A REGIONALIZED CONTENT-BASED IMAGE RETRIEVAL FRAMEWORK

Stefan Uhlmann, Serkan Kiranyaz and Moncef Gabbouj[1]

Institute of Signal Processing, Tampere University of Technology, Finland
{stefan.uhlmann, serkan.kiranyaz, moncef.gabbouj}@tut.fi

ABSTRACT

Utilizing regionalized features in Content-based Image Retrieval (CBIR) has been a dynamic research area over the past years. Several systems have been developed using their specific segmentation and feature extraction methods. In this paper, a strategy to model a regionalized CBIR framework is presented. Here, segmentation and local feature extraction are not specified and considered as "black-boxes", which allows application of any segmentation method and visual descriptors. The proposed framework further adopts a grouping approach in order to "correct" possible over segmentation faults and a spatial feature called region proximity to describe regions topology in a visual scenery by a block-based approach. Using the MUVIS framework the proposed approach is developed and tested as feature extraction module, and its retrieval performance is compared against two frame-based color-texture descriptors. Experiments are carried out on synthetic and natural image databases and results indicate that a promising retrieval performance can be obtained if the segmentation quality is reasonable; however texture descriptors in general are degraded whenever applied on arbitrary-shape regions.

1. INTRODUCTION

Nowadays digital media technologies and compression standards combined with significantly cost-effective hardware technologies that are readily available on computers, several digital storage peripherals, broadcast systems and Internet lead us to the widespread exchange of multimedia information. This, however, brings the problem of retrieval, handling, and accessibility of such a massive media load. In order to overcome this problem several content-based indexing and retrieval techniques and applications have been developed and their performance mainly depends on descriptors and their capabilities to characterize the visual content. Two major approaches exist to tackle this challenge. The first uses global (frame-based) properties, such as color, texture, and shape information, to characterize an image. The main drawback of this approach is that it is not in accordance with the human visual system (HVS) since local visual properties are mixed and the overall description can be noisy and severely degraded. To model the HVS better the second approach uses the same properties locally, i.e. images are segmented into homogeneous regions and features are then extracted from those regions. The use of regions also introduces a new feature to CBIR, the so-called spatial layout [15], which describes the location and orientation information among image regions. Regionalized CBIR and Region-based Image Retrieval (RBIR) employ the second approach. However, they differ in their retrieval scheme. RBIR systems such as Blobworld [2] and Netra [8] usually consider the retrieval of a user-selected region whereas regionalized CBIR systems such as Windsurf [1], Walrus [11], and Simplicity [14] include all regions into retrieval process. This implies an integration of a region matching scheme to evaluate the overall image similarity.

This paper presents a strategy to model a regionalized CBIR framework. It includes four main stages: *segmentation*, *grouping*, *local* and *spatial features* extraction. *Segmentation* and *local features* are considered as black-boxes, which means that any segmentation algorithm and descriptor can be used within the framework. The assumptions are made that the black-boxes within the framework should perform reasonably well because if either of them fails nothing can be done to correct them. In the grouping operation, a correction is only possible if the applied segmentation method produces over-segmentation. The spatial feature stage integrates a similar approach to spatial layout, the so-called *region proximity*. Common region properties in spatial layout approaches are Minimum Boundary Rectangle (MBR) or Centre of Mass (CoM). Due to their lack of describing arbitrary regions efficient enough, such as crescent-shaped and ring-shaped, a block-based region representation and an average distance computation is introduced rather than the Hausdorff distance [6].

For evaluation of the proposed framework we used a quad-tree region splitting [5] due to its speed and simplicity and further employed a few modifications. MPEG-7 dominant color descriptor (DCD) is applied since HVS primarily uses *dominant colors* (i.e. the few colors prominent in the scenery) to judge similarity and this also makes the feature vector size quite limited as desired. As a texture descriptor Local Binary Pattern (LBP) [12] is chosen. It provides good texture discrimination with a reasonable vector size and it can be easily applied to regions. However, we do not employ any shape descriptor over regions since neither segmentation algorithm yet exists which can produce reliable boundary or contour information for regions so as to justify the utilization of shape information. As the region matching a many-to-one scheme is applied where each region tries to maximize its similarity with another region.

The proposed framework implementation is integrated as a Feature eXtraction (FeX) module for indexing and retrieval into the MUVIS framework [10]. Performance will be evaluated using a synthetic database and natural images from Corel collection [3]. The rest of the paper is organized as follows. Section 2 presents an overview about the proposed framework. Experimental results are given in Section 3 and Section 4 concludes the paper and suggests topics for future research.

2. REGIONALIZED CBIR FRAMEWORK

The framework is divided in the two parts of a CBIR system: indexing and retrieval. The following subsections describe the roles for each part.

2.1 Indexing Part

During the indexing phase, CIE- Luv is used as the color space due to its similarity in uniform perception and distance measure of the colors. It is fixed in the sense that conversion from RGB is applied once at the beginning and then CIE-Luv color space is used through out all stages. As illustrated in Figure 1, this part of the proposed framework consists of four main stages, namely *segmentation*, *grouping*, extraction of the *local features*, and *spatial features*.

At first, segmentation is applied to the image where, generally, any algorithm ranged from basic to advanced can be used. The main goal in this stage is to produce homogeneous regions in terms of color and texture. For testing and evaluation purposes we applied a modified color S+M algorithm applying a bilateral filtering [13] pre-processing

[1] This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence Program (2006 - 2011).

[1] The research leading to this work was partially supported by the COST 292 Action on Semantic Multimodal Analysis of Digital Media.

step where regions are smoothed but edges stay intact. This suppresses high frequencies such as noise and textural details to speed up S+M algorithm. The modification to the original S+M employs a post-processing phase in a way that the number of regions produced by S+M is further reduced. This is achieved by merging similar or small regions where the inter-similarity or size are below certain thresholds, e.g. regions are merged if their area is smaller than one per cent of the image size. The *grouping* stage follows up *segmentation* , which should be tuned beforehand to yield over-segmentation rather than under segmentation since grouping can eventually correct this by merging similar adjacent regions based on their local features, color and texture; however under-segmentation results cannot be corrected in any way. For grouping, similarity scores between all adjacent regions are calculated compared, and regions with the highest similarity are grouped together. This is carried out until maximum similarity thresholds for color, thr_{sc}, and texture, thr_{st}, are reached. Thresholds represent similarity values until regions are considered similar based on their local features. Note that segmentation results are expected to be similar for images representing similar content with similar features. Otherwise this might have a great influence on the feature description, region matching, and also on retrieval performance.

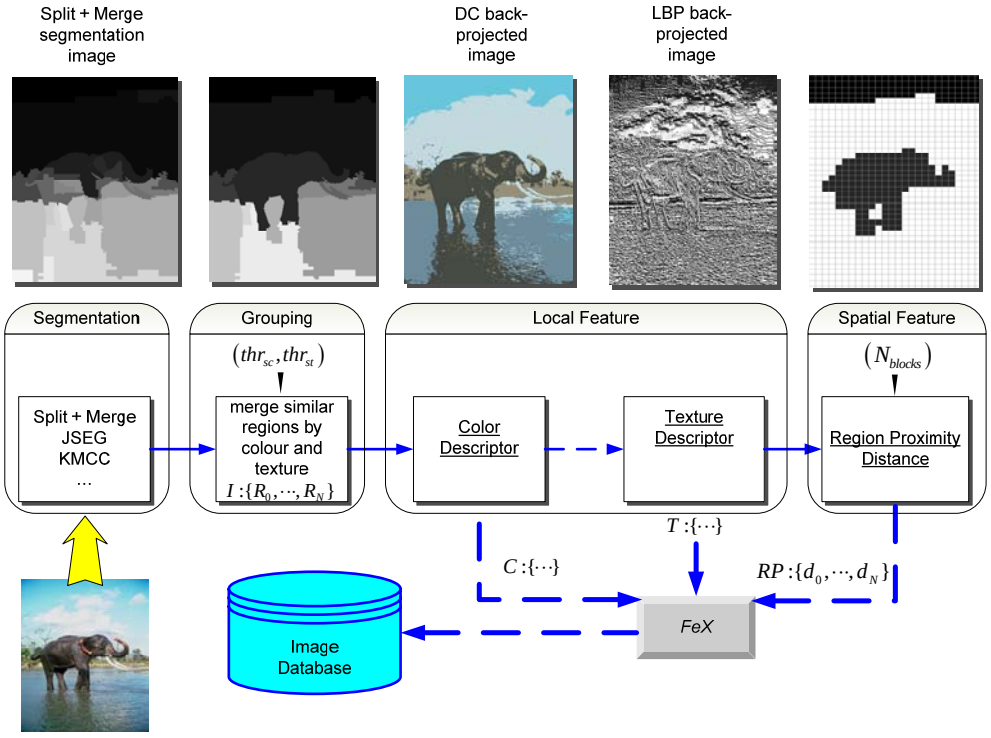

Figure 1 - Overview of the regionalized CBIR framework.

After the segmentation phase, local features such as color and texture are extracted from those regions. We used DCD and LBP as color and texture descriptors, respectively. MPEG-7 DCD, with the maximum number of DCs set to 8, is extracted from the entire image. In order to obtain the DCDs per region, the DCs extracted over the image are back projected onto the regions. The region color can then be represented by a set of DC classes where each class holds the DC centroid and weight information, i.e. $DC_i:\{c_i, w_i\}$. Generally speaking, the maximum number of DCs appearing in homogeneous regions shall not be higher than two. The extraction is similar for region-based texture description where the descriptor is first extracted over the entire image and then back projected onto the regions. The sample texture descriptor used in the proposed framework is LBP, which works directly on the luminance values of a center pixel with its neighborhood as shown in Figure 2 (a). The neighboring pixels are then thresholded by the center pixel and binomial factors are multiplied to the neighboring positions that are greater than or equal to the center pixel as shown in Figure 2 (b), (c), and (d) respectively. Finally, the sum of the binomial factors yields in the LBP value being assigned to the center pixel. This procedure is repeated for each pixel in the image. The feature vector is represented as a texture histogram.

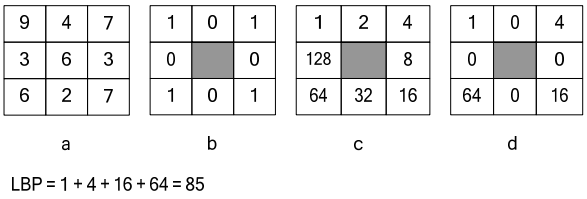

Figure 2 - Calculation of LBP value over 3x3 neighborhood.

The last stage is the extraction of the spatial feature, region proximity, which exploits and describes the spatial properties among regions. Thus, it shows a certain similarity to spatial layout descriptor. The basic difference is that in region proximity distances are calculated between the regions rather than describing directional and topological relationships. In order to approximate the distance between two regions ¨the image is divided into a block grid and each region is then described by its building blocks where a block belongs to the region if at least one region pixel lies within that block, as illustrated in Figure 3A. The number of blocks in our implementation is set as, N_{blocks} = 1024, representing a grid of size 32x32, which is mainly a trade-off between complexity and effectiveness. The distance between two regions is then estimated as,

$$h(A,B)=\frac{\sum_{a\in A}\min_{b\in B}(|a-b|)}{N}$$
$$H(A,B)=\min\{h(A,B),h(B,A)\} \quad (1)$$

where A and B represent image regions, a and b represent blocks describing those regions, respectively, and $|\cdot|$ represents the underlying L-norm. In our case the L_∞–norm is employed to compute the distance between single blocks due to its low complexity. Eq. 1 expresses an average distance where $h(A, B)$ is the direct distance between to regions and $H(A,B)$ is then the final distance between two regions A and B for symmetry reasons. The difference between our proposed and the original Hausdorff distance is that whenever regions have an arbitrary shape, touching or enclosing regions, the original Hausdorff distance becomes biased by shape and region size because it mainly considers the region boundary for the distance due to its maximization functions. This is the reason why the single block distances for a region are averaged to enhance the effects of shape and size. The advantage of the proposed distance calculation is illustrated in Figure 3B and Figure 3C. It can be seen in Figure 3C that MBR and CoM are rather poor region descriptions for a spatial distance calculation especially for ring-shaped regions. In this specific case of Figure 3B, CoM would return a zero distance whereas with MBR it would be difficult to calculate the direct distance without special handling. The difference between our proposed and the original Hausdorff distance is shown in Figure 3C.

After this stage each region holds its proximity (distance) description to any other region within the image. Once the extraction of region features is completed an encapsulating feature vector can be formed and indexed into a database.

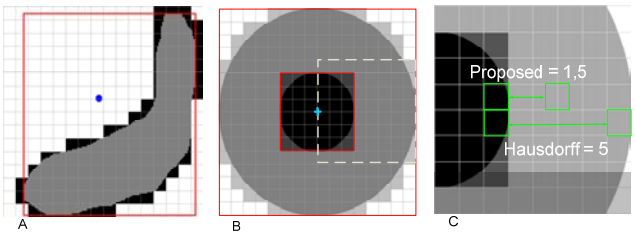

Figure 3 – A: Region description by MBR (red square), CoM (blue point), and building blocks. B: Hausdorff and the proposed distance computations between two circles. C: zoom-in for B

2.2 Retrieval Part

The retrieval approach for regionalized CBIR differs from the common frame-based approaches. Here, a direct comparison cannot be applied since it is unknown *a priori* which regions are to be compared. Hence, a region matching approach should be performed. The goal with the employed region matching is to maximize the overall image similarity. This is achieved by so-called region-based similarity maximization where each region tries to maximize its similarity to other regions based on local and spatial features. The total image similarity yields in the maximum value for equally sized regions with similar features. The matching process is integrated in the following way. Generally, the local and spatial features are unit normalized and we skip details due to page limit. Total image similarity TS between two images is defined as,

$$D_Q(C_i,C_j)=(C_i-C_j)^T A(C_i-C_j)$$
$$G(LBP^i,LBP^j)=2\sum_{b=1}^{B} LBP_b^i \log\frac{LBP_b^i}{LBP_b^j}$$
$$S_C(i,j)=1-D_Q(C_i,C_j)$$
$$S_T(i,j)=1-G(LBP^i,LBP^j)$$
$$S_L(i,j)=[\alpha S_C(i,j)+(1-\alpha)S_T(i,j)]$$
$$S_S(i,j)=[S_{RP}(i,j)]$$
$$Sim(i,j)=\frac{[S_L(i,j)+S_S(i,j)]}{2}$$
$$RAF(i,j)=\min(w_i,w_j)$$
$$TS_{I_0I_1}=\sum_{i\in I_0}^{N}\max_{j\in I_1}\{RAF(i,j)\times Sim(i,j)\}$$
$$SD=1-\left(\frac{TS_{I_0I_1}+TS_{I_1I_0}}{2}\right)\le 1 \quad (2)$$

where each region i in query image I_0 is compared to each region j in target image I_1. Region similarity score $Sim(i,j)$ between two regions is divided into two parts: local, $S_L(i,j)$, and spatial, $S_S(i,j)$, feature similarity scores. $S_L(i,j)$ merges similarities $S_c(i,j)$, which is obtained from quadratic distance formulation, D_Q, [4] over DC sets C_i and C_j, and $S_T(i,j)$, which is obtained by G-Statistic measure, G, [12] over LBP histograms LBP^i and LBP^j The variable $0<\alpha\le 1$ is the weighting factor between color and texture similarity scores. This is due to the fact that if there is no texture present in a region only color should be used to determine the region similarity via $\alpha=1$. In general, α is automatically determined by the approach presented in [7] on the Luminance part of CIE-Luv color space where the window size is decreased to a 5-by-5 pixel neighborhood. This is due to the reason that a large window might be biased by edges and contours in the image and, furthermore, it might produce undesired results for non-textured regions when it comes to the region borders. To reduce the effects of noise and compression artifacts especially for non-textured regions such as *sky* the luminance image is quantized to multiples of 5. This also assures if there is any dominant texture within the regions then it will be detected even after quantization. The spatial similarity $S_S(i,j)$ tries to match similar surrounding of a region by using the region proximity feature. Basically, the following assumption is exploited. Two regions which do not match by local features can still be considered similar if their surroundings are similar (i.e. neighbor region proximities and their local features match). Figure 4 illustrates a simplified scenario of such a region (spatial) similarity where a *white* and a *black horse* are present in the same scenery having similar surrounding regions. This similarity will yield in a significant S_S score and thus a spatial similarity contribution to the regions of *horses*. Computational-wise, regions are only considered if there are similar (matching) regions in I_0 and I_1. Hence, if two images have no similar regions at all, the region proximity does not bring any contribution to $Sim(i,j)$. Furthermore, the region proximity similarity $S_{RP}(i,j)$ favors large regions in small distances which is described in the following equation:

$$S_{RP}(i,j)=S_{RP}^{ij}(k,l)=$$
$$\left(\frac{\min(w_k,w_l)}{\max(w_k,w_l)}\right)\times\left(\frac{(1+\min(w_k,w_l))}{2}\right)\times\left(\frac{\text{minDist}_{(k,l)}}{\text{maxDist}_{(k,l)}}\right)\times\left(1-\text{maxDist}_{(k,l)}\right) \quad (3)$$

where i and j are the regions which are to be matched (e.g. the *horses* in Figure 4), k and l represent the surrounding regions, and w_k and w_l are the region weights for k and l, respectively. $minDist_{(k,l)}$ and $maxDist_{(k,l)}$ provide the minimum or maximum of the region proximity distances between region i to k and j to l. Term 1 in Eq. 4 favors regions with similar sizes, term 2 punishes small region sizes, term 3 favors similar region proximity distances, and term 4 punishes large region distances. Generally speaking, if two regions have large similar regions in their closer proximity, higher similarity scores are contributed to the overall image similarity (TS). $Sim(i,j)$ is computed for each region i and weighted by region area factor $RAF(i,j)$to measure the influence (importance) of $Sim(i,j)$ for TS. $RAF(i,j)$ is calculated by region area weights. TS is then computed as in Eq. 2 where only the

best match (maximum *Sim(i.j)*) for each region is considered. In order to obtain retrieval symmetry between two images, *TS* is calculated for I_0 and I_1 where the overall similarity distance is then averaged over both *TS*. Unit normalized similarity distance, *SD*, (as required by MUVIS) is calculated as given in Eq. 2.

Figure 4 - Two similar sceneries with a white and black horse.

3. EXPERIMENTAL RESULTS

The database indexation (FeX) and retrieval experiments are performed using the MUVIS framework, which provides a development and test-bed environment to implement third party modules in form of Dynamic Linked Libraries (DLLs). Furthermore, it offers two applications each of which is dedicated to feature extraction and multimedia retrieval. A particular module where a certain descriptor is implemented can dynamically be linked to the application and the user can perform extraction and retrieval tasks via using it during the runtime. For a comparative evaluation three FeX modules have been employed: frame-based modules for DCD and LBP (texture) along with the one for proposed framework.

The retrieval process in MUVIS is based on the traditional query by example (QBE) scheme. The features of the query item are used for (dis-) similarity measurement among all features of the database items. Ranking the database items according to their similarity distances yields the retrieval result. For the frame-based case this simply means comparing the feature(s) of the query item with the feature(s) of each database item using the aforementioned metrics. For all retrieval tasks, performance is evaluated by using Average Normalized Modified Retrieval Rank (ANMRR), which is defined in MPEG-7 as the retrieval performance criterion. Besides the traditional recall and precision measures, ANMRR also includes rank information to measure the retrieval efficiency. For a detailed description it is referred to [9]. The first part of experiments is conducted on synthetic databases with similar regions in different layouts where no segmentation faults are present and regions do not contain any texture information. These experiments are performed to test the region proximity feature as well as showing its importance in content description. Here, a few queries are particularly chosen for visual demonstration. The latter part covers retrieval experiments over a general-purpose image database such as Corel, which is generated of 1000 (*Corel1K*) images including natural images of various categories such as *beaches, buildings, busses, dinosaurs, flowers, horses, mountains*, and etc. The retrieval performance of the proposed framework is compared against frame-based color descriptor (DCD) and its combination with texture descriptor (DC+LBP). For this database five images are randomly selected for each category and queried in *Corel1K* database.

3.1 Performance Evaluation

At first results are shown for synthetic database unbiased by segmentation. Figure 5 and Figure 6 display retrieval results for two different queries for the proposed region proximity feature compared to only color feature. It can be seen that the region proximity senses different distances between the regions and better retrieval results can be achieved based on the improved content description. Furthermore, the same similarity distances are obtained for the 1st to 4th and 5th to 8th rank, indicating that they are identical, which is only true for their global color proportions.

Table 1 - ANMRR results for *Corel1k*

	DC	DC+LBP	Proposed
Corel1k	0,3499	0,2817	0,2909

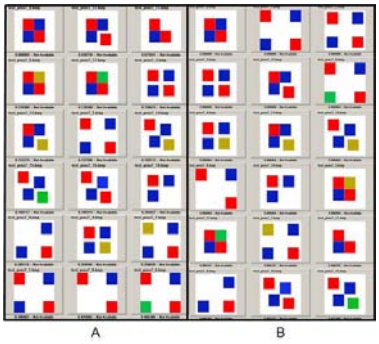

Figure 5 - Results for two queries on synthetic database; query image is in the top-left position. Left side shows results by region proximity feature and right side shows results by color only.

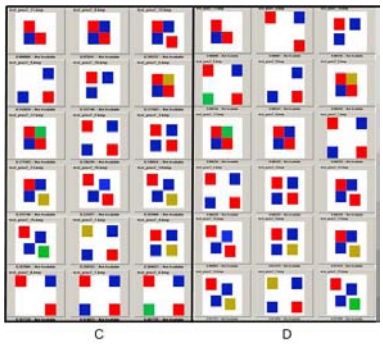

Figure 6 - Results for two queries on synthetic database; query image is in the top-left position. Left side shows results by region proximity feature and right side shows results by color only

Table 1 presents the ANMMR results for *Corel1k* using the three FeX modules. It can be seen that DC + LBP and the proposed region-based method have a close performance whereas DC feature performs the worst. Hence, combination of color and texture as well as regionalized features improves results compared to DC feature.

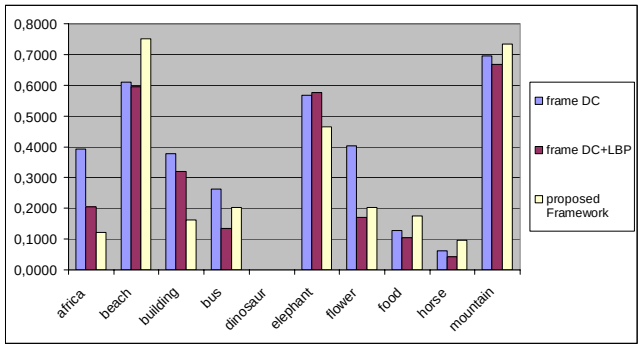

Figure 7 - Results for *Corel1k* categories for three features. ANMRR value represented on y-axis and categories on x-axis.

Figure 7 shows performance for the individual categories of the selected images where y-axis represents ANMRR values and x-axis shows the different categories. Here, it can be noticed that proposed approach performs better than both frame-based modules for the categories *africa*, *buildings*, and *elephants*. Further, it achieves higher ANMMR values than DC FeX module for *bus* and *flower* classes. For classes such as *beach, food, horses, and mountain* frame-based modules perform slightly better (except for *beach* class where the gap broadens significantly). This might be due to irrecoverable faults of segmentation and/or the selected regionalized descriptors. An example for the influence of different segmentation is illustrated in Figure 8. Two similar images are shown with their corresponding segmenta-

tion masks. Due to this difference the image on the right will appear in lower rank (35^{th}) when querying the image in the left side. Here, the major influence comes from different segmentation of the background regions, which is represented as a single region on the left image whereas it is split into two regions on the right. Furthermore, frame-based texture description by applied LBP does not significantly improve the retrieval performance when combined with frame-based color description except for *africa, bus, and flower* images. Since frame based texture description is a mixture of all textures in an image, the assumption can be made that frame-based texture description mainly catches strong edges and contours in the image such as in *flower* and *bus* classes. *Flower* images mainly have a light colored flower in the center of a dark background and bus images contain major running lines due to the object shape. This would support the aforementioned observations. Additionally, it has been observed that regionalized texture has only minor effects on retrieval probably due to low discrimination in such natural images. What the ANMMR values in Table 1 and Figure 7 do not show are the subjected evaluation of the retrieved images. This means retrieval for items in the same class is good but what about their content. Figure 9 presents such an example where "two brown horses" were queried. Frame-based retrieval returns relevant matches for the *horses* class but only one out of the first eleven retrievals is with "two brown horses". Whereas the region-based approach returns 5 to 7 out of the first eleven with "two brown horses" depending on the observer. Generally speaking, it has been witnessed that images with similar subjective content were retrieved in earlier ranks.

Figure 8 - Two similar images with their segmentation masks.

Figure 9 - Results for query "Two brown horses" frame-based (left) and region-based (right) results.

4. CONCLUSIONS

This paper presents the modeling of a regionalized CBIR framework where segmentation and local feature extraction are considered as black-boxes: any segmentation and visual descriptor can be applied. Yet the proposed framework has a dedicated grouping approach after the segmentation phase in order to correct over segmentation faults and a spatial feature so-called region-proximity that describes the relationship between regions. Results for synthetic images show the promising performance related to the image content due to the region proximity feature. Performance of frame-based and region-based color/texture description on natural images is similar based on ANMRR whereas differences occurred in subjective evaluation. Even though we used both local and spatial features from regions; the results are not yet superior to frame-based methods due to the effect of segmentation faults and the negligible contribution of regionalized texture in these images. Since robustness of segmentation is essential over the retrieval performance and the influence of applied regionalized texture description is minimal, the future work will concentrate on their improvements. Furthermore, investigation will be conducted to evaluate the effects of different segmentation methods with different descriptors.

References

[1] S. Ardizzoni, I. Bartolini, M. Patella, "Windsurf: region-based image retrieval using wavelets", *Tenth International Workshop on Database and Expert Systems Applications*, pp. 167-173, 1999.

[2] C. Carson, S. Belongle, H. Greenspan, and J. Mailk, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on PAMI*, vol. 24, pp. 1026-1038, 2002.

[3] "Corel Clipart and Photos", http://www.corel.com/products/clipartandphotos/

[4] Z. Dengsheng, L. Guojun, "Evaluation of Similarity Measurement for Image Retrieval", *In Proceedings of the International Conference on Neural Networks and Signal Processing*, vol. 2, pp. 928-931, 2003.

[5] S. Horowitz and T. Pavlidis. "Picture segmentation by a tree traversal algorithm", *Journal of the ACM*, vol. 23, pp.368-388, 1976.

[6] D. Huttenlocher, G. Klauderman, W. Rucklidge, "Comparing images using the Hausdorff-distance", *IEEE Transactions on PAMI*, vol. 15, pp. 850-863, 1993.

[7] K. Karu, A.K. Jain, R.M. Bolle, "Is there any texture in the image?", *Proceedings of the 13th International Conference on Pattern Recognition*, vol. 2, pp. 770-774, Aug 1996.

[8] W.Y. Ma, and B.S. Manjunath, "NeTra: A Toolbox for Navigating Large Image Databases", *IEEE Int. Conf. On Image Processing*, vol. 1, pp. 568-571, Santa Barbara, USA, Oct 1997.

[9] B. S. Manjunath, J-R. Ohm, V.V. Vasudevan, and A. Yamada, "Color and Texture Descriptors", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp.703-715, Jun 2001.

[10] MUVIS, http://muvis.cs.tut.fi/

[11] A. Natsev, R. Rastogi, K. Shim, "WALRUS: a similarity retrieval algorithm for image databases", *IEEE Trans. on Knowledge and Data Engineering*, vol.16, pp. 301-318, 2004.

[12] T. Ojala, M. Pietikainen, D. Harwood, "A comparative study of texture measures with classification based on feature distributions", *Pattern Recognition*, vol. 29, pp. 51-59, 1996.

[13] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images", *Proceedings of the IEEE International Conference on Computer Vision*, Bombay, India, 1998.

[14] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLcity: Semantics-sensitive integrated matching for picture libraries", *IEEE Transactions on PAMI*, vol. 23, no. 9, pp. 947-963, 2001.

[15] Y. Wang, "Image indexing and similarity retrieval based on spatial relationship model", *Information Sciences—Informatics and Computer Science*, vol. 154, pp. 39-58, Aug 2003.