ELSEVIER

Pattern Recognition Letters 21 (2000) 1193–1198

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

Content-based query of image databases: inspirations from text retrieval

David McG. Squire [b,*], Wolfgang Müller [a,1], Henning Müller [a], Thierry Pun [a]

[a] *Computer Vision Group, Computing Centre, University of Geneva, Geneva, Switzerland*
[b] *Computer Science and Software Engineering, Monash University, Clayton, Vic. 3168, Australia*

Abstract

This paper reports the application of techniques inspired by text retrieval research to content-based image retrieval. In particular, we show how the use of an inverted file data structure permits the use of an extremely high-dimensional feature-space, by restricting search to the subspace spanned by the features present in the query. A suitably sparse set of colour and texture features is proposed. A weighting scheme based on feature frequencies is used to combine disparate features in a compatible manner, and naturally extends to incorporate relevance feedback queries. The use of relevance feedback is shown consistently to improve system performance. © 2000 Published by Elsevier Science B.V.

Keywords: Content-based image retrieval; Inverted files; Relevance feedback

1. Introduction

In recent years the use of digital image databases has become common, both on the web and in general publishing. Consequently, the efficient querying and browsing of large image databases is ever more important. Content-based retrieval from large text databases has been studied for decades, yet the insights and techniques of text retrieval (TR) have largely been ignored or reinvented by content-based image retrieval (CBIR) researchers. The utility of *relevance feedback* (RF) is long-established (Salton and Buckley, 1990), yet its application in CBIR systems (CBIRSs) is very recent (Wood et al., 1998). Similarly, many term-weighting approaches have been investigated, both empirically and theoretically (Salton and Buckley, 1988). Performance evaluation has also been thoroughly studied (Salton, 1992), yet *precision* and *recall*, the usual performance measures, are not widely used in CBIR.

TR systems usually treat each possible term as a search space dimension: $O(10^4)$ dimensions are thus typical. Crucially, in such systems both queries and stored objects are sparse: they have only a small subset ($O(10^2)$) of all possible attributes. Search can thus be restricted to the subspace spanned by the query terms. The data structure which makes this efficient is the *inverted file* (IF) (Squire et al., 1999). Conversely, CBIR researchers have devoted considerable effort to the search for compact image representations (choosing the "right" features), and the use of dimensionality

* Corresponding author.
E-mail address: david.squire@CSSE.monash.edu.au (D.McG. Squire).

[1] Supported by Swiss National Science Foundation grant No. 2000-052426.97.

0167-8655/00/$ - see front matter © 2000 Published by Elsevier Science B.V.
PII: S0167-8655(00)00081-7

reduction techniques such as factor analysis (Pun and Squire, 1996).

We present a CBIRS which uses an IF with more than 80 000 possible features ($\approx O(10^3)$ features per image). A feature weighting scheme based on feature frequencies in both the query image and the entire collection, commonly used in TR, is employed. RF is also used. Evaluation using precision and recall demonstrates a clear improvement over a previously reported system using a smaller feature set and nearest-neighbour search.

2. Related work

CBIR researchers generally acknowledge that semantic retrieval remains impossible. The usual approach is to attempt to capture image similarity using some function of a small set of low-level features. Most systems employ features based on colour, texture or shape. Features are often computed globally, and contain no spatial information. Some systems allow the user to influence the relative weights of these classes of features.

Features. The use of colour features, usually calculated in a space thought to be "perceptually accurate" (e.g., HSV or CIE), is almost universal (Niblack et al., 1993; Smith and Chang, 1996; Sclaroff et al., 1997). The usual representation is the colour histogram, with histogram intersection used as the similarity measure. This usually takes no account of perceptual similarities between bins. A matrix of similarity coefficients can be used (Niblack et al., 1993), but the coefficients must be be determined, and the cost is quadratic.

Many systems use texture to improve image characterization. A great variety of texture features has been employed: hierarchies of Gabor filters (Ma and Manjunath, 1996); the Wold features used in Photobook (Pentland et al., 1996); the coarseness, contrast, and directionality features used in QBIC (Niblack et al., 1993); and many more.

Shape features are often computed assuming that images contain only one shape, and are thus best applied to restricted domains. Shape features include: modal matching, applied, for example, to isolated machine tools (Sclaroff, 1997); histograms of edge directions, applied to trademarks (Jain and Vailaya, 1996); and matching of shape components such as corners, line segments or circular arcs (Cohen and Guibas, 1997).

Global features are often inadequate for CBIR: spatial layout and individual objects are frequently important. Features which retain spatial information, such as wavelet decompositions (Ze Wang et al., 1997), may be used. Alternatively, the image can be segmented. Features such as color and texture are extracted for each region, as well as spatial properties such as size, location and relationships to other regions (Smith and Chang, 1996; Carson et al., 1997). This turns CBIR into a labeled graph matching problem.

Similarity. The meaning of similarity in CBIR is rarely addressed, even though human similarity judgments vary greatly (Mokhtarian et al., 1996; Squire and Pun, 1997). Image similarity is typically defined using a metric on a feature space. This implies that, if the "right" features are chosen, proximity in feature space will correspond to perceptual similarity. There are several reasons to doubt this, the most fundamental being the *metric assumption*. There is evidence that human similarity judgments do not obey the requirements of a metric: "[Self-identity] is somewhat problematic, symmetry is apparently false, and the triangle inequality is hardly compelling" (Tversky, 1977, p. 329). The lack of symmetry is the most important issue: the features which are significant depend on which item is the query.

Some attempts have been made to address these problems. Self-organizing maps have been used to cluster texture features according to class labels provided by users (Ma and Manjunath, 1996). A set-based technique has been applied to learn groupings of similar images from positive and negative examples provided by users (Pentland et al., 1996).

Relevance feedback. There are two basic approaches to RF. According to the RF, a system can create a composite query from relevant and non-relevant images (Huang et al., 1997), or it can adjust its similarity metric (Sclaroff et al., 1997). Some use the variances of features in the relevant set as a weighting criterion (Rui et al., 1998). Whilst related to the variance-based approach, the

technique presented here can cope with multimodal distributions of relevant features, and with much greater numbers of possible features.

3. *Viper* system overview

In this section, a brief overview of the *Viper* system is presented. [2] A more detailed account may be found in (Squire et al., 1999). *Viper* employs more than 80 000 simple colour and spatial frequency features, both local and global, extracted at several scales. These are intended to correspond (roughly) to features present in the retina and early visual cortex. The fundamental difference between traditional computer vision and image database applications is that there is a human "in the loop". RF allows a simple classifier to be learnt on the fly, corresponding to the user's information need.

3.1. Features

Viper uses a palette of 166 colours, derived by quantizing HSV space into 18 hues, 3 saturations, 3 values and 4 grey levels. Two sets of features are extracted from each image. The first is a colour histogram, with empty bins discarded. The second represents colour layout. Each block in the image (starting with the image itself) is recursively divided into four, at four scales. The mode color of each block is treated as a binary feature, meaning that there are 56 440 possible colour block features. Each image has 340.

Gabors have been applied to texture characterization, as well as more general vision tasks (Ma and Manjunath, 1996; Jain and Healey, 1998). We employ a bank of real, circularly symmetric Gabors:

$$f_{mn}(x,y) = \frac{1}{2\pi\sigma_m^2} e^{-(x^2+y^2)/2\sigma_m^2} \cos(2\pi(u_{0_m} x \cos\theta_n + u_{0_m} y \sin\theta_n)), \quad (1)$$

where m indexes filter scales, n their orientations, and u_{0_m} is the centre frequency. The half peak radial bandwidth is chosen to be one octave, which determines σ_m. The highest centre frequency is chosen as $u_{0_1} = 0.5$, and $u_{0_{m+1}} = u_{0_m}/2$. Three scales are used. The four orientations are: $\theta_0 = 0$, $\theta_{n+1} = \theta_n + \pi/4$. The resultant bank of 12 filters gives good coverage of the frequency domain, with little filter overlap. The mean energy of each filter is computed for each of the smallest blocks in the image, and quantized into 10 bands. A feature is stored for each filter with energy greater than the lowest band. Each image has at most 3072 of the 27 648 such possible features. Histograms of these features are used to represent global texture characteristics.

3.2. Similarity computation and relevance feedback

In a CBIR application, RF offers two advantages. First, augmenting the query with features from relevant images produces a better representation of the user's desires. The second advantage is unique to CBIR. In TR, feature extraction is free: the documents' component words are the features. This is not the case in CBIR. We envisage a system in which expensive features are extracted off-line, even though they may be too costly to evaluate for a new query image. Once some images are retrieved using a subset of cheap, simple features, potentially relevant, complex features can be introduced via RF.

Features are combined according to Eq. (2). For a query q containing N images i with relevance levels $R_i \in [-1,+1]$ and features j with frequencies df_{ij},

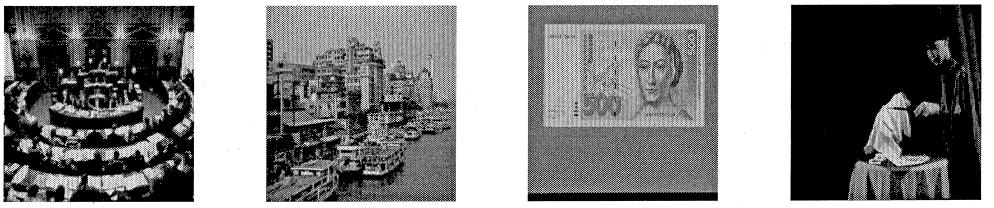
$$df_{qj} = \frac{1}{N} \sum_{i=1}^{N} df_{ij} R_i. \quad (2)$$

For each feature j, the images containing j are added to the result pool. For non-histogram features, the score s_k of each image k is updated according to Eq. (3), where cf_j is the frequency of the feature j in the database.

$$s_{k_{\text{new}}} = s_{k_{\text{old}}} + df_{qj} df_{kj} \log cf_j^{-1}, \quad (3)$$

$$s_{k_{\text{new}}} = s_{k_{\text{old}}} + \text{sign}(df_{qj}) \min(|df_{qj}|, df_{kj}) \log cf_j^{-1}. \quad (4)$$

[2] http://viper.unige.ch/



"Parliament" "Cityscape" "Banknote" "Magician"

Fig. 1. Sample images from the *Viper* test database.

The motivation is very simple: features which are common in an image characterize that image well; features which are common in the collection do not distinguish well between images (Salton and Buckley, 1988). For histogram features Eq. (4) is used, which is a weighted variant of standard histogram intersection.

4. Experiments

Viper performance was evaluated using a set of 500 heterogeneous colour images provided by Télévision Suisse Romande (see Fig. 1). Ten images were selected as queries. Five users examined all 500 images to determine their relevant sets for each query. [3] These relevant sets varied greatly in size, and the degree of visual similarity within each set also varied greatly. *Viper* returned the top 20 ranked images for each query. Using a "consistent user" assumption, the relevant set for each user for each query was inspected, and the set of relevant images present in the top 20 was then submitted as a second, RF query.

The performance of *Viper* was compared with that of a low-dimensional vector space system of the sort commonly used in image retrieval. The system uses a set of 16 colour, segment, arc and region statistics (Squire and Pun, 1997). System performances are compared using precision P and recall R,

$$P = \frac{r}{N}, \qquad R = \frac{r}{N_{\text{rel}}}, \tag{5}$$

where N is the total number of images (documents) retrieved, r the number of relevant images retrieved, and N_{rel} is the total number of relevant images in the collection. In general, precision decreases as more images are retrieved. An ideal P vs. R graph has $P = 1\ \forall R$.

Precision and recall data are often presented in the form of a precision vs. recall graph, which shows, in general, how precision decreases as increasingly large fractions of the collection are retrieved. An ideal precision vs. recall graph has precision = 1 for all values of recall: all the relevant images are retrieved before any irrelevant ones. The closer precision stays to 1, the better.

Fig. 2 shows the performances of the systems averaged over all users and queries. Two plots are shown for the *Viper* system, indicating performance before and after the RF step. It should be remembered that only the top 20 ranked images

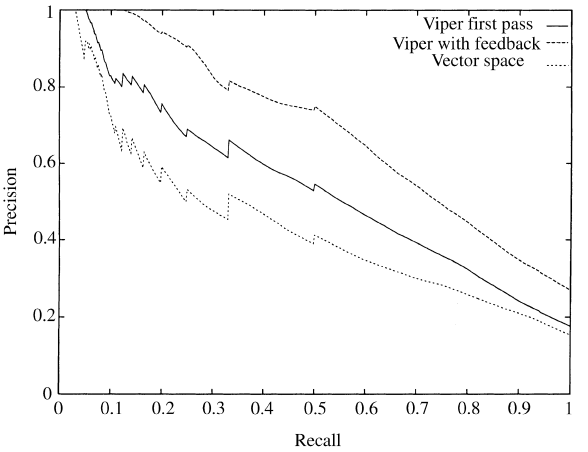


Fig. 2. Precision vs. Recall graphs, averaged over all users and queries.

[3] All users were computer vision researchers, so some bias can be expected.

from the first pass were used: it was thus not, in general, possible to include all relevant images in the feedback query.

The plots clearly indicate the value of RF. The averaged precision of the RF queries remains higher than that of either the first pass of *Viper* or that of the vector space system at all recall values. The average first pass performance of the very high-dimensional *Viper* system is also better than that of the low-dimensional vector space system.

5. Conclusion

In this paper we have shown how techniques inspired by text retrieval can be applied to the content-based query of image databases. We believe that there is much to be learnt from the decades of research in text retrieval, despite the fact that the terms of text queries (words) are much closer to the semantic level than the simple features usually used for image retrieval.

The use of inverted files, coupled with an appropriate choice of discrete features, allows feature spaces of extremely high dimensionalities to be searched efficiently. We have demonstrated the application of this technique to an image retrieval system with more than 80 000 possible features.

The use of precision and recall graphs provides a standard means of comparing system performances. Experiments using 10 queries on a test database of 500 images demonstrated that the *Viper* system, using frequency-based weights, performed better than a vector space system even without relevance feedback. One iteration of relevance feedback always improved performance, often dramatically.

References

Carson, C., Belongie, S., Greenspan, H., Malik, J., 1997. Region-based image querying. In: Proc. 1997 IEEE Conf. Computer Vision and Pattern Recognition (CVPR'97), San Juan, Puerto Rico.

Cohen, S.D., Guibas, L.J., 1997. Shape-based image retrieval using geometric hashing. In: Proc. ARPA Image Understanding Workshop, pp. 669–674.

Huang, J., Kumar, S.R., Mitra, M., 1997. Combining supervised learning with color correlograms for content-based image retrieval. In: Proc. 5th ACM Internat. Multimedia Conf. (ACM Multimedia 97), Seattle, WA, USA, pp. 325–334.

Jain, A., Healey, G., 1998. A multiscale representation including opponent color features for texture recognition. IEEE Trans. Image Processing 7 (1), 124–128.

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

Ma, W., Manjunath, B., 1996. Texture features and learning similarity. In: Proc. 1996 IEEE Conf. Computer Vision and Pattern Recognition (CVPR'96), San Francisco, CA, pp. 425–430..

Mokhtarian, F., Abbasi, S., Kittler, J., 1996. Efficient and robust retrieval by shape content through curvature scale space. In: Image Databases and Multi-Media Search. Amsterdam, The Netherlands, pp. 35– 42.

Niblack, W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G., 1993. QBIC project: querying images by content, using color, texture, and shape. In: Storage and Retrieval for Image and Video Databases. In: SPIE Proc., Vol. 1908, pp. 173–187.

Pentland, A., Picard, R.W., Sclaroff, S., 1996. Photobook: tools for content-based manipulation of image databases. Internat. J. Comput. Vision 18 (3), 233–254.

Pun, T., Squire, D.M., 1996. Statistical structuring of pictorial databases for content-based image retrieval systems. Pattern Recognition Letters 17, 1299–1310.

Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S., 1998. Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. Circuits Systems Video Technol. 8 (5), 644–655.

Salton, G., 1992. The state of retrieval system evaluation. Inf. Process. Manage. 28 (4), 441–450.

Salton, G., Buckley, C., 1988. Term weighting approaches in automatic text retrieval. Inf. Process. Manage. 24 (5), 513–523.

Salton, G., Buckley, C., 1990. Improving retrieval performance by relevance feedback. J. Am. Soc. Inf. Sci. 41 (4), 287–288.

Sclaroff, S., 1997. Deformable prototypes for encoding shape categories in image databases. Pattern Recognition 30 (4), 627–642.

Sclaroff, S., Taycher, L., La Cascia, M., 1997. ImageRover: a content-based browser for the world wide web. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, San Juan, Puerto Rico, pp. 2–9.

Smith, J.R., Chang, S.-F., 1996. Tools and techniques for color image retrieval. In: Storage and Retrieval for Image and Video Databases IV. In: IS&T/SPIE Proc., San Jose, CA, USA, Vol. 2670, pp. 426–437.

Squire, DM., Müller, W., Müller, H., Raki, J., 1999. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In: The 11th Scandinavian Conf. Image Analysis (SCIA'99), Kangerlussuaq, Greenland, pp. 143–149.

Squire, D.M., Pun, T., 1997. A comparison of human and machine assessments of image similarity for the organization of image databases. In: The 10th Scandinavian Conf. Image Analysis (SCIA'97), Lappeenranta, Finland, pp. 51–58.

Tversky, A., 1977. Features of similarity. Psychol. Rev. 84 (4), 327–352.

Wood, M.E., Campbell, N.W., Thomas, B.T., 1998. Iterative refinement by relevance feedback in content-based digital image retrieval. In: Proc. 5th ACM Internat. Multimedia Conf. (ACM Multimedia 98), Bristol, UK, pp. 13–20.

Ze Wang, J., Wiederhold, G., Firschein, O., Xin Wei, S., 1997. Wavelet-based image indexing techniques with partial sketch retrieval capability. In: Proc. 4th Forum on Research and Technology Advances in Digital Libraries, Washington, DC, pp. 13–24.