposal that try to solve CBIR when images are represented in a vectorial data model is VISTO (Vector Images Search TOol). VISTO has been introduced and developed in [9-11] by considering an initial application domain represented by a 2D animation production environment supporting cartoon episodes management. VISTO was initially developed to meet the cartoonists requirements, and also with the aim of having a system that can be tuned to satisfy the requirements of other application domains. In fact, other application domains utilizing vector images (e.g., Clip-art, and CAD systems) share similar requirements.

The main characteristics of VISTO from the engine and the interface point of view have been described in [9, 11], while a preliminary experimental evaluation was

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 259–268, 2009.

carried out in [10], whose purpose was to evaluate the effectiveness and the efficiency of VISTO by studying the so-called Precision versus Recall curves (see, e.g., [7, 15]). The effectiveness of VISTO was demonstrated by the fact that the behavior of these curves was always descendent. The contribution of the paper is twofold:

We present a new interface developed for VISTO. The first prototype [9] of VISTO interface was developed only for tuning purposes of the system and hence it appeared quite raw. The new interface is more sophisticated and it has been developed having in mind the end-users and their retrieval requests. The main characteristics of the new interface are described in Section 2.

We provide a deeper evaluation of the effectiveness and the efficiency of VISTO using the Blissymbolic images application domain. In the evaluation process it is important to follow a consolidated evaluation methodology. To the best of our knowledge, a shared evaluation methodology for CBIR systems is not known in the literature. To this aim, we first studied evaluation methodologies in the area of Multimedia Retrieval. As a result, we derived a set of reasonable choices that can be applied to the evaluation of CBIR systems (described in Section 3). After that we applied the derived rules to the evaluation of VISTO in the Blissymbolic images application domain. The outcome of the experiments is described in Section 4 and can be summarized as follows: (1) the effectiveness and the efficiency of VISTO have been confirmed also in the new application domain; (2) for each category of images, we determined the more appropriate engine among those of VISTO; (3) we determined the image category on which VISTO has the best performance.

## 2 The VISTO System

Similarly to CBIR systems for raster images (see, e.g., [16]), VISTO was initially bound to the application domain and it uses, as feature representation, moments representing visual features of images. Differently from the CBIR systems proposed in the literature, VISTO uses the shape, and not the color, as the main visual feature, since in the context of vectorial images the shape is more important and representative than the color, and allows independence from affine transformations. Moreover, VISTO gives the possibility of interactively setting parameters of the retrieval process; it allows performing queries by sketch and queries by example. These design choices lead to a system supporting application domain users in searching tasks and researchers in domain-oriented tuning tasks. In what follows we concentrate on the aspects related to the engines and to the interface.

### 2.1 The VISTO Collection of Engines

Engines currently available in VISTO follow the classical architecture of CBIR systems (see, e.g., [10]). Given a query image, database images are ranked based on the similarity with the input image, so that more relevant images are re-turned first in the query result vector. The processing hence requires a Feature Description Processor to extract visual features and to create a vector containing a proper descriptor of each image, and a Comparison Processor to create a ranking vector representing the query image result using distances between descriptors. The similarity between any two images is computed as the similarity between the two corresponding descriptors.

Concerning the Feature Description Processor, the image is considered as an inertial system which is obtained by discretizing the vectorial image, and by associating material points with basic elements obtained by the discretization process. The origin of the inertial system is then moved to the center of mass, to which transformation can be applied. Once an image has been transformed into an inertial system, the natural way to represent image shape is to exploit the first four central moments: average, variance, skew and kurtosis. These moments are indices of distribution providing useful information about the image. In our context, the average represents the dimension of image: low average means image poor in strokes, high average means image rich in strokes. The variance represents how image center of mass area is composed: low variance connotes image center of mass area poor in strokes, high variance connotes image center of mass area rich in strokes. The skew suggests the symmetry of images: high value of skew means low symmetry of image, low value of skew suggests high symmetry of image. Finally, kurtosis represents how image is composed, high kurtosis means image poor in empty areas, low kurtosis means image rich in empty areas. In the literature, different invariant central moments sets have been proposed, differing in the way the moments are computed (see, e.g, [17]). The moment sets supported by VISTO are those of Hu [5], Zernike [1], and Bamieh [17].

Concerning the Comparison Processor, our approach is to use metrics well consolidated in the literature, that is: Cross Correlation [1], Discrimination Cost [1], and Euclidean [17] distances.

### 2.2 The New VISTO Interface

Differently from the first prototype [9], the new VISTO interface is organized in Tabs, each being dedicated to a specific task. In detail, the Basic and the Advanced Search tabs are designed to retrieving tasks, the Testing and the Clustering tabs are designed to tuning tasks. Moreover, three new types of results visualization have been included in the interface that well supports users in browsing results. The new interface supports both query-by-example and query-by-sketch. Result images may be selected as target images in a new search, in an incremental querying process. The new interface has been designed to help both application domain users and researchers in retrieving images and in tuning the engine in an interactive way, based on system feedback. In order to support users in these tasks, the new interface provides a Basic Mode for application domain users, and an Advanced Mode for researchers. The Basic Search Tab supports the Basic Mode, the Advanced Search Tab, the Testing Tab and the Clustering Tab support the Advanced Mode.

- The Basic Search Tab is designed to handle users input actions, and it is composed of two windows, the query-selection window that is always displayed, and the query-result window that is invoked only when the results of the query are ready to be visualized.

  - The query-selection window is shown in the left part of Figure 1 and it is composed of two panels, the query-input panel (left part of the window) that accepts user input actions, and the query-view panel (right part of the window) that displays the query image selected. The query-input panel requires the user to provide an image, either by sketching it, or by selecting a file containing it. The image selected is automatically visualized in the query-view panel as shown in the right part of the Basic Search Tab.
  - The query-result window is composed of two panels; the result-view panel that displays the retrieved images ranked by similarity and the re-query-input panel that allows users to perform an incremental querying process. To better use the display space, the re-query-input panel is visualized by simply clicking on the "Show query panel" button, and the result-view panel use tabs to support different types of visualization. When displayed, the re-query-input panel has the same query-input panel form of the query-selection window; the Tabular visualization, the Detailed visualization and the 3D visualization are the different supported types of results visualization. Users can also just point and click on an individual image result to select it as target image in a new search.
- The Advanced Search Tab is composed of two windows, the query-selection window and the query-result window.
  - The query-selection window is composed of four panels (see the right part of Figure 1); in the high part of this window, there are the query-input panel that accepts user input actions, the query-view panel that displays the query image selected, and the search-engine panel that contains engines supported by VISTO; users can select an engine by browsing different tabs of this panel. The fourth panel, containing the selected engine setting parameters, the available indexing and the folder indexing is automatically zoomed on when users click on the "Hide other parameters" button. The search button and the progression bar appear in the low part of this window.
  - The query-result window, is composed of three panels; the first is the re-query-input panel that allows users to perform queries incrementally; the second is the result-view panel that displays query results in tabular, detailed and 3D visualization type; the third is the analysis-panel visualized when clicking on the "Hide statistics" button, it provides an initial indication on the search effectiveness.
- The Testing Tab is dedicated to researchers to favor both an in-depth analysis of the tool effectiveness. It is composed of two panels: the test-input panel that supports users in selecting a new image file to be added in the test set, and the test-view panel that shows the path of selected image files added in the test set. The "Add to query list" button, in the low part allows the adding operation. It is worth noting that only after that a testing session made using this tab is concluded, the "Hide statistics" button in the query-result window of the Advanced Search Tab is able to work.
- The Clustering Tab supports the clustering process, which creates an optimized indexing. The objects of this tab are spatially organized according to the same philosophy used in the design process of the others tabs. This tab supports researchers in the clustering creation process used in the test sessions.

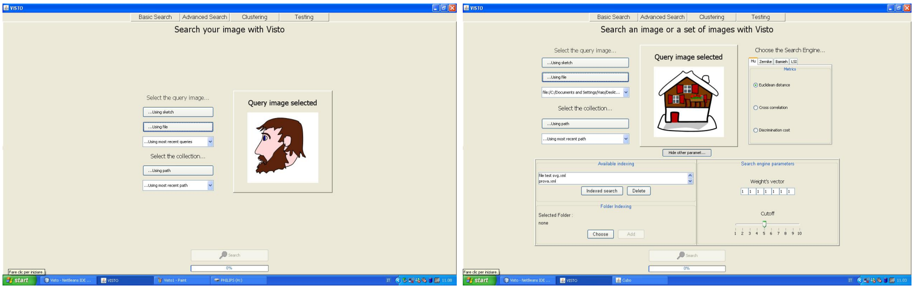

**Fig. 1.** The Basic Tab (left) and Advanced Search Tab (right) query-selection

# 3 Evaluation Methodologies

In this section we summarize our study of the literature concerning the evaluation of Multimedia Retrieval system.

## 3.1 State of the Art

Interesting results on evaluation methodologies have been proposed in the text, video and image retrieval areas. Among them we considered the systems of Table 1 and we studied their main features.

**Table 1.** Considered Systems

| | Test set | | | Query set | | Ground truth | Parameters | Results Analysis | |
|---|---|---|---|---|---|---|---|---|---|
| | ♯im. | st. | ♯cat. | ♯que. | st. | | | st. | devices |
| QBIC [12] | 1000 | - | - | 10 | - | - | AVRR, IAVRR | - | - |
| Chabot [13] | 500000 | - | - | 2 | - | - | P, R | - | |
| MARS [6] | 286 | - | - | 13 | - | real users | P, R | - | - |
| STAR [2] | 500 | - | - | 1 | - | real users(10) | - | - | - |
| CORE [2] | 100 | - | 10 | - | - | - | - | - | - |
| VisualSeek [14] | 3100 | - | - | 1 | - | - | - | - | P vs R curves |
| VideoQ [14] | 200 | - | - | 4 | - | - | - | - | - |
| Jacob [2] | 1500 | - | 5 | - | - | categories | - | - | - |
| MiAlbum [8] | 10009 | - | 79 | 200 | - | real users(9) | - | - | Cartesian |
| Artisan [3] | 10745 | - | - | 24 | - | real users | P, R, LPR | - | - |
| VISTO | 400 | AT | 12 | 12 | AT | real users(5) | P, R, G, F | AT | P vs R curves |

Each column of Table 1 represents a feature we considered relevant for defining a proper methodology. The Test set column represents the images set used in the experiments: the size (♯im.), the number of categories (♯cat.), and the statistic consistency (st.). The Query set column represents the set of benchmark queries used to evaluate the system. The Query set features we considered is cardinality (♯que.) and statistic consistency (st.). The Ground-truth column represents relevance judgments. The evaluation Parameters column allows evaluators to observe the retrieval process and to discover where systems are weak. The Results Analysis column analyzes data obtained during experiments, in order to have correct conclusions about evaluation.

The Test sets studied in the literature are often small in size; a large number of images is necessary to assure good results in recognizing differences or analogies among images. To guarantee the Test set goodness, it should be divided in categories to easily verify the statistic consistency. Not all evaluation methodologies of Table 1 consider categories and the statistical analysis lacks. We can conclude that the evaluation methodologies we studied: (1) use different test sets, (2) test set cardinality is often inadequate, (3) do not often use division into categories, (4) do not consider a statistic analysis of test sets.

The Query set elements have to represent, as much as possible, all kinds of requests users submit to the system. Also in this case, it is important to check the statistic consistency. The methodologies we studied use a small number of queries. Also in the case of the Query set, statistic consistency analysis completely lacks. According to these considerations we can conclude that in the evaluation methodologies we studied: (1) query set size in often inadequate, (2) statistic consistency analysis of query set is not considered.

The Ground-truth is not easy to define since it involves several aspects: environmental aspect (it depends on users' present needs), dynamic aspect (it changes frequently), subjective aspect (it depends on users' judgment), and cognitive aspect (it depends on users' behavior and perception). To define the Ground-truth, it can be helpful to involve real users, in order to make more realistic relevance definition. Some of the considered systems involve real users (10 users in STAR, 9 users in MiAlbum); in some evaluations study the ground truth is not defined. We conclude that (1) real users help is not always used; (2) the ground truth is not always defined for queries.

The evaluation Parameters are very important since they determine the system efficiency and effectiveness. In the literature several evaluation parameters exist [15], the more frequently used are Precision (P) and Recall (R) or parameters derived from them. Different parameters are used for example in QBIC [12], denoted as AVRR and IAVRR, and depending on the order of the relevant images and on the ideal order of the relevant images, and in Artisan [3], denoted as LPR, and depending on the position of the last relevant image found.

To appreciate experiments results, devices (e.g., tables containing parameters values, graphics device histograms, and Cartesian graphics) are used as support in the Results analysis. During this study, some unpredictable and out of control errors occur; for them it is impossible performing a deterministic analysis and, in order to estimate if a relation between a variable and the observed effect exists, it is necessary a statistical test, as for example the well known ANOVA test (AT). The statistical inaccuracy has to be always included in the Results analysis; Table 1 refers that in VisualSeek and in MiAlbum evaluations graphical devices are used, but statistical tests are never performed.

### 3.2 Choices

According to the considerations made in the previous section, we can derive the following reasonable choices:

1. A good Test set has to be large and made of heterogeneous images. It is important to define categories to test statistical consistency of the chosen Test set. To obtain a realistic classification it is advisable to involve real users to analyze images and to decide categories and their elements.
2. To define a Query set we can choose the set size and the set structure. We can define the Query set choosing a random sequence of images from Test set or we can choose Query set elements more carefully: we can consider an element (or more) from each category of the Test set.
3. To define the Ground truth it is necessary to define the relevance as follow: for a given query, all images belonging to the same category of the query are relevant. The Ground-truth is automatically defined: for each query the Ground-truth is the category of the query.
4. The evaluation Parameters to be chosen are Precision and Recall; in fact they are intuitive, easy to use and to be graphically elaborate.
5. To obtain good Results analysis, it can be helpful to organize the evaluation process by defining work sessions and operative work sheets. After executing experiments it is necessary a statistical analysis with ANOVA tests and logical analysis to find out data relations.

We applied the above choices to the evaluation of VISTO as follows: (1) Test set: we consider a set of 400 Blissymbolic images in SVG format. Thanks to real users (♯real users: 5), the Test set is divided into 12 categories (see Table 2). We use the Indexing VISTO functionality and then we prove statistical consistency of the query set with the ANOVA test. (2) Query set: we consider one element of each Test set category (♯im.: 12); we also prove statistical consistency of Query set with the ANOVA test. (3) Ground truth: using VISTO functionalities for the ground truth definition, we save relevance judgments using the Testing Tab (see Section 2.2). (4) Parameters: VISTO offers charts depicted in the query-result window of Advanced Search Tab (see Section 2.2) to analyze Precision vs Recall curves. (5) Results analysis: tables containing evaluation parameters values, graphics device histograms, and Cartesian graphics are used for the analysis, and ANOVA tests are performed to statistically validate consistency.

## 4 Evaluation Experiments

The goals of our evaluation experiments are the following:

- **Goal A:** to evaluate the effectiveness of our system in terms of retrieval performance of all engines of VISTO over all queries listed in Table 2.
- **Goal B:** to study the retrieval efficiency of each engines.
- **Goal C:** to evaluate the best retrieved category.

**Table 2.** Blissymbolic images classification

| Category | Cardinality | Query |
|---|---|---|
| houses and buildings | 14 | bank.svg |
| circles | 58 | candle.svg |
| hearts | 10 | afraid.svg |
| curves | 13 | mind.svg |
| narrows | 29 | crane.svg |
| letters | 4 | france.svg |
| mixtures | 132 | noun.svg |
| not classified | 12 | caterpillar.svg |
| numbers | 15 | second.svg |
| question points | 16 | answer.svg |
| squares | 13 | all.svg |
| segments e points | 98 | addition.svg |

- To these aims, we follow the methodology described in the previous section:

1. Test set: we have used a set of 400 Blissymbolic images in SVG format;
2. Query set: we have randomly selected the 12 queries listed in Table 2;
3. Ground-truth: following the real users judgments, an image j of the Test set is relevant for a query Q on an image i, denoted as Q(i), if and only if j and i belong to the same category. We denote as $GT_{Q(i)}$ the ground-truth set of query Q(i). For instance, for query france.svg in Table 2, the ground-truth set $GT_{Q(france)}$ is the set containing all images in the letters category;
4. Parameters: we used the well-known Precision and Recall measures (see, e.g., [7]); Precision is the fraction of the retrieved images which are relevant and Recall is the fraction of the relevant images which have been retrieved.
   The tests proceed in steps as follows:

- **Step 1:** 12 * 9 * 96 = 10368 executions are issued to the system, in fact, 12 are the queries, 9 are the VISTO engines, and 96 are the couples cut-off and generality (from now denoted as (k, g)) chosen to evaluate the effectiveness of the system. Given a query Q(i) on a collection $C_{Q(i)}$, we define $g = |GT_{Q(i)}|/|C_{Q(i)}|$ and $k = |A_{Q(i)}|$, where $|C_{Q(i)}|$ is the cardinality of the collection and $A_{Q(i)}$ is the set containing all images of $C_{Q(i)}$ ranked by similarity with respect to i. According to the cardinality of the Test set categories, we chose fixing $k \in \{3, 6, 9, |GT_{Q(i)}|\}$ (k = 6 means that 6 are the retrieved images) and $g \in \{0.3, 0.5, 0.6, 0.9\}$ (g = 0.5 means that the collection contains 50% of the relevant images).
- **Step 2:** in order to perform executions described in the previous step, for each query Q(i), the collection set $C_{Q(i)}$ must be composed according to the g value we considered; then $C_{Q(i)}$ will contain all images relevant for Q(i) (all images $\in GT_{Q(i)}$) plus a number of images, randomly selected from the Test set, such that $|C_{Q(i)}| = |GT_{Q(i)}|/g$.

- **Step 3:** for each query Q(i), the Precision and Recall values, denoted as $PR_{Q(i)}$ and $RC_{Q(i)}$ respectively, are computed as follows:

  - $PR_{Q(i)} = |GT_{Q(i)} \cap A_{Q(i)}| / k$ (1)

  - $RC_{Q(i)} = |GT_{Q(i)} \cap A_{Q(i)}| / |GT_{Q(i)}|$ (2)

Formulas (1) and (2) highlight that the Precision and Recall values depend on $GT_{Q(i)}$ and $A_{Q(i)}$ . For the query Q(i), in our experiments, while $A_{Q(i)}$ vary with the chosen engine, $|GT_{Q(i)}|$, representing the cardinality of the category which i belongs, is fixed.

- **Step 4:** using formulas (1) and (2), for each of the 12 queries in Table 2, for each of the 9 engines of VISTO, and for each couple (k, g), different graphics are calculated for inspection. In particular:

– Set I: contains the 10368 PR vs RC curves defined in step 1;

– Set II: contains 12 histograms, one for each category. Each histogram contains, the average values of P $R_{Q(i)}$ and $RC_{Q(i)}$ , for each couple (k, g).

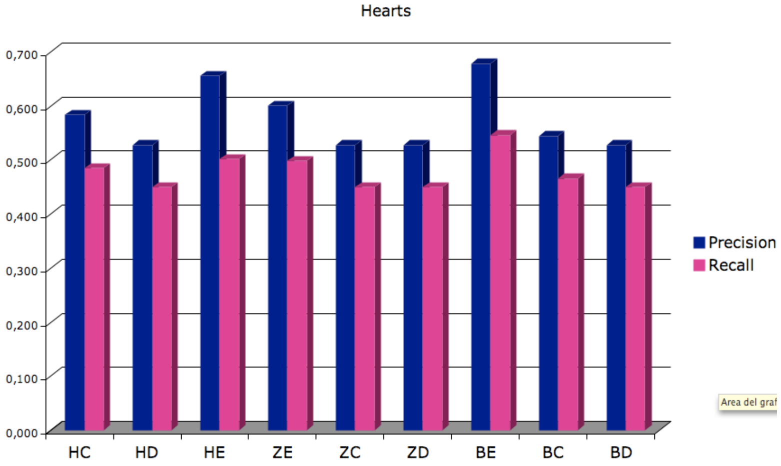

**Fig. 2.** Histogram for the hearts category

In relation to goal A, studying curves of Set I, we observed that all curves are descendent. This behavior demonstrates that, as well described in [4], given a query Q(i), all engines well retrieve images in the same category of image i, independently from (k, g). In relation to goal B, studying histograms of Set II, we individuated the best and the worst engines for each category (the best engine is such that has the higher value bars, see for example Figure 2). In relation to goal C, we can conclude that the best retrieved category in terms of Recall is the Letters category, and the best retrieved category in terms of Precision is the Numbers category. In conclusion, by the above described experiments we have demonstrated that all VISTO engines well works, independently from g (when g increases, the relevant images decreases). We also discovered that BE is the best engine for 4 categories (Hearts, Question points, Squares and Segments and points), ZC for 3 categories (Houses and Buildings, Circles and Mixtures), BD for 2 categories (Curves, and Letters ), HC for one category (Arrows), BC for one category (Not classified ), and HD for one category (Numbers ). Finally, we discovered that the best retrieved category is Letters.

## References

1. Chim, Y.C., Kassim, A.A., Ibrahim, Y.: Character recognition using statistical moments. Image and Vision Computing 17, 299–307 (1997)
2. de Vries, A.P.: The role of evaluation in the development of content-based retrieval techniques
3. Eakins, J.P., Boardman, J.M., Graham, M.E.: Similarity retrieval of trademark images. IEEE, Multimedia 5(2), 53–63 (1998)
4. Heesch, D., Ruger, S.: Combining features for content-based sketch retrieval: a comparative evaluation of retrieval performance. IEEE Transaction on Image Processing
5. Hu, M.K.: Visual pattern recognition by moments invariants. IRE Transactions on Information Theory 8, 179–187 (1962)
6. Huang, T., Mehrotra, S., Ramchandran, K.: Multimedia analysis and retrieval system (MARS) project. In: Data Processing Clinic (1996)
7. Koskela, M., Laasonen, J., Laakso, S., Oja, E.: Evaluating the performance of content based imag retrieval system. In: Laurini, R. (ed.) VISUAL 2000. LNCS, vol. 1929, pp. 430–441. Springer, Heidelberg (2000)
8. Liu, W., Su, Z., Li, S., Zhang, H.: A performance evaluation protocol for content-based image retrieval algorithms/systems (2001)
9. Di Mascio, T., Francesconi, M., Frigioni, D., Tarantino, L.: Tuning a CBIR system for vector images: The interface support. In: Proceedings of Working Conference on Advanced Visual Interfaces (AVI 2004), pp. 425–428. ACM, New York (2004)
10. Di Mascio, T., Frigioni, D., Tarantino, L.: Evaluation of VISTO: the new vector image search tool. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 836–845. Springer, Heidelberg (2007)
11. Di Mascio, T., Frigioni, D., Tarantino, L.: A visual environment for tuning content-based vector image retrieval. In: Lawrence Erlbaum Associates (ed.) Proceedings of HCI 2005, Adjunctive Proceedings (2005)
12. Niblack, W., Barber, R.: The qbic project: Querying images by content using color, texture and shape. In: Proceedings of Conference on Storage and Retrieval for Image and Video Databases, pp. 173–187 (1993)
13. Ogle, V.E., Stonebraker, M.: Chabot: Retrieval from a relational database of images. IEEE Computer 28(9), 40–48 (1995)
14. Smith, J.R., Chang, S.-F.: Visualseek: A fully automated content-based image query system. ACM Multimedia, 87–98 (1996)
15. Smith, J.R.: Image retrieval evaluation. In: IEEE WorkShop on Content-Based Access of Image and Video Libreries (June 1998)
16. Veltkamp, R.C., Tanase, M.: A survey of content-based image retrieval systems. In: Proc. of Content-based image and video retrieval, pp. 47–101. Kluwer Academic Publishers, Dordrecht (2002)
17. Yang, L., Albregtsen, F.: Fast computation of invariant geometric moments: a new method giving correct results. In: Proceedings of IEEE International Conference on Pattern Recognition, pp. 201–204 (1994)