Ground-Truth-Less Comparison of Selected Content-Based Image Retrieval Measures

Rafał Fraczek[1], Michał Grega[1], Nicolas Liebau[2], Mikołaj Leszczuk[1], Andree Luedtke[3], Lucjan Janowski[1], and Zdzisław Papir[1]

[1] AGH University of Science and Technology
[2] Technische Universitaet Darmstadt
[3] Universitaet Bremen

Abstract. The paper addresses the issue of finding the best content-based image retrieval measures. First, the authors describe several image descriptors that have been used for image feature extraction. Then, the detailed description of a query by example psycho-physical experiment is presented. The paper concludes with the analysis of the results obtained.

1 Introduction

Several content-based image QbE (Query by Example) techniques have been presented during the last years. While developing real QbE applications, a question arises: which image retrieval method should be applied? The available benchmarks are commonly incomplete in terms of the number of image similarity measures. Furthermore, re-executing a benchmark usually imposes possession of a well-annotated database of images (ground-truth). In this paper the authors present a ground-truth-less comparison of several content-based image retrieval measures. Results of the comparison are applicable for several usage scenarios. Below, one of them is briefly presented: a QbE search system for a Web portal being a gateway to archives of media art.

QbE systems are based, in most cases, on features extracted from the media. Sets of features are generally referred to as "descriptors" and their instances are called "descriptor values". The descriptor values are the meta-data of the media. Some of the descriptor extraction methods are standardised in the MPEG-7 (Moving Picture Experts Group) standard [6].

The experiments described in this paper have been carried out in the context of the European project GAMA (Gateway to Archives of Media Art). The main goal of this project is to give public and multidimensional access to European collections of media art. One of the objectives within the GAMA project is to "provide sophisticated multilingual query performance and implement advanced search functionality". Obviously today we cannot imagine advanced search without QbE functionality. Nevertheless, there is not standard answer to the question what best results mean in QbE, neither for images nor video sequences. Moreover, we cannot find a ground truth since asking the question of the definition of similarity to different users we will get different answers. Therefore, we tried to find out what kind of distance metric is the best in the most general case.

P. Daras and O. Mayora (Eds.): UCMedia 2009, LNICST 40, pp. 101–108, 2010.

The approach to the problem was to perform a set of psycho-physical experiments in order to allow the subjects to vote for the best QbE method. "Best" is here understood as the most satisfying and closest to the users expectations. A set of QbE retrieval methods was designed and implemented and the results were presented to the subjects. The experiment was designed to be as simple and uncomplicated as possible in order to allow it to be performed on inexperienced subjects. The task of the subject was just to choose from a set of available result images the one most similar to the presented query image.

The QbE techniques are nowadays utilised by numerous applications and widely researched. The VICTORY project can be referenced [8] as a good example of advanced research in the area of QbE – in this case focusing on search for 3D objects. As mentioned before, executing an image retrieval benchmark usually imposes possession of an image ground-truth. Most of available databases contain a rather limited number of images (e.g. [7]). Another problem with a ground-truth databases is related to the way the images are described. Some databases contain free-text descriptions only, meaning that there are no straightforward metrics allowing for measuring computational similarity between two annotated images (e.g. TRECVid [10]). Finally, for vast image databases that are annotated by the community, the quality of annotations is moderate (e.g. Flickr). Consequently, the authors decided to carry out an experiment that would not require the possession of a ground-truth database. The image database used in the experiment consists of more than 33.000 images downloaded from the Flickr service with the accompanying user-generated keywords.

There is evidence that some subjective measures such as, for example, AAMRR (Average Normalised Modified Retrieval Rate) [3] coincide linearly with the subjective evaluation results [9]. The authors however decided to create and perform their own subjective psycho-physical experiment in order to avoid any error introduced by objective measures.

The rest of the paper is organised as follows. Section 2 presents the similarity measures included in the subjective experiment. Section 3 describes the methodology of the subjective experiments and the results are presented and discussed in Section 4. Section 5 concludes the paper and gives and insight into the further work on the topic.

2 Similarity Measures

The experiment considered several image retrieval metrics. The predominant group were metrics based on MPEG-7 visual descriptors. Cross-combinations of MPEG-7 descriptors have been considered as well. Furthermore, the PictureFinder content-based image retrieval algorithm has been applied. Moreover, the authors included a "Tag Metric", specifying the image-to-image distance, based on the degree of overlapping tags. Finally, a virtual "Random Metric" has been used in order to see how the real metrics actually differ from the totally random selections. The metrics (and the corresponding MPEG-7 descriptors, if applicable) have been described in detail below.

2.1 VS (PictureFinder)

VS [5] is a fast image retrieval library software for image-to-image matching. Similarity is measured based on spatial distribution of colour and texture features. It is especially optimised for fast matching within large data-sets.

VS applies a hierarchical grid-based approach with overlapping grid cells on different scales. For every grid cell up to 3 colours (in a quantised 10-bit representation in CIELab[1] colour space) and a texture energy feature are stored in the descriptor.

2.2 Metrics Based on MPEG-7 Visual Descriptors

MPEG-7 is an ISO/IEC standard for multimedia description defining a set of descriptors that are designed to extract specific information from the given content. This description allows for efficient multimedia content indexing and searching. In the experiment, the following MPEG-7 descriptors have been used.

DC (Dominant Colour) addresses the issue of finding major colours in the image. The descriptor quantises all colours present in the image and then the percentage of each quantised colour is calculated correspondingly.

SC (Scalable Colour) describes an image in terms of a colour histogram in HSV (Hue, Saturation, and Value) space. The descriptor representation is scalable in terms of both, bit representation accuracy and bin number. This feature makes the descriptor convenient for image-to-image matching.

CL (Colour Layout) is designed to efficiently represent the spatial colour distribution. The descriptor clusters the image into 64 (8×8) cells and the average colour of each block is derived. Finally, a DCT (Discrete Cosine Transform) is applied. The representation is very compact.

CS (Colour Structure) captures both, colour and structure information. The algorithm retrieves colour structure by analysing all colours in an 8×8 window that slides over the image. In consequence, the descriptor is able to distinguish between two images in which a given colour is present in identical amounts but where the structure of the groups of pixels having that colour is different.

EH (Edge Histogram) represents the spatial distribution of edges present in the image. These are four directional edges (vertical, horizontal, 45°, 135° and one non-directional edge. Then the image is divided into 16 (4×4) blocks and a five-bin histogram for each block is generated.

[1] Comission Internationale de l'Eclairage Lab.

2.3 TAG (Tag Metric)

The question "what does a similar image mean?" is difficult and very subjective. Nevertheless, instead of similarity definition one can ask what does a user see in the image. On the basis of such description obtained for two different images the similarity can be approximated since if the descriptions are similar the images should be similar too. Therefore, we decided to use the TAG, i.e. a metric based on a set of tags in order to know if other metrics have similar accuracy.

The TAG is not a perfect one mainly for two reasons. The first reason is the requirement of having the images tagged. Since we used images from the Flickr service our images were accompanied by the user-provided tags. The entered tags are far from perfect but it is almost impossible to obtain a database that is large and correctly described. The second problem is how to compute a distance between two different sets of tags. We decided to use the Jaccard similarity [2] denoted $\mathcal{J}_s(A, B)$, and defined as:

$$\mathcal{J}_s(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where, A and B are sets of tags of images a and b respectively and $|A|$ is the cardinality of A set.

Jaccard similarity has an interesting property that will be explained by an example. Let us assume we have three images. The first one a has 10 tags with "tree" tag in it. The second b has only 3 tags with "tree" tag also. We are wondering which of a or b image is closer to c image with only one tag "tree"? Jaccard similarity will show b image as closer since $\mathcal{J}_s(C, B) = 1/3$ and $\mathcal{J}_s(C, A) = 1/10$. The obtained result is correct since in A tree is one of 10 objects and in B one of three.

The TAG is not perfect since synonymous tags change the obtained results and different people can tag different numbers of objects in the same picture. Nevertheless, our main goal was to examine whether the TAG is more accurate than other considered metrics.

3 Description of Psycho-Physical Experiments

We performed tree different experiments on different groups of people and with two different scenarios. All results obtained were very similar. Therefore, we are presenting the final result without a detailed description of each experiment.

The final result presented in this paper has been obtained by analysis of two last experiments. The first experiment had a different scenario and was used to calibrate the user interface. Since the two other experiments where slightly different we are presenting results obtained in the last two experiments.

3.1 General Assumptions

QbE interfaces show some results, i.e. n images that are the most similar to the query image, where n is determined by the user interface since a subject has

to be able to see the results. The results order is determined by the comparing metric. We make two assumptions. The first one is that if a subject cannot find a (subjectively) similar image in the first n of them than the metric results are not correct. The second one is that $n = 10$, chosen based on our experience with existing QbE interfaces. Therefore, as a metric result we considered the set of the 10 most similar images. An important fact is that we did not consider if a picture was first or tenth, the only important property was to be in the first 10.

We analysed 12 different metrics:

1. 5 metrics based on MPEG-7 descriptors, introduced in Section 2.2
2. VS described in Section 2.1
3. TAG described in Section 2.3
4. Sum of ranks[2] obtained for EH and each MPEG-7-based metric (without EH) (4 different combinations)
5. Sum of logarithm of ranks obtained for each MPEG-7-based metric

The 13th metric was a random metric (i.e. a random image). We add it just to check subjects' reliability. On the other hand, we present only 7 images at a time since such an interface was the best in terms of the visual layout at all screen resolutions. Therefore, a subject can choose one of 7 different images (see Section 3.2). Each presented image is a result of two draws. First, the metric is drawn (for example the EH-based one) than from 10 images marked as 10 the most similar images one is drawn (in this case it is a random image from 10 the most similar images obtained for the EH-based metric). Since we considered 13 different metrics and each time we show 7 imagines not all of them were visible at once. Nevertheless, more than 2500 queries where answered therefore all possible combinations where properly represented.

Note that it is possible that there is no similar image in the set of 7 presented images. Therefore, we added an image that a subject can click if he/she cannot find a similar image among the presented. We added it as an image to make this answer identical to a similar image answer.

3.2 Experiment Setup

The user interface was implemented as a Web page in a form of a PHP (PHP: Hypertext Preprocessor) script. This allows for the execution of the experiment outside of the laboratory via Internet and gave access to a larger and more varied in the terms of age, occupation and nationality, group of subjects. The subjects were given a URL (Uniform Resource Locator) which directed them to the experiment.

3.3 Experiment Execution

The experiment was performed at two stages. First the subject was given the instructions and the general purpose of the experiment was explained. Information about the subject was collected, such as age, gender and nickname. Also

[2] Rank is a metric giving 1 to the most similar image, 2 to the second one etc.

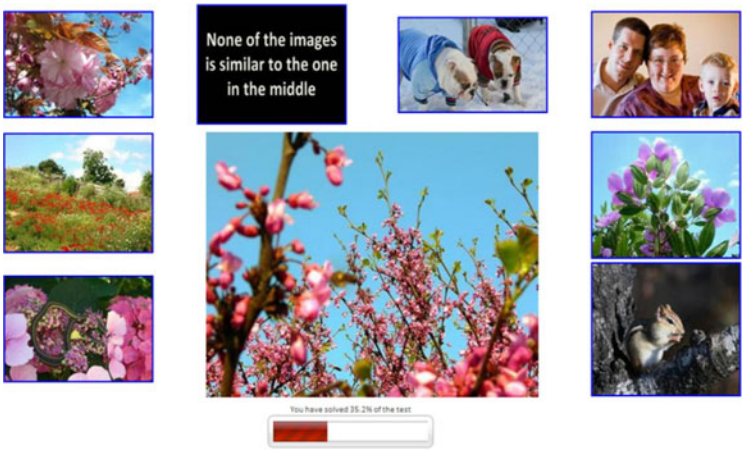

Fig. 1. The web-based interface for the psycho-physical experiments

a 6-step colour blindness test was performed in order to identify colour-blind subjects.

The second stage was the experiment itself. 7 randomly chosen images were presented to the subject (Fig. 1 and Section 3.1). The subject's task was to choose from the small images the one most similar to the middle, large one. Subjects could also chose "no similarity" answer (see Fig. 1). Step five was repeated 300 times, but the subject was free to end the experiment at any time.

The middle image was the query and the surrounding images were the query results obtained with the different QbE techniques. So, the subjects were performing a vote for the QbE method that was most satisfying.

4 Analysis of Results

As the experiment could be terminated by a subject any time, we obtained a different number of answers from each subject. Therefore, for each subject a distribution of answers (i.e. probability of choosing any metric) was computed and analysed. Additionally we removed all subjects answering less than 50 queries. We collected 31 results and computed confidence intervals with $\alpha = 0.05$ [1]. The results obtained are shown in Figure 2.

The probability of the "no similarity" answer was the highest and reached 31%. Note that if lots of answers are "no similarity" probably the database was too small to have a similar picture. Since we are interested in comparing different descriptors and metrics quality and not the database quality we are not showing this value in the plot.

We considered two additional metrics. The first one is random metric i.e. a random image was shown. The random metric enables to conclude if the other metrics are better than random. The second one is the TAG which enables to

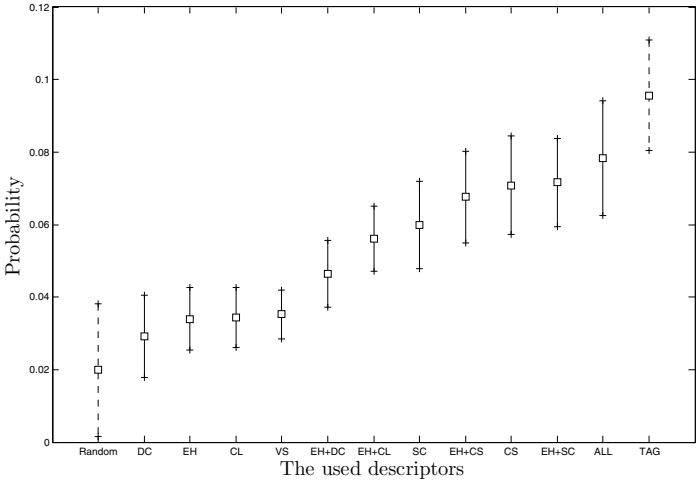

Fig. 2. The obtained results with the confidence intervals, solid lines — descriptors' combinations; dashed lines — the random metric and the TAG

compare MPEG-7-based metrics with descriptive metric used in classic search systems.

We analysed all MPEG-7-based metrics with the random metric and TAG by t-test [1] with $\alpha = 0.05$. The results show that metrics based on EH, DC, CL descriptors as well as the metric based on VS, are not statistically different from the randomly chosen image. On the other hand, only one metric (based on all MPEG-7 descriptors) is not statistically different from the TAG. Therefore we chose it as the best and implemented in GAMA project. However, CS is just slightly worse. Moreover, the CS value is computed on the basis of just one descriptor and thus computationally cheaper.

5 Conclusions

The paper addressed the issue of finding the best content-based image retrieval measures. First, the authors described several image descriptors that have been used for image feature extraction. Then, a detailed description of query by example psycho-physical experiment has been presented. The results show that the metric based on MPEG-7 CS Descriptor is the most commonly chosen among metrics based on single descriptors. Nevertheless, it is outperformed by the metric being a combination of various MPEG-7 descriptors as well as by manual tagging.

In the GAMA project the metric based on CS descriptor will be used as a prefilter for the image and video QbE systems. This will allow for fast and accurate search in the vast repository of media art.

As a further work, the results will be used in a project related to the application of the QbE media search in P2P (Peer-to-Peer) overlays. In order to create a QbE system for the P2P overlay a decision has to be made on the selection of a descriptor or a set of the descriptors. It is planned then to implement the QbE mechanism in the unstructured and structured P2P overlays on the example of the Gnutella and CAN (Content Addressable Network) overlays. The implementations will be done in the environment of the PeerfactSim.KOM simulator [4].

Acknowledgements

The work presented in this paper was supported by the European Commission, under the projects: "CONTENT" (FP6-0384239) and "GAMA" (ECP-2006-DILI-510029). The authors thank panel of subjects for their efforts.

References

1. Nist/sematech e-handbook of statistical methods
2. Arasu, A., Ganti, V., Kaushik, R.: Efficient exact set-similarity joins. In: Proceedings of VLDB 2006 (2006)
3. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface. John Wiley and Sons Ltd., Chichester (2002)
4. Graffi, K., Kovacevic, A., Steinmetz, R.: Towards an information and efficiency management architecture for peer-to-peer systems based on structured overlays. Technical report, Multimedia Communications Lab KOM, Technische Universitaet Darmstadt (2008)
5. Hermes, T., Miene, A., Herzog, O.: Graphical Search for Images by PictureFinder. Multimedia Tools and Applications. Special Issue on Multimedia Retrieval Algorithmics (2005)
6. ISO/IEC. Information technology – multimedia content description interface. ISO/IEC 15938
7. Li, Y.: Object and concept recognition for content-based image retrieval. PhD thesis, University of Washington (2005)
8. Mademlis, A., Daras, P., Tzovaras, D., Strintzis, M.G.: 3d volume watermarking using 3d krawtchouk moments. In: VISAPP (1), pp. 280–283 (2007)
9. Ndjiki-Nya, P., Restat, J., Meiers, T., Ohm, J.R., Seyferth, A., Sniehotta, R.: Subjective evaluation of the MPEG-7 retrieval accuracy measure (ANMRR). Technical report, ISO/ WG11 MPEG Meeting, 200
10. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR 2006: Proceedings of the 8th ACM international workshop on Multimedia information retrieval (2006)