PERCEPTUAL COLOR CLUSTERING FOR COLOR IMAGE SEGMENTATION BASED ON CIEDE2000 COLOR DISTANCE

Begoña Acha[1], Carmen Serrano[1], Irene Fondón[1]
[1] Departamento de Teoría de la Señal y Comunicaciones. University of Seville, Spain

ABSTRACT

In this paper, a novel technique for color clustering with application to color image segmentation is presented. Clustering is performed by applying the k-means algorithm in the $L^*a^*b^*$ color space. Nevertheless, Euclidean distance is not the metric chosen to measure distances, but CIEDE2000 color difference formula is applied instead. K-means algorithm performs iteratively the two following steps: assigning each pixel to the nearest centroid and updating the centroids so that the empirical quantization error is minimized. In this approach, in the first step, pixels are assigned to the nearest centroid according to the CIEDE2000 color distance. The minimization of the empirical quantization error when using CIEDE2000 involves finding an absolute minimum in a non-linear equation and, therefore, an analytical solution cannot be obtained. As a consequence, a heuristic method to update the centroids is proposed. The proposed algorithm has been compared with the traditional k-means clustering algorithm in the $L^*a^*b^*$ color space with the Euclidean distance. The Borsotti parameter was computed for 28 color images. The new version proposed outperformed the traditional one in all cases.

Keywords: color segmentation, clustering, CIEDE2000

CONTACT

bacha@us.es

INTRODUCTION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a dataset into subsets (clusters), so that data in each subset (ideally) share some common trait, often proximity according to a pre-defined distance measure. When each element in the set represents a pixel, clustering means grouping of pixels, each cluster representing an object within the image. In color image segmentation using clustering technique, each pixel is usually characterized by its three color components.

To apply a clustering technique, one has to define: 1) The feature space, which will be employed to represent data. In image segmentation based on color, the feature space must be one color space. 2) The metric to measure distances between pixels. 3) The algorithm to partition the feature space according to the distance measure previously defined.

The scope of this paper is to optimize both the feature space and the distance measure for color clustering. Images are usually stored and displayed in the *RGB* space. Therefore, first works on color clustering[1,2] utilized this space to represent pixels in an image. Each sample in this feature space was a pixel in the *RGB* color coordinates. As regards to the distance, the Euclidean distance is the preferred distance in the literature. Also in the nineties the *HSI* color space was employed.
Nevertheless, MacAdam demonstrated[3] that none of these spaces is uniform, in the sense that perceived differences among color are not exactly related to Euclidean distances in those color spaces. Consequently, it is not appropriate to do the clustering in the above color spaces.

In 1960 *CIE* defined the Uniform Chromaticity Scales (*UCS*) diagram, in an effort to make

chromaticity diagrams more perceptually uniform. In this sense, it is more exact to use this color representation for color clustering. Nevertheless, although *UCS* 1960 diagram is a good linear transformation from *RGB*, this model still does not provide equal distances throughout its chromaticity diagram. Thus, McAdam proposed a non-linear transformation from *UCS* called geodesic chromaticity. Kehtarnavaz et al proposed a color image segmentation algorithm that performs a clustering in this chromaticity diagram. But the use of chromaticity diagrams has been made largely obsolete by the advent of the *CIE* 1976 color spaces, $L^*u^*v^*$ and $L^*a^*b^*$. The main aim in the development of these spaces was to provide uniform practices for measurement of color differences, something that cannot be done reliably in tristimulus or chromaticity space. Modern color clustering algorithms utilized the $L^*u^*v^*$ and the Euclidean distance in this space[4,5,6].

However, to ensure the isotropy of the feature space a uniform color space, where perceived color differences can be measured by Euclidean distances, should be used. $L^*u^*v^*$ and $L^*a^*b^*$ color spaces have been particularly designed to closely approximate perceptually uniform color spaces. Therefore traditionally $L^*u^*v^*$ and $L^*a^*b^*$ spaces with Euclidean distances are the feature spaces chosen in the literature. In this paper, we propose a new method that performs clustering in $L^*a^*b^*$ color space. But instead of using the Euclidean distance, we propose to employ *CIEDE2000* color distance, for it corrects the non-uniformity of $L^*a^*b^*$ color space.

METHODOLOGY

As stated in Section 1, in clustering algorithms, one has to define the feature space which will be employed to represent data, the metric to measure distances between pixels, and the algorithm to partition the feature space according to the distance measure previously defined.

Feature space and distance measure

To perform color segmentation a uniform color space is required. That is, in the chosen color space, distance measures must be correlated with perceived color differences. In 1976, *CIE* proposed two color spaces that approximately possessed this property: $L^*a^*b^*$ and $L^*u^*v^*$. Euclidean distances in those spaces were believed to be approximately correlated with perceptual color differences. But later on it was demonstrated that this goal was not strictly achieved. To improve the uniformity of color difference measurements in $L^*a^*b^*$, an empirical modification of the Euclidean distance was proposed in 1995. This distance measure is abbreviated as *CIE94*. More recently, the *CIE* has established the *CIEDE2000* color difference equation that extends the concept of *CIE94* with further complexity. It has been demonstrated that *CIEDE2000* performs better than *CIE94*[7] and Euclidean distance when a 10 degrees observer is considered as we are.

Perceptual k-means

K-means algorithm was first proposed by Lloyd[8] in 1957. Let $X = \{x_1, \ldots, x_n\}$ be a data set where each $x_i \in R^3$ represents a pixel in the color space. The codebook V is defined as the set $V = \{v_{1,\ldots,} v_k\}$, whose elements are the codevectors or centroids. The Voronoi set π_i of the codevector v_i is the subset of X for which the codevector v_i is the nearest vector: $\pi_i = \{x \in X | i = \text{argmin}_j\, d(x, v_j)\}$. Starting from the finite data set X, this algorithm moves iteratively the k codevectors to the centroids of their Voronoi sets and recalculates the Voronoi sets. The codebook V is chosen to minimize the empirical quantization error:

$$E(V) = \frac{1}{2n} \sum_{i=1}^{k} \sum_{x \in \pi_i} d^2(x, v_i) \qquad (1)$$

where d is the distance measure and n the number of pixels. In the case of the Euclidean distance, this error is minimized when the codevectors are chosen as:

$$v_i = \frac{1}{|\pi_i|}\sum_{x\in\pi_i} x \tag{2}$$

Codebook determination for CIEDE2000

The *CIEDE2000* color distance between two pixels of color values $\left(L_1^*, a_1^*, b_1^*\right)$ and $\left(L_2^*, a_2^*, b_2^*\right)$ is calculated as:

$$\Delta E_{00}^* = \left[\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C_{ab}'}{k_C S_C}\right)^2 + \left(\frac{\Delta H_{ab}'}{k_H S_H}\right)^2 + R_T\left(\frac{\Delta C_{ab}'}{k_C S_C}\right)^2\left(\frac{\Delta H_{ab}'}{k_H S_H}\right)^2\right] \tag{3}$$

where

$$L'_{1,2} = L^*_{1,2}; a'_{1,2} = (1+G)a^*_{1,2}; b'_{1,2} = b^*_{1,2}; C'_{ab} = \sqrt{a'^2 + b'^2}; h'_{ab} = \tan^{-1}(b'/a'); G = 0.5\left(1 - \left(\frac{\overline{C}^{*\,7}_{ab}}{\overline{C}^{*\,7}_{ab} + 25^7}\right)^{\frac{1}{2}}\right); \overline{C}^*_{ab} = \frac{C^*_{ab1} + C^*_{ab2}}{2}$$

$$\Delta L' = L'_1 - L'_2; \Delta C'_{ab} = C'_{ab1} - C'_{ab2}; \Delta H'_{ab} = 2\sqrt{C'_{ab1}C'_{ab2}}\sin\left(\frac{\Delta h'_{ab}}{2}\right); \Delta h'_{ab} = h'_{ab1} - h'_{ab2}; \overline{L'} = \frac{L'_1 + L'_2}{2};$$

the values of the parametric factors and the weighting functions can be found in[7]. As already mentioned, the codebook V is chosen to minimize the error (1). To this purpose the distance formula must be derivated respect to L_1^*, a_1^*, and b_1^* and then equaled to zero. The results obtained are transcendental equations that must be solved with numerical methods. To solve this optimization problem we propose a method based on the 2-D-log-search[9,10] method extended to 3-D. It has a recursive structure and, starting from a particular point in the feature space, it computes the objective function in a set of neighboring points and it selects the one with the minimum value in the objective function among all the points analysed. Then the old centroid is updated to the point with this minimum value.

Accordingly, first we have to define which points in a particular neighborhood of the old centroid will be considered as candidates to centroid. In this method, the set of neighboring points to be considered is shown on Fig. 1. As observed in this figure, each step tests 19 points around the old centroid in a sphere arrangement. In the next step, the search is repeated with the centroid moved to the best matching point resulting from the previous step and the radius of the sphere reduced to half its former value.

In order to avoid a local minimum, if none of the new points is better than the old one, the algorithm does not finish. Instead, the neighborhood is decreased until its size is reduced two magnitude orders. If the best centroid does not change, the algorithm finishes. Although this stopping condition may be thought to limit the accuracy of the algorithm, experimentally it have been demonstrated that the centroid does not change when the radius reduces further than its octave part..

Some parameters must be chosen in order to implement this simple algorithm. The first one is the initial neighborhood of search, defined by the initial radius of the sphere around the first centroid, that is, the maximum distance between the first centroid and its candidates. If the centroid position is not modified in the first iteration, this initial
could be excessive.... On the other hand, the shape of the neighborhood is significant because it must contain the new centroid. We have chosen a shape that gives equal probabilities to any direction of the space, that is, a sphere.

Regarding to the number of points that must be defined to assure the convergence of the algorithm in an efficient way, we need a compromise solution between efficiency and efficacy. As the computational cost of each iteration grows factorially with the number of test points, if it is high the

algorithm will converge in a few number of high computational cost iterations. On the contrary, if the number of test points is low, the algorithm will need many low computational cost iterations. The problem is that, as the radius decrease half at a time, the convergence will be probably reach in the infinite. Experimentally we have determined that the best quantity is 18 points, one for each extreme of the coordinate axe, and one more extra point for the bisect between each of them. We also include the old centroid because if the configuration is stable, that is, only a few number of points is moved from its previous position, the new centroid will be very near the old one, sometimes the displacement could be considered null and the new centroid could be taken as the old one. The algorithm convergence will be reached faster. An additional condition has been also added to assure the

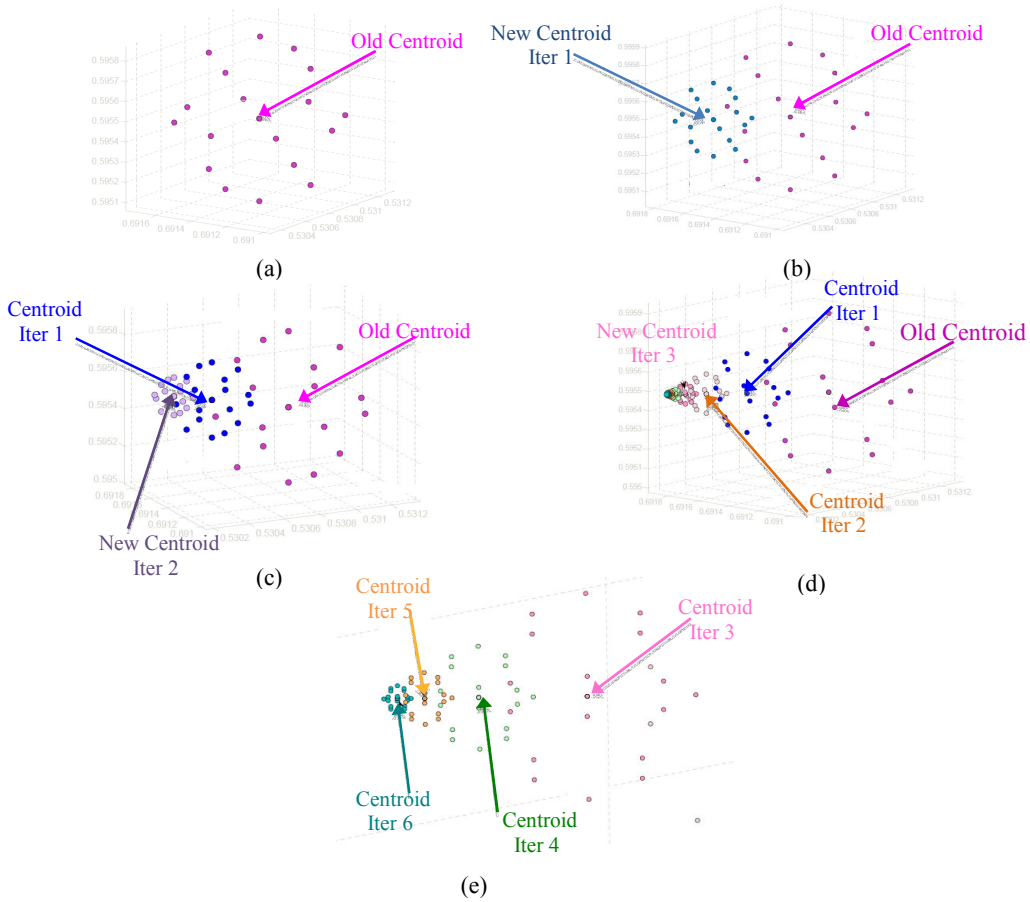

Fig 1. Example of centroid updating. (a) Initial distribution. In the first iteration objective function is computed and the point that provides the minimum value is taken as the new centroid (b). The process is repeated until none of the candidates is better than the previously selected centroid. Results of iteration 2 (c) and iteration 6 (d). (e) Centroids from iteration 3 to 6.

convergence in relation to the precision. If the radius is $\leq 10^{-6}$, the old centroid is the chosen one. Therefore the maximum number of iterations is 54. Besides, the position of the first centroids is randomly chosen because this new algorithm is not application dependent.

EXPERIMENTAL RESULTS

We have performed some experiments to evaluate the performance of the two proposed modified k-means algorithms. The image database consists of 14 natural images especially interesting because the variety of colors presented on them or because the similarity among some of the colors, making the segmentation process more difficult and the k-means comparison more interesting. We include histological images, cellular images and some well-known images as Lenna or Peppers.

General images

- **Subjective validation**

Selecting a color space for the segmentation process is an effort to approximate the segmentation result to the way humans perceives different colors. Therefore, a human validation of the images obtained is obviously needed. To this purpose, we have presented the segmentation result for the 14 images in the database to 8 experts. The images were segmented with the traditional k-means algorithm and the Euclidean distance on different color spaces. Then, they had to score the segmentation quality with punctuation from 0 to 4. 4 means excellent segmentation results: suitable color selection, expected clusters, visually coherent segmentation without failures, etc. 0 means bad segmentation results: Clusters without any logic in a perceptive level almost always related to failures in the selection of the number of clusters. The results are shown on Fig. 2.

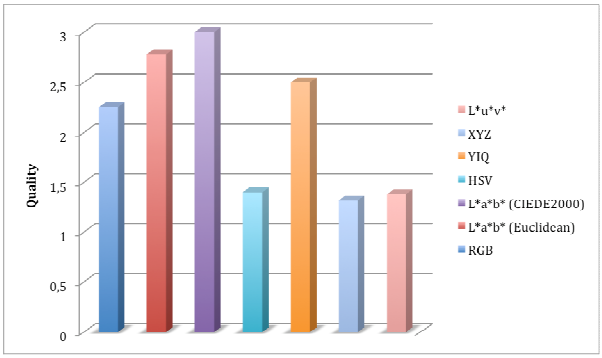

Fig 2.The 8-expert assignments of the segmentation quality on each of the color spaces is shown on the graphic. A value of 4 corresponds to perfect segmentation and 0 is a bad segmentation

Segmentation in $L^*a^*b^*$ space gives the best results regardless of the distance measure used. In any case, when using *CIEDE2000* color distance the results are slightly better. In *XYZ* or $L^*u^*v^*$ the images are over segmented in some way so they are not good to separate objects. Two segmentation results are shown on Fig. 3.

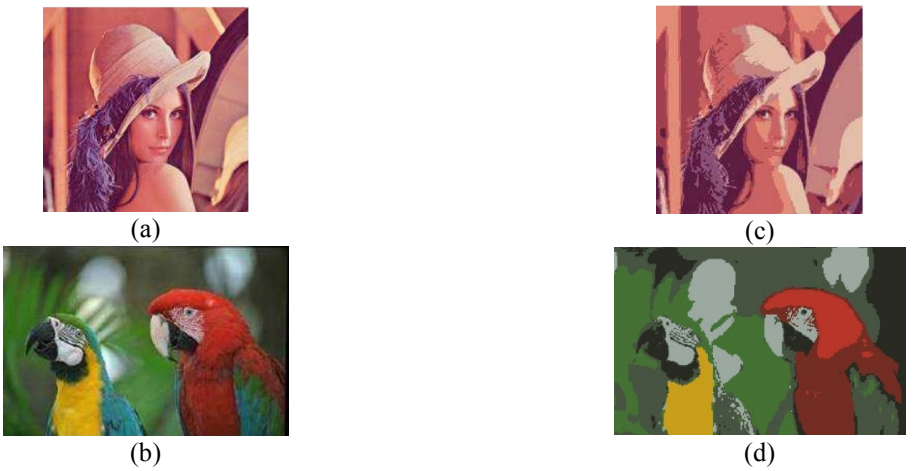

Fig 3. (a) and (b) original images; (c) and (d) images segmented with k-means and *CIEDE2000* color distance.

- **Borsotti parameter**

Once we have visually evaluated the segmentation results concluding that $L^*a^*b^*$ color space is one of the best spaces to perform the segmentation, we have evaluated numerically which distance in

$L^*a^*b^*$ color space gives the best results with the computation of Borsotti parameter[11]:

$$Q(I)=\frac{1}{1000(N\cdot M)}\sqrt{R}\sum_{i=1}^{R}\left[\frac{e_i^2}{1+\log A_i}+\left(\frac{R(A_i)}{A_i}\right)^2\right] \quad (4)$$

where for an image I of dimension NxM, R is the number of segmented regions, A_i is the number of pixels in region i, $R(A_i)$ represents the number of regions having an area equal to A_i and e_i is the error calculated as the sum of the appropriated distances between color components of pixels of region and components of average color, which is an attribute of this region in the segmented image[11] . The idea of using this parameter is: the smaller the value of $Q(I)$ the better is the segmentation result. On Table 1 we summarized the results for the segmentations performed. All the results must be multiplied by a factor of 10^{-7}, as the pixels values are in the [0, 1] range. The values of the Borsotti parameter achieved with the proposed algorithm are smaller and sometimes the difference is two magnitudes order. We can conclude that the new algorithm outperforms the k-means algorithm with Euclidean distance.

Table 1. Borsotti parameter for 14 images. The higher the parameter the worse the segmentation

Distance	Average	Standard Deviation
Euclidean distance	22,015	34,879
CIEDE2000 distance	7,933	3,723

Dermatoscopic images

We have also applied the algorithms to 14 dermatoscopic images to prove the generality and applicability of both algorithms. These images present two dominant colors: the one of the healthy skin and the one of the lesion or possible melanoma. They also present illumination variability, different camera resolution, hair, brightness, low contrast, etc. so they are especially difficult to segment.

- **Subjective evaluation**

Once again 25 experts evaluated the segmentation results and their opinion is shown in Fig. 4. Each of the 14 database images were segmented with k-means algorithm in the specified color space and with the selected distance (Euclidean by default). The 25 experts are requested to choose the color space which provide the best segmentation results.

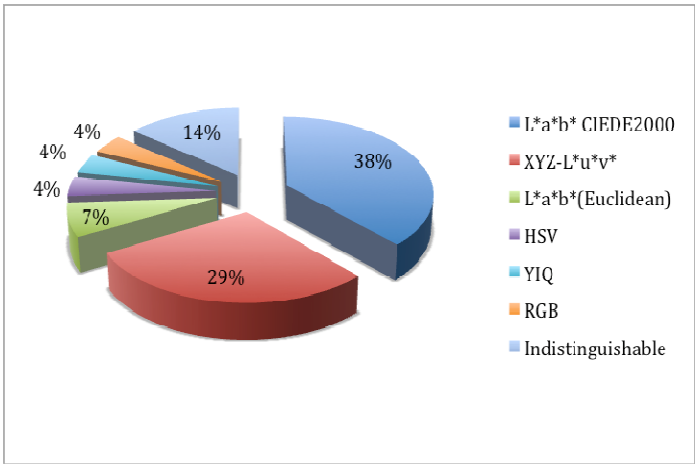

Fig 4. Preferred color space for the segmentation based on the opinion of 25 experts.

We can observe that $L^*a^*b^*$ color space is desirable with both distances, while RGB, YIQ and HSV are

the least valued.

- **Borsotti parameter**

We have calculated Borsotti parameter again with this special kind of images. Results are summarized in Table 2. Once again, the new algorithm outperforms the traditional one achieving smaller values for *Q(I)*. In this case, the differences in performance are smaller.

Table 2. Borsotti parameter for 14 melanoma images.

Distance	Average	Standard Deviation
Euclidean distance	14,150	1,401
CIEDE2000 distance	13,94	1,369

CONCLUSION

There is an increasing interest in the development of automatic segmentation algorithms related to human color perception. In this sense, many color spaces have been considered although none of them is perceptually uniform and therefore the correlation with human perception is low. Among all of them, $L^*a^*b^*$ color space is the most approximately uniform, although Euclidean distances in this color space are not exactly correlated with perceived differences . To correct its non-uniformity, *CIEDE2000* color distance formula was proposed. Then it is desirable to adopt $L^*a^*b^*$ and *CIEDE2000* when an algorithm to process images is developed. Therefore an approach to adapt the well-known k-means segmentation algorithm to the *CIEDE2000* color distance formula will be desirable.. In this paper, we propose to use $L^*a^*b^*$ as feature space and *CIEDE2000* as distance to measure color differences in this space. The method has been extensively tested with general and medical images. In order to assess the algorithm, two subjective experiments have been performed. In the first one, 8 experts have evaluated the quality of the segmentation algorithm. The proposed algorithm obtained 2.9 out of 4, whereas $L^*a^*b^*$ and Euclidean distance obtained 2.7. On the other hand, when 25 expertshad to choose the best segmentation result, the preferred algorithm was the proposed one for 38% of the experts. In addition, our algorithm outperforms the traditional version of k-means method providing better results, as Borsotti's parameter shows.

REFERENCES

1. D. Comaniciu, P. Meer, "Robust Analysis of Feature Spaces: Color Image Segmentation." *Proc. Of CVPR'97*, 750-755, 1997.
2. W.P. Berriss and S.J. Sangwine, "A Colour Histogram Clustering Technique for Tissue Analysis of Healing Skin Wounds." *Proc. of 6 t8u Int'l Conf. on Image Proc. and Its Appl.* **Volume 2**, 693-697, 1997.
3. D. MacAdam, *Color Measurement***,** Springer-Verlag: Berlin, 1981**.**
4. L. Lucchese, S.K. Mitra, "Unsupervised Segmentation of Color Images Based on K-means Clustering in the Chromaticity Plane." *Content-Based Access of Image and Video Libraries (CBAIVL '99) Proceedings* 74-78, 1999.
5. W. Tao, H. J. and Y. Zhang, "Color Image Segmentation Based on Mean Shift and Normalized Cuts." *IEEE Transactions on Systems, Man, and Cybernetics—part b: Cybernetics* **Volume 37**, No. 5, 1382-1389, 2007.
6. Dorin Comaniciu, "An Algorithm for Data-Driven Bandwidth Selection." *IEEE Transactions On Pattern Analysis And Machine Intelligence*, **Volume 25**, No. 2, 281-288, 2003.
7. M. R. Luo and G. Cui, B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000." *Color Research & Application* **Volume 26,** Issue 5, Special Issue: Special Issue on Color Difference, 340–350, 2001.
8. S. Lloyd, "Last square quantization in PCM's." *Bell Telephone Laboratories Paper*, 1957.
9. M. Avriel, *Nonlinear Programming: Analysis and Methods*. Dover Publishing, 2003.
10. J. A. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*. Springer Publishing: New York, USA, 2005.
11. M. Borsotti, P. Campadelli and R. Schettini, "Quantitative evaluation of color image segmentation results." *Pattern Recognition Letters* **Volume 19**, Issue 8, 741– 47, 1998.