1-1-2007

Perceived similarity and visual descriptions in content-based image retrieval

Yuan Zhong
University of Wollongong, yz505@uow.edu.au

Lei Ye
University of Wollongong, lei@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Philip Ogunbona
University of Wollongong, philipo@uow.edu.au

Perceived similarity and visual descriptions in content-based image retrieval

Abstract

The use of low-level feature descriptors is pervasive in content-based image retrieval tasks and the answer to the question of how well these features describe users' intention is inconclusive. In this paper we devise experiments to gauge the degree of alignment between the description of target images by humans and that implicitly provided by low-level image feature descriptors. Data was collected on how humans perceive similarity in images. Using images judged by humans to be similar, as ground truth, the performance of some MPEG-7 visual feature descriptors were evaluated. It is found that various descriptors play different roles in different queries and their appropriate combination can improve the performance of retrieval tasks. This forms a basis for the development of adaptive weight assignment to features depending on the query and retrieval task.

Keywords

perceived, similarity, content, image, visual, descriptions, retrieval

Disciplines

Physical Sciences and Mathematics

Publication Details

Zhong, Y., Ye, L., Li, W. & Ogunbona, P. (2007). Perceived similarity and visual descriptions in content-based image retrieval. The IEEE International Symposium on Multimedia (pp. 173-180). IEEE Computer Society Press.

Perceived Similarity and Visual Descriptions in Content-Based Image Retrieval

Yuan Zhong, Lei Ye, Wanqing Li and Philip Ogunbona
School of Computer Science and Software Engineering,
University of Wollongong,Australia
Email: {yzhong, lei, wanqing, philipo}@uow.edu.au

Abstract

The use of low-level feature descriptors is pervasive in content-based image retrieval tasks and the answer to the question of how well these features describe users' intention is inconclusive. In this paper we devise experiments to gauge the degree of alignment between the description of target images by humans and that implicitly provided by low-level image feature descriptors. Data was collected on how humans perceive similarity in images. Using images judged by humans to be similar, as ground truth, the performance of some MPEG-7 visual feature descriptors were evaluated. It is found that various descriptors play different roles in different queries and their appropriate combination can improve the performance of retrieval tasks. This forms a basis for the development of adaptive weight assignment to features depending on the query and retrieval task.

1 Introduction

Current text-based image search services do not offer the user the ability to provide their query in a manner that describes the content of the image or target images they have in mind. Content-based image retrieval systems promise to solve this problem through the use of descriptors based on visual features extracted from an example image or images [1]. Several questions are raised by this paradigm. What features are effective in describing image content? How well do these features mimic the human perception of image content? On what level is there a similarity between the descriptors and human description? Indeed, how many of such descriptors are required and how should they be combined to retrieve images close to what the user desires? There is also the question of how many example images are required to provide a description of the desired target images.

These questions need to be answered in a principled manner if the exponential growth of images and videos on the World Wide Web is to be turned into viable search and retrieval business. Despite the unavailability of complete answers to these questions, current content-based retrieval schemes can perform well for some categories of images and poor performance have been recorded in other cases depending on the nature of the image database and the specific visual features used.

In this paper we shed some light to provide better understanding of some of the questions and provide results of the experiments we have conducted. In Section **??** a concise overview of current thinking on human visual perception is provided. Human subjects were invited to participate in a perception experiment that tries to elicit how humans perceive similarity in images. These experiments are described in Section **??**. In order to understand and evaluate the degree of correlation between human description of similarity and the extent to which visual feature descriptors capture this description, we use some MPEG-7 descriptors [2] on a dozen categories of images. The experiment and the results are described in Section 3. The effect of appropriate weighted combination of the features is explored in Section 4. Conclusions are offered in Section 5

2 Psychological Experiments on Perceived Similarity of Images

The human mind can be considered as an information processing system [3] that interacts with the external world, thus making us active processors of information. Pre-acquired knowledge and other neural processes are combined to interpret the sensations impinging on the mind. Thus, when humans pose an example image as a query in a retrieval task, there is the understanding that the given image embodies the description of some previously sensed imagery. Any computer system that will aid the human user in a retrieval task must of necessity consider the human visual perception when modelling the image retrieval process. The modelling process is made complicated because, for a given image, it is expected that users will respond and gain different perception, especially from an interpretation or description viewpoint, because of differences in pre-acquired knowledge and overall perception.

0-7695-3084-2/07 $25.00 © 2007 Crown Copyright
DOI 10.1109/ISM.Workshops.2007.38

Perception is part of human intelligence underlined by a hierarchical structure of attentional stages. In this model of the human visual perception, people will "capture" the low-level features of an image in a "bottom-up" processing and then combine them into objects through the help of the *attentive* process. The bottom-up processing begins with external input and travels "up" through the cognitive system. In bottom up processing, the fundamental units of perception is at the level of features, not at the level of objects. Features are combined into more complex objects by attention. However, Hochstein and Ahissar [4] proposed that explicit vision advances in reverse hierarchical direction, as shown in perceptual learning. They argued that conscious perception begins at the top of the hierarchy, gradually returning downward as needed. In this model, one first sees unified whole images, then the features are perceived through attention.

In order to retrieve images based on human visual perception, we need to determine which features best match the cues used in human perception. Processes in the pre-attentive stage are responsible, chiefly, for the perception of colour and edges. In experiments reported in [5], Treisman used a visual search task to show which features were important at a perceptual level. In other words, which features formed the building blocks of human visual perception. The result showed that the features include the so called primitive features, namely, colour, orientation, curvature and line intersections. Individual features are combined into objects and it is recognized that feature combination requires attention to bind the features together. Thus, from a computational viewpoint we expect that extracted features that encode colour, texture and shape will mimic cues used in aspects of the human visual perceptual process.

A computer-based experiment is designed to evaluate the human perception of similarity. Apart from being computer platform agnostic, care was taken to ensure that the interface presented to participants is identical on all screen sizes available in the laboratories used to conduct the experiment. There are 12 people who participated in these experiments.

Each participant is presented with an application window partitioned appropriately for each experiment. The participant is shown one example image at the upper part of the application window and $\mathbf{N}$ candidate images at the lower part of the window. They are required to select $\mathbf{K}$ images from the candidate images which they think are similar to the example image by clicking on them. Due consideration was given to the optimum values of $\mathbf{N}$ and $\mathbf{K}$. It was found that values of $\mathbf{N} = \mathbf{20}$ and $\mathbf{K} = \mathbf{5}$ were adequate and did not provide cognitive overload to participants. The experiments are repeated 10 times for each participant. Each time, an example image and a set of candidate images are presented to the participant. The sample image and the candidates are selected from a categorized image database. The categories

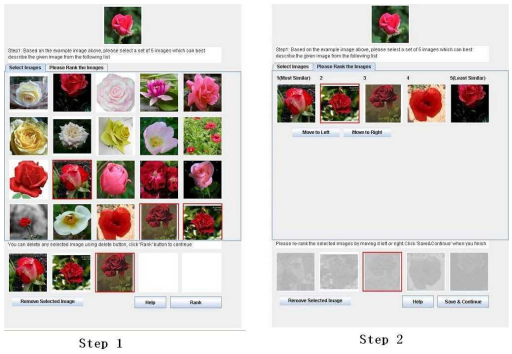


Figure 1. User Interface of the Experiment System

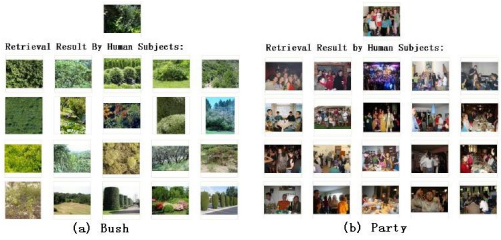


Figure 2. Experimental Results: Bushes

include Beach, Bushes, Cars, Flowers, Horses, Mountains, Opera House, Party, Ships in the Ocean and Sunset.

Each set of candidate images contains 20 images from one category, which may have various visual features. For example, the 20 images in the category of "Car" may contain cars with different shapes, colours and sizes. The participants are required to choose 5 images from them that are believed to be similar to the example image. Figure 1 shows the user interface of the computer-based experiment system.

The participants produce their results in 2 steps. In Step 1, participants are required to select 5 images that they consider to be similar to the example. In Step 2, they can reorder (or rank) the 5 selected images in order of similarity to the example, so that the first (or leftmost) image is the most similar.

Figure 2(a) shows an example of the results for the category "Bushes". The image order is determined by the number of participants that have selected the image as being similar to the example image. The most selected image by all subjects is displayed first. Therefore, the displayed image mosaic in the specific order is considered as the intended retrieval results when the example image is used as the query image by humans.

In order to compare the results to the results by computer retrieval systems and evaluate the visual descriptors in following sections, a set of relevant images to the query image

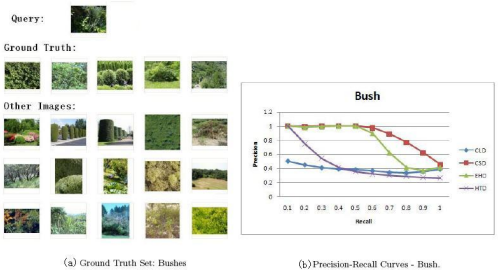


Figure 3. Ground truth set for "Bush" and the retrieval result by different descriptors

is determined by the majority votes. The rest are considered as irrelevant images. The relevant images determined by human subjects are used as ground truth sets. As an example, the ground truth set of Bushes is shown in Figure 3(a).

3 Perceived Similarity and Similarity Measured by MPEG-7 Descriptors

Generally speaking, people search for similar images based on the perceived visual content of images and using a so called perceived similarity. However, visual features perceived and used by people to judge the similarity are subjective and their nature not completely understood. Some low-level visual descriptors are used in content-based image retrieval (CBIR) systems to represent image features including colour, texture and shape. In particular these feature descriptors, for example as specified in the MPEG-7 standard [2], may not be the same as the features used by human beings. The ultimate goal of a CBIR system design is to rank the target images so that they have similar ranking as would a human user.

Common visual features used by machines, such as MPEG-7 descriptors, are designed to describe aspects of the visual characteristics of the image content. The effectiveness of these descriptors is to be evaluated against the results by human subjects. The ground truth sets for all queries are created as described in the previous section. Similar images in the ground truth sets are ranked according to perceived similarity. In this section, similar images ranked according to visual similarity based on visual descriptors are evaluated against the ground truth images. Four MPEG-7 visual descriptors, namely, CLD, CSD, EHD and HTD are used individually and in combinations as the features for similarity measurement. Precision and recall graphs are used to evaluate the performance.

Two results from various image categories are presented and discussed as follows.

Figure 4. Retrieval Results by CLD, CSD, EHD and HTD for "Bush"

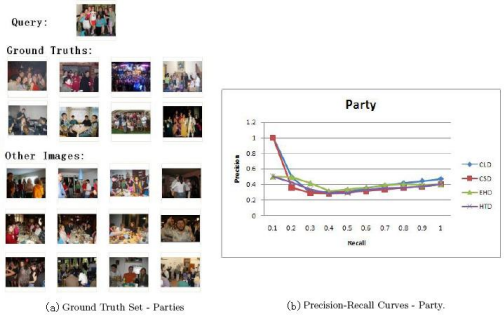


Figure 5. Ground truth set for "Parties" and retrieval results for different descriptors

3.1 Bush

Figure 3(b) shows the curves for category "Bush". The performances of descriptors CSD and EHD are much better than that of the other two descriptors. Therefore, features described by CSD and EHD can be used to rank the images in this category. They are considered more important than others for this category.

Figure 4 shows the retrieval results using the four descriptors. The ground truth has been shown in Figure 3(a)

3.2 Party

Figure 5(b) shows the performances for the "Party" image category. All the four descriptors do not give a good performance for this category. No descriptor could describe images from this category well enough. Note that it is difficult for humans to judge the similarity in this case, as shown in Figure 2(b). The retrieval results by human participants

Figure 6. Retrieval Results by CLD, CSD, EHD and HTD for "Party"

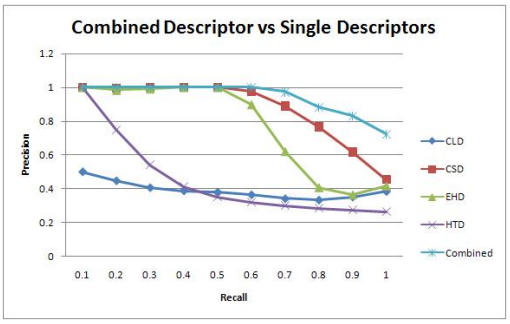


Figure 7. Combine the features by putting greater weights on more important features in category "Bush".

have diverse selections of images with perceived similarity, which results in a larger number of ground truth images as shown in Figure 5(a). The retrieval results from the four descriptors are shown Figures 6. They not only result in poor performances but also very different rankings. This query is considered difficult for both humans and machines.

4 Effect of Weights in Combining Visual Descriptors

In a CBIR system using multiple features, it is important to understand the role played by individual descriptors in various queries and how they would impact the overall retrieval result.

As presented in Section 3, it is clear that the performance of individual descriptors is limited. This section will investigate the effects of weights of individual descriptors on the retrieval performance. It is expected that a proper combination of features used in retrieval systems could result in improved performances. The general idea is to assign a higher weight to a feature that is more important to the query. The question arises as to how the importance will be assessed and its relevance to the query.

Experiments are designed to evaluate the effects of weight assignments based on the performance of individual descriptors.

We tested the system performance by combining all the four MPEG-7 descriptors. It is straightforward to use equal weights under the assumption that the different descriptors take on the same importance for retrieval purpose. However, in most cases, the features play different roles depending on the nature of the image used as query. Given a single query image it is difficult to ascertain, in any objective manner, the relative importance of the features. In an image retrieval scheme based on multiple query images it is possible to devise means of assigning proper weights to each feature.

In our experiment, we assign higher weight to descriptors which are considered more important for the category (from the previous single descriptor experiments), and lower weights to the descriptors which do not play important roles during the retrieval. The system performance is also evaluated using precision and recall curves. These curves are compared with the curves generated by using single descriptors. Different weighting methods are tested and the result curves are shown in the following paragraphs.

The curves in Figure 5(b) show that the descriptors CSD and EHD are considered more important than CLD and HTD for the category "Bush", which means the CSD and EHD contribute more than the other two descriptors. Based on the discussion above, if we combine the four descriptors together, they should be assigned more weights than the others. Since we need to calculate the distance between query and the database images, we assign the weights to the distances of different descriptors and use the combined distance to rank the images. We set the weights of CSD and EHD distances to 0.3 and the weights of CLD and HTD distances to 0.2, which makes the sum of all the weights be 1. Then we perform the retrieval using the combined distances. The precision and recall curve of the new retrieval are compared with the curves obtained by using single descriptors in Figure 7. As we can see from the figure, the retrieval using combined features has a higher performance than using any of the single descriptors.

This demonstrates that proper combination of descriptors can improve the performance. Figure 8 shows the retrieval result and this can be compared to the ground truth set (in Figure 3(a)) to see the improvement gained in the retrieval results.

Experiment has been conducted on the same category to evaluate the case of equal weight assignment. The preci-

Figure 8. Retrieval results with combined descriptors for “Bush”

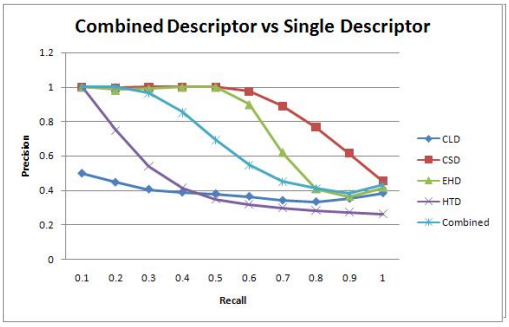


Figure 9. Combine the features by putting equal weights to the descriptors in the category ”Bush”.

sion and recall curves are compared in Figure 9. The curve of the retrieval using equally weighted combined features is not indicative of superior performance. Although it is better than using a single descriptor(CLD or HTD), it is not as good as using the supposedly better single descriptors (CSD or EHD). This suggests that using equally weighted combination of features may not improve the performance over what is achievable by single good descriptor.

In another experiment the effect of using the wrong weight is investigated. We set the weights of CSD and EHD to be 0.2 and the weights of CLD and HTD to be 0.3. In other words we assign the higher weights to the less important features. The curves showing the performance are in Figure 10. The result indicate that if we combine the features in a manner that assign higher weights to less important features, the retrieval performance may be inferior to the cases of using single descriptors.

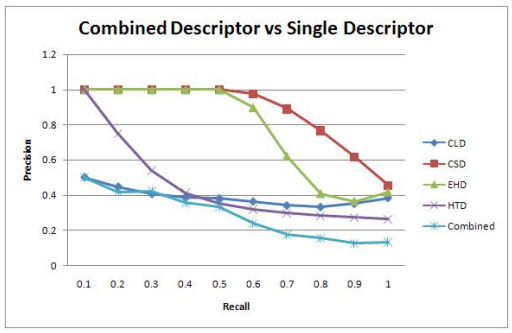


Figure 10. Combine the features by putting greater weights on less important features in category ”Bush”.

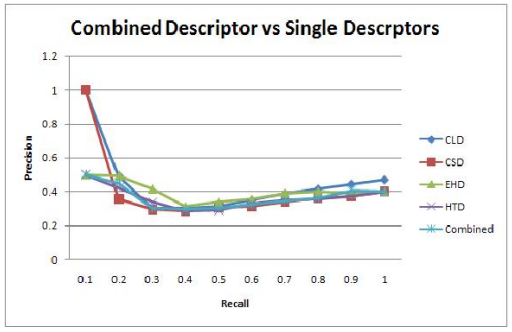
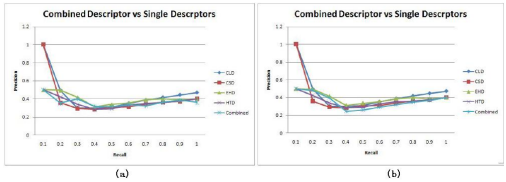


Figure 11. Combine the features by assigning equal weights to all descriptors in category ”Party”.

The discussion above focuses on the situations where there are significant features to describe the images in the category. However, for some of the categories, there are no significant features that can be easily captured to describe the images. For example, from the standard deviations of the distances calculated for each image category, it is found that the standard deviation value for category ”Party” is relatively very small for all the four descriptors. This indicates that irrespective of the feature space, images in category ”Party” are very similar to each other. This will explain why the “similar” images selected by the human participants are very different. In this case, using any of the single descriptors could not have resulted in a good performance.

We combine the descriptors to perform the retrieval in the category ”Party”, and compare the result with retrieval using single descriptors. As shown in previous results, we could not find any descriptor that is more significant that the others. Thus, we assign equal weights to all the four descriptors (0.25 each). The curves are shown in Figure 11. It is clear that using combined features in category ”Party”



(a) (b)

Figure 12. In category "Party", (a)Weights of CSD and EHD are set to 0.3 and weights of CLD and HTD are set to 0.2 (b)Weights of CSD and EHD are set to 0.2 and weights of CLD and HTD are set to 0.3

does not give a good result as well. It may be concluded that in categories that do not have significant features, combining the features does not provide any apparent improvement. It is worth noting that if any of the selected features had captured a description of human faces and the number of such faces detected, perhaps some level of discrimination could be achieved.

We also assigned the same two sets of weights as was used in "Bush" category, to features in this category. The system performance of using combined features are compared with that of using single descriptors in Figure 12.

As we can see from these curves, assigning different weights to different features does not help to improve the performance of the image retrieval system in this category. This also demonstrates that for the categories without significant features, the performance of retrieval system can hardly be improved by using weights on the features. It is important to select an appropriate feature that captures the salient characteristic of the images under consideration. We also note that although assigning higher weights to important features can improve the system performance to a certain degree, there is no fixed set of weights which is suitable for all situations. The weights of features varies from one category to another as is to be expected. To improve the performance of an image retrieval system, a weighting method which can assign different weights to the features based on categories is needed. Unfortunately, a single query image can hardly provide all the information about the image category. However, the use of multiple images as the query can reflect the attributes of target image category.

5 Conclusions

A series of psychological experiments have been conducted to collect data on how human subjects judge image similarity. Based on the experimental results, some common visual descriptors are evaluated against the results by human subjects. It is found that various descriptors play different roles in different queries and their appropriate combination can improve retrieval performance. There are no fixed weight assignments for all queries or categories. The weights of different descriptors change from one category to another.

In order to improve the performance of image retrieval systems, we can combine appropriately weighted features for each individual query. The use of multiple images as the query will facilitate the derivation of the weights. In this paper, analytically derived weight assignment has not been considered, however, some efforts in this direction are reported in [6].

Acknowledgment

The authors would like to thank Dr. Stephen Palmisano and Dr.Simone Favelle, School of Psychology, University of Wollongong, for valuable discussions and their guidance in the design of the psychological experiments.

References

[1] S. Deb and Y. Zhang, "An overview of content-based image retrieval techniques," in *International Conference on Advanced Information Networking and Applications*. Los Alamitos, CA, USA: IEEE Computer Society, 2004, pp. 59–64.

[2] T. Sikora, "The MPEG-7 visual standard for content description - an overview," *IEEE Trans on Circuits and System for Video Technology*, vol. 11, no. 6, June 2001.

[3] J. Andrade and J. May, *Cognitive Psychology*, 2nd ed. London and New York: BIOS Scientific Publishers, 2004.

[4] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies," *Neuron*, vol. 36, pp. 791–804, December 2002.

[5] A. Tresisman, "Features and objects: The fourteenth bartlett memorial lecture," *Quarterly Journal of Experimental Psychology*, vol. 40A, pp. 201–237, 1988.

[6] Y. Zhong, "A weighting scheme for content-based image retrieval," Master's thesis, School of Computer Science and Software Engineering, University of Wollongong, 2007.