

Similarity-based Ordering of Images in Search Engines

Klaas Bosteels

Supervisor(s): Etienne Kerre, Valérie De Witte, Stefan Schulte

Abstract—Search engines such as Google and Yahoo allow internet users to search through huge collections of images. These existing search engines are all text-based, i.e. the searching is based on the comparison of keywords with textual annotations. Content-based image retrieval (CBIR) is a more natural approach in which the searching is based on the actual content of the images. However, CBIR systems are still not capable of searching through collections consisting of millions of images. Therefore they are not suitable yet as a basis for a search engine on the internet. As a solution to that problem, we introduce a hybrid approach that combines text-based image retrieval (TBIR) with CBIR, without sacrificing the scalability of the system. More precisely we propose an extension of TBIR that adds the ability to sort the search results according to their similarity with a given set of examples. Such an extension requires measures that are able to express the similarity between two images. Because similarity is a gradual concept, we use fuzzy set theory to model these measures.

Keywords—similarity measures, colour images, fuzzy sets, content-based image retrieval

I. INTRODUCTION

MOST CBIR systems use the query-by-example paradigm, i.e. the user supplies an example image and the system searches for images that are similar to the example. The approach introduced in this paper adds comparable functionality to a traditional TBIR system. A query is still initiated by providing one or more keywords, but as soon as the results are available, it is possible to mark some of them as examples. The collection of results is then reordered according to the similarity with the examples. By rearranging the results this way, we can avoid the need for manual inspection of each result. Moreover the similarity-based arranging of search results has an extra advantage: the user does not need to have an appropriate example.

II. PRELIMINARIES

A. Fuzzy Sets

A fuzzy set A in a universe X is a $X \rightarrow [0, 1]$ mapping that associates with each element x from the universe X a degree of membership $A(x)$ [1]. We use the notation $\mathcal{F}(X)$ for the class of fuzzy sets in X .

For two fuzzy sets A and B in X , the classical set theoretic operations co (complement), \cap (intersection) and \cup (union) can be generalized as follows [1]: $(co A)(x) = 1 - A(x)$, $(A \cap B)(x) = \min\{A(x), B(x)\}$ and $(A \cup B)(x) = \max\{A(x), B(x)\}$ for each $x \in X$. The support of a fuzzy set A in X is given by $supp A = \{x \in X \mid A(x) > 0\}$. For a fuzzy set A in X with finite support, the sigma count, defined as $|A| = \sum_{x \in X} A(x)$, is a generalisation of the crisp concept cardinality.

Fuzzy sets can be further generalized to L -fuzzy sets. Such a L -fuzzy set in a universe X is a $X \rightarrow L$ mapping, with (L, \leq) a complete lattice for some partial order relation \leq [1].

B. Fuzzy Similarity Measures

Fuzzy similarity measures are fuzzy sets in $\mathcal{F}(X) \times \mathcal{F}(X)$ that express the similarity between each pair of fuzzy sets in X , i.e. the degree of membership $M(A, B)$ of $(A, B) \in (\mathcal{F}(X))^2$ denotes the similarity between A and B . If A and B are completely similar then $M(A, B) = 1$, otherwise $M(A, B) < 1$. The fuzzy similarity measures that we considered are a selection from the measures discussed in [2].

III. SIMILARITY MEASURES FOR IMAGES

A. Construction and Properties

If we are able to identify objects with fuzzy sets, then we can use fuzzy similarity measures to compare these objects. We have used this idea to construct several similarity measures for images. All these measures are reflexive and symmetric, i.e. each measure M satisfies $M(A, A) = 1$ and $M(A, B) = M(B, A)$ for every pair of images (A, B) .

B. Restricting the Execution Time

When identifying images with fuzzy sets, it can happen that the supports of the sets contain far less elements than the universe. In such cases we can restrict the execution time dramatically by rewriting the fuzzy similarity measures. Consider for instance M_5 [2]:

$$M_5(A, B) = \frac{\min\{|A|, |B|\}}{\max\{|A|, |B|\}} = \frac{\min\{||A||, ||B||\}}{\max\{||A||, ||B||\}}$$

where $||C|| = \sum_{x \in X'} C(x)$ with $C \in \mathcal{F}(X)$ and $X' = supp A \cup supp B$. The old form requires $2 \cdot (|X| - 1)$ additions to calculate $|A|$ and $|B|$, followed by one comparison to determine $\min\{|A|, |B|\}$ and $\max\{|A|, |B|\}$ and finally one division. For the rewritten form we need $2 \cdot (|X'| - 1)$ additions, one comparison and one division. Hence the calculation of the new form will indeed be faster, since $|X'| \ll |X|$ implies that $2 \cdot (|X'| - 1) + 2 \ll 2 \cdot (|X| - 1) + 2$. Furthermore the execution times are the same if $X' = X$, which means that the new form is always at least as fast as the old one.

C. Pixel-based Similarity Measures

A digital greyscale image can be identified with a fuzzy set in a straightforward manner:

$$\begin{aligned} A(p) = 1 &\iff p \text{ is a white pixel} \\ A(p) = 0 &\iff p \text{ is a black pixel} \\ A(p) \in]0, 1[&\iff p \text{ is a greyscale pixel} \end{aligned}$$

for all $p \in P$, with P the universe of pixels ($P \subset \mathbb{N}^2$). Thereby applies: the greater the greyscale value, the lighter the pixel. We

call such measures pixel-based, because the universe consists of pixels.

On the internet, however, most images are colour images. We consider three approaches to construct similarity measures for colour images: (1) convert the images to greyscale images and use the aforementioned representation, (2) interpret the images as three greyscale images — one for each colour component — and use an aggregation operator [3] to combine the three resulting similarities, and (3) identify the images with $[0, 1]^3$ -fuzzy sets and generalize the fuzzy measures in order to make them applicable to these $[0, 1]^3$ -fuzzy sets [4].

Another problem that needs to be solved, is that the measures need to be able to cope with images of different resolutions. Measures based on the preceding representations lack that feature because the universe of the resulting fuzzy set depends on the resolution of the image, and only fuzzy sets in the same universe can be compared using a fuzzy similarity measure. We can get rid of this shortcoming by constructing similarity measures that use intermediate images of the same resolution, e.g. rescaled versions of the original images. An alternative solution consists of partitioning the images in parts of equal resolution, calculating the similarity between certain parts and aggregating the resulting similarities to one value [3].

D. Colour-based Similarity Measures

Colour-based similarity measures for images use fuzzy sets in a universe that contains colours instead of pixels. They can be constructed by considering image histograms. We define the histogram h_A for an image A as

$$h_A(c) = \sum_{p \in P} \delta(\text{bin}(c) - \text{bin}(A(p)))$$

for or all $c \in C' \subset C$, with C the universe of colours, δ the Dirac function ($\delta(0) = 1$ and $\delta(x) = 0$ for all $x \in \mathbb{Z} \setminus \{0\}$) and bin a mapping from C to $\{1, 2, \dots, |C'|\}$. The histogram h_A for an image A can be converted to a fuzzy set by normalizing it [2], [4], [5]:

$$H_A(c) = \frac{h_A(c)}{\max_{c \in C'}(h_A(c))}$$

for all $c \in C'$. We call the fuzzy sets obtained in this way pseudo-fuzzy histograms.

In addition to the pseudo-fuzzy histograms, which are fuzzy interpretations of the classical histogram, we also consider a fuzzy histogram. We define the fuzzy histogram \tilde{H}_A for an image A as follows [5]: $\tilde{H}_A = \bigcup_{c \in C_A} \tilde{C}_c$ where C_A is the set consisting of the colours appearing in A and

$$\tilde{C}_c(c') = \begin{cases} 1 & \text{if } d'_{c,c'} = \frac{d(\text{lab}(c), \text{lab}(c'))}{2.3} \leq 1 \\ \exp \frac{-(d'_{c,c'} - 1)^2}{2 \cdot \lambda^2} & \text{else} \end{cases}$$

for all $c' \in C$, with d the Euclidian distance, lab the function that maps each colour to its $L^*a^*b^*$ -coordinates and λ an adjustable parameter.

Various extensions of the classical histogram have been proposed in the literature (e.g. smoothed histograms [5]). Many

of these extensions are special cases of the extended histogram, which we define as follows:

$$\mathring{h}_A(c) = \sum_{c' \in C'} \sum_{p \in P} \delta(\text{bin}(c') - \text{bin}(A(p))) \cdot w_{A,c}(p)$$

for each c in C' . If $w_{A,c}(p) = \delta(\text{bin}(c) - \text{bin}(A(p)))$, $\forall p \in P$, then $\mathring{h}_A = h_A$. Hence the classical histogram is also a special case of the extended histogram.

E. Experimental Observations

We have constructed and implemented several pixel-based and colour-based similarity measures for images. In order to test and compare all these measures, we used each measure to arrange a subcollection of the Columbia Object Image Library (COIL) [6] according to the similarity with an example. Then we evaluated the resulting orderings using the Normalized Average Rank (NAR) [7]. For every measure we calculated multiple NARs, each corresponding to another example. We call the arithmetic mean of the NARs for a certain measure the Global Normalized Average Rank (GNAR) for that measure. The lower the GNAR, the better the performance of the corresponding measure.

In our experiments we observed very good results for the similarity measure that employs M_{I_3} [2] for comparing pseudo-fuzzy histograms in the Irb colour space ($I = (R + G + B)/3$, $r = R/(3 \cdot I)$ and $b = B/(3 \cdot I)$) using

$$\text{bin}(c_1, c_2, c_3) = \left[\sum_{i=1}^3 \left(\prod_{j=1}^{i-1} N_j \right) fl(c_i, N_i) \right] + 1$$

for all $(c_1, c_2, c_3) \in C$, with $N_1 = 4$, $N_2 = N_3 = 8$ and

$$fl(x, n) = \begin{cases} \lfloor x \cdot n \rfloor & \text{if } x < 1 \\ n - 1 & \text{if } x = 1 \end{cases}$$

for each $x \in [0, 1]$ and $n \in \mathbb{N}$.

IV. CONCLUSIONS

Fuzzy set theory proves to be very convenient for the modeling of similarity measures for images. Such measures can be used as a basis for an extension of traditional TBIR that incorporates content-based aspects without sacrificing the scalability of the system. Our prototype that implements this extension is available at <http://imilarity.berlios.de>.

REFERENCES

- [1] E. E. Kerre, "Vaagheids- en onzekerheidsmodellen," Course text, 2004.
- [2] D. Van der Weken, *Het gebruik en de constructie van similariteitsmaten in beeldverwerking*, Ph.D. thesis, Ghent University, 2004.
- [3] M. Detyniecki, "Numerical aggregation operators: State of the art," in *International Summer School on Aggregation Operators and their Applications*, 2001.
- [4] D. Van der Weken, V. De Witte, M. Nachtegael, S. Schulte, and E. E. Kerre, "The construction of quality measures for colour images: two approaches," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2006.
- [5] C. Vertan and N. Boujemaa, "Using fuzzy histograms and distances for color image retrieval," in *Challenge of Image Retrieval*, 2000.
- [6] S. Nene, S. Nayar, and H. Murase, "Columbia object image library: Coil-100," Tech. Rep. CUCS-006-96, Columbia University, 1996.
- [7] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593–601, 2001.