Human-Centered Object-Based Image Retrieval

Egon L. van den Broek[1,4], Eva M. van Rikxoort[2,4], and Theo E. Schouten[3]

[1] Department of Artificial Intelligence, Vrije Universiteit Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
egon@few.vu.nl
http://www.few.vu.nl/~egon/

[2] Image Sciences Institute, University Medical Center Utrecht,
Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
eva@isi.uu.nl
http://www.isi.uu.nl/

[3] Institute for Computing and Information Science, Radboud University Nijmegen,
P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
T.Schouten@cs.ru.nl
http://www.cs.ru.nl/~ths/

[4] Nijmegen Institute for Cognition and Information, Radboud University Nijmegen,
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Abstract. A new object-based image retrieval (OBIR) scheme is introduced. The images are analyzed using the recently developed, human-based 11 colors quantization scheme and the color correlogram. Their output served as input for the image segmentation algorithm: agglomerative merging, which is extended to color images. From the resulting coarse segments, boundaries are extracted by pixelwise classification, which are smoothed by erosion and dilation operators. The resulting features of the extracted shapes, completed the data for a <color, texture, shape>-vector. Combined with the intersection distance measure, this vector is used for OBIR, as are its components. Although shape matching by itself provides good results, the complete vector outperforms its components, with up to 80% precision. Hence, a unique, excellently performing, fast, on human perception based, OBIR scheme is achieved.

1 Introduction

More and more, the world wide web (www), databases, and private collections are searched for audio, video, and image material. Subsequently, As a consequence, there is a pressing need for efficient, user-friendly, multimedia retrieval and indexing techniques. However, where speech and handwriting recognition algorithms are generally applicable, image and video retrieval systems are only successful in a closed domain. These techniques have in common they are computational expensive and their results are judged as non-intuitive by its users.

In this paper, these drawbacks are tackled, for the field to content-based image retrieval (CBIR). An object-based approach on CBIR is employed: object-based image retrieval (OBIR), inspired by the findings of Schomaker, Vuurpijl, and De Leau [1], who showed that 72% of the people are interested in objects when searching images.

S. Singh et al. (Eds.): ICAPR 2005, LNCS 3687, pp. 492–501, 2005.

Moreover, a human-centered approach is chosen, based on the 11 color categories used by humans in color processing, as described in Section 2. These 11 color categories are also utilized for texture analysis, as discussed in Section 2.1, and for image segmentation, done by agglomerative merging (see Section 3.1). From the resulting, coarse image segments, the shape of the object is derived using pixelwise classification (Section 3.2). Next, erosion and dilation operations are applied on the boundary in order to smooth it, as described in Section 3.3. Section 3 introduces the shape matching algorithm. OBIR is conducted using four query schemes (see Section 5): two of them are based on color and texture, one on the object boundaries, and one on their combination. The results are presented in Section 6 followed by a discussion in Section 7.

2 Color and Texture in 11 Categories

As mentioned by Forsyth and Ponse [2]: "It is surprisingly difficult to predict what colors a human will see in a complex scene." However, it is known that humans use 11 color categories (red, green, blue, yellow, orange, brown, pink, purple, black, white, and gray) when processing color. These 11 color categories are considered universal and optimal [3,4]. These categories should, therefore, be: (i) generic, (ii) computationally cheap, and (iii) can be expected to yield results that are intuitive for users. Then these advantages support the aim of tackling the computational burden of CBIR (cf. QBIC uses a scheme with 4096 colors [5]) and to provide intuitive results for the users [6]. Therefore, we adopted the 11 color quantization scheme [7]: a unique color space segmentation, based on data gathered through experiments in which subjects categorized colors into the 11 color categories. So, the color distribution of images is characterized by a color vector with 11 color values.

Besides color, texture is an important feature for the human visual system [8]. Texture analysis can be done based on intensity differences, but nevertheless, color is important in texture recognition of color image material. With respect to color representation, Fujii, Sugo, and Ando [8] stated that "considering the effective computational strategy in our visual system, it is quite possible that not all the information carried out by the high-dimensional sensory representation is preserved for rapid judgments of natural textures." Taken this into account, the 11 color category quantization scheme should perfectly fit the job, and is, therefore, applied to color-based texture analysis.

For the analysis of texture, various methods are available, such as: statistical methods (e.g., co-occurrence matrices and autocorrelation features), geometrical methods (e.g., Voronoi tessellation features and structural methods), model based methods (e.g., random field models and fractals), and signal processing methods (e.g., spatial domain filters, Fourier domain filtering, Gabor models, and Wavelet models). Originally, they were developed for gray-value images but some of them have recently been adapted to fit texture analysis on color images.

2.1 The Color Correlogram

For the current research, one of the most intuitive texture analysis methods is applied: the color correlogram, as suggested by Huang, Kumar, Mitra, Zhu, and Zabih [9], which is constructed from an image by estimating the pairwise statistics of pixel color. In order

to (i) provide perceptual intuitive results and (ii) reduce the computational cost, the 11 color scheme for quantization of color is chosen.

The color correlogram $C_{\bar{d}}(i,j)$ counts the co-occurrence of pixels with colors i and j at a given distance $\bar{d}$. The distance $\bar{d}$ is defined in polar coordinates (d,α), with discrete length and orientation. In practice, α takes the values $0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$, and $315°$. The color correlogram $C_{\bar{d}}(i,j)$ can now be defined as follows:

$$C_{\bar{d}}(i,j) = \Pr(\mathrm{I}(p_1) = i \wedge \mathrm{I}(p_2) = j \mid |p_1 - p_2| = \bar{d}), \tag{1}$$

where Pr is probability and p_1 and p_2 are positions in the color image I. Let N be the number of colors in the image, then the dimension of the color correlogram $C_{\bar{d}}(i,j)$ will be $N \times N$, which is in our scheme 11×11. This algorithm yields a symmetric matrix. Hence, only angles up to $180°$ need to be considered. A direction insensitive color correlogram can be defined for each distance (d) by averaging the four color correlograms of the different angles (i.e., $0°$, $45°$, $90°$, and $135°$).

From the color correlogram, a large number of textural features can be derived, such as: energy, entropy, correlation, inverse difference moment, inertia, Haralick's correlation, cluster shade, and cluster prominence, which characterize the content of the image. Based on previous research [10], the combination of entropy, inverse difference moment, cluster prominence, and Haralick's correlation, with distance $d = 1$ is used, resulting in a vector of four texture features.

3 Shape Extraction

The shape extraction phase is divided in three stages: (i) coarse image segmentation, (ii) pixelwise classification, and (iii) smoothing. The coarse image segmentation uses only texture information to segment the image in texture regions. In the pixelwise classification phase, only color information is used because the regions are too small for our texture descriptor to be informative. The complete process of shape extraction is illustrated in Figure 1.

3.1 Segmentation by Agglomerative Merging

Segmentation is applied by agglomerative merging, as described by Ojala and Pietikäinen [11]. Their algorithm was introduced for gray-scale images but is extended to color images, using a color texture descriptor. The algorithm is applied using the color correlogram as texture descriptor based on the 11 color quantization scheme.

At the initial state of the agglomerative merging algorithm, the images are divided in sub blocks of size 16×16 pixels. At each stage of the merging phase, the pair of blocks with the lowest merger importance (MI) is merged. This merger importance is defined by the distance measure MI [9]. For two images I and I', the MI distance measure is defined as follows:

$$MI = |I - I'| = \sum_{i,j=0}^{m-1} |C_{\bar{d}}(i,j) - C'_{\bar{d}}(i,j)|, \tag{2}$$

where m is the number of bins used and $C_{\bar{d}}(i,j)$ and $C_{\bar{d}}(i,j)$ are the average color correlograms of images I and I' (see Equation 1), and $\bar{d}$ is set to 1 (see Section 2.1).

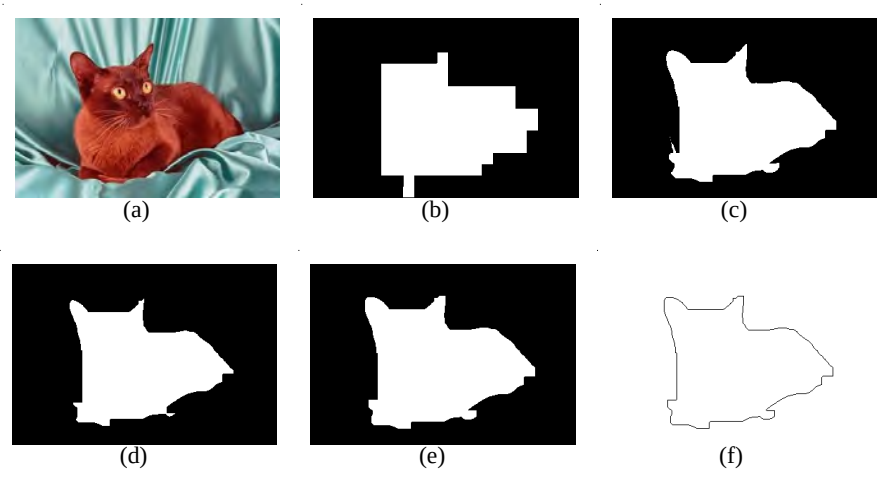

Fig. 1. (a) The original image (b) The coarse segmentation (c) The object after pixelwise classification (d) The object after erosion (e) The object after dilation (f) The final shape.

The closer MI is to zero, the more similar the texture regions are. When two regions are merged, the MI-values between this region and all adjacent regions are computed. The agglomerative merging phase continues until the experimentally determined stopping criterion (Y), given in Equation 3 is met:

$$MI_{stop} = \frac{MI_{cur}}{MI_{max}} < Y, \tag{3}$$

where MI_{cur} is the merger importance for the current best merge, MI_{max} is the largest merger importance of all preceding merges. For the current dataset, Y is determined to be 0.700. When the coarse segmentation phase is complete, the center segment of the image is selected to be the object of interest for OBIR.

3.2 Pixelwise Classification Based on the 11 Colors

After the center object has been identified in the coarse segmentation phase, pixelwise classification [11] is applied to improve localization of the boundaries of the object. In pixelwise classification, each pixel on the boundary of the center object is examined. A disk with radius r is placed over the pixel and the 11 color histogram is calculated for this disk and all adjacent segments. Next, the distance between the disk and the adjacent segments is calculated, using the intersection distance measure [7] based on the 11 color histogram. The pixel is relabeled if the label of the nearest segment is different from the current label of the pixel. This process is repeated as long as there are pixels that are being relabeled.

The radius r of the disk determines how smooth the resulting boundaries are: a small radius will produce ragged regions, a larger radius will produce smoother boundaries but may fail in locating the boundaries accurately. In order to tackle these problems we used a two-step approach: In the first iterations, a relatively small radius of 5 is used,

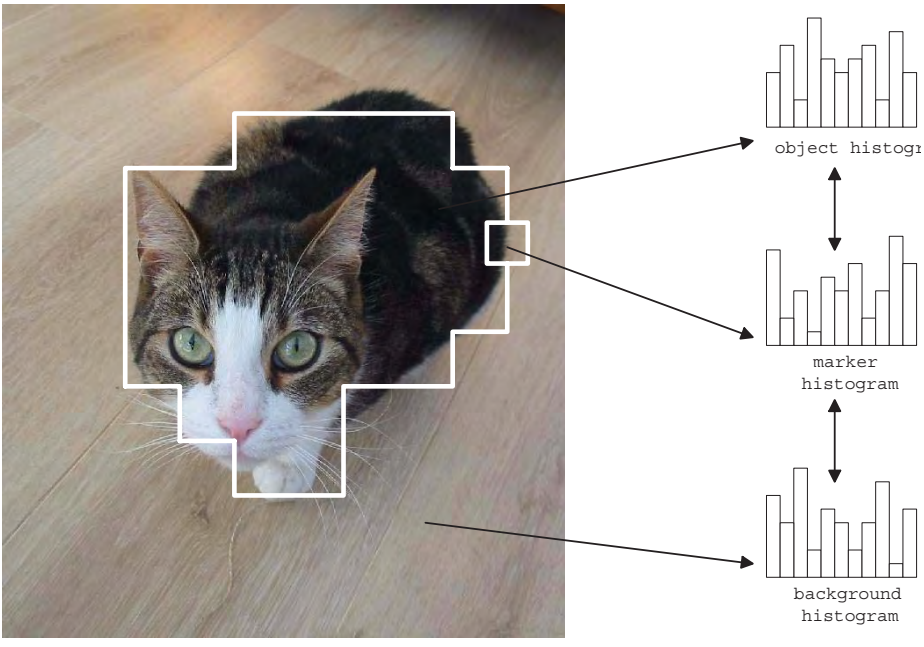

Fig. 2. The process of pixelwise classification illustrated. A pixel at the boundary is selected and a marker is placed over it. Next, the color histogram over this marker is calculated as well as the histograms of the center segment and the background. The histogram over the marker is compared to the other histograms and the pixel is assigned to the area with the most similar histogram (of the background or the object).

in order to locate the boundaries correctly. Secondly, a radius of 11 is used to produce more stable segments.

3.3 Smoothing

Although the pixelwise classification phase produces correct object boundaries, the shapes are smoothed to optimize for the shape matching phase. Smoothing is done using two fundamental operations: dilation and erosion.

Given two sets A and B in $\mathbb{Z}^2$, the dilation of A by B is defined as:

$$A \oplus B = \{x \mid (B)_x \cap A \neq \emptyset\}, \tag{4}$$

where $(B)_x$ denotes the translation of B by $x = (x_1, x_2)$ defined as:

$$(B)_x = \{c \mid c = b + x, \text{for some } b \in B\} \tag{5}$$

Thus, $A \oplus B$ expands A if the origin is contained in B, as is usually the case.

The erosion of A by B, denoted $A \ominus B$, is the set of all x such that B translated by x, is completely contained in A, defined as

$$A \ominus B = \{x \mid (B)_x \subseteq A\} \tag{6}$$

Thus, $A \ominus B$ decreases A.

The smoothing starts with two iterations of erosion with a square erosion marker (B) of size 3×3 pixels. Next, two iterations of dilation are applied with the same marker.

4 Shape Matching

Shape matching has been approached in various ways. A few of the frequently applied techniques are: tree pruning, the generalized Hough transform, geometric hashing, the alignment method, various statistics, deformable templates, relaxation labeling, Fourier and wavelet transforms, curvature scale space, and classifiers such as neural networks [12].

Recently, Andreou and Sgouros [12] discussed their: "turning function difference", as a part of their G Computer Vision library. It is an efficient and effective shape matching method. However, Schomaker et al. [1] introcuded a similar approach five years before. In the current research, the latter, original approach is adopted. This "outline pattern recognition", as the authors call it, is based on three feature vectors containing: (i) x and y coordinates, normalized using the center of gravity of the shape and the standard deviation of all radii, (ii) the running angle (θ) along the edge of the segment $(cos(\theta), sin(\theta))$, which contains more information on the local changes of direction, and (iii) the histogram of angles in the shape: the probability distribution $p(\theta)$ [1].

The algorithm proved to be translation, scale, and rotation invariant. Based on this algorithm, the outline-based image retrieval system Vind(X) was developed and has been used successfully since then. Vind(X) relies on outline-outline matching: the user draws an outline, which is the query. This outline is matched against the outlines of objects on images, present in its database. Subsequently, the images containing the best matching outlines are retrieved and shown to the user.

The Vind(X) system provides excellent retrieval results. However, in order to make its techniques generally applicable, automatic shape extraction techniques had to be developed. Moreover, these techniques had to be computationally cheap in order to preserve its fast retrieval, as much as possible. The latter was already achieved by the techniques as described in the previous sections. In combination with the matching algorithm of Vind(X), unsupervised OBIR was applied.

5 Method

In Sections 2 and 3, color, texture, and shape features are defined. They are combined and used in four distinct query schemes for object matching, using four vectors:

1. color and texture, for object versus complete images
2. color and texture
3. shape
4. color, texture, and shape combined

Feature-based and shape-based image retrieval was employed by two separate retrieval engines, connected to the same database, both using the intersection distance

Fig. 3. Sample images from the database used

measure for ranking their results. For both engines, the number of retrieved images (n) could be chosen by the user. All query schemes performed an object - object comparison, except scheme 1 for which object features are matched with the features of the complete images in the database. For query scheme 4, for each image its ranks on both engines are summed and divided by two.

In total, the database used, consists of 1000 images gathered from the Corel image database, a reference database for CBIR applications, and from the collection of Fei-Fei [13]. Since we are interested in objects, the six categories chosen represent objects: cats, leaves, revolvers, motorbikes, pyramids, and dinosaurs.

Adopted from the field of Information Retrieval, the performance of CBIR systems can be determined by the measures recall and precision. Recall signifies the proportion of relevant images retrieved from the database in response to the query. Precision is the proportion of retrieved images that is relevant to the query.

6 Retrieval Results

Recall and precision are calculated for each of the four different query schemes, as defined in Section 5, using a variable number of images retrieved. The precision of the retrieval results for the four schemes are plotted in Figure 4(a), for 5–25 images retrieved. The recall of the retrieval results for the four schemes are plotted in Figure 4(b), for the complete dataset.

All four schemes performed well, as shown in Figure 4(a) and 4(b). However, note that with the combined approach, four of the top five images are relevant; i.e., an average precision of 80% was achieved. Moreover, the recall achieved with the combined approach converges much faster to 100% than with the other approaches.

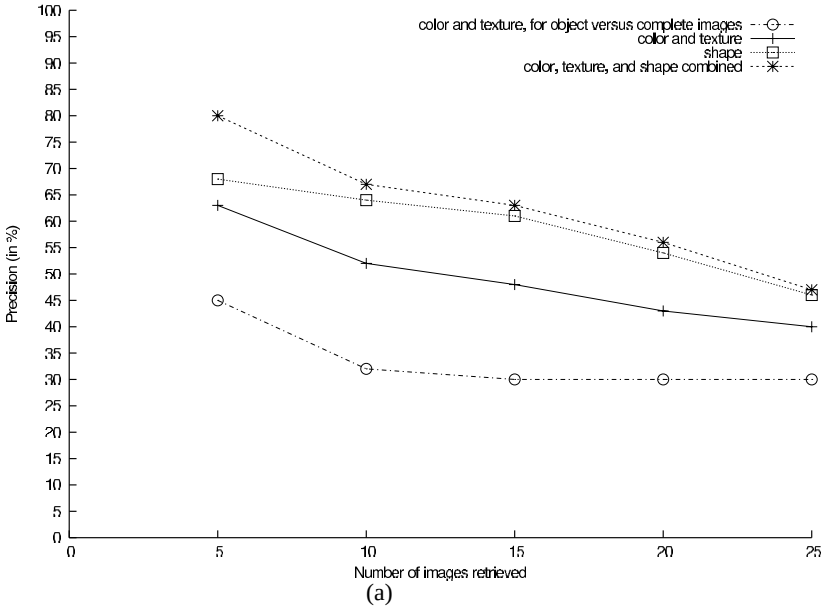

Fig. 4. Average precision (a) and recall (b) of retrieval, with global and local color&texture features, outline of extracted objects from images, and their combination.

7 Discussion

The rationale of the CBIR approach presented in this paper is that it be human centered. This is founded on two principles: (i) CBIR should be object-based and (ii) it should utilize the 11 color categories, as used by humans in color processing [7]. Both principles contribute to efficient CBIR, providing intuitive results for users. It was shown that the 11 color categories work well for describing color distributions, for the extraction of texture descriptors [10], and for object segmentation, as illustrated by the recall and precision of the retrieval results.

The success of matching the 2D shapes of segmented objects with each other is striking. This can, at least partly, be explained by the fact that "photographers generate a limited number of 'canonical views' on objects, according to perceptual and artistic rules" [1]. Moreover, even in the most recent research still (computationally expensive) gray-scale techniques are applied [14]. In contrast, we are able to extract shapes from color images. This is very important, since most of the image material available on the www and in databases is color.

In contrast with the reality on the www, the images in our database all contain images of objects against a rather uniform background, as illustrated in Figure 3. With our database, a first step is made toward processing real world images, where in comparable, recent work [15], object images are used that lack a background.

Despite the success of the current approach on real world images, it also has some drawbacks. First, it should be noted that the number of categories and its members were limited and follow-up research should be conducted with a larger database, incorporating a large number of categories. Second, in further developing the engine, the segmentation parameter should be set dynamically; i.e., setting the parameter to a minimum value and resetting it dynamically during the merging phase, based on the texture differences between the remaining blocks. This would obviate the current dependency on a good pre-defined parameter setting. Third, the ultimate goal would be to identify all objects in an image, instead of one, as is currently the case. Fourth, we expect that the use of artificial classifiers can improve the results, compared to the distance measures, used in the current research. When these drawbacks have been overcome, the resulting CBIR engine can be applied to real-world images instead of only to object-classes.

In this paper, a highly efficient scheme for the extraction of color, texture, and shape features is introduced. Combined with the intersection distance measure, it forms the basis of a unique, excellently performing, fast object-based CBIR (OBIR) engine, which provides results intuitive for its users.

Acknowledgments

Lambert R.B. Schomaker, Louis G. Vuurpijl, and Edward L. de Leau are honored, for their work on the initial version of the Vind(X) system. Without their effort, the NWO ToKeN project Eidetic (project-number: 634.000.001), which funded this research, would not have been undertaken. In addition, the authors are grateful to the anonymous reviewers and to Eduard Hoenkamp for their valuable comments.

References

1. Schomaker, L., Vuurpijl, L., Leau, E. de: New use for the pen: outline-based image queries. In: Proceedings of the 5th IEEE International Conference on Document Analysis, Piscataway (NJ), USA (1999) 293–296
2. Forsyth, D.A., Ponse, J.: Computer Vision: A modern approach. Pearson Education, Inc., Upper Saddle River, New Jerssey, U.S.A. (2002)
3. Berlin, B., Kay, P.: Basic color terms: Their universals and evolution. Berkeley: University of California Press (1969)
4. Derefeldt, G., Swartling, T., Berggrund, U., Bodrogi, P.: Cognitive color. Color Research & Application **29** (2004) 7–19
5. Hafner, J.L., Sawhney, H.S., Equitz, W., Flickner, M., Niblack, W.: Efficient color histogram indexing for quadratic form distance functions. IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995) 729–736
6. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
7. Broek, E.L. van den, Kisters, P.M.F., Vuurpijl, L.G.: Content-based image retrieval benchmarking: Utilizing color categories and color distributions. Journal of Imaging Science and Technology **49** (2005) [in press]
8. Fujii, K., Sugi, S., Ando, Y.: Textural properties corresponding to visual perception based on the correlation mechanism in the visual system. Psychological Research **67** (2003) 197–208
9. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In Medioni, G., Nevatia, R., Huttenlocher, D., Ponce, J., eds.: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (1997) 762–768
10. Broek, E.L. van den, Rikxoort, E.M. van: Parallel-sequential texture analysis. In Singh, S., Perner, P., Apte, C., eds.: Proceedings of the 3rd International Conference on Advances in Pattern Recognition (ICAPR2005). (2005) [conditionally accepted]
11. Ojala, T., Pietikäinen, M.: Unsupervised texture segmentation using feature distribution. Pattern Recognition **32** (1999) 477–486
12. Andreou, I., Sgouros, N.M.: Computing, explaining and visualizing shape similarity in content-based image retrieval. Information Processing & Management **41** (2005) 1121–1139
13. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In Pece, A.E.C., ed.: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, D.C., USA (2004)
14. Shih, M.Y., Tseng, D.C.: A wavelet-based multiresolution edge detection and tracking. Image and Vision Computing **23** (2005) 441–451
15. Gevers, T., Stokman, H.M.G.: Robust histogram construction from color invariants for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 113–118