Spatial color image retrieval without segmentation using thumbnails and the Earth Mover's Distance

Thomas Hurtut, Haroldo Dalazoana, Yann Gousseau, Francis Schmitt;
Ecole Nationale Superieure des Telecommunications, TSI, Paris, France
Ecole Polytechnique de Montreal, LIV4D, Montreal, Canada

Abstract

We introduce a spatial color image retrieval method which does not include any segmentation step. This method relies on small image thumbnails and the Earth Mover's Distance (EMD). We then derive an unsupervised matching criterion using an a contrario *approach. Experiments are performed on a database of illuminated manuscripts.*

Introduction

Color image retrieval methods that do not take into account any spatial layout miss important cues of human visual perception. Hence spatial color indexing retrieval has become a very active research area. Some approaches augment histograms with pixel localization or spatial color correlation, see e.g. [1, 2, 3, 4]. Other methods use points of interest in the image to include spatial features [5, 6]. Region-based approaches first perform a segmentation of images which is then used for image comparison. The *Blobworld* image representation, for instance, uses multi-dimensional Gaussian mixtures as the image model [7]. Many other unsupervised segmentation methods have been used to index images, see e.g. [8, 9, 10]. A matching step follows the segmentation and indexing steps, generally relying on a distance or a similarity measure. One of the most efficient distance to compare sets of features is the Earth Mover's Distance (EMD) [11], see e.g. [12, 8].

We propose a retrieval method that rely on the spatial organization of colors. The plan of the paper as follows. We first define the features used to represent the spatial and color contents of images. We then introduce a distance derived from optimal transportation problems, the Earth Mover's Distance, and discuss its use for the comparison of images signatures. Next we derive an automatic criterion, enabling to decide whether two images should be matched or not, using an *a contrario* approach. We conclude with numerical experiments on a database of illuminated manuscripts.

Features

The proposed method uses coarsely sampled thumbnails to represent images (Fig. 1). The total number of pixels of these thumbnails is fixed to n, depending on the database content and on computing constraints. Image subsampling is achieved by averaging the image pixels. We choose the psychometric CIELab color space to represent thumbnails, in order to use the Euclidean distance to compare their color components. The signature for an image is composed of the n thumbnail pixel features $f^i = \{L^i, a^i, b^i, X^i, Y^i\}$, where L, a, b are the color coordinates and X, Y the spatial positions.

Distance

To measure the distance between two pixel features f_Q^i and f_T^j of a query image Q and a target image T respectively, a weighted exponential distance is used:

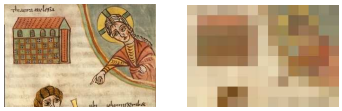

Figure 1. *A coarsely sampled thumbnail. Left: source image (*3200×2200 *pixels). Right: the thumbnail (*10×15 *pixels).*

$$d_e(f_Q^i, f_T^j) = \alpha * \left(1 - e^{-\left(\frac{(L_Q^i - L_T^j)^2 + (a_Q^i - a_T^j)^2 + (b_Q^i - b_T^j)^2}{\delta_c^2}\right)}\right) + (1-\alpha) * \left(1 - e^{-\left(\frac{(X_Q^i - X_T^j)^2 + (Y_Q^i - Y_T^j)^2}{\delta_s^2}\right)}\right). \quad (1)$$

The parameter α balances color and spatial contributions. This weighted exponential distance supports the idea that beyond a certain distance, two different images are just considered as not similar. Parameters δ_c and δ_s are chosen according to the color and spatial dynamics. Typical values that will be used in this paper, are $\delta_c = 15$ and δ_s a quarter of the thumbnails diagonal. This nonlinear distance is adapted to an image retrieval task, where a target image is close to a query, in which case the similarity is measured, or is simply different and penalized.

We then use EMD [11] to measure the distance between the signature of a query image and the signature of a target image. EMD is defined as the minimum amount of work needed to map one signature onto the other. The notion of work is defined as the product of the ground distance d_e and the feature weight. In our approach, each feature is given the same weight (since thumbnail pixels correspond to a fixed area in the original image). It may then be shown that each pixel of the query image is assigned to exactly one pixel of the target image, see Figure 2. In this particular case, the EMD distance between two images Q and T is equal to

$$d(Q,T) = \min_{\phi} \sum_{i=1}^{n} d_e\left(f_Q^{\phi(i)}, f_T^i\right), \quad (2)$$

where ϕ belongs to the set of permutations of $\{1, \ldots, n\}$. Computation of the EMD is therefore simplified and can be achieved by solving an optimal assignment problem. This problem is solved by the Hungarian method, also called the Kuhn-Munkres algorithm [13], in $O(n^3)$.

In contrast, when using EMD with region based methods, each region is usually weighted according to its area. In this case, EMD breaks apart the weights of the features when optimizing the cost of the mapping. Therefore, the usefulness of a

segmentation step before EMD is questionable since pixels that have been grouped into regions can be mapped separately to different target regions by EMD. Indeed no constraints are included in EMD that take into account the fact that regions represent pixels groups. Besides, segmentation methods require complex tunings of the parameters. Furthermore, undersegmentation can yield regions that are not well represented by, e.g., their mean color, affecting the performances of the matching step. Our approach is simpler, more robust and takes full advantage of EMD : we somehow oversegment the image into small and regularly spaced regions, and then let the EMD distance resolve the matching problem. The number of regions that are considered is only limited by computation constraints.

Figure 2. *Two images (left) and the associated thumbnails (middle), made of $n = 10 \times 15$ pixels. Each pixel of a thumbnail is associated to a feature in 5 dimensions (L,a,b,X,Y). All of them have the same weight. In this context, EMD is equivalent to an assignment problem. Each pixel of the query image (top left) is assigned to exactly one pixel of the target image. To illustrate the EMD operation, the thumbnail on the right represents the optimal permutation of the query pixels to match the target. The query pixels of the red region are moved towards their location in the target image. The white column of paper is also correctly moved from right to left.*

Unsupervised matching criterion using an *a contrario* approach.

In this section, we present a method to answer the question of whether two color images should be matched or not. Our approach is similar to the one in [14, 15] : we will match two images as soon as their proximity is unlikely to be due to chance. This is an application of the general principle of *a contrario* methods: we do perceive events that are unlikely in a white noise situation. Such methods have been successfully applied to several tasks in computer vision following the work of [16], see [17]. We now detail how this principle is used in the framework of the comparison of images.

Our setting is as follows. We write Q for a color query image, and $\mathscr{B} = \{T_1, \ldots, T_m\}$ for a database composed of m color images. For two images Q and T, their EMD distance (as detailed in the previous section) is denoted as $d(Q,T)$. For each image T_i, we want to develop a statistical test for the hypothesis $H_1 = \{T_i \text{ is similar to } Q\}$ and we choose to rely on the quantity $d(Q,T_i)$. A usual approach to this problem would be to have a probabilistic model for the images that are similar to the query and, for example, to perform a Bayesian test. This implies the development of specific models for each category of color images that we deal with. An *a contrario* approach to this test consists in relying on a *background model* $\mathscr{M}$ of generic images, and then to fix the number of false alarms when testing H_1 against $H_0 = \{T_i \text{ follows the background model }\}$. The general idea is that if an image has been generated by the model $\mathscr{M}$, then it should not be paired with Q. Assume that we are able to compute the probability $Pr_{H_0}(d(Q,T) \leq C)$, for any $C \geq 0$. We will then say that

$$T_i \text{ is an } \varepsilon\text{-meaningful match of } Q \text{ if } d(Q,T_i) \leq C_B,$$

where C_B is such that $Pr(d(Q,T) \leq C_B) \leq \varepsilon m^{-1}$, and where T is distributed according to $\mathscr{M}$. It can be shown that if all images T_i from the database $\mathscr{B}$ follow the model $\mathscr{M}$, then the expectation of the number of ε-meaningful matches is less than ε (the proof of this fact relies on the linearity of the expectation, see [15]). In practice, we will always choose $\varepsilon = 1$, that is we will fix the expected number of false matches to one. In what follows, we use the term meaningful match instead of 1-meaningful match.

Of course, for this to be feasible, we need to choose a background model $\mathscr{M}$. Since the distance d only take thumbnails into account, we seek a background model for thumbnails. The first model we tried was a uniform white noise model, in which all pixels of a thumbnail are drawn independently and uniformly in the *Lab* space. This was not satisfactory, since too many images were then meaningful matches of the query. Our interpretation of this fact is that, even when they are perceptually different, images share some common structure (e.g. the presence of homogeneous regions). Similar results were obtained with a model in which thumbnails are white noise with color marginal learned on the database $\mathscr{B}$. We therefore need a more structured model, with geometric features similar to the ones encountered in real images, such as homogeneous regions and edges. The model we consider is a dead leaves model ([18]), consisting in the superimposition of random objects, together with a power law distribution for the size of objects. This model has been experimentally proved to be well adapted to the structure of natural images, [19, 20]. For the sake of simplicity, we restrict ourselves to a model where objects are simply disks. The radius of these disks are random variable with density $f(r) \propto r^{-\gamma}$. The model is then characterized by three parameters : the scaling parameter γ, an dr_0, r_1 the minimal and maximal sizes of objects[1]. We chose $\gamma = 3$, a typical value for natural images, $r_0 = 1$ and r_1 of the same magnitude as the dimensions of thumbnails. The distribution of the color of objects is learned from the color distribution of pixels from thumbnails of the database. Samples of the model (using color marginals from the database of illuminated manuscripts used in the experimental section) are displayed in Figure 3.

Figure 3. *Samples of the background model $\mathscr{M}$ using color marginals from the illuminated manuscript database used in the experimental section.*

The last point, in order to be able to compute meaningful matches of a given query image, is to compute the constant C_B such that $Pr(d(Q,T) \leq C_B) \leq \varepsilon/m$ where T follows $\mathscr{M}$. Since this quantity would be very tricky to compute exactly, we rely on numerical simulations. That is, we sample r realizations F_j of the model and then approximate C_B by the empirical quantile of order $m^{-1}\varepsilon$ of $\{d(Q,F_j)\}_{j=1,\ldots,r}$. The choice of r will be addressed in the experimental section.

Let us stress that an advantage of this approach to the matching problem is that the threshold automatically adapts to both the

[1] One can get rid of r_0 by considering a limit model yielding the same small scale structure as natural images, [21], but this model is too involved for the modeling of coarse representations of images.

query and the database. In particular, the threshold is expected to get more conservative as the database gets larger, which can be crucial when dealing with very large databases. Also observe that the threshold is driven by parameter ε, which has the intuitive meaning of a number of false detections when submitting a query and is therefore easier to choose than a bound on distances. In all experiments to be performed in this paper, we choose $\varepsilon = 1$.

Experimental evaluation

Tests have been performed on an illuminated manuscript database provided by the *Institut de Recherche et d'Histoire des Textes* (IRHT)[2]. This database contains 1500 high quality and color-calibrated illuminations. Experiments use $n = 150$ (10×15 or 15×10 thumbnails). Tests are performed using a query image. We then compute the distances between the query thumbnail and all the thumbnails in the database and return the nearest matches. Except where mentioned, all tests are performed using $\alpha = 0.5$, $\delta_c = 15$ and $\delta_s = 5$. Distances between the query thumbnail and the background model are also computed in order to estimate the matching criterion. The parameters of the background model we use are $\gamma = 3$, $r_0 = 1$ and $r_1 = 30$. A point that is rarely discussed and that concerns all spatial color methods is how to manage different image ratios. Various choices are available to handle this problem. In the experiments, we chose to treat separately portrait and landscape orientations. However EMD has no constraints on signature length and total weight, and therefore we could also compare images with different ratios. A query on a PC Pentium IV 4.3 gHz has a runtime of approximately 10 seconds. More results are visible on website [22].

Distance

We use EMD to measure the distance between signatures of images. The weighting parameter $\alpha \in [0,1]$ used in the distance between two features (see Eq. 1) allow us to support either spatial or color constraint. If the user wish to be severe on spatial organization (pixels cannot move freely on the image), α must be set to a low value. The higher the value, the more objects can move and be split apart. The extreme case where $\alpha = 1$ is equivalent to a color histogram method, with EMD similarity measure. Two results are displayed on Fig. 4, using different α values.

Comparison with blobworld + EMD

Newt we compared our method with a region-based method using EMD for the matching step. We chose Blobworld segmentation[3] [7] that represents regions as Gaussian mixtures. The original method uses texture and location to extract *blobs*. We only use color and location information to be directly comparable with our method. Nevertheless we observe similar retrieval results on the illuminations database when also using texture. Blobs tend to represent objects in the image. Hence the number of blobs to find in an image is an important parameter. This parameter is set by an MDL method (see [7] for details). For the sake of coherence, we also use EMD to measure the distance between two blob signatures. In this case, the weight of each blob is chosen as its area and EMD breaks apart the weights of the features when optimizing the cost of the mapping.

Two comparative results are displayed on Figure 5 and Figure 6. Among all our tests, the *blobs+EMD* method performs well on queries where simple objects stand on a homogeneous background. In this case our method will perform similarly because, as mentioned earlier, blobs will be split apart by EMD.

[2]IRHT, CNRS, 40 avenue d'Iena, 76116 Paris.

[3]Code available at URL: http://elib.cs.berkeley.edu/blobworld/

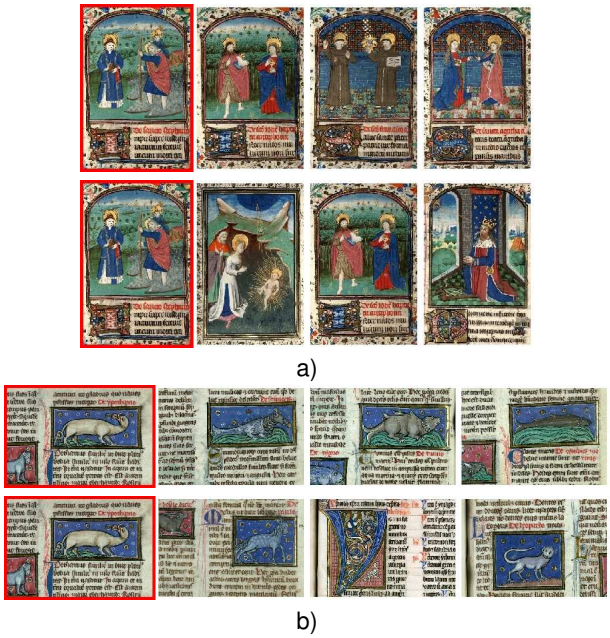

Figure 4. **Color spatial weighting using parameter α.** *A query image (red framed on the left) is followed by its three nearest matches on first lines of figure a) and b), with $\alpha = 0.5$. The same query is visible on the second line of figure a) et b) with $\alpha = 1$ (no spatial constraint). This situation is equivalent to color histogram retrieval methods with EMD distance. We clearly see on these last examples that the color content remains constant, but the spatial organization is lost.*

Such an situation is displayed on Fig. 5. On the other hand, we observed that our method perform better on images where segmentation methods generally fail, i.e. quite complex and heterogeneous images. Indeed, wrong segmentation yield regions that are not well represented, highly affecting the performances of the matching step. Figure 6 illustrates this drawback.

Statistical framework

We analyze in this section the robustness of the proposed unsupervised matching criterion. We then present more image retrieval results using this criterion. Recall that C_B the threshold on distances enabling to define meaningful matches, is estimated thanks to samples of the background model. In order to investigate the effect of the number of samples that are generated, we performed the following experiment. Background databases of $m \times N_{bg}$ realizations of the background model $\mathscr{M}$ (m being the number of images in the database $\mathscr{B}$) are synthesized, for $N_{bg} = 1, \ldots, 9$. Then, for a query image I_Q, we count how many images from $\mathscr{B}$ are meaningful matches of I_Q. Figure 7 shows the standard deviation of this number of meaningful images as a function of N_{bg}, when considering ten background databases for each N_{bg}. This experiment suggests that, despite the high variability of realizations of the dead leaves model, $N_{bg} = 3$ is enough to obtain consistent results with different background databases, and is therefore the numerical value retained in this paper.

Retrieval results are displayed on Fig. 8 to Fig. 13. For each query we display the nearest target images and indicate the automatic threshold by a grey thumbnail. Let us stress that the threshold value can greatly vary from one query to the other. The number of returned results vary from 0 (second result of Figure 12) to 23 (Figure 10) according to the query and the content of the database. The criterion is not perfect: it sometimes stops beyond

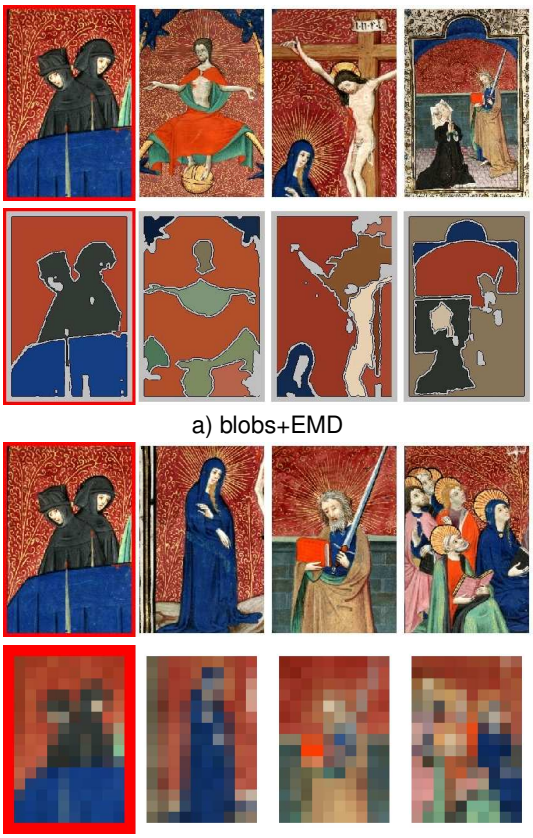

Figure 5. *Comparison between a blobs + EMD method, and our method. A query image (red framed) followed by its three nearest matches are shown on first lines of figure a) and b), according to each method. The blob representations and the thumbnails (10×15) of respective images are shown on the second line of figure a) and b) respectively. Here the blob segmentation is relatively satisfactory and our method performs similarly.*

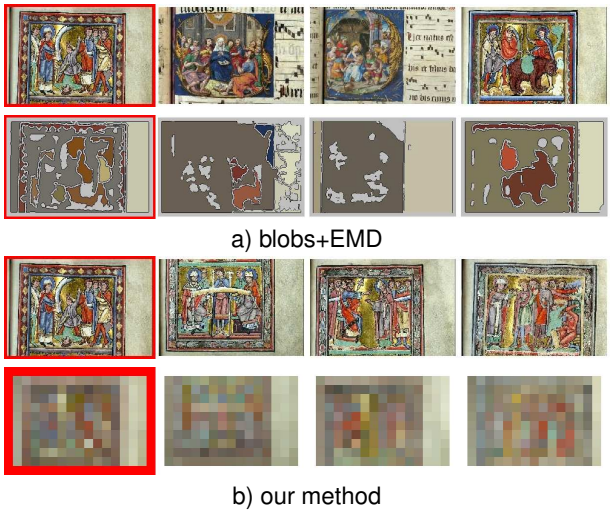

Figure 6. *Same layout as on Figure 5. Here the blob segmentation fails to give a satisfactory representation of the color spatial organization of each image. Images are clearly undersegmented, yielding large regions with average colors of different patterns, leading to unsatisfactory matching.*

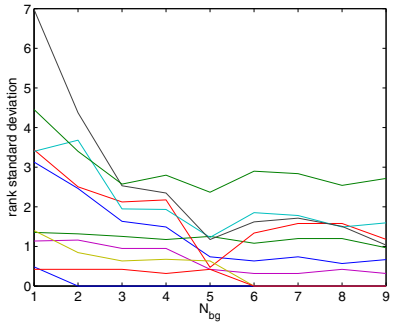

Figure 7. *Variability of the automatic threshold. Each colored line correspond to a different query image. The retrieval is made on a database of m images. The (empirical) standard deviation of the rank of the last meaningful match is shown for background databases of size $m \times N_{bg}$, for $N_{bg} = 1, \ldots, 9$. For each N_{bg}, ten background databases have been generated in order to estimate the standard deviation.*

or before where a human would exactly stops the list of results. It also moderately depends on the background model occurrence, see Fig. 7. But results are relevant most of the time. The criterion is a good indication of where the results should stop, and also inform the user on the quantity of relevant images present in the database. Here, we use a small database, and results seldom exceed a hundred, but this type of information can be crucial with much bigger databases.

Conclusion and future work

We propose a spatial color image retrieval method without initial segmentation, based on thumbnails and EMD. Its main interest is its robustness and its ability to fully use the EMD efficiency compared to region-based approaches. Experimental comparisons with a classic spatial color method is promising and confirm that complex segmentation methods are not necessary at best, misleading at worst. Future work will use more elaborated features for each thumbnail pixel including local vector quantization and texture characteristics. One drawback of our method is that EMD computing cost is high, and that querying can be slow for large databases. A possible solution could be to adapt a recent EMD approximation proposed in [23] which claims a two order magnitude speedup. This would allow for fast retrieval in databases having several hundreds thousands images.

We also propose an automatic matching criterion relying on *a contrario*. Future work will study the behavior of the threshold on much larger databases, and investigate in more details the way it adapts to the query and database specificities.

Acknowledgments

We thank Gilles Kagan from IRHT for providing us the IRHT database and for his helpful comments. We also thank Farida Cheriet for her suggestions. This work was supported by a CNRS ACI supervised by Anne-Marie Edde and Dominique Poirel from IRHT, and by a grant from the LIV4D laboratory.

References

[1] P. Lambert, N. Hervey, H. Grecu, "Image retrieval using spatial chromatic histograms," *CGIV*, pp. 343–347, 2004.

[2] G. Ciocca, R. Schettini, L. Cinque, "Image indexing and retrieval using spatial chromatic histograms and signatures," *CGIV*, pp. 255–258, 2002.

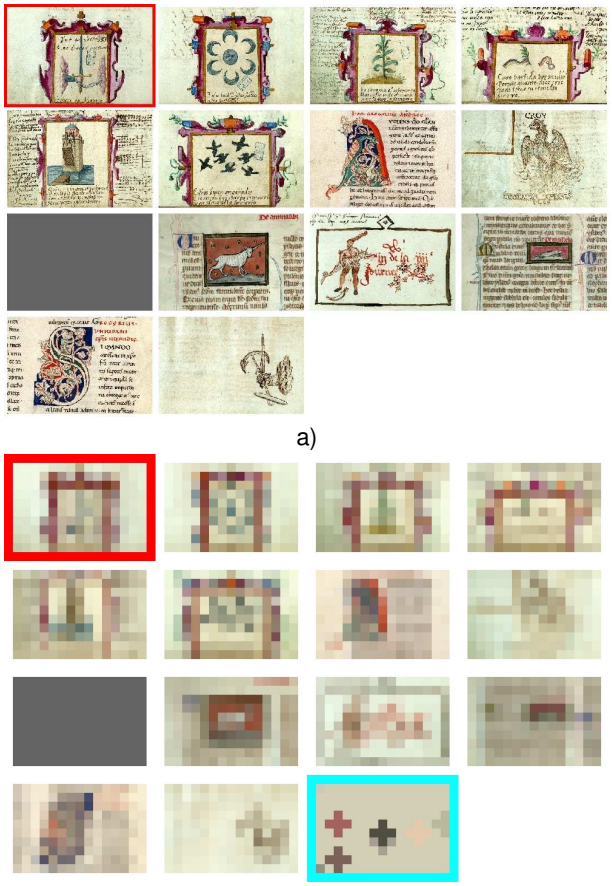

Figure 8. *The query image is red framed on top left in figure a) and is followed in scanline order by the nearest retrieved images. Meaningful matches are displayed before the grey image which corresponds to the automatic criterion. Beyond this limit, the next five nearest retrieved images are shown. Except the second one, their content is clearly not relevant according to the query. The corresponding features (10×15 thumbnails) are shown on figure b). The last thumbnail (blue framed) is the background model thumbnail found and used as a criterion for this query. As the query, this background model thumbnail is quite homogeneous.*

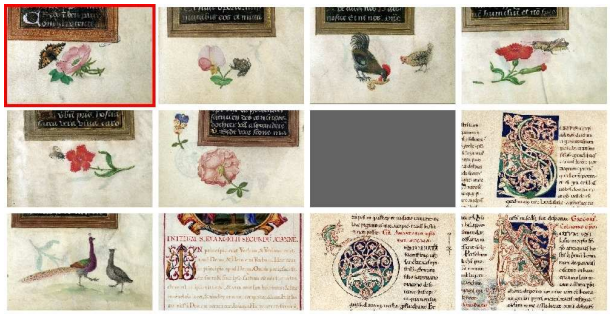

Figure 9. *The query image is red framed on top left and is followed in scanline order by the nearest retrieved images. As in Figure 8, meaningful matches are obtained before the grey image corresponding to the unsupervised criterion and lress relevant images beyond it (except image in line 3 first column).*

Figure 10. *Same layout as on Figure 9. We see here that the automatic criterion allows very variable rank limit, according to the database content and the query. Spatial organization is also important here, and one could wonder if the database contains only* 23 white illuminations with a blue rectangle. *Actually, the database roughly contains a hundred of this kind of illumination. On others, the blue rectangle has a different size and/or is not at the same location as the query.*

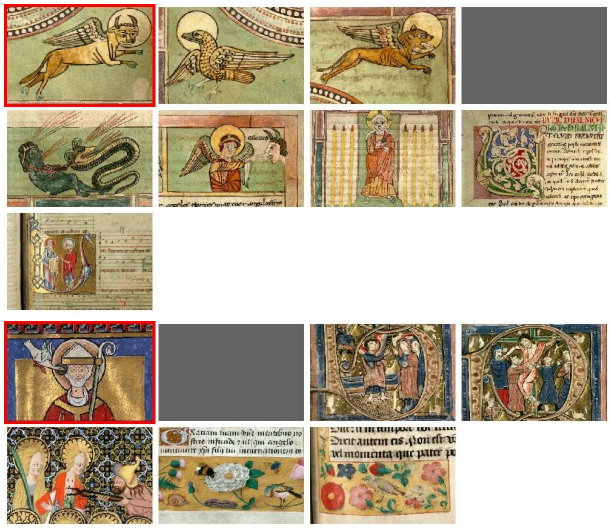

Figure 11. *Same layout as on Figure 9. Two examples where the unsupervised matching criterion provides very few results according to the database content and query. On the second example, it rightly gives no result. We see that the five closest results are not relevant according to the query.*

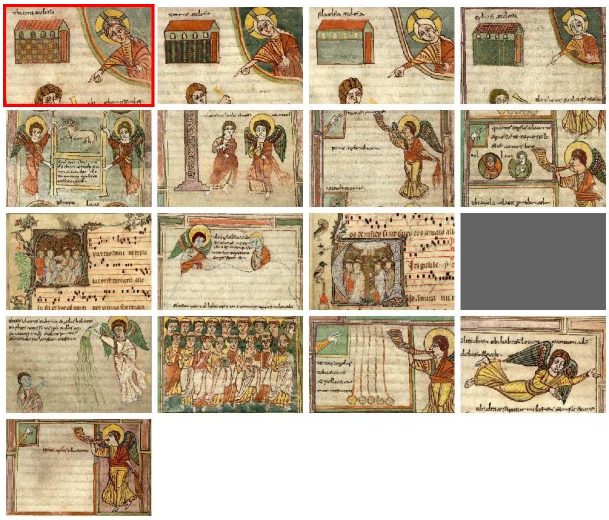

Figure 12. *Same layout as on Figure 9.*

Figure 13. *Same layout as on Figure 9.*

[3] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, R. Zabih, "Image indexing using color correlograms," *CVPR*, p. 762, 1997.

[4] Greg Pass and Ramin Zabih and Justin Miller, "Comparing images using color coherence vectors," *ACM Multimedia*, pp. 65–73, 1996.

[5] K. Grauman and T. Darrell, "Efficient image matching with distributions of local invariant features," in *Proc. CVPR*, pp. 627–634, 2005.

[6] G. Heidemann, "Combining spatial and colour information for content based image retrieval," *Computer Vision and Image Understanding*, vol. 94, pp. 234–270, 2004.

[7] C. Carson, S. Belongie, H. Greenspan, J. Malik, "Blobworld: image segmentation using EM and its application to image querying," *PAMI*, vol. 24, no. 8, pp. 1026–1038, 2002.

[8] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, "Region-based image retrieval with high-level semantic color names," *IEEE Int. Multimedia Modelling Conference*, pp. 180–187, 2005.

[9] J. D. Rugna, H. Konik, "Color coarse segmentation and regions selection for similar images retrieval," *CGIV*, pp. 241–244, 2002.

[10] B.G. Prasad, K.K. Biswas, S.K. Gupta, "Region-based image retrieval using intergrated color, shape, and location index," *Computer Vision and Image Understanding*, vol. 94, pp. 193–233, 2004.

[11] Y. Rubner, C. Tomasi, L.J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[12] G. Dvir, H. Greenspan, Y. Rubner, "Context-Based image modelling," *ICPR*, pp. 162–165, 2002.

[13] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[14] Y. Gousseau, "Comparaison de la composition de deux images, et application la recherche automatique.," in *proceedings of GRETSI 2003*, (Paris, France), 2003.

[15] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel, "An *a contrario* decision method for shape element recognition," *International Journal of Computer Vision*, 2006.

[16] A. Desolneux, L. Moisan, J-M Morel, "Meaningful alignments," *Int. Journal of Computer Vision*, vol. 40, pp. 7–23, 2000.

[17] A. Desolneux, L. Moisan, and J.-M. Morel, Gestalt Theory and Image Analysis : A Probabilistic Approach. Lecture Notes in Mathematics, Springer, 2006. To appear.

[18] G. Matheron, "Modèle séquentiel de partition aléatoire," tech. rep., CMM, 1968.

[19] L. Alvarez, Y. Gousseau, and J.-M. Morel, "The size of objects in natural and artificial images," *Advances in Imaging and Electron Physics, Academic Press*, vol. 111, pp. 167–242, 1999.

[20] A. Lee, D. Mumford, and J. Huang, "Occlusion models for natural images: A statistical study of a scale invariant dead leaves model," *International Journal of Computer Vision*, vol. 41, pp. 35–59, 2001.

[21] Y. Gousseau and F. Roueff, "Modeling occlusion and small scales behavior in natural images." submitted, 2006.

[22] T. Hurtut, "More results web page." http://www.tsi.enst.fr/recherche/cbir/.

[23] P. Indyk, N. Thaper, "Fast image retrieval via embeddings," *3rd Intl. Workshop on Statistical and Computational Theories of Vision*, 2003.