Using Regions of Interest for Adaptive Image Retrieval

Michael Springmann and Heiko Schuldt

Database and Information Systems Group, University of Basel, Switzerland
{michael.springmann, heiko.schuldt}@unibas.ch

Abstract. Content-based image retrieval mainly follows a Query-by-Example approach and therefore requires well selected examples to start an initial search. This position paper describes how Regions of Interest (ROI) can be used to better adapt the system to the user’s information needs. In particular, it highlights how novel input devices such as Interactive Paper or TabletPCs can be used to capture much more details about what the user is precisely looking for already at the time the query is defined or when relevance feedback is specified.

1 Introduction

In traditional keyword-based text retrieval, the user poses a query by choosing the subset of terms that are relevant for his information need. Each document in a collection to be searched, in turn, is expected to contain significantly more words than the query. Therefore, the user already provides a good starting point by defining what he considers relevant – and can gradually refine his search by either adding new terms or adjusting the weights of individual terms by means of relevance feedback.

In content-based image retrieval (CBIR), the user starts by selecting one or more images to pose a query by example (QbE). The example may contain several regions, and it is rather natural to assume that some of them may be more relevant for the user’s information need than others (e.g., the foreground or the center of the image might be more relevant than the background or elements towards the boundary of the image). Systems like Blobworld [1] allow the user to select the region of interest from the automatically segmented image for formulating a query. However, automatic segmentation does not always give desired results for CBIR, e.g., if the object is partially occluded or consists of several parts. The latter is quite common, for instance in medical CBIR, where the region of interest might be the complete fracture of a bone and thus encompasses two disconnected regions as well as the part in between.

Relevance feedback has been proposed to overcome the well-known semantic gap between high-level semantics intended by the user and low-level features used for retrieval (e.g., in MARS [2]). However, little attention has been put so far on the perspective to use relevance feedback also to lower the misalignment of relevant query parts to irrelevant parts of the retrieved images.
One reason for this is that selecting appropriate regions from images can be more time-consuming and therefore less user friendly. In the context of the Query-by-Sketching (QbS) project [3], we aim at overcoming this limitation by using input devices like Interactive Paper [4] and Tablet PCs. Although these devices, in particular Tablet PCs, are available since several years, most applications do not exploit their full potential – in particular, they are rather used as a replacement for mouse and keyboard and no particular attention is paid to the question which tasks can be made more user-friendly with the special support offered by these devices. We believe that visual tasks, especially image retrieval, will be candidates for significant improvements.

In this paper, we present first results of an ongoing project aiming on providing adaptive image retrieval using regions of interest. In order to make use of the special character of novel input devices in a best possible way, the overall retrieval process is performed using several steps (which are also reflected in the structure of this paper):

1. The user poses a query by selecting one or more regions of interest (ROI) in a way tailored to exploit Interactive Paper or Tablet PCs. This step is further detailed in Section 2.
2. Corresponding regions of images in the database are determined. Automatically derived classifications and constraints may be used to reduce the search space and therefore reduce the overall execution time (see Section 3).
3. The ranking of results is based on the similarity of image regions that have been identified in the previous step. Customized distance measures are used to ensure that only relevant parts of the image contribute to the computed score (see Section 4).
4. Relevance feedback gives the user the chance to refine the search. This iterative process might affects all steps of the retrieval. The related parts are therefore described directly in each of the sections.

The paper concludes with a summary in Section 5.

2 Selecting Regions of Interest

Conventional approaches to region-based image retrieval assume that the user’s PC is equipped with keyboard and mouse. Therefore the two selections that can easily by performed are (a) select regions by clicking on a pre-segmented area of an image or and (b) to draw rectangular bounding boxes. Proper image segmentation on-the-fly is non-trivial and time consuming, even if tools like magic wand or magnetic lasso are used. Thus, existing solutions trade segmentation quality for speed of execution and convenience for the user.

In our approach, we assume that the user is equipped with a digital pen as input device that resembles a regular pen. The digital pen can determine its positions on the interactive paper and perform the actions defined for that position (see [3, 4] for more details on this technology). In our case, this is used for selecting regions on printed images and issuing queries. It is therefore straightforward
Fig. 1. An example query image (a), a rectangular bounding box of the cat (b), a manually segmented region containing the same cat (c) and the car in the background (d).

to select individual regions of interest (ROI) in the same easy and flexible, yet possibly imprecise way as one would do that on regular paper. At the same time, this is still likely to be more accurate than using an ill-segmented image or the rectangular selection of non-rectangular objects. Fig. 1 shows examples of such selections. If needed, the initial selection can be refined using semi-automatic segmentations like color invariant snakes [5] or GrabCut [6].

For relevance feedback, the digital pen is not only used to select relevant or irrelevant result images, but also to select the regions that make them relevant. Notice, that this might be much easier and intuitive to end users than assigning numeric values or preference judgements to images, since the task can be easily described as: “Select all relevant regions of the presented result images.”

Retrieval using region selection can be enhanced by annotations (for instance also automated image analysis algorithms such as face detection) in order to use additional features which might help in determining similarity and adding constraints on extracted image attributes. The advantages are twofold: Firstly, users do not need to specify all relevant concepts in the query which can be quite challenging (e.g., in [7] medical annotations with 116 distinct classes are used). Secondly, class membership of images determined at insertion time can be stored and evaluated very efficiently and therefore reduce the search space for nearest neighbors based on visual features.

3 Matching Regions in Images

Since our approach does not rely on pre-segmented images in the database, it is more flexible in adapting to the user’s information need. The user can, for instance, also select areas high contrast in itself, but low contrast from the background like the dark parts of the fur of the cat in Fig. 1(c) or partly covered objects like the car in Fig. 1(d). The latter is the common case where automatic image segmentation fails as already mentioned in [1]. On the downside, more processing time is needed during query execution, since the regions of images of the database corresponding to the ROI still need to be identified. The following section describes a generic solution used in our approach. However, it should be
noted that for particular image domains, there might exist much more sophisticated solutions that incorporate more knowledge about the objects shown in the images, e.g., model-based image registration for medical images or face detection/recognition. Such techniques should be preferred whenever available and applicable. Nevertheless, in the general CBIR setting, arbitrary images might be stored in the system and user queries may also vary significantly. Therefore such generic approaches are still of great importance.

For images in the database we extract salient keypoints and descriptors using SIFT [8], which has been proposed for object identification and is invariant to scale, rotation, and to a certain extent to variations of the illumination or viewpoint. For every such keypoint SIFT extracts a 128-dimensional descriptor [9]. In addition to this vector, the main orientation of the gradient and the scale at which it was detected are computed and stored with the keypoint. Fig 2 shows an arrow for each of the 1481 keypoints that have been extracted from the 682 × 512 pixel image in Fig. 1(a).

Matching keypoints in the ROI with corresponding keypoints in the database uses the squared Euclidean distance of the SIFT descriptors. For each keypoint in the query image the best match in the compared image needs to be determined and also the second best match, if the ratio between the distance of the two is used as quality measure as proposed in [9]. This computation is quite intensive – in the particular example when matched with the image shown in Fig. 2(b) it requires 1481 × 2536 computations of the distance between 128-dimensional vectors. Therefore this takes about 924 ms on a Intel Core 2 Duo with 2.33 GHz. Taking a closer look at the keypoints in our query image, we see that many of them are located outside the region of interest. In fact, the area containing the cat as in Fig. 1(c) contains only 684 keypoints and matching only these takes 438 ms. The car shown in Fig. 1(d), despite of covering an area significantly bigger in size, contains only 536 keypoints and matching takes 350 ms. Notice also, that both image, Fig. 1(a) and the original image of 2(b), have exactly the same size and very similar content, yet SIFT detected approximately 70% more keypoints in Fig. 2(b). Many of these additional keypoints are located in the lower part of the image, where there is additional road visible, which is in our case of little interest. When the rectangular selection shown in Fig. 1(b) is
Fig. 3. Result of determining the corresponding regions. Lines connect keypoints which are matched, color of lines indicate cluster. The red lines have been used to determine the affine transformation which maps the query image inside the red bounding box.

used, also additional keypoints are contained, 900 for total, and matching takes 576 ms and therefore more than 30% longer than our region of interest.

In order to speed up finding the matches, we apply an early termination strategy to the computation of the squared Euclidean distance similar to [10]. This reduces the computation time to 629 ms for the entire image, 311 ms for the cat region, 236 ms for the car region and 401 ms for the rectangular bounding box without changing the result. We did not yet exploit multi-threading or any kind of indexing. For the time being, our main focus has been on the exploitation of keypoints not only for object recognition where SIFT was originally used and a correct match is defined as matching the same point in the same or identical object, but apply a more relaxed rule, where also points in similar objects can be identified and thus, make also use of it in more general retrieval settings.

Using not only the number of matched keypoints, but also their location and the resulting rotation, scaling and translation of the image, an affine image transformation can be constructed, which maps the query region to the relevant region of the image from the database. When the query image is provided by the user and not part of the database of known images, it might not be possible to define a single threshold on distance or nearest neighbor ratio as proposed in [8] for the selection of matched keypoints. In particular, if the threshold is tight to find only very accurate matches, the system may return few or not even a single match for the region of interest. Instead, an ordered list of matches can be computed such that it always contains enough entries even in unfavorable cases, e.g., where illumination is very different or only similar objects are present. These matches can be clustered based on the scale and rotation in which this transformation would result. Selecting only the cluster with the most consistent transformation and filtering with RANSAC, the transformation gets more robust to many mismatched keypoints as long as at least three good ones remain [11]. The results of this matching step are displayed in Fig. 3. Notice if entire images containing several objects like in Fig. 1(a) were used, either only one region can be matched or both regions just a little – with the user having no control on what will happen.

As a result, the matching region of the ROI in the images of the database can be determined. Using regions in relevance feedback on result images, the
corresponding keypoints in the query can be identified and weighted or excluded. Also certain types of transformations can be excluded, e.g., if the user does not want to allow matches of the ROI being upside-down or scaled to small size. Such constraints can already be enforced within the loop over each keypoint since we have stored their orientation. If the orientation of a keypoint in the matched image differs too much from the query keypoint, we can already reject it and do not even need to compute the distance between the descriptors. This further reduces the matching time, in our experiments with an allowed angle of 10 degrees for the cat region to 152 ms without noticeable changes in the result, because the determined transformation would stay within that limit anyway. At this point, further improvements might also consider optimization of the clustering and RANSAC, since those two steps together require approximately the same time after matches have been determined. Furthermore, any available knowledge about the desired content should be evaluated to limit the number of images for which correspondence needs to be evaluated. The similarity between the corresponding regions itself does not need to be based on SIFT, but can be any appropriate feature(s) and distance measure computed on regions.

4 Similarity Between Corresponding Regions

Many image retrieval systems focus on features that are to a certain extent robust to transformations, e.g., global color histograms are invariant to rotation and not very sensitive to slight variations of the viewpoint or displacement of the observed object. But the tradeoff for that robustness is that such features may not be specific enough to capture the user’s perception of similarity. Since our approach can start computing the similarity with the assumption that we have a transformation for corresponding regions, we can evaluate similarity down to the pixel level. We can therefore apply comparison directly on the pixels or areas of reduced resolution and allow small deformations, for instance, with an image distortion model (IDM) of appropriate size [10].

Simple extensions of this model would also allow enforcing strict spatial constraints between several unconnected regions by transforming the two regions into a single one with a mask or alpha channel defining the weight of each pixel within similarity computation. All pixels within a ROI have a weight set to 1, pixels outside a weight of 0 and are ignored when distortion is evaluated. The
same extension can also be used to limit the effect on the computed similarity to pixels that are inside the ROI rather than the entire query image. This is therefore relevant even when a single region is selected. In Fig. 4, all white pixels would be assigned a weight of 0. Notice that the transformed images do not perfectly fit since SIFT and the derived affine transformation cannot compensate for viewpoint changes in 3D and movement and deformation of the objects in itself, like the cat not holding exactly the same pose. However, IDM allows for small deformations which will also reduce such effects. In comparison to Fig. 1(b), many background pixels are ignored which otherwise would impact similarity score, e.g., the mirror just behind the right ear of the cat.

For using several regions without strict spatial constraints, e.g., both regions must be present but no particular spatial arrangement is enforced, regions have to be evaluated separately. Similar to [12], distance combining functions based on Fuzzy-And, Fuzzy-Or, and Weighted Average can be used whenever the retrieved image should contain at least one region similar to each query region (And semantics), to at least one query region (Or semantics) or a weighted combination in case of different importance. In case of even more sophisticated semantics, e.g., if two query regions may not be matched to the same region in the image of the database as it is enforced in [13], these constraints would need to be evaluated already in the matching phase, in order to avoid overlapping regions.

By the means of relevance feedback, the parameters for similarity and allowed transformations as well as weights for keypoints and regions can be refined. Since our approach expects feedback on regions, differences in background or other unrelated parts of the image which otherwise could affect query refinement negatively, will not have an impact.

5 Summary and Conclusion

This paper presents an approach that allows a user in CBIR to specify his information need in a user-friendly, yet powerful way: i.) the user can select (almost) free-form regions of interest inside the query; ii.) the user can add new regions to the query and refine previously selected ones; iii.) the user can give a detailed feedback by selecting such regions, which also gives the opportunity to re-adjust weights. The approach will provide users the expressiveness for queries, similar to what he is used to from text retrieval. In addition, this will enable far more expressive relevance feedback.

Novel devices for interacting with the system reduce the burden of defining regions in images significantly. In the context of the ongoing QbS project, we have identified and implemented the basic components for making use of a digital pen for region selection. In particular, this includes a component for matching regions of interest to other images based on keypoints. Defining such regions and therefore excluding unrelated parts of the image also reduce the computation costs during matching and can enhance the quality of similarity computation.

Subsequent clustering and filtering make SIFT applicable to user-specified query images, i.e., where no pre-set threshold can be used. First experiments
have shown that matching of two images takes several milliseconds. If applied naively to searching within large collections thousands of images, this clearly can sum up to overall retrieval times exceeding the criterion for interactive response significantly. Yet, the approach can be seamlessly parallelized, either on a multi-core machine or in a cluster or Grid environment. In addition, sophisticated pre-selection (e.g., combination with high-level features or limitations on transformations) will further speed up the retrieval process. This makes the approach highly appropriate for use in an interactive mode where the query can already be refined based on the first images retrieved. Detailed evaluations and user studies will have to show how much regions of interest can improve image retrieval quality and how well the interaction modalities and patterns are accepted by users.

References

2. Rui, Y., Huang, T., Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS. Proc. of Int. Conf. on Image Processing (ICIP 1997) 2 (Oct 1997) 815–818