Multimodal Image Retrieval Based on Keywords and Low-Level Image Features

Miran Pobar(✉) and Marina Ivašić-Kos

Department of Informatics, University of Rijeka, R. Matejčić 2, Rijeka, Croatia
{mpobar,marinai}@inf.uniri.hr

Abstract. Image retrieval approaches dealing with the complex problem of image search and retrieval in very large image datasets proposed so far can be roughly divided into those that use text descriptions of images (text-based image retrieval) and those that compare visual image content (content-based image retrieval). Both approaches have their strengths and drawbacks especially in the case of searching for images in general unconstrained domain. To take advantage of both approaches, we propose a multimodal framework that uses both keywords and visual properties of images. Keywords are used to determine the semantics of the query while the example image presents the visual impression (perceptual and structural information) that retrieved images should suit. In the paper, the overview of the proposed multimodal image retrieval framework is presented. For computing the content-based similarity between images different feature sets and metrics were tested. The procedure is described with Corel and Flickr images from the domain of outdoor scenes.

Keywords: Image retrieval · Multimodal query · Content-based similarity

1 Introduction

To help dealing with a huge number of images produced daily, different approaches for image search and retrieval have been proposed that can be roughly divided into those that use text descriptions of images (text-based image retrieval) [1, 2] and those that compare visual content (content-based image retrieval) [3, 4].

In content-based image retrieval approach images are retrieved and ranked based on visual similarity to a query image. The similarity between images is commonly computed based on low-level features so the consequence is that the similarity of semantics actually relies on the similarities of colors and other low-level features. On the specific domains such as criminalistics and medical diagnostics when search is performed among images that all have the same semantics, e.g. chest x-rays this approach gives excellent results because user is searching for exactly that query image or the most similar ones. When searching for images in general, it is more likely that the user is looking for images that are similar to the query image but differ in some aspect, i.e. the user is not actually looking for images that are as similar as possible to the query image but those images that semantically match the query image.

© Springer International Publishing Switzerland 2015
J. Cardoso et al. (Eds.): KEYWORD 2015, LNCS 9398, pp. 133–140, 2015.
DOI: 10.1007/978-3-319-27932-9_12

In most everyday cases where image semantics matter, image retrieval based on text has appeared to be easier and more suitable for image search and retrieval. This is because it is always possible to write a keyword-based query, and image examples are not always available. However, to be able to retrieve images using text, they must be labeled or described in the surrounding text, and most images are not.

Still, in some cases content-based image retrieval can be preferred to keyword-based search, especially when searching for images with very specific visual appearance that may be difficult to describe with a few keywords. Multimodal retrieval that uses both keywords and visual properties of images appears as a solution [4]. The approach in [5] uses either complex text queries describing relative configurations of objects in images, or use queries of different modalities (text, sketches and images) that are converted into a common semantic representation used for retrieval.

In this paper, we propose a multimodal image retrieval framework that integrates keyword-based image search with content-based ranking according to the visual similarity to a query image. In this way, users can provide both a reference example image to which the results should be similar, and keywords to specify the desired semantics of retrieved images, which can be different than in the example image. Visual similarity between candidate images and the example image is computed based on low-level visual features extracted from images. To present the perceptual information about the image, pixel-based and structure-based feature descriptors are used.

The overview of the proposed multimodal image retrieval framework is presented in Sect. 2. Section 3 presents the feature sets used for computing the content-based similarity between images and the content-based similarity measures are introduced in Sect. 4. The details about the experiment with examples from the outdoor image domain are given in Sect. 5 with the conclusion in Sect. 6.

2 Overview of Multimodal Image Retrieval Framework

The proposed pipeline for multimodal image retrieval is shown in Fig. 1. The user provides a query to the system consisting of a keyword and an example image. The expected results are images that match the given keyword because they contain a particular word in the description or annotation and visually resemble the example image. The system first retrieves all the images from the image database that satisfy the query keyword. Then, low-level visual features are extracted from image regions of the example and retrieved images. Low-level visual features are then used to compare the example and all retrieved images by computing a similarity score. The retrieved images are then ranked by the similarity to the example image, yielding the final results that are presented to the user.

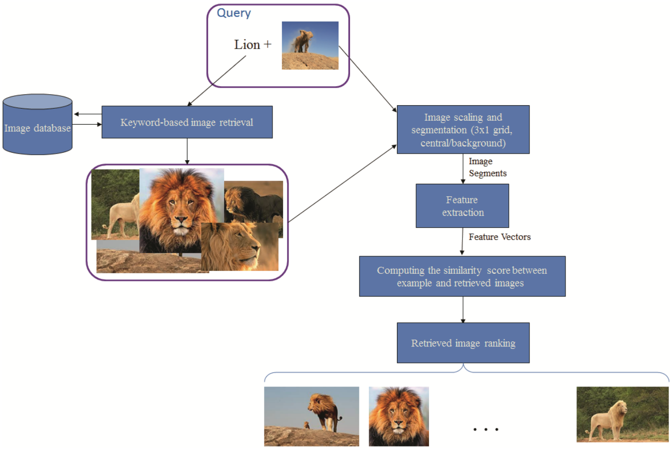
Query
Lion +
Image database
Keyword-based image retrieval
Image scaling and segmentation (3x1 grid, central/background)
Image Segments
Feature extraction
Feature Vectors
Computing the similarity score between example and retrieved images
Retrieved image ranking
. . .

Fig. 1. Multimodal image retrieval pipeline.

3 Features

Most systems for content based image retrieval and image annotation perform feature extraction as a preprocessing step of presenting perceptual information about the image, obtaining global and local image features like dominant color or color histogram, structure, etc.

Here we have considered color histograms as pixel-based descriptors and GIST [6] as structure-based descriptors for computing the visual similarity of images during content-based image retrieval. Pixel-based descriptors are used because they are robust in position, translation and rotation changes and are useful for rapid detection of objects in image databases. Here, they are computed on the whole image, but also on two centrally symmetric regions to capture the information about the possible central image object and regions obtained by applying a 3×1 grid, to preserve the information about the color layout of an image.

To calculate the color histograms, the images were converted from RGB color space to indexed color, where the palette consists of 256 colors obtained by uniformly quantizing the RGB color space. The choice of number of colors in the palette depends on the number of desired bins in each histogram. Too few colors will lose accuracy by overestimating the overlap region, and too many colors will lose accuracy by creating individual bins with no values except in a densely populated sample space.

To represent coarse spatial information the GIST image descriptor was used. It is a structure-based image descriptor [7] that refers to the dominant spatial structure of the scene characterized by properties of its boundaries (e.g., the size, degree of openness, perspective) and its content (e.g., naturalness, roughness) [5]. The spatial properties are estimated using global features computed as a weighted combination of Gabor-like multi-scale oriented filters. In our case, we used 4×4 encoding samples in the GIST descriptor within 8 orientations per 4 scales of image components, so the GIST feature vector has 512 components.

For the content-based similarity calculation for image retrieval, subsets of features containing color histograms, and GIST descriptor is used.

The features were extracted from images that were sized 128 × 192 pixels or 192 × 128 pixels in the case of the Corel dataset. For the Flickr dataset, the images were rescaled to the width of 256 pixels before feature extraction.

4 Content-Based Similarity Ranking

To present to the user the most visually similar images to the query image, content-based similarity ranking is performed. Visual features are extracted from the query image and are compared with the low-level feature vectors extracted from all images obtained as results of keyword-based query. A similarity score between each retrieved image and the query image is computed as the distance between the corresponding feature vectors. The obtained images are then ranked by the similarity to the query image.

To compare the visual similarity of images, we used different features and suitable metrics. For histogram comparison we used the Bhattacharyya distance [8], histogram intersection [9], and the chi-squared histogram matching distance [8] and for distance between GIST features the Euclidean distance. Other metrics for comparing visual similarity can also be used, depending on the feature set, see [8] for a comprehensive review.

The Bhattacharyya distance is an appropriate distance measure for discrete probability distributions or normalized histograms p and q over the same histogram range, and it is defined as:

$$D_B(p,q) = -\ln(BC(p,q)), \tag{1}$$

where:

$$BC(p,q) = \sum\nolimits_{i=1}^{n} \sqrt{p_i q_i}, \tag{2}$$

is the Bhattacharyya coefficient, and for histograms p and q, n is the number of bins and p_i and q_i are the i-th bin values of histograms p and q.

The Bhattacharyya coefficient is a measurement of the amount of overlap between two histograms and can be used to determine the similarity of the two sample images being considered.

For Bhattacharyya distance, low scores indicate good matches, with perfect match being 0, and high scores indicate bad matches, with infinite value for total match.

The intersection of two histograms is the same as the minimum misclassification or error probability, which is computed as the overlap between two probability density functions [10]:

$$\text{histint}(p,q) = \sum\nolimits_{i=1}^{n} \min\left(p_i, q_i\right), \tag{3}$$

where p and q are the compared normalized histograms, and n is the number of histogram bins.

To use histogram intersection as a distance measure, the inverse is computed:

$$D_I(p,q) = 1 - \text{histint}(p,q)\,. \tag{4}$$

If both histograms are normalized to 1, then 0 indicates perfect match and 1 a total mismatch.

The chi-squared distance is defined as:

$$X^2(p,q) = \frac{1}{2}\sum\nolimits_{i=1}^{n} \frac{[p_i - q_i]^2}{p_i + q_i}, \tag{5}$$

where p, q, and n have the same meaning as in (3). For the chi-square distance, a perfect match is 0 and a total mismatch is unbounded and depends on the size of the histogram.

5 Experiments

The experiments of the keyword-based, content-based and multimodal image retrieval were performed on the Corel image dataset [11] and on a set of images from the Flickr website. We used images in the Corel dataset related to outdoor scenes, labeled with one or more keywords from a vocabulary of 27 keywords pertaining to natural and artificial objects, such as ‘airplane’, ‘bird’, ‘lion’, ‘train’ etc. The Flickr images were obtained by using the same set of 27 keywords as in the Corel dataset to query the Flickr website. For each of the chosen keywords, 100 most relevant image results were collected, resulting in a dataset of 2700 images belonging to 27 classes. Each of the Flickr images was annotated with the query keyword, but possibly also with other keywords or text descriptions. These labels are used for image retrieval when a user provides a keyword query.

An example of keyword based retrieval results from the Flickr image database is shown in Fig. 2. It shows the top nine images obtained for the keyword “tiger”.

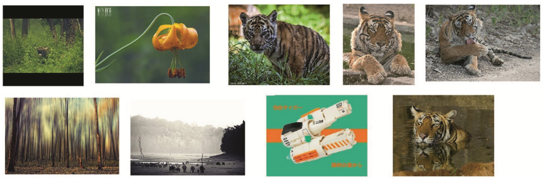

Fig. 2. Top nine results for keyword “tiger” obtained by text-based image retrieval.

To perform content-based similarity ranking and image retrieval, low-level visual features are extracted from Corel and Flickr images, as detailed in Sect. 3. The considered feature descriptors and combinations of appropriate distance measures were tested in isolation by querying the image databases using only the query images, and ignoring the text labels. As an example, Fig. 3 shows the results of visual similarity ranking of images when the color histogram is used as feature descriptor with histogram intersection distance measure and no keyword is specified. Similar results are obtained with other features and measures.

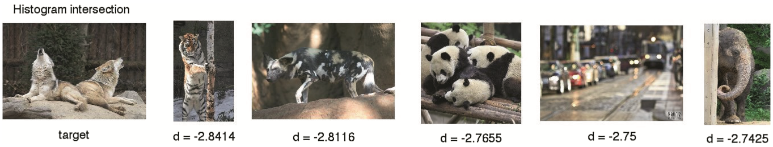

Fig. 3. Top five most similar images to the set target image (content-based retrieval).

Image retrieval is improved in comparison to text-based and content-based case by using multimodal query where both keywords and target image is used. Some examples of multimodal image retrieval are shown in Figs. 4 and 5 with different feature sets and distance measures. In Fig. 4, top three images are shown for the specified target image (wolf scene, same as in Fig. 3) and keyword "tiger" (as in Fig. 2), so the expected results should look like the target image, but a tiger should appear in the results. In this case, all obtained images correspond to the desired semantics (a tiger appears in the image), and the visual impression of the target image is preserved. This cannot be simply achieved with either text-based or content-based retrieval alone, although some of the images could appear among the results. In this example, color histograms were used as features with distance measures described in Sect. 4.

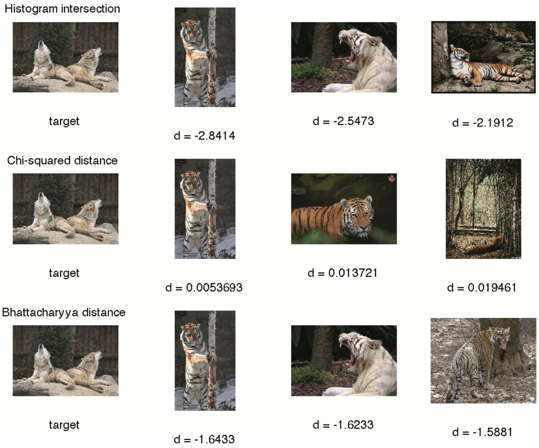

Fig. 4. Top three results for multimodal image retrieval using the target query image (left) and keyword "tiger", using color histograms for content-based ranking.

In Fig. 5, top three images are shown for the specified elephant scene as the target image and the keyword "lion". The obtained images have preserved the visual impression of the target image in both cases.

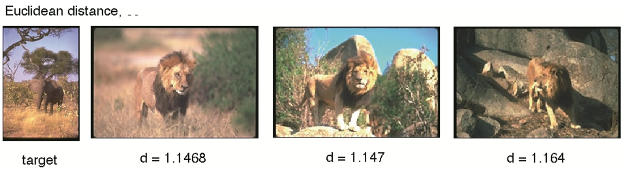

Fig. 5. Top three results for multimodal image retrieval using the target query image (left) and keyword "lion", using GIST descriptor with Euclidean distance measure for content-based ranking.

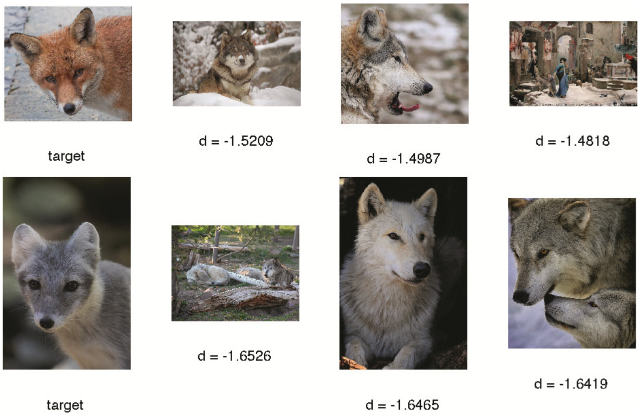

Fig. 6. Top three results for multimodal image retrieval using the target query image (left) and keyword "tiger", using color histograms for content-based ranking.

The influence of the target image in the multimodal image retrieval can be illustrated with an example shown in Fig. 6. The figure shows the results of two multimodal queries, both with the same keyword (wolf), but with different target images. Due to different feature values of the target images, the obtained results show different visual appearances of wolves. In both queries, the Bhattacharyya distance with color histograms was used.

6 Conclusion and Future Work

In this paper, a multimodal image retrieval framework was proposed, integrating keyword-based image search with content-based ranking according to the visual similarity to a query image. The semantics of the retrieved images are specified by keywords that are used to retrieve candidate images according to their textual annotations. The retrieved images are then ranked and sorted according to the visual similarity to the query image and presented to the user.

For content-based image ranking, different visual features extracted from images and distance measures were tested on image retrieval tasks on Corel and Flickr image databases of outdoor scenes. All tested measures and features have proven useful for improving the image retrieval. To choose the most suitable features and measure for the task of image retrieval in general domain, more formal evaluation will be performed since their performance in our experiment was similar to one another with many appearances of the same images in top 10 results.

The experiments have shown that multimodal image retrieval gives the user the opportunity to specify his query more easily and accurately, in terms of visual appearance and structure than with keywords alone. Simultaneously, the semantics are specified with keywords, so the query image does not have to be semantically related to the image the user is looking for. Thus, it is more likely that the user actually has a usable query image, making the proposed multimodal retrieval more user-friendly than the traditional content-based retrieval.

Multimodal image retrieval narrows the search results among images corresponding to the query keyword and can thus help when dealing with huge image databases.

In the future work, we plan to improve the results of multimodal image retrieval by exploring other types of image features that might be more appropriate for visual image comparison, as well as the corresponding similarity metrics. In case when images are not labeled nor described in the surrounding text, we plan to integrate automatic image annotation with multimodal image retrieval.

References

1. Eakins, J., Graham, M.: Content-based image retrieval. Technical report JTAP-039, JISC, Institute for Image Data Research, University of Northumbria, Newcastle (2000)
2. Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the gap: another look at the problem of the semantic gap in image retrieval. In: Multimedia Content Analysis, Management and Retrieval. IS&T/SPIE, Bellingham (2006)
3. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1349–1380 (2000)
4. Datta, R., Joshi, D., Li, J.: Image retrieval: ideas, influences, and trends of the new age. ACM Trans. Comput. Surv. **20**, 1–60 (2008)
5. Siddiquie, B., White, B., Sharma, A., Davis, L.S.: Multi-modal image retrieval for complex queries using small codes. In: Proceedings of International Conference on Multimedia Retrieval, p. 321. ACM (2014)
6. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
7. GIST. http://people.csail.mit.edu/torralba/code/spatialenvelope/
8. Cha, S.H., Srihari, S.N.: On measuring the distance between histograms. Pattern Recogn. **35**(6), 1355–1370 (2002)
9. Swain, M.J., Ballard, D.H.: Color indexing. Int. J. Comput. Vis. **7**(1), 11–32 (1991)
10. Pass, G., Zabih, R., Miller, J.: Comparing images using color coherence vectors. In: Proceedings of the 4th ACM International Conference on Multimedia, pp. 65–73. ACM (1997)
11. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)