IMAGE REPRESENTATION AND RETRIEVAL WITH TOPOLOGICAL TREES

C. GRANA[†], G. PELLACANI[‡], S. SEIDENARI[‡], R. CUCCHIARA[†]
[†]*Dipartimento di Ingegneria dell'Informazione*
[‡]*Dipartimento di Dermatologia*
Università di Modena e Reggio Emilia, Italy

Typical processes of image representation comprehend initial region segmentation followed by a description of single regions' feature and their relationships. Then a graph model can be exploited in order to integrate the knowledge of the specific regions (that are the attributed relational graph's (ARG) nodes) and the regions' relations (that are the ARG's edges). In this work we use *color* features to guide region segmentation, *geometric* features to characterize regions one by one and *topological* features (and in particular *inclusion*) to describe regions' relationships. Guided by the inclusion property we define the *Topological Tree* (TT) as an image representation model that exploiting the transitive property of inclusion, uses the adjacency and inclusion topological features. We propose an approach based on a recursive version of fuzzy c-means to construct the topological tree directly from the initial image, performing both segmentation and TT construction. The TT can be exploited in many applications of image analysis and image retrieval by similarity in those contexts where inclusion is a key feature: we propose an applicative case of analysis of dermatological images to support the melanoma diagnosis.In this paper describe details of the TT algorithm, including the management of not ideality and an approximate measure of tree similarity in order to retrieve skin lesion with a similar TT-based description.

1. Introduction

A fruitful representation of the image content, often exploited in many tasks of understanding, recognition, and information retrieval by similarity, is based on region segmentation; a richer description adds to the region's attributes some relationships between regions, spatial and topological, that describe the way we perceive the mutual relations between parts of the image. To this aim, graph-based description is a power formalism to model the knowledge extracted from the images of the regions of interest and their relationships.

Moreover, the management of large volumes of digital images has generated additional interest in methods and tools for real time archiving

and retrieval of images by content[3]. Several approaches to the problem of content-based image management have been proposed and some have been implemented on research prototypes and commercial systems[4,5]. In some works, Attributed Relational Graphs (ARGs) have been introduced as a mean[6,1] to describe the spatial relationships and indexing techniques have been proposed to speed up the matching based on the edit distance[6] approach. In Petrakis' papers[1,2]ARGS and edit distance are used for image retrieval in medical image databases. Accordingly, we defined [9] the Topological Tree , a rich description model that can be constructed a posteriori, after the region segmentation step, for each type of image. However, in some applicative contexts, in which the inclusion is a key feature, the inclusion property can be exploited for segmentation too. Thus we propose an approach called Recursive-FCM (fuzzy c-means) that exploits both color and inclusion to perform segmentation and at the same time the TT construction. This algorithm has a general formulation but is meaningful in applications that search for inclusion and color: typical examples are dermatological images of skin lesions that appear as skin's zones darker than the normal skin, with many nuances of skin color. Many techniques have be proposed for color segmentation: among them, many have been adopted for skin lesion segmentation, as grayscale thresholding and color clustering[7]. *Fuzzy c-means* (FCM) color clustering has been successfully adopted in the work of Schmid[8] that adds Principal component Analysis (PCA)to FCM: a FCM segmentation over the first two principal components of the color space is tested to be meaningful and robust for skin lesion images. In a recent work[9] we described a recursive extension of that approach and here we will show further improvements that take into accounts not idealities. Moreover, in the second part of the paper we propose an approximate measure of tree similarity that can be exploited to search similarities between skin lesions in a image retrieval system.

2. Topological Relations

Given an image space and an 8-connection neighborhood system, that for each point x_i defines the neighbor set N_{x_i}, segmentation by color clustering aims to partition the image into a set of regions $\boldsymbol{R} = \{R_1, \ldots, R_k\}$ such that $\bigcup R_i = I$ and $\bigcap R_i = \varnothing$. To this aim, a clustering process that groups pixel w.r.t their color, should embed or be followed by a pixel connectivity analysis, according with the given neighborhood system.

Then a graph-based representation describes spatial and/or topological

relations between regions. An example is the *adjacency graph*, a graph $G(V, E)$ whose vertexes are the image regions ($V \equiv R$) and whose arcs show the adjacency property, that is a neighborhood system at region level. In this context adjacency is defined as follows:

Def.1: A region R_i is *adjacent* to $R_j \Leftrightarrow \exists\ x_i \in R_i,\ x_j \in R_j$: $x_j \in N_{x_i}$.

In addition to connectivity intra-region and adjacency inter-regions, we aim to evaluate inclusion of a region into another, thus we need to formally define inclusion between regions. First, we consider an "extended" set of image regions $\overline{\boldsymbol{R}} = \boldsymbol{R} \cup \{R_0\}$, being R_0 a dummy region representing the external boundary of an image. Then, we define the inclusion property as follows:

Def.2: A region $R_i \in \boldsymbol{R}$ is *included* in $R_j \in \overline{\boldsymbol{R}} \Leftrightarrow \nexists \boldsymbol{P} = \{R_1, \ldots, R_N\}$: $R_0 \cup R_i \cup \bigcup_{n=1}^{N} R_n$ is a connected region $\wedge R_j \notin \boldsymbol{P}$.

This definition means that is not possible to draw a path of connected points between region R_i and the end of the image space (R_0) that doesn't include points of R_j. The transitive property holds for inclusion: if R_i is included into R_j and R_j into R_t, than R_i is included into R_t. Thus a tree model is a natural representation for inclusion.

This ideal definition must be relaxed for implementation purposes in real images: thus we use a*FCH- inclusion* definition that substitutes in Def.2 the *filled convex hull* of R_j to R_j itself. In this mode also a not exact inclusion in a topological sense is accepted in real image description.

3. Construction of the Topological Tree

Using the FCH-inclusion propriety we developed an algorithm for providing segmentation and tree description. In previous works[9] we detailed the color based segmentation with recursive-FCM. Here we add improvements to deal with exceptions found in particular cases. The algorithm for TT construction of Fig. 1 can be summarized saying that:

(1) it carries out a color based segmentation in two clusters, using the PCA and FCM algorithm[9];
(2) while segmenting it builds the corresponding tree;
(3) it recursively applies the segmentation to the regions of interest created by the previous steps of the algorithm.

In particular the algorithm finds the presence of a region that contains all the others, inserts it in the tree and then continues to apply the algorithm

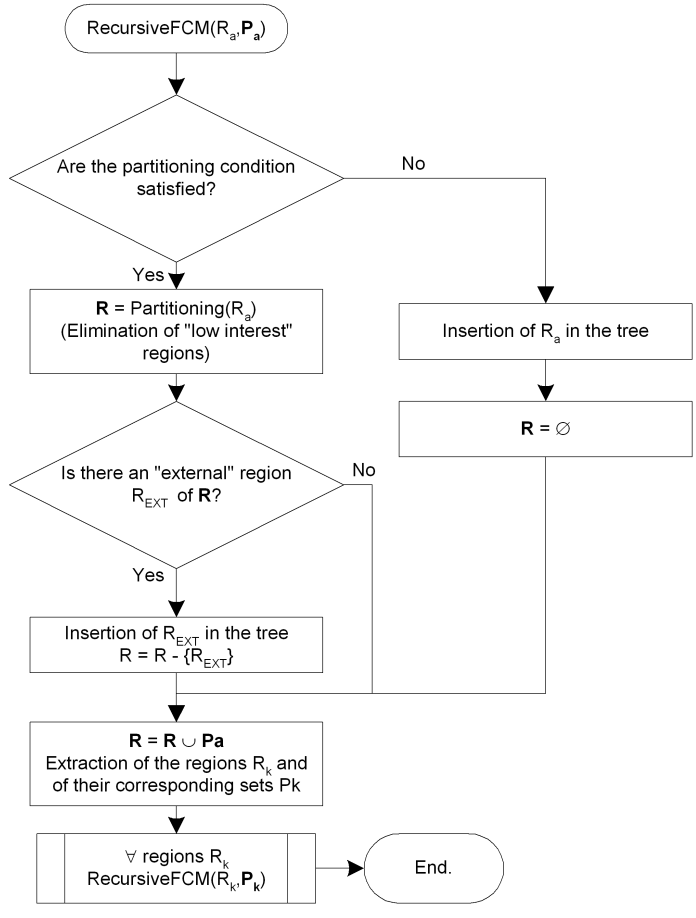

Figure 1. Algorithm flow chart

to the other regions, obtaining a further partitioning.

In Fig. 1 is possible to see the recursive structure of the algorithm *Recursive FCM* for the construction of the Topological Tree. Starting from a region R_a (initially equivalent to the whole image I) is verified if it is possible to further partition R_a. To obtain regions of interest of uniform color and significant area, the limits for the partitioning conditions shall be given by the variance of the first two components of PCA and from the size of the extracted regions. If the partitioning condition is not verified, R_a is inserted in the tree. Otherwise, R_a is clustered in more regions, the not

significant areas are erased and it is searched for the presence of an *external* one. The remaining regions are organized in a structure that allows for a correct recursion step. In particular, in the ideal case (which generates a TT with a single child for each node), the FCM algorithm creates two clusters and should create two regions, one including the other. In real images, often many regions are created. If one of these can be chosen as "external" it becomes a new node, parent of the others; all the other regions are further inspected in the recursion. However, some of these regions could be also present mutual inclusions and thus not allow a correct tree generation. We call these regions *suspended*, since they need a specific management.

3.1. *Search for an "external" region of $\mathcal{R}$*

The external region R_{EXT} of $\mathcal{R}$ is the region that FCH-includes all other regions of $\mathcal{R}$. This is the region that is searched for and added to the tree. In formulae $R_{EXT} \in \mathcal{R} : \forall R_i \in \mathcal{R}, R_i \neq R_{EXT} \Rightarrow R_i$ is included in the filled convex hull of R_{EXT} (is FCH-included).

Since is much easier and fast to check the the inclusion between the *extents* (or bounding box) of two regions (extent-inclusion), it is possible to use the observation that FCH-inclusion implies extent-inclusion to search for R_{EXT}. This is accomplished searching a region R^*_{EXT} such that all remaining regions are extent-included in it; if such a region exists, is necessary to check if all the other regions are also FCH-included in R^*_{EXT}.

3.2. *Use of "low interest" and* suspended *regions*

The decomposition of a region R_a can cause the generation of regions with negligible size. Such regions are considered as *low interest* for the interpretation of images. We will use a parameter to select the minimum area that a region can assume to be interesting. The regions that have to be eliminated are collected, during the tree construction, in a specific structure for later integration in the tree, after its complete construction.

After obtaining the set of regions $\mathcal{R}$, (in Fig. 1 from $R = partitioning(R_a)$), eliminating "low interest" regions and inserting an external region to the tree, is not possible to call the algorithm for all the extracted regions. In fact the possible presence of inclusion between the regions of $\mathcal{R}$ could lead to the construction of a wrong tree, with a loss of inclusion relationships between children produced by different clusters. Because of this, the concept of "suspended" regions has been introduced,

indicating with this term the set of all the regions that cannot be immediately analyzed, but must wait for the including one.

We thus consider the set $\mathcal{R}$ after the elimination of low interest regions and the possible external one. From $\mathcal{R}$, we distinguish between regions R_k not included in others and sets $\mathbf{P}_k$ of regions included in R_k . Now, for each region the algorithm is recursively called along with its set of suspended regions.

```
R_NI = ∅
P = ∅
∀ R_a ∈ R
{
      if (∃ R_k ∈ R_NI : R_a is included in R_k)
            P_k = P_k ∪ {R_a}
      else
      {
            R_NI = R_NI + {R_a}
            P_a = ∅
            ∀ R_k ∈ R_NI : R_k is included in R_a
            {
                  P_a = P_a ∪ {R_k} ∪ P_k
                  R_NI = R_NI - {R_k}
                  P = P - {P_k}
            }
            P = P + {P_a}
      }
}
```

Figure 2. Pseudo-code for the integration of "low interest" regions

The process for finding all suspended regions is described in Fig. 3.2.

It is to note that a reduction of the search space is obtained, by ignoring all regions of P, in fact the external region should contain not only all regions of $\mathcal{R}$ but also all the suspended regions, but for the transitive property of inclusion this is guaranteed by the fact that they are included in regions of $\mathcal{R}$.

4. Tree matching

The construction of TT is the basis of a retrieval approach searching fro tree similarities. An interesting non exact tree matching uses the *edit distance* to compare two-trees. It measure the cost of operations such as adding or eliminating nodes to transform a tree into another. Unfortunately the

edit distance based approach problem has proved to be computationally too expansive to be used without modifications in a search for similarities context. We tested this approach over our databases using linear assignment to explore all the search space but we found unacceptable response time. Moreover, it can be difficult to describe the cost of an operation in order to use it together with an inter-node, feature based, similarity. These reasons lead us to produce a quick and sub-optimal algorithm that heavily relies on two assumptions:

(1) we can match only nodes on the same level of the tree;
(2) given two sets of nodes, taken from two trees, we match one against the other without solving the associated linear assignment problem, but considering a sorting of the two sets and letting greater importance nodes have first choice on the other set.

The first assumption strong limit the search space: it is acceptable in dermatological context and is motivated by the observation that in our images each level tends to represent a specific feature as the skin, the lesion or its colored areas and an inter-level matching doesn't always make great sense. The second one is a simplification that quickly produces good results, without any assurance of reaching an optimum. An observation that qualitatively justifies this choice is the fact that higher importance nodes are weighted more in their contribution to the matching function, so guaranteeing that they get a better match leads towards an higher match direction.

The algorithm works recursively comparing two sub-trees according to the following steps:

(1) The roots are compared in an Euclidean feature space by the distance d of the feature vector. It can comprehend color, area, symmetry, texture and whichever other information of each region. An *equivalence* measure is obtained as

$$E = \frac{1}{1+d}. \tag{1}$$

(2) Children equivalence is evaluated:
    (a) Let us call it T_1 the tree with more nodes and T_2 the other one;
    (b) The nodes of T_1 are considered in order of importance (evaluated on the feature vector);

(c) Each children of T_1 is matched against all not assigned children of T_2;
(d) After evaluating the equivalence of all nodes, not assigned children of T_1 are matched against the null vector and produce a negative match.

(3) Total equivalence is given by

$$E_{tot} = \frac{E_{root}}{2}\left(\frac{\sum_i I_i E_i^*}{\sum_i I_i} + 1\right), \tag{2}$$

where E_i^* is the signed equivalence of each node (with a matching node or with the null vector), I_i is the importance of the node and E_{root} is the equivalence of the roots, as previously defined.

The equivalence measure E is bounded between 0 and 1 and this guarantees that E_i^* is in range $[-1, 1]$ so the the weighted sum can give -1 in case of total mismatch of the tree structure or 1 in case of perfect match. This value is converted by the equation to the interval $[0, 1]$ and used as a reduction factor for the matching value of the roots. The interval shift has the implicit property of reducing the influence of a mismatch at lower levels of the tree.

In this way, given an image represented by its TT we are able to find in a image database other images with a similar TT on the basis of the previous algorithm. Obviously the TT representation cannot be the unique approach to support query-by-example system and in melanoma diagnosis a number of other features[10] on the whole lesion or their part should be considered in an integrated way. Nevertheless, this is a powerful representation method that integrated with proven dermatological criteria can give interesting results in retrieval.

5. Experimental Results

Experimental results have been conducted on synthetic ad on real images, to first test the correct response of the adopted algorithm and then to verify its applicability to real world images.

5.1. *Synthetic Images*

In Fig. 3 we report an example of test over synthetic images: we want search the similarity between S1 and the whole set. Images, the TT and the match score is indicated. Obviously the matching value of 1.000 in S1 means a

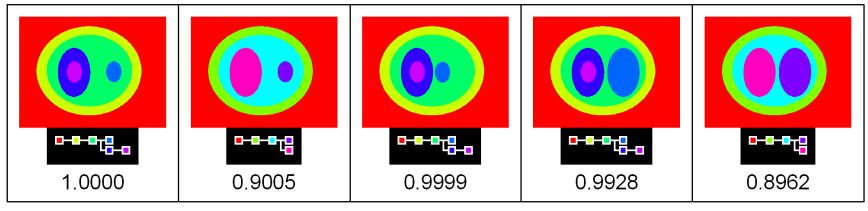

Figure 3. Synthetic images(from left to right S1,S2,S3,S4,S5)

perfect match. All other images present not so significant variation from the original image. The colors were ignored in this evaluation and the feature vector distance is provided computing only the distance of the center of mass from the parent one and the percentage of parent area occupied by the region. Thus S3 and S4 are very similar to S1 while S2 and S5 do not have a node of S1. The results follow a correct evaluation, giving the ability to order the images by the similarity from the first one.

5.2. *Dermatological Images*

Our application context is the analysis of dermatological images for melanoma diagnosis and this family of images present a natural partition of color regions included one into the other because of their usual growth process; moreover the position and size of inner areas are significant (as a diagnostic feature). In Fig. 4 some results of a query by example research are shown and an overall good retrieval was observed. In particular ideal non-melanoma skin lesion have a TT described by a list (as the last one image in Fig. 4), while melanomas typically present a more complicated structure. Unfortunately we still didn't have the possibility of including a complete diagnostic set of features in the retrieval algorithm, so a quantitative measure is not still available.

6. Conclusions

We showed a segmentation technique able to extract the inclusion-adjacency structure of the image, accompanied by a low computational cost matching technique that enables a flexible feature search over trees, instead that over the whole images. Visual comparison over synthetic and real images have been shown to assess the promising opportunity of this methodology. We would like to thank Fabio Zanella and other students for the code generation and the tests performed.

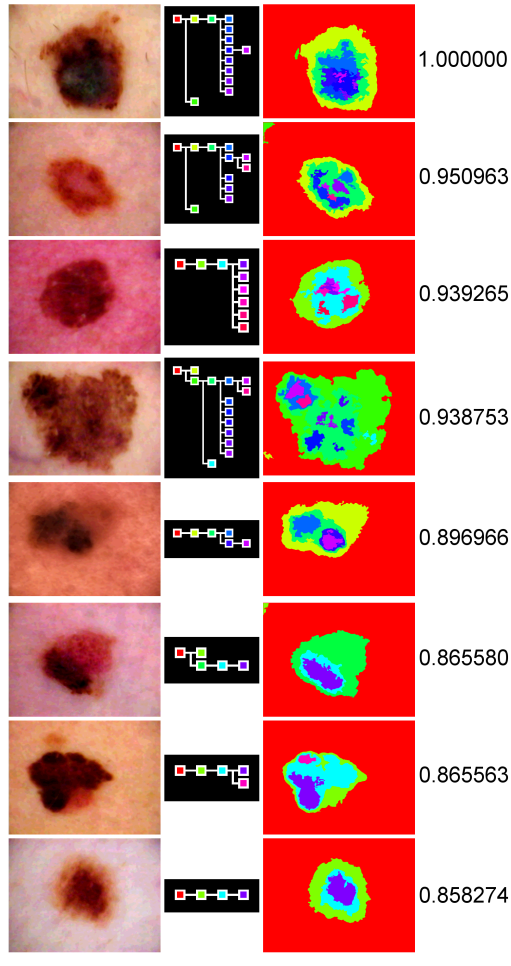

Figure 4. Experiments on real images

References

1. E.G.M. Petrakis *et al.*, Image Indexin Based on Spatial Similarity, Technical Report MUSIC-TR-01-99, Multimedia Systems Institute of Crete (MUSIC), 1999.
2. E.G.M. Petrakis *et al.*, Similarity Searching in Medical Image Databases *IEEE Trans. Knowl. Data Eng.* **9**, 435–447 (1997)
3. A.W.M. Smeulders *et al.*, Content-Based Image Retrieval at the End of the Early Years, *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1349—1380 (2000).
4. M. Flickner *et. al.*, Query By Image and Video Content: The QBIC System,

Computer **28**, 23—32 (1995).
5. A. Pentland *et al.*, Photobook: Content Based Manipulation of Image Databases, *Int. J. Comput. Vis.* **18**, 233—254 (1996).
6. B.T. Messmer, Efficient Graph Matching Algorithms. PhD thesis, Univ. of Bern, Switzerland, 1995.
7. S.E. Umbaugh *et al.*, Automatic Color Segmentation Algorithms: With Application to Skin Tumor Feature Identification, *IEEE Engineering in Medicine and Biology* **12**, 75–82 (1993).
8. Ph. Schmid, Segmentation of Digitized Dermatoscopic Images by Two-Dimensional Color Clustering, *IEEE Transactions on Medical Imaging* **18**, 164–171 (1999)
9. R. Cucchiara *et al.*, Exploiting Color and Topological Features for Region Segmentation with Recursive Fuzzy c-means, *Machine Graphics and Vision* **11**, 169–182 (2002)
10. C. Grana *et al.*, A New Algorithm for Border Description of Polarized Light Surface Microscopic Images of Pigmented Skin Lesions, in press on *IEEE Transactions on Medical Imaging* , (2003)