Object Based Image Retrieval Based on Multi-Level Segmentation

Y.Xu [(1)] ,P.Duygulu[(2)], E.Saber[(1),(3)], A.M. Tekalp[(1)], F.T.Yarman-Vural[(2)]
(1). Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627
E-mail: {yaxu, saber, tekalp}@ece.rochester.edu
(2)Department of Computer Engineering, Middle East Technical University,Ankara,Turkey,TR-06531
E-mail:{duygulu,vural}@ceng.metu.edu.tr
(3). Xerox Corporation, 800 Phillips Road, Webster, NY 14580
E-mail: eli.saber@usa.xerox.com

ABSTRACT

Currently, image retrieval systems are based on low level features of color, texture and shape, not on the semantic descriptions that are common to humans, such as objects, people, and place. In order to narrow down the gap between the low level and semantic level, object-based content analysis, which segments the semantically meaningful object on images, is an essential step. In this study, we propose a learning process in order to perform effective automatic off-line analysis on a multi-level segmented image stack. Meaningful objects are extracted given certain user search patterns and interest profiles. Color and/or shape information of the objects is stored in the hierarchical content representations of the images. This information is utilized by a hierarchical matching scheme to improve the retrieval speed in the subsequent searches

1. INTRODUCTION

Efficient image retrieval in digital visual database systems has been of great interest over the last decade. Recent developments in digital imaging technology, broadband networking and digital storage devices have set the stage for the generation, transmittal, manipulation and storage of large numbers of digital images and documents. To access these images automatically and on demand requires the ability to segment, index, store, and retrieve visual information effectively and efficiently and as such offers unprecedented challenges in the development of these technologies. These challenges have generated significant interest in the development of content based image indexing and retrieval algorithms and systems. (see: [1], [2], [3] for recent surveys on image and video indexing and retrieval technologies).

In everyday life, humans are accustomed to utilizing high level concepts, like objects, people, places, etc., to help us navigate through our daily quests. While these concepts come naturally to a human observer, they pose a significant challenge to computer systems that are attempting to perform content based image indexing and retrieval in an automatic fashion. Researchers have utilized various features such as color, shape, texture, and motion in an attempt to develop a "semantic" level of understanding for image content. Although the progress in this area has been steady and forthcoming with many papers published on a yearly basis, the research is still in its infancy stages with many breakthroughs yet to be made.

In order to narrow down the gap between the low level features and semantic level, object based content analysis, which perform semantically meaningful object segmentation on images, is an essential step. However, meaningful and accurate object segmentation has proven to be a difficult task although identification of object boundaries comes naturally to a human observer.

Segmentation of an image requires a similarity measure to identify regions, which are "homogeneous" for the pixels of the same object. The degree of homogeneity of a region depends on the threshold value selected for each similarity measure. Selection of an appropriate threshold for extraction of meaningful objects is a highly difficult and application dependent problem. Therefore, it is not possible to design a generic segmentation scheme with a generic threshold method that extracts objects from a large class of images.

In this study, a multi-level segmentation scheme [7], which generates a stack of images with "homogeneous" regions that are ordered in a tree, is proposed. At the bottom leave of the tree, there are the "most similar" pixels of the same neighborhood, in ***Lab*** color space. At the root of the tree the image is reduced to "one-color" plane. Each layer in the tree corresponds to an image in the stack. The more we approach to the bottom layer, the more details we reserve in the image.

We propose a hierarchical content description, where elementary nodes that represent regions of image with uniform color or texture form the lowest level, and composite nodes that represent semantic objects formed by learning from user profiles and search patterns form the higher levels. The similarity of content is established by searching this hierarchical representation in a top-down fashion, i.e., by first matching composite nodes if they have already been formed, to reduce search time, and thereafter, matching combinations of elementary nodes if no match has been found at higher levels. Furthermore, we present a procedure to automatically construct composite nodes by combining sets of elementary nodes according to user profiles and search patterns.

2. MULTI-LEVEL SEGMENTATION

Given an image, two neighboring pixels in 4-neihborhood forms a homogeneous region R, if they have the "closest colors" in the image palette with respect to the Euclidean distance in *lab* color space. Mathematically speaking,

If $E(k,l) < E(m,n)$, $\forall(k,l) \in \eta$, then $(k,l) \in R$,

where η indicates all he pixels in the image in 4-neighborhood. Two neighboring pixels are merged into a homogeneous region, when the Euclidean distance between the colors has the smallest possible value in the color palette of the image.

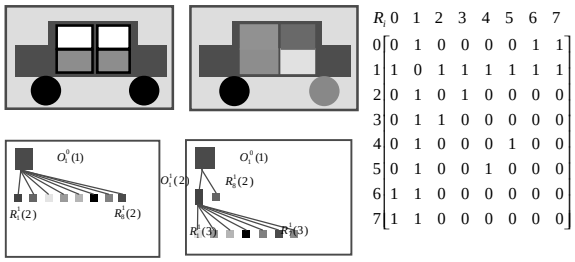
The above approach is used to generate a stack of images $I_{1,\ldots\ldots} I_M$ starting from the original image and iteratively growing the regions in the image I_j, from the homogeneous regions of image I_{j-1}. In the initial image, the neighboring pixels with the same color automatically forms the homogeneous regions. Then, as the iterations continue, the homogeneous regions grow in size and the Euclidean distance increases for the "closest color" in the same neighborhood.

Let, $\{R_{i,j}\}$, i=1,....,N, be a set of homogeneous regions in an image in level j. Then, an image at this level can be represented as mutually exclusive and collectively exhaustive set of homogeneous regions as

$$I_j = \bigcup_{i=0}^{N} R_{i,j} \qquad \text{for } j=1,.....,M.$$

Notice that a homogeneous region consists of a single pixel if its neighboring pixels have colors that are not most similar in the palette. Also, note that Ri,j-1 Ri,j for j=1,.....,M. In this study the concept of homogeneity is restricted to color.

3. HIERARCHICAL REPRESENTATION & MATCHING



R_i

	0	1	2	3	4	5	6	7
0	0	1	0	0	0	0	1	1
1	1	0	1	1	1	1	1	1
2	0	1	0	1	0	0	0	0
3	0	1	1	0	0	0	0	0
4	0	1	0	0	0	1	0	0
5	0	1	0	0	1	0	0	0
6	1	1	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0

Figure 1: Hierarchical object description. a) Original image; b) Final segmentation Map; c) Adjacency matrix; d) Parent- child relationship tree with only elementary nodes, where the big square represents the whole image and small squares represent elementary nodes; and (e) Parent-child relationship tree with an intermediate semantic-layer (composite nodes), where rectangle represents the composite node

We propose a region-based hierarchical image description represented by a "scene graph" which consists of an *object-region tree* that indicates the parent-child relationships between high-level objects and low-level image regions (henceforth regions), and an *adjacency matrix* ([4], [5]) that captures the spatial relationships between the regions. Let *N* and *L* denote the number of regions in the image segmentation map, and the number of levels within the tree respectively. The root node *I* of the tree, at level $l = 1$, corresponds to the whole image. Each leaf node $R_i^p(l)$ (also called elementary node) represents a homogeneous image region with uniform color as indicated by the image segmentation map, where *p*, *l* and *i* denote the parent node index, the level of the tree *{l = 1, …, L}*, and the leaf node index *{i = 1, …, N}* within level *l*, respectively. Furthermore, an intermediate node (also called composite node), denoted by $O_k^p(l)$ represents a semantic object, where $k = 1,\ldots, M_l$ is the index for objects in the l^{th} level and M_l denotes the number of nodes within level *l*. On the other hand, the adjacency matrix $\mathbf{A}=\{a_{ij}\}$ is such that

$$a_{ij} = \begin{cases} 1 & \text{if } R_i^p(l) \text{ is a neighbor of } R_j^p(l) \ \ \forall i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

The matching of the query template with any database image proceeds in a top-down fashion, where the template is matched against all highest-level (composite) nodes (denoting objects) in the description of the database image first. If no match has been established at the highest-level, then combinations of all next lower level nodes (which may still be composite nodes corresponding to parts of an object that consist of multiple homogeneous regions) are considered. The search terminates when all combinations of elementary (lowest-level) nodes (i.e., homogeneous regions) are considered. That is, the color and shape of regions represented by the node or combination of nodes would be matched with those of the template yielding an overall match score. The matching at each level of the content hierarchy is accomplished by utilizing the appropriate similarity measures at that level. Clearly, if database images do not contain composite nodes, as they initially do not, then only low-level searching can be performed. We now discuss computation of color and shape similarity measures.

Our approach then attempts to match the query template with the hierarchical description of all database images using a top-down approach. If the database has not been searched using this query template before, then, the search must consider all combinations of neighboring elementary nodes to determine if a match satisfying the user specified threshold (t_c or t_s, for color or shape) can be identified. If a match is established, then we form a composite node containing the matching combination of elementary nodes, the query template, and the similarity measure between them. This composite node would also contain, from this point on, the appropriate color and shape attributes of the "red car." This process constitutes a learning step that would not have taken place had we not performed the match. As a result, subsequent searches using the same query template and a similarity threshold t_c or t_s which is tighter than the stored similarity measure would immediately identify the composite node as a match using its shape and/or color attributes without

the need to process all lower level nodes. This concept of learning from a user query can also be extended to learning from user preferences, where the learning is done off-line as the image in introduced into the database. This assumes that the user possesses an interest profile consisting of frequently searched templates.

The hierarchical content description and object formation by learning through the use of composite nodes provides significant advantages. First and foremost is the notion of semantic object segmentation (formation) without user intervention. Another advantage of the proposed composite node structure is a significant reduction in the number of combinations to be tested during subsequent searches using the same template.

4. SEARCH STRATEGIES BASED ON MULTI-LEVEL SEGMENTATION

In the above described methods, several difficulties associated with image analysis have been identified:

1. **Over-segmentation or Under-segmentation**: Problem with over-segmentation is that it leads to too many small regions. As we have known that the complexity with the first search of certain object is $O(e^n)$, where every possible combination of the regions has to be checked. Problem with under-segmentation is the wrong labeling of certain amount of pixels. Resulting regions may be too big that small objects have been accidentally hidden in large region and can not match to the object.
2. **Objects are in different scale:** It is very obvious that the objects in certain images can be in different scales, while in Current image segmentation procedure, using single size criteria to remove small regions could lead to that smaller objects with small region size are "eaten" by big objects with large region size.
3. **Computation Complexity**: In the current matching scheme, the first search of an object in the database is very expensive, since all neighboring combination of regions in every image will be matching to the object using "shape matching" or "color matching".

For overcoming the above problems, the multi-level image segmentation is proposed. Instead of generate only one segmentation map for one given image, multiple segmentation map are generated for any color image, which can classified as from coarse level, where the image is only segmented to two separated regions, to the possible finest level, where each pixel in the image is a single region. (Practically, instead of being segmented to single pixels, images are only segmented to the level that somehow over-segmented.)

5. EXPERIMENTAL RESULTS

Figures 2 and 3 demonstrate the result of multi-level segmentation, on real life "car" and "face" images respectively. The advantages of the "hierarchy" and "learning" mechanisms are clearly highlighted. The queries utilized are "shape" and "color" respectively. In each Figure, the original image is indicated in (a), the lowest level segmentation obtained by utilizing the multi-level segmentation scheme is given in (b), the initial hierarchy with elementary nodes are in (c), the "semantic" segmentation obtained as a result of the query is in (d), the modified hierarchy is in (e), and the ground truth image is in (f). The "query" template is also shown with each set.

The proposed algorithm provides: 1) a "semantic" segmentation (see (d)) that can be utilized, at the very least, in future queries of the image, and 2) an updated hierarchy with composite level nodes (see (e)) that can be employed to reduce future "car" and "face" search and retrieve queries respectively by reducing the number of combinations to be tested.

Table 1 provides a summary of the performance of our proposed algorithms. In the table, we show: a) the level of segmentation utilized in our top-down search to match the query template to the appropriate combination of regions resulting in the semantic segmentation, b) the number of regions found in the image at that segmentation level, c) the number of valid combinations at the first and subsequent searches, and d) the similarity measure. The main point demonstrated by Table 1 is the significant advantage gained by the multi-level segmentation in reducing the number of combinations to be tested at the initial search due to its ability to provide "coarse" to "fine" segmentation levels that can be searched in a top-down fashion , thereby reducing the number of combinations to be tested. The alternative would be to provide a single segmentation layer at the "fine" level with multiple regions to be tested while accepting a significant increase in the number of combination to be tested at the initial search.

6. CONCLUSION

In this paper, we proposed a learning procedure that captures object-region relationship in searching the user profiles or search patterns. Visual objects are formed in a learning process and represented by composite nodes in the hierarchical scheme, in which the elementary nodes correspond to the homogeneous image regions found in a multi-level segmentation process. The object-region relationships are further used in a hierarchical matching scheme to improve the image retrieval speed.

There are two primary advantages of the proposed scheme: First of all, representing the image by a multi-level stack removes the threshold problem in the segmentation, thus the objects in any size or detail can be extracted in one of the image stacks. Secondly, this image stack significantly improves the speed of retrieval for subsequent searches as a consequence of the introduction of composite nodes. The superiorities of the proposed scheme are clearly shown in the tables and figures of the experimental results.

Acknowledgment: The authors would like to thank Mr. Nafiz Arica for his many helpful comments and suggestions.

7. REFERENCES

[1] Yong Rui, Thomas S. Huang, and Shih-Fu Chang, "Image retrieval: current techniques, promising directions and open issues", *Journal of Visual Communication and Image Representation*, Vol. 10, no. 4, pp. 39-62, April 1999

[2] F. Idris and S. Panchanathan, "Review of image and video indexing techniques", *Journal of Visual Communication and Image Representation*, vol. 8, no. 2, pp. 146-66 June 1997.

[3] M. De Marsicoi, L. Cinque, and S. Levialdi, "Indexing pictorial documents by their content: a survey of current techniques", *Image and Vision Computing*, vol. 15, no. 2, pp. 119-41, Feb. 1997.

[4] M. Swamy and K. Thulasiraman, "Graphs, Networks, and Algorithms", *Wiley, NY, 1981*.

[5] R. Sedgewick, "Algorithms in C", Addison-Wesley, Reading, MA, 1990

[6] Y.Xu, E.Saber, A.M.Tekalp, "Hierarchical Content Description and Object Formation by Learning", *Computer Vision and Pattern Recognition Workshop*, Colorado, 1999

[7] P. Duygulu and F.T. Yarman-Vural, Multi-Level Image Segmentation Based on a Simple Color Descriptor, submitted to ICASSP2000

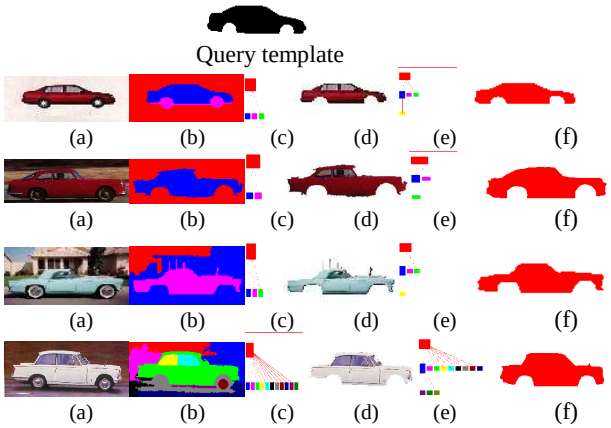


Figure 2: Results on car images using shape matching: a) The original image; b) Final low-level segmentation using multi-level segmentation scheme; c) Parent-child relationship tree with elementary nodes only; d) Semantic segmentation matching the query template; e) Parent-child relationship tree with an intermediate semantic-layer, where the rectangle (i.e. the composite node) represents the segmented car. (f)ground truth image

Table 1 Performance evaluation of retrieval rate before and after learning .

Image name	Car 1	Car 2	Car 3	Car4	Face 1	Face 2	Face 3	Face 4
Segmentation level	2	1	2	10	2	3	2	10
Number of regions	3	2	3	11	3	4	4	11
Number of valid combinations at first search	7	3	6	735	7	12	12	824
Similarity	3.7	7.6	11.7	5.1	0.9	0.7	0.5	0.5
Number of valid combinations at subsequent searches	3	2	3	60	3	3	6	41

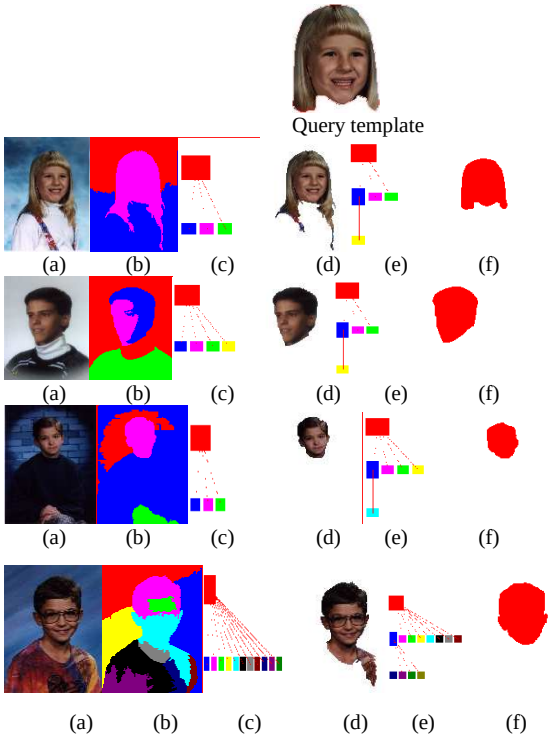


Figure 3: Results on face images using color matching: a) The original image; b) Final low-level segmentation using multi-level segmentation scheme; c) Parent-child relationship tree with elementary nodes only; d) Semantic segmentation matching the query template; e) Parent-child relationship tree with an intermediate semantic-layer, where the rectangle (i.e. the composite node) represents the segmented face. (f)ground truth image