An Unsupervised Method for Clustering Images Based on Their Salient Regions of Interest

Gustavo B. Borba and Humberto R. Gamba
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial
Universidade Tecnológica Federal do Paraná
Curitiba - Paraná - Brasil
{gustavo, humberto}@cpgei.cefetpr.br

Oge Marques and Liam M. Mayron
Department of Computer Science and Engineering
Florida Atlantic University
Boca Raton, FL – USA
{omarques, lmayron}@fau.edu

ABSTRACT

We have developed a biologically-motivated, unsupervised way of grouping together images whose salient regions of interest (ROIs) are perceptually similar regardless of the visual contents of other (less relevant) parts of the image. In the implemented model cluster membership is assigned based on feature vectors extracted from salient ROIs. This paper focuses on the experimental evaluation of the proposed approach for several combinations of feature extraction techniques and unsupervised clustering algorithms. The results reported here show that this is a valid approach and encourage further research.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Algorithms, Human Factors, Performance.

Keywords

Visual Attention, Image Retrieval, Clustering.

1. INTRODUCTION

The dramatic growth in the amount of digital images available for consumption and the popularity of inexpensive hardware and software for acquiring, storing, and distributing images has fostered considerable research activity in the field of content-based image retrieval (CBIR) [6]. In spite of the large number of related papers, prototypes, and several commercial solutions, the CBIR problem has not been satisfactorily solved.

Chen et al. [1] have shown that clustering and ranking of relevant results is a viable alternative to the usual approach of presenting the results in a ranked list format. The results of their experiments motivated the cluster-based approach taken in our work.

We have developed a CBIR solution [4] in which results from two different computational models of visual attention (VA) are combined to extract ROIs in an unsupervised way.

In [4] we present a complete evaluation of the ROI extraction algorithm as well as performance measures for the entire system. In this paper we focus on testing the proposed model for a combination of feature extraction and clustering algorithms. In doing so, we are using classical clustering evaluation techniques – such as measures of purity and entropy – as indirect measures of success of the overall approach.

2. THE PROPOSED MODEL

This section presents an overview of the proposed model and explains its main components in detail.

2.1 Overview

We present a biologically-plausible model that extracts ROIs using saliency-based visual attention models, which are then used for the image clustering process.

The visual attention models used are those proposed by Itti and Koch [3] and Stentiford [7]. The Itti-Koch model of visual attention considers the task of attentional selection from a purely bottom-up perspective, although recent efforts have been made to incorporate top-down impulses [3]. The model generates a map of the most salient points in an image, the *saliency map*. The model of visual attention proposed by Stentiford [7] is also a biologically inspired approach to CBIR tasks. It functions by suppressing areas of the image with patterns that are repeated elsewhere. As a result flat surfaces and textures are suppressed while unique objects are given prominence. Regions are marked as high interest if they possess features not frequently present elsewhere in the image. The result is a visual attention map that is similar in function to the saliency map generated by Itti-Koch.

There are several key aspects that our model adheres to. It is biologically-inspired. The Itti and Stentiford models are both biologically-inspired while the biological-plausibility of clustering the results is verified by Draper et al. [2]. Our model is unsupervised and content-based: it is able to function without the intervention of a user, producing clusters of related images at its output. We limit our model to incorporating only bottom-up knowledge. Finally, our model is modular. While we rely on the Itti-Koch model of visual attention, our model allows for a variety of other models of visual attention to be used in its place. Similarly, the choice of feature extraction techniques and descriptors as well as clustering algorithms can also be varied. This allows a good degree of flexibility and fine-tuning (or customization) based

on results of experiments, such as the ones described in Section 3.

2.2 Components

Our model consists of the following four stages: early vision, ROI extraction, feature extraction, and, clustering.

2.2.1 Early vision

The first stage models early vision. Its purpose is to indicate what the most salient areas of an image are. The input to this stage is a source image. The output is the long-range saliency map generated by the Itti-Koch model of visual attention [3].

2.2.2 Region of interest extraction

The second stage of our model generates ROIs that correspond to the most salient areas of the image. It is inspired by the approach used by Rutishauser et al. [5]. Our model appreciates not only the magnitude of the peaks in the saliency map, but the size of the resulting salient regions as well. The extracted ROIs reflect the areas of the image we are likely to attend to first. Only these regions are considered for the next step, feature extraction.

The algorithm for extracting one or more regions of interest from an input image described in this paper combines the saliency map produced by the Itti-Koch model with the segmentation results of Stentiford's algorithm in such a way as to leverage the strengths of either approach while minimizing the impact of their shortcomings. More specifically, two of the major strengths of the Itti-Koch model – the ability to take into account color, orientation, and intensity to detect salient spots (whereas Stentiford's is based on color and shape only) and the fact that it is more discriminative among potentially salient regions than Stentiford's – are combined with two of the best characteristics of Stentiford's approach – the ability to detect entire salient regions (as opposed to Itti-Koch's peaks in the saliency map) and handle regions of interest larger than the 5% ROS limit mentioned in [5].
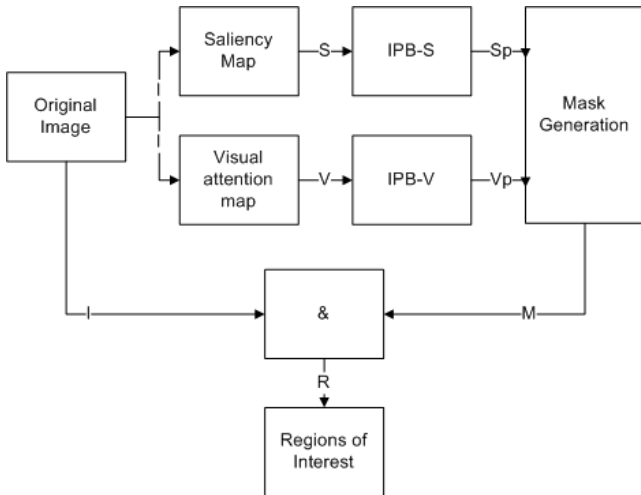

Figure 1: General block diagram of the ROI extraction algorithm.

Figure 1 shows a general view of the whole ROI extraction algorithm. The basic idea is to use the saliency map produced by the Itti-Koch model to start a controlled region growing of the potential ROIs, limiting their growth to the boundaries established by Stentiford's results and/or a predefined maximum ROS. The first step is to extract the Saliency (S) and VA (V) maps from the original image (I). While the saliency map returns small highly salient regions (peaks) over the ROIs, the VA map returns high VA score pixels for the entire ROIs, suggesting that a combination of S and V could be used in a segmentation process. In figure 1, the IPB-S (Image Processing Box) block takes S as input and returns a binary image S_p containing small blobs that are related to the most salient regions of the image. The IPB-V block takes V as input and returns a binary image V_p, containing large areas with high VA scores, instead of blobs. Images S_p and V_p are presented to the Mask Generation block, that compares them and uses the matching regions as cues for selection of the ROIs into V_p. The result is a near perfect segmentation of the ROIs present in the original image I.

2.2.3 Feature extraction

The proposed system allows using any combination of feature extraction algorithms commonly used in CBIR applied on a region-by-region basis. Each independent ROI has its own feature vector. An image may be associated with several different feature vectors.

The current prototype implements two color-based feature extraction methods: 3×3×3, 5×5×5 and 6×6×6 quantized RGB histogram (27, 125 and 216 bins) and a 32-, 128-, and 256-cell quantized HMMD (MPEG-7-compatible) histogram (32, 128 and 256 bins).
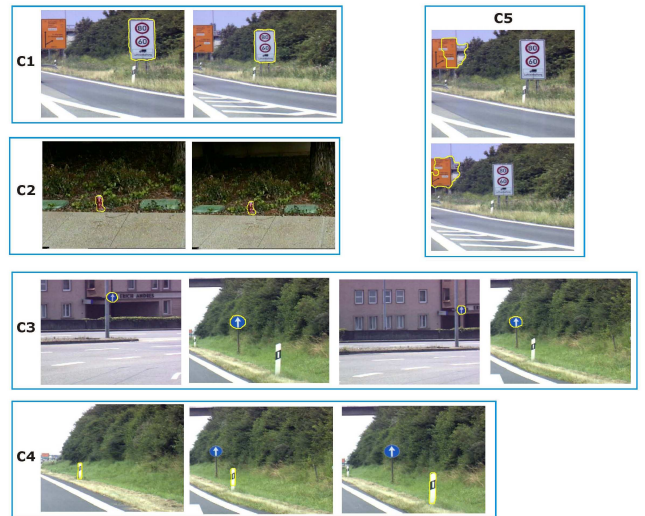

Figure 2: Examples of clustering based on ROIs for a small dataset. The extracted ROIs are outlined. The images can be found at http://ilab.usc.edu/imgdbs/.

2.2.4 Clustering

The final stage of our model groups the feature vectors together using a general-purpose clustering algorithm. The current prototype implements four clustering algorithms,

namely: k-means, partitioning around medoids (PAM), fuzzy c-means, and hierarchical.

Just as an image may have several ROIs and several feature vectors it may also be clustered in several different, entirely independent, groups. This is an important distinction between our model and other cluster-based approaches, which often limit an image to one cluster membership entry. The flexibility of having several ROIs allows us to cluster images based on the regions (objects) we are more likely to perceive rather than only global information.

Figure 2 shows the results of clustering nine images containing five ROIs with possible semantic meaning, namely: road marker, blue/white road sign, orange/black road sign, white/red road sign, and Coke can. It can be seen that the proposed solution does an excellent job grouping together all occurrences of similar ROIs into the appropriate clusters. Among the resulting clusters, cluster C3 captures an essential aspect of the proposed solution: the ability to group together similar ROIs (blue/white road signs in this case) in spite of large differences in the background.

3. EXPERIMENTS AND RESULTS

This section contains representative results from our experiments and discusses the performance of the proposed approach on a representative dataset.

3.1 Methodology

We built a database containing images with semantically well-defined ROIs (regions that are *salient by design*), made of photographs of scenes with a combination of naturally occurring and artificial objects. Further details of the dataset used may be found in [4].

For each image the corresponding saliency map was extracted and used to compute the relevant ROIs using the algorithm described in Section *2.2.2*. Each ROI was encoded using either an RGB color histogram or a quantized HMMD color descriptor as the feature vector. The resulting feature vectors were clustered using different clustering algorithms.

3.2 Results

The ROI extraction algorithm was assessed in [4], presenting the following rates: true positive (TP) = 77%, false negative (FN) = 23% and false positive (FP) = 30%.

Quantitative evaluation of the clustering stage was performed on raw confusion matrices obtained for each relevant case. The analysis was done from two different angles: (i) we used measures of purity and entropy (defined in equations 1 and 2 below) to evaluate the quality of the resulting clusters; and (ii) we adopted measures of $F1$ (which is described in equation 5 as a function of precision (p) and recall (r), equations 3 and 4, respectively) to capture how well a certain semantic category was represented in the resulting clustering structure.

Given a number of categories c, we can define purity as:

$$p(C_j) = \frac{1}{|C_j|} \max_{k=1,\dots,c} |C_{j,k}| \qquad (1)$$

while entropy can be defined as:

$$h(C_j) = -\frac{1}{\log c} \sum_{k=1}^{c} \frac{|C_{j,k}|}{|C_j|} \log \frac{|C_{j,k}|}{|C_j|} \qquad (2)$$

Where: $|C_j|$ is the size of cluster j, and $|C_{j,k}|$ represents the number of images in cluster j that belong to category k.

Purity values may vary between $1/c$ and 1 (best), whereas entropy values may vary between 0 (best) and 1.

In the context of clustering:

$$p = \frac{|C_{j,k}|}{|C_j|} \qquad (3)$$

$$r = \frac{|C_{j,k}|}{|C_k|} \qquad (4)$$

$$F1 = \frac{2 \times p \times r}{(p + r)} \qquad (5)$$

Where: $|C_j|$ is the size of cluster j, $|C_{j,k}|$ represents the number of images in cluster j that belong to category k, and $|C_k|$ represents the total number of images that belong to category k.
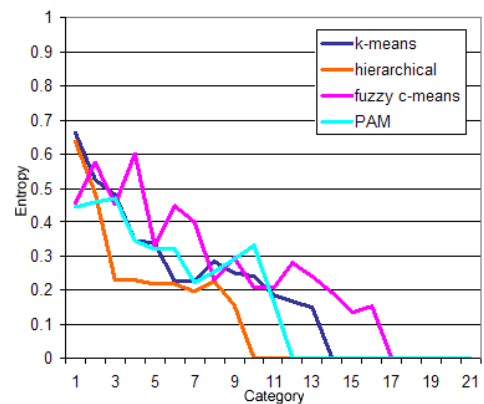

Figure 3: Measure of entropy for all clustering algorithms.

Figure 4 shows the variation in the measure of purity for all four clustering algorithms, whereas Figure 3 shows the corresponding plot for measures of entropy. In both cases, the values have been sorted so that best results appear on the right-hand side of each figure. For both cases, the hierarchical clustering method emerges as the best of all four.

Figure 5 shows the variation in the measure of maximum value of F1 for two different feature extraction methods (216-bin RGB color histogram and 256-bin quantized HMMD histogram) and the best clustering algorithm. The HMMD descriptor outperforms the RGB histogram in almost all clusters.

Figure 6 shows the variation in the measure of maximum value of F1 for three different feature vector sizes (32, 128, and 256 bins quantized HMMD histogram) and the best clustering algorithm. Results for the three cases are comparable.

3.3 Discussion

Experiments with four different clustering algorithms (K-means, Fuzzy C-means, PAM, and Hierarchical) have shown that for a certain feature vector (HMMD, 256 bins), the
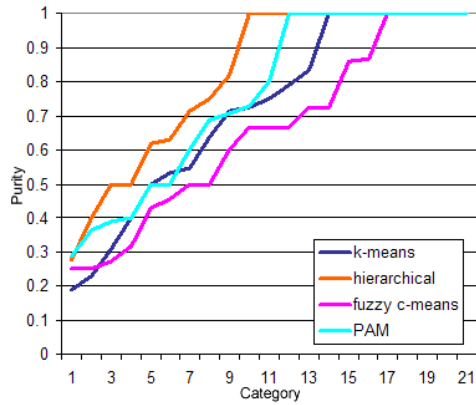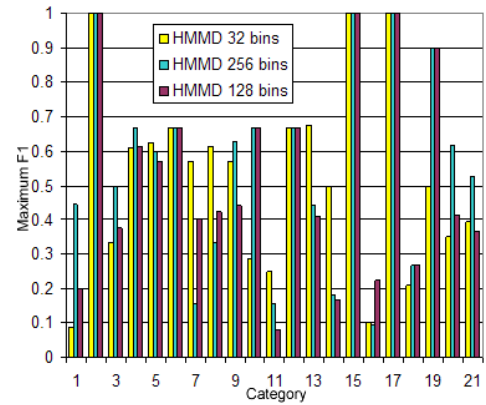
Figure 4: Measure of purity for all clustering algorithms.

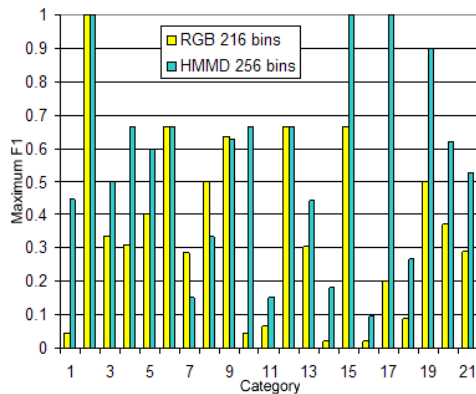**Figure 6: Measure of maximum value of F1 for three different feature vector lengths.**

4. CONCLUSION

This paper presented a model for grouping images based on their salient regions. It uses a biologically-inspired model of visual attention to detect the most salient points within an image. These are then used as seeds for extracting regions of interest. Regions are processed by a feature extraction module. The results are used to assign cluster membership. Images containing perceptually similar objects are grouped together, regardless of the number of occurrences of an object or distracting factors. Results of our experiments using standard feature descriptors and clustering algorithms are encouraging and suggest that the approach should be extended and improved. Future work includes further study of user needs as well as the incorporation of relevance feedback into the existing implementation.

Figure 5: Measure of maximum value of F1 for two different feature extraction methods.

Acknowledgment

hierarchical method outperforms the others both in terms of purity and entropy (Figures 4 and 3).

Since purity and entropy provide a measure of the quality of the clusters which is somewhat independent of the intended cluster distribution, we chose the maximum F1 figure of merit to compare two different color-based feature vectors (HMMD 256 bins versus RGB 216 bins) for the best clustering method (hierarchical). Results (Figure 5) show that the HMMD feature vector outperforms RGB in almost all cases. These results are probably due to the chosen color space (which is closer to a perceptually uniform color space than the RGB counterpart) and to the non-uniform subspace quantization that it undergoes.

Since feature vector size is of primary concern for large databases, the next comparison looked at how the quality of clustering was impacted by using different number of bins (32, 128, 256) for the same (HMMD) feature extraction technique. Figure 6 show that the results are comparable, which suggest the use of the most compact (32 bins) of the three.

5. REFERENCES

[1] Y. Chen, J. Z. Wang, and R. Krovetz. CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE Trans. on Image Processing*, 14(8):1187–1201, Aug 2005.

[2] B. Draper, K. Baek, and J. Boody. Implementing the expert object recognition pathway. In *International Conference on Vision Systems, Graz, Austria*, 2003.

[3] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259, Nov 1998.

[4] O. Marques, L. M. Mayron, G. B. Borba, and H. R. Gamba. An attention-driven model for grouping similar images with image retrieval applications. *Eurasip Journal on Applied Signal Processing (submitted)*.

[5] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *IEEE Conf. on CVPR*, pages 11–37, 2004.

[6] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22(12):1349–1380, Dec. 2000.

[7] F. Stentiford. An attention based similarity measure with application to content-based information retrieval. In *Proceedings of the Storage and Retrieval for Media Databases Conference, SPIE Electronic Imaging*, Santa Clara, CA, 2003.