Coarse Adaptive Color Image Segmentation for Visual Object Classification

A. Pujol and L. Chen
LIRIS
Ecole Centrale de Lyon
Batiment E6, 15 Avenue Guy de Collongue, 69134 Ecully Cedex
Phone: (33) 04-7218 6459 Fax: (33) 04-7218 6443 E-mail: alain.pujol@ec-lyon.fr

Keywords: Image processing, color, segmentation, clustering.

Abstract - This paper deals with perceptually inspired image segmentation for the purpose of generic image classification or object detection. Indeed, in our algorithm we will try to stay true to human perception and more specifically Gestalt Theory. Input images are processed in a three-step framework: a pre-processing step where the image is filtered and where perceptually similar colors are grouped as per color constancy law, a clustering step where we also determine an optimal number of quantized colors and a post processing step where we add spatial information and merge smaller regions as per "good continuation" and proximity laws. Another major feature of our algorithm is that it adapts to image dynamics and doesn't require image-specific parameter tuning. Application on a 10,000 image dataset shows the algorithm succeeds in producing large coarse regions that can be used for feature extraction.

1. INTRODUCTION

Visual object classification or detection within unconstrained images is very challenging problem with a broad range of potential applications. Current successful approaches in classification problems such as the Pascal VOC Challenge [1] are mainly based on a "bag of features" kind of approach [2], [3] which captures statistical properties of local independent features. Our basic hypothesis, on the other hand, is that effective visual object classification or detection should be driven by visual perception principles. For instance the well known Gestalt laws of Perceptual Organization suggest both the grouping of pixels into homogeneous regions as well as the interaction between regions. We feel that lacking these principles, "bag of features" approaches deprives themselves of meaningful information.

In this paper we propose a color segmentation scheme designed to follow basic Gestalt principles (namely color constancy, vicinity, similarity and good continuation laws), remove the necessity of image dependant thresholds, as well as emphasize robustness by ensuring the output remains usable for feature extraction. This results in perceptually significant partial gestalts for further visual object analysis. After a brief introduction of Gestalt theory and an overview of the existing work in section 2, we will describe our proposed algorithm in section 3. Section 4 presents some experimental results and future work while we will have some concluding remarks in section 5.

2. RELATED WORK

A.Desolneux, L.Moison and J.M.Morel have given in [4] a comprehensive introduction to Gestalt theory in an image analysis perspective. Gestalt theory starts with the assumption of active grouping laws in visual perception which recursively cluster basic primitives into a new, larger visual object, a gestalt. These grouping laws follow criterion such as spatial proximity, color similarity. These laws also highlight the interaction between regions.

The principle of our image segmentation algorithm is to segment an image into partial gestalts for further visual object recognition. We thus made use of the following Gestalt basic grouping laws in our gestalt construction process: The color constancy law stating that connected regions where color does not vary strongly are unified; the similarity law leading to group similar objects into higher scale object; the vicinity law suggesting grouping close primitives with respect to the others; and finally good continuation law saying that reconstructed amodal object should be as homogenous as possible. Because those laws are defined between regions and their context, at each step we assess the possibility to merge according to global information. Through this we also adapt to image contents.

Most visual object analysis systems found in the literature do not perform any image segmentation and instead apply a statistical computation of local features on some "points of interest". However, if we want to carry out visual object recognition according to a Gestalt Theory-based approach, we first need to recursively merge by Gestalt grouping laws basic primitives, for instance pixels, into more and more composite gestalts, which amounts at its first level to segment an image into regions.

There exists a very abundant literature on image segmentation. We can roughly classify existing approaches into two major categories [5]: pixel color classification based and spatial relationship based methods. Pixel color classification based methods are basically clustering problems and as such present the usual difficulties: with about tens of thousands of colors clustering is complex and very slow; Moreover the determination of the optimal number of colors mostly is an image dependent parameter to be tuned. We therefore need either adaptive algorithms like in [6] or algorithms that evaluate an optimal number of colors. Spatial relationship based methods work within image space rather than color space. We can again distinguish three subcategories : edge based methods such as geodesic active contours [7], region-based methods including region merging, region splitting and split and merge methods [8], and finally variational segmentation methods such as Edgeflow [9].

All these techniques are more or less application or image content dependent as region of interest is typically application dependent. For the purpose of generic visual object classification problem, the nature of images is very complex and gives birth of a large variability in lighting conditions, scale, background, pose, etc. Our conjuncture is that reliable generic visual object recognition needs to rely on coarse bigger regions, thus making use of the recursive application of Gestalt basic grouping principles to segment image content into perceptually significant “objects”. Moreover, our segmentation process should avoid image dependant parameters, having in mind the diversity of images and the size of the dataset.

3. PROPOSED ALGORITHM

3.1 Clustering Preparation

As we said clustering methods suffer from the vast amount of color data within an image. Regrouping identical colors and weighting their values using their population still leaves us with tens of thousands of colors; which remains computationally very expensive. After applying vector median filtering [10] to add some robustness to noise, we first reduce image color depth. On the other hand, adding a preclustering step is dangerous, as chaining processes sums up their respective flaws and result in increased approximations. To limit these problems, we propose a simple preclustering step that only groups colors which are perceptually similar according to an appropriate metric. We therefore chose to agglomerate colors using an accumulator array within the CIELab color space as it shows acceptable perceptual homogeneity. Given a fixed grid size, colors falling in a same cuboid are agglomerated and replaced by the barycenter of all the colors within the cuboid. An important factor here is obviously the size of the grid: in order not to harm segmentation accuracy, we choose to select a grid size that only allows merging colors which are perceptually similar.

To measure perceptual similarity, we use the CMC distance ([5], [11]) which is one of the most accurate metrics according to human perception. Equation (1), show the expression of this distance.

$$\Delta E = \sqrt{\left(\frac{\Delta H}{Sh}\right)^2 + \left(\frac{\Delta L}{l \cdot Sl}\right)^2 + \left(\frac{\Delta C}{c \cdot Sc}\right)^2} \qquad (1)$$

Sh, Sl and Sc are determined, $\Delta H, \Delta L$ and ΔC are differences computed in a transformation of Lab Color space (Hue, Chroma and Luminance).

As we can see it expresses the ellipsoidal shape of perceptual similarity areas described in the works of Wright and Mc. Adam. We then compute distances between all possible colors coded using RGB values which, once converted to CIELab, fall into the same cuboid. According to CMC metrics, two colors are perceptually similar provided the distance between them is inferior to 1. We thus seek for cuboid dimensions that would give a maximal CMC distance inferior to 1. This was achieved with a 0.33 x 0.66 x 0.66 cuboid (in Lab space). Such an agglomeration guarantees that two merged colors will be perceptually similar. However it does not provide a sufficient reduction of the amount of processed data. Therefore we loosened the constraint and chose cuboid dimensions which guaranteed that 90% of agglomerated colors would be perceptually similar. This produces a 1.5 x 3 x 3 cuboid size and allows performing a quick agglomeration by replacing colors that fall within a same cuboid by their barycenter. This process reduces the number of colors by an average of 89% (computed on a sample diversified set of 600 images). Maximal possible distance is 4 and 98% of possible distances are below 1.5. Experiments revealed this approximation was reasonable. Cuboid size may, of course, be adjusted as a part of a speed vs. accuracy tradeoff.

3.2 Adaptive Clustering

In order to evaluate an optimal number of color clusters we observe a relationship between the number of quantized colors and the Mean Square Error (MSE) between quantized and original colors. Coarse centroids are obtained using Neural Gas [12], which is quite fast and less sensitive to initialization than the more commonly used K-Means algorithm. The principle of this algorithm is simple: after a random initialization of the centroid set A, at each iteration t, a random sample color ξ is selected and all centroids w_i converge towards the color by Δw_i as per (2); closer centroids converge more than distant ones. Δw_i decays over time.

$$\Delta w_i = \varepsilon(t) \cdot h_\lambda\left(k_i(\xi, A)\right) \cdot (\xi - w_i) \qquad (2)$$

With the following time dependencies:

$$h_\lambda(k) = e^{\left(\frac{-k}{\lambda(t)}\right)} with \ \lambda(t) = \lambda_i(\lambda_f / \lambda_i)^{\frac{t}{t_{\max}}}$$

$$\varepsilon(t) = \varepsilon_i(\varepsilon_f / \varepsilon_i)^{\frac{t}{t_{\max}}}$$

With ε_i, ε_f, λ_i and λ_f being parameters allowing adjusting convergence speed. $k_i(\xi,A)$ represent the rank of the i^{th} centroid when the centroid set A is sorted from the centroid closest to ξ to the furthest.

For each image, we perform several clustering tasks for different target cluster counts study the evolution of the MSE according. Results are presented on *Fig. 1*.

We can see that there is a steep increase in MSE starting for a number clusters. Depending on the picture, this evolution is more or less marked. We use this information to dynamically determine an information-loss threshold

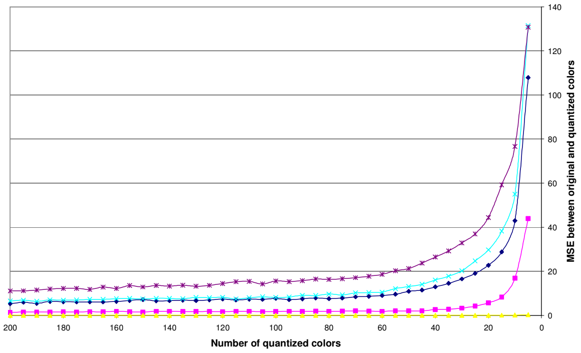


Figure 1: Evolution of Mean Square Error (MSE) between quantized colors and original colors different curves represent images with different color variety

which we chose as a percentile value (set empirically) from the initial clustering MSE. Our method thus adapts to image dynamics and avoids resorting to image dependant thresholds. We use a simple "divide and conquer" strategy to limit the number of neural gases to perform. The process can also be speeded up using parallelization by executing a neural gas task on each available processing unit. Moreover, the clustering does not have to be very accurate: experiments revealed that down to a certain limit the use of less iterations did not impact the detection of an appropriate number of clusters. Once this optimal number of clusters is reached we use an agglomerative hierarchical clustering algorithm to build the appropriate clusters.

The decision regarding adaptive clustering (using an algorithm such as competitive agglomeration [13] which was done in [6]) vs. regular clustering + dynamic determination of the number of clusters was based on experiments that revealed two drawbacks in adaptive clustering. The first is that the algorithm sometimes fails to adapt and leads to either a highly oversegmented image or a single region for the whole image. We were also unable to set the parameters to obtain a low number of regions without noticeably increasing the number of those extreme cases. The second is the lack of consistency: when an image is slightly altered, the regions outside the altered area should not change. While both algorithms are affected with this problem, our experiments revealed this was more the case with adaptive clustering. Our preference therefore went to the more costly but also more consistent approach of dynamically determining a number of clusters.

3.3 Post Processing

Having agglomerated colors within the CIELab color space we still need to separate clusters that are not spatially connected. This last step is performed in a single pass on the original image and allows mapping pixels to regions and respectively. As we said earlier, smaller regions do not interest us because they produce feature vectors from few to no data, inducing noise within the image representation. Thus, during this step we also control the size of regions.

As per [14] we choose to merge smaller regions according to a statistical similarity test. We use the popular squared Fisher's distance (3) for that purpose.

$$D(R_1, R_2) = \frac{(n_1 + n_2)(\mu_1 - \mu_2)^2}{n_1\sigma_1^2 + n_2\sigma_2^2} \qquad (3)$$

Where n_i, μ_i, σ_i^2 are respectively the number of pixels, the average color and the variance of colors within region i. This information is computed at the time of region separation, although variance computation requires another pass. This distance fits our stance of keeping our independence towards image dynamics as it involves intra-cluster distance vs. inter-cluster distances. The threshold therefore only represents a region coarseness parameter. However, we need to address the problem of consecutive merging (for instance in gradients) which could lead to abusive merging with respect to the original regions. To prevent that we use binary tree structures to store region merging history and enforce the condition that a parent region can only merge with a target region if all its child regions could have merged with this target region.

Finally, because we absolutely wanted to avoid small regions we add an optional step which merges regions smaller than 100 pixels with their closest neighbor (with respect to Fisher's distance). We also generate an adjacency graph during the spatial separation step which can be used to obtain neighborhood information for the automated indexing task.

This algorithm provides us with a coarse segmentation producing rather big regions consistently and independently of the conditions as the following section will illustrate.

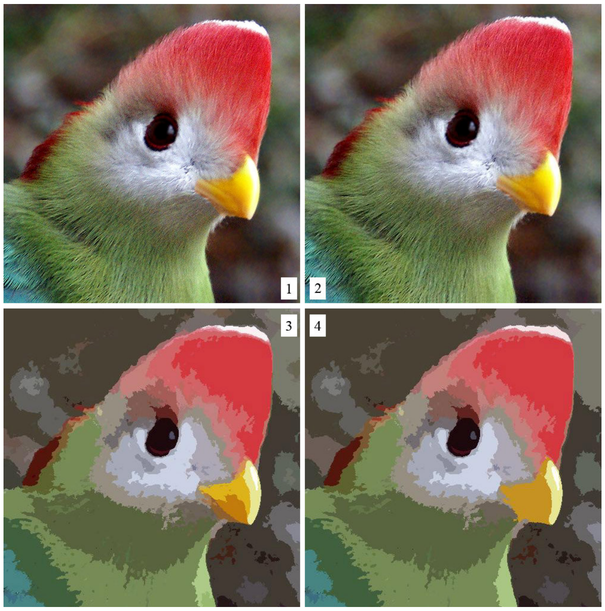
4. Some Results And Comments

4.1 Sample results

We first show the results of segmentation on a sample image in order to illustrate the various steps of the algorithm (*Fig. 2.*). This is an image with a huge color palette (120,526 unique colors). We can see the first step does not alter the image although the difference is slightly perceptible on the real-size image. Indeed while the colors agglomerated and the original color are perceptually similar, color gradients are altered in this operation which makes the difference perceptible. We can then see the third step merges various smaller regions, with the most noticeable change visible in the beak region. The merging of regions smaller than 100 pixels was not applied here.

We then show some sample images drawn from the 2007 Pascal Challenge dataset. The first sample showed results for an image with lots of colors, the following samples (*Fig. 3.*) have been selected in order to show the behavior of the algorithm under various image color depth and lighting circumstances. The algorithm was actually applied to the whole dataset (roughly 10,000 images) giving in every case an adequate output of coarse big regions without having to change any parameter.

As we can see the algorithm adapts to a wide range of situations, the same set of parameters was used for the segmentation of all the images. Situations encompass high color images, low color images with both high and low



1 is the original image, **2** is the result of our fast perceptual quantization, **3** is after clustering and spatial separation and **4** is after color merging using Fisher's distance.

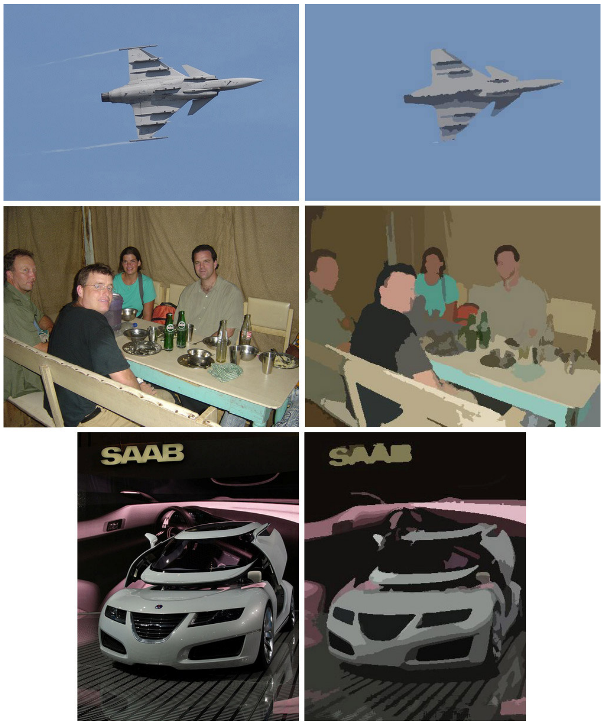


Figure 3: Sample image segmentation results

contrast. In all cases the algorithm adapts and performs consistently without having to change any parameter. While we can regret regions such as the ones created by shadows below the plane, the contrast is simply too strong. Moreover, in most cases we would not like to have an object merge with its shadow. The second image shows problems with reflections and transparencies in items on the table which lead to inaccurate merging. Last sample poses the problem of color gradients: in this case it would be better to merge the different parts of the car's chassis.

These issues show the limits of our algorithm because the most appropriate segmentation is clearly image dependant. Sometimes it will be appropriate to merge different regions from a color gradient, sometimes doing so will lead to merging very different areas.

4.2 Future work

The algorithm is being put to use in its current version for our scene categorization and object detection algorithms. Tests regarding classification performance should lead to a more efficient post-processing step with more efficient region merging criterion like adapting the threshold to region size rather than handling small regions separately. We also plan to improve the determination of the threshold which determines the number of colors by approximating the derivative of the MSE curve.

Also, while performance is not a real issue for our automated indexing platform, it remains an area which we need to work on as the image processing part has not been properly optimized. Current implementation uses quite inefficient high level image processing API. Optimizing computational efficiency could make this algorithm suitable for more applications.

Further work would involve more in-depth study on the image features in order to identify troublesome regions with reflections, cast shadows, etc.

5. CONCLUSION

In this paper, we proposed a hybrid color/spatial information-based segmentation algorithm.

It is inspired by Gestalt theory and centered on three parts: perceptual color reduction, dynamic determination of the number of clusters and spatial post processing. This algorithm allows the production of big coarse color regions (partial gestalts) which are suitable for feature extraction. A major feature of this algorithm is also the lack of image-dependant threshold. Use on the 2007 Pascal Challenge 10,000 image dataset showed its consistent performance.

REFERENCES

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results", http://www.pascal-network.org/challenges/VOC/voc2007/workshop/

[2] S. Ullman, E. Sali and M. Vidal-Naquet, "A Fragment-Based Approach to Object Representation and Classification", *4th International Workshop on Visual Form*, Springer-Verlag, p. 85-102, 2001.

[3] F. Rothganger, S. Lazebnik, C. Schmid and J. Ponce, "Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints", *International Journal of Computer Vision*, 66(3), 2006.

[4] A. Desolneux, L. Moisan and J.M. Morel, "From Gestalt Theory to Image Analysis: A Probabilistic Approach", Springer, 2008

[5] A. Tremeau, C. Fernandez-Maloigne and P. Bonton, "Digital Color Imaging - From acquisition to Processing (in French)", Dunod, janvier, 2004. (in French)

[6] J. Fauqueur and N. Boujemaa, "Region-based retrieval: coarse segmentation with fine color signature", *ICIP (2)*, p. 609-612, 2002.

[7] V. Caselles, R. Kimmel and G. Sapiro, "Geodesic Active Contours", *Int. J. Comput. Vision*, Kluwer Academic Publishers, Hingham, MA, USA, 22, 1, p. 61-79, 1997.

[8] L. Priese and V. Rehrmann, "A Fast Hybrid Color Segmentation Method", *DAGM-Symposium*, p. 297-304, 1993.

[9] W. Ma and B. Manjunath, "EdgeFlow: a technique for boundary detection and image segmentation", *IEEE Trans. on image processing*, 9-8, p. 1375-1388, 2000.

[10] J. Astola, P. Haavisto and Y. Neuvo, "Vector median filters", *Proceedings of the IEEE*, 78, 4, p. 678-689, 1990.

[11] X. Haisong and Y. Hirohisa, "Visual evaluation at scale of threshold to suprathreshold color difference", *Color Research and Application*, 30, 3, p. 198-208, 2005.

[12] T. Martinetz and K. Schulten, "A Neural-Gas Network Learns Topologies", *Artificial Neural Networks*, I, p. 397-402, 1991.

[13] H. Frigui and R. Krishnapuram, "Clustering by Competitive Agglomeration", *PR*, 30, 7, p. 1109-1119, July 1997.

[14] S.C. Zhu and A.L. Yuille, "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation", *PAMI*, 18, 9, p. 884-900, 1996.