Color Quantization For Image Processing Using Self Information

Alain Pujol
LIRIS, UMR 5205 CNRS
Ecole Centrale de Lyon
36 avenue Guy de Collongue
69134 Ecully Cedex - France
Email: alain.pujol@ec-lyon.fr

Liming Chen
LIRIS, UMR 5205 CNRS
Ecole Centrale de Lyon
36 avenue Guy de Collongue
69134 Ecully Cedex - France
Email: liming.chen@ec-lyon.fr

Abstract— **A digital picture generally contains tens of thousands of colors. Therefore, most image processing applications first need to apply a color reduction scheme before performing further sophisticated analysis operations such as segmentation. While a lot of color reduction techniques exist in the literature, they are mainly designed for image compression and are unfortunately not suited for many image processing operations (e.g. segmentation) as they tend to alter image color structure and distribution. In this paper, we propose a new color reduction scheme (SICR), using probabilities and information theory elements to balance between the information provided by the selected colors and the necessity to accurately represent the selected colors. We also advocate for the use of perceptually accurate metrics for evaluation. Experimental results on a diversified dataset of images selected from the internet show that our technique performs well compared to other color reduction schemes.**

I. Introduction

Nowadays, colour processing is a very important tool in image processing and is widely used as a low level feature or as decomposition tool through image segmentation. However the huge number of colors involved in high resolution images induce prohibitive computation costs which make color depth reduction algorithms a necessity not only for compression but also for image processing ([1]).

Various color reduction algorithms have been proposed in the literature. While achieving good results, many of these algorithms have been designed for compression purpose and rather aim to optimize a global perceptual similarity between an original color image and the compressed one. As such, these techniques unfortunately either alter local color structure of an image (as illustrated by the technique of halftoning) or modify the original color palette too much, resulting in inaccurate image approximation often harmful for further image processing operations such as segmentation or image content characterization.

In the field of content-based image indexing, we are rather interested by the extraction of various descriptors for image content characterization. Therefore, low level features extracted by image processing techniques from a quantized color image are expected to be as accurate as possible compared to the original image. For this purpose, a color reduction scheme has to respect as faithfully as possible color distribution of the original image while preserving local color structure.

In this paper, we propose a new color reduction scheme, named SICR (Self Information color reduction), a two step color quantization algorithm, which tries to provide the most informative color quantization by balancing color information and color representativity. We wanted an algorithm that quantized colors while preserving color dynamics (extreme color variations within an image) because this seemed an interesting feature of an image. Existing algorithms tend to damp this information as they seek to minimize mean square error. Our algorithm aims at keeping this information and also avoid creating colors within the quantized image. Experimental results on a diversified dataset of images selected from the Internet show that our technique still performs well when compared to other color reduction schemes when evaluated according using standard similarity measurements (Mean Square Error).

The rest of this paper is organized as follows: section 2 overviews related work in the literature. Section 3 studies color spaces, discusses about color similarity metrics and about how to evaluate color similarity quality of a color quantization scheme. Section 4 details our two step color quantization scheme. Section 5 discusses and compares experimental results. Section 6 contains conclusion and gives some hints on future work.

II. Related Work

Many color quantization techniques have been proposed in the litterature ([1], [2]), mainly for compression purpose. The process of halftoning (or dithering) typically illustrates this: [3] proposes an efficient technique for color quantization based on the priciple of dithering. While the resulting image preserves a good global perceptual similarity as compared to original color image, the technique deeply alters local color structures and hinders further image processing operations, for instance segmentation.

In [2], two classes of color quantization algorithms are studied. The first class contains methods based on splitting algorithms which are generally quite fast but tend to introduce substantial alteration on original color palette. The well known median cut algorithm (Heckbert, 1982) and the variance based algorithm (Wan, 1990) are cited examples of this approach. The Principal Analysis Algorithm (Wu, 1992 [4]) is a more advanced algorithm of this category, which obtains interesting results and will be used for comparison in this paper. A recent work proposed in [5] aiming at a better processing time by sacrificing some accuracy can also be classified in this category.

The second class consists of clustering based methods with, for instance, the commonly used K-Means as well as Fuzzy C Means. It is, however, well known that these techniques are very sensitive to initialization. Another problem of most clustering approaches is that algorithms repeatedly compute distances between colors as well as between colors and clusters, resulting in polynomial complexity which proves taxing with large data sets. Unfortunately, color reduction scheme is precisely needed and applied to images usually having with 40 000 to 80 000 colors. Thus, a preliminary color reduction step may be necessary prior to these clustering based color quantization methods.

More recently, unsupervised clustering methods (mostly Self Organizing Maps) have become incrasingly popular as they produce good results using common evaluation techniques and allow the use of spatial constraints which also tend to produce visually good results. For instance, [6] and [7] propose to integrate local color features and rely on unsupervised learning to extract optimal color prototypes. Classes are then handled by a split and merge process. The inclusion of spatial features brings visually very good results. While the implementations proposed in [6] and [7] significantly improve computational efficiency as compared to other clustering based approaches, the resulting algorithm remains quite complex.

In our work, we propose a new simple and efficient color reduction scheme which falls in none of these two categories. Our color reduction scheme proceeds in two steps. Firstly, image color space is divided into cubes of perceptually identical colors. Secondly we iteratively select the most appropriate colors according to a criterion mixing color population and color information.

III. Color Spaces And Metrics, Quality Assessment

Many color spaces and metrics are used in the literature for color quantization purpose. The basic idea is that we should choose a color space and an appropriate metric which would avoid merging dissimilar colors or separating perceptually close colors. In this section, we discuss and compare various color spaces in order to set the most appropriate color space for our color reduction scheme. We will study the most widespread color metrics and more specific perceptually accurate metrics within their respective color spaces, we also emphasize distortion between perceptual color similarity and distances produced by classical metrics and then propose the use of a perceptually accurate color similarity metrics for Lab color space usually used in textile industry.

A. Color spaces

It is well known that RGB color space is far from perceptual homogeneity and, as such, its quantization produces perceptually redundant bins and leaves perceptual holes. Therefore, any ordinary distance function defined in this space will be unsatisfactory. The HSL, color space shows much better results for ordinary distances. Its was designed to be much more intuitive and a very interesting point is that, as stated in [8], it is easier to perform channel specific processing (either application specific or for correcting color distortions). For example, lighting and shading artifacts will typically be addressed in the lighting channel. It, however, introduces various flaws. First, the Lightness (or Brightness) is usually calculated as L = (R + G + B) / 3, versus a more realistic implementation that involves hue-dependent computations. It also shows a new problem common to many color spaces: a perceptually redundant representation of very dark or very bright colors: indeed when a color becomes too dark or too bright hue information gradually loses of its importance. This is especially true in HSL color space where hue becomes meaningless for low saturation values. HSL color space also has a discontinuity for its cyclic Hue component (which has to be handled by a specific distance measure as in [9]).

CIE color spaces provide better approximations of perceptual homogeneity when using simple metrics: in many applications those color spaces are used in combination with Euclidean distance with satisfactory results. As we will see later these color spaces are still subject to some distortions.

B. Color metrics and quantization assessment

There are a lot of various color similarity metrics used in color processing, ranging from euclidean distance to custom color metrics such as complex but perceptually more accurate measures used in the textile industry such as CIE94, CMC, etc. In color quantization, it is important to search appropriate color space and metrics which translate as faithfully as possible perceptual color similarity, i.e. that produce distances avoiding inaccurate values which may lead to merging dissimilar colors or to separating perceptually really close colors. It is noteworthy that, whichever the color space, those errors tend to happen either in quite specific conditions or in some occasional cases. As such the choice of an appropriate color space along with a measure respecting color perceptual simalarity will sacrifice processing time to reduce the probability of incoherent color reduction. Unfortunately, given the quantity of color distance measures to be computed within a color image, we need to make a tradeoff between processing speed and the respect of perceptual similarity.

To assess computaional cost, we benchmarked several color distances, including CIE94, CMC, CIE2000 and euclidean distance (as a reference), on a set of pixelwise distance measures between two different images of the same size (both converted to CIELab as previousely stated). This experiment reveals that these distances have significant different computational time, with CIE2000 being significantly slower than the two other advanced metrics, which are themselves quite slower (about twice as slow) than the euclidean distance. Therefore any choice of an advanced color metrics has to be motivated either by performance and/or by a guarantee that a limited number of measurements will be taken.

On the other hand, as shown by a study on perceptual color distances [10], the CMC distance formula, shows convincing results on its property to better characterize perceptual color similarity : it is only beaten by the much more complex CIE2000 distance and as such represents an interesting compromise between accuracy and complexity. We will not go into detail for the computation of this distance, we just mention the base formula.

For two colors of respective CIELab components (L_1, a_1, b_1) and (L_2, a_2, b_2), CMC metrics define three components for the distance measure as follows:

Chroma difference:

$$\triangle C = \sqrt{a_1^2 + b_1^2} - \sqrt{a_2^2 + b_2^2}$$

Lighting difference :

$$\triangle L = L_1 - L_2$$

Hue perceptual difference :

$$\triangle H = \sqrt{(\triangle a)^2 + (\triangle b)^2 - (\triangle C)^2}$$

With the global distance given by :

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

l and c are application dependent coefficients. Typical respective values are 1:1 for perceptual thresholding and 2:1 for acceptability thresholding. The $\triangle E$ equation basically corresponds to an ellipsoid shape of equal perceptual distances What this equation translates is the ellipsoidal shape of perceptually homogeneous colors in CIELab color space. It is noteworthy that $\triangle H$ as well as the various weights take into account the aforementioned problem of the pertinence of hue information when chroma is close to 0 and/or in extreme lighting conditions which tend to occur in consumer photographs where colors are frequently "burnt" in overexposed pictures and where underexposed pictures are also common).

We benchmarked several classical distance formulae, including CieLab euclidean distance, RGB euclidean distance, HSL eudlidean/angular distance by computing a lot of distances and seeking mismatches between visual observations and obtained distances. While no distance achieves perfect accuracy, mismatches were far more common and more significant on RGB and HSL samples. Table 1 highlights the better accuracy of the CMC distance over three classical color distances on four significant cases among our tested samples. The first and the second ones compare two completely different colors (blue and green) with respectively strong and very low brightness. The third one illustrates the case with two very similar colors normally perceived as teal. The last case compares two poorly saturated but quite different colors (a tone of brown and a tone of purple). Therefore, perception

TABLE I

SAMPLE TROUBLESOME COLOR COMPARISONS USING VARIOUS METRICS

color space/metrics	1	2	3	4
CIELab Euclidean distance :	11.6	13.5	9.1	14.0
RGB Euclidean distance :	10	13	29.5	11.4
HSL Euclidean/angular distance :	33.4	68.9	67.8	39.0
CMC distance :	14.1	17.6	5	22.7

1) Distance between light blue and light green (RGB 240/255/255 and 234/255/234)
2) Distance between dark blue and dark green (RGB 0/0/12 and 1/15/0)
3) Distance between two very close teal tones (RGB 0/128/128 and 64/128/128)
4) Distance between two different poorly saturated colors (RGB 131/114/131 and 140/113/108)

suggests that the first, second and fourth distances should be significantly higher than the third. As we can see on this table, CIELab euclidean distance respects expected results but produces close values. RGB euclidean distance typically fails on the third case (redundancy). The same drawback is witnessed by HSL euclidean/angular distance. The CMC distance performs adequately in these cases and, globally, shows few (although sometimes noticeable) distortion between human perception and measured distances.

It is to be noted that Lab color metrics are accurate in most cases and as such our experiments concludes that cmc metrics could be used for evaluation and offline tasks where we need as much accuracy as possible while Lab remains an acceptable choice when computation time is an issue.

Quantization evaluation will be performed using standard MSE, we will provide MSE values for both euclidean Lab and CMC metrics. We think that CMC distances make more sense as a distance below 1 which represents a non-noticeable difference will not really impact MSE as a Lab euclidean distance would. While we encourage the use of CMC distance for MSE we must not forget visual assessment as no matter how perceptually accurate our metrics are, more than often we found that the images that best looked like the original were not always showing very good MSE scores. Sample images comparisons (using images extracted from the Berkeley Image Segmentation Dataset) are available at the author's website at *http://liris.cnrs.fr/%7Eapujol/colorQuant.shtml*

IV. OUR COLOR REDUCTION ALGORITHM

In this section we introduce our color quantization scheme, SICR, which tries to agglomerate colors according to a criterion balancing color population representativity and color self information to cover color space as much as possible. According to our scheme, we proceed in three steps. Firstly, the color space is partitioned according to perceptual color similarity. Then, we perform an intial selection of quantized colors according to pixel population and distance between color clusters. Finally, this selection of quantized colors is further refined using remaining colors (this third step being optional).

A. Color space partitioning

According to our previous discussion on color spaces and similarity measures, an input color image to quantized is first coverted into CieLab space. The first step of cour color quantization scheme then consists of partitioning CieLab color into cuboids which group all colors of an image into clusters. The size of these cuboids was chosen so that all colors within it are percetually similar. As previously stated, two colors in CieLab are considered perceptually similar when their CMC 2:1 distance is lower than 1. We also note that lighting ranging from 1 to 100 while a and b range approximately from -100 to 100, cuboid should be twice as accurate on the L scale. Experiments revealed that the size of these cuboids should be chosen smaller than 1.5 x 3 x 3 to ensure that any color within the cuboid would have a distance below 1 when compared to the centre of the cuboid. This size may be adapted for speed vs accuracy balance.

Once this partitioning has been completed, we perform a first color reduction by replacing each color in a given cuboid by the centroid of all the colors in it, each color being ponderated by its population within the original image. We bechmarked this simple color reduction scheme on a 600 image dataset including images as diversified as possible (see evaluation section for details). The average number of colors within our data set was 55 759 and this pre-quantization step produces an image with an average of 5 973 colors and an average CMC distance of 0.816. Therefore while keeping color similarities below a perceptual threshold of 1, an average color reduction rate of 89% is achieved by this simple color space partitioning.

B. Representative color slection

This reduced set of colors is still too important for many image processing applications. In our color quantization scheme, we therefore further reduce it by a selection of representative colors according to a balanced criterion of population and color information.

The basic idea is to iteratively select the most representative colors until we reach a chosen number of quantized colors. The inputs of this selection process are all the color centroids along with the population associated with each color cuboid from the previous step. The first quantized color is then simply chosen to be the color barycenter having the most population. Now let C1, C2, ..., Ci-1 be the next (i-1)th selected colors. We select the i-th color among the remaining color barycenters as folows. We first compute a probability which is the population associated to the barycenter divided by the total number of colors within the image. We then compute color self-information I(c) as per Shannon's information theory: $I(c) = -log(P_2(c))$ where P_2 represents the probability of observing a similar color within the selected color set. This both translates our aim at preserving image dynamics as much as possible, as well as a general strategy to avoid selecting highly probable but similar colors. To define this similarity, we set a neigborhood in Lab color space within which colors are considered similar. As we first aim at validating our approach, we simply defined it as a hard threshold. Colors are then evaluated through a combination of these two criterion. Various ways of combining them were evaluated (simple linear combination, ...) and the best results were obtained by emphasizing self information during the first iterations and gradually giving more importance to color probability, we thus obtain the following formula for color evaluation:

$$\alpha.e^{-\frac{i}{\tau}}.\frac{1}{1-\sum_{j=0}^{i} P(c_j)}.I(c_i) + P(c_i) \tag{1}$$

Where i is the current iteration, P(c) represents the probability of observing color c within the pre-quantized image and I(c) is the self information of color as defined above. α and τ are parameters.

C. Reduced color set refinement

When the appropriate colors are selected we may opt to minimize the approximation error by using a clustering alogrithm such as K Means or Fuzzy C-Means to refine the position of barycenters. While offering a significative gain regarding MSE, this approach shows two drawbacks: firstly it is quite computationally intensive (though few iterations are necessary to reach a good result), secondly it creates colors. While the initial quantization phase also creates colors, the approximation is bounded by the size of the cuboids.

D. Time complexity analysis

For quantizing an image containing p pixels into t target colors, we proceed as described above. The first color partitining step consists mainly of distrubuting pixels of the input color images into predefined color cuboids, thus has a linear complexity compared to p. This step produces c colors. The second quantized color selection step selects the barycenter representing the highest color probability which can actually be done during the barycenter computation process. After that, the selction of i-th quantized color mainly needs browsing unselected colors and computing the associated self information which mean c - i oprations. Therefore the computation of all the remaining colors takes $\frac{(t+1).(c+c-t)}{2}$ operations. The final complexity of our color quantization scheme is therfore given by $\frac{(t+1).(c+c-t)}{2}$.

V. Experimental results And evaluation

For comparison purpose, we benchmarked our color quantization scheme and two other representative schemes on a dataset of 600 images selected from the Berkeley image segmentation dataset as well as a lot of various sources with diversified exposure conditions. Furthermore, all these images belong to different categories (cityscapes, forests, night pictures, etc.), this results in very different color palettes and thus tests the quantization process in many different configurations.

TABLE II

RESULTS OF FINAL COLOR QUANTIZATION ON TEST SET

	Average MSE					
	109 colors		135 colors		162 colors	
	CMC	euclidean	CMC	euclidean	CMC	euclidean
ACR	20.2	46.1	17.3	39.1	16.1	36.5
Wu's quantization	27.6	67.7	24.6	60.5	22.7	55.7
K-Means Clustering	35.3	83.9	31.2	74.82	28.4	67.29
SICR	27.1	67.6	24.1	57.4	22.5	55.4

A. Color quantization results and accuracy evaluation

As initial evaluation revealed, state of the art methods used for compression that include halftoning and other spatial processes did not perform very well (as expected) and thus are not included in the evaluation. The evaluation consists in computing MSE between quantized image and each of the 600 original images using both CMC and euclidean distances as a base for MSE. Table 3 presents the results on our dataset along with those of two other well known methods: Wu's quantization[4] and adaptive color reduction ([6]). These methods are all-around reference methods that perform very well using MSE evaluation criterion in almost any situation. We also show the results of a K-Means clustering approach starting from a random selection among our pre-quantized color set. Our method was evaluated with a refinement step of 10 K-Means iterations. We note that our method performs a little bit better than wu's quantization although the more complex ACR color reduction scheme achieves the best MSE measures. It is also important to repeat that even CMC measure, while quite accurate, is just an approach of visual similarity. Visual samples can be found at the author's website (*http://liris.cnrs.fr/%7Eapujol/colorQuant.shtml*). We may also note that improving MSE scores globally hurt image dynamics (see website for illustrations).

B. Performance

The algorithm was implemented in C# without specifically optimised code. It takes less than 1s to process a 51 000 color 480*320 image (with a single K-Means iteration for MSE optimization). After collecting image colors, computation depends only on color palette depth. Therefore, image size has only indirect impact on processing time: the resulting increase in the number of colors making agglomeration take a little bit more time. It is also worth noticing that this is the first part of the algorithm that consumes the most time. When used, iterative quantized color improvement using clustering also consumes significant processing time. Average measurements show that the first part consumes about 25% of the total processing time (66% if iterative improvement is not used) while the iterative improvement takes about 66% of the total processing time for 10 iterations. All performance experiments were conducted on a laptop using a Pentium core 2 duo 2ghz processor with 2gb of RAM. There was no dual core optimizations although this is a definite possibility for speeding up color space partitioning, color evaluation and other parallelizable tasks. As for visual similarity assessment, we used calibrated CRT displays to be as objective as possible.

VI. CONCLUSION

Given our task of image segmentation and characterization, we adressed the image quantization problem with the intent of minimizing the loss of color information during this step as well as preserving image dynamics. We proposed an simple two step algorithm using self-information to respect color diversity and which can avoid creating new colors. Regarding conventional MSE, this algorithm performs well, getting results as good as Wu's algorithm although not matching the results of the more costly algorithms based on unsupervized clustering.

As this initial work shown interesting results, future work involves determining more advanced criterion to combine color probability with color self information as well as refining self information computation by adapting the neighborhood size to the image color palette and improving computational efficiency using optimized c++ code as well as implementing parallelization.

REFERENCES

[1] J. P. Braquelaire and L. Brun, "Comparison and optimization of methods of color image quantization," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 1048–1052, 1997.

[2] P. Scheunders, "A comparison of clustering algorithms applied to color image quantization," *Pattern Recogn. Lett.*, vol. 18, no. 11-13, pp. 1379–1384, 1997.

[3] J. Ketterer, J. Puzicha, M. Held, M. Fischer, J. M. Buhmann, and D. W. Fellner, "On spatial quantization of color images," in *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I*. Springer-Verlag, 1998, pp. 563–577.

[4] X. Wu, "Color quantization by dynamic programming and principal analysis," *ACM Transactions on Graphics*, vol. 11, no. 4, pp. 348–372, 1992.

[5] K. Kanjanawanishkul and B. Uyyanonvara, "Novel fast color reduction algorithm for time-constrained applications." *J. Visual Communication and Image Representation*, vol. 16, no. 3, pp. 311–332, 2005.

[6] N. Papamarkos, A. Atsalakis, and C. P. Strouthopoulos, "Adaptive color reduction," *IEEE Transactions on systems man and cybernetics*, vol. 32, no. 1, pp. 44–56, 2002.

[7] A. Atsalakis, N. Papamarkos, N. Kroupis, D. Soudris, and A. Thnailakis, "Colour quantization technique based on image decomposition and its embedded system implementation," *IEE Proceedings Vision, Image and Signal Processing*, vol. 151, no. 6, pp. 511–524, 2004.

[8] J. R. Smith and S.-F. Chang, "Single color extraction and image query," in *Proc. IEEE Int. Conf. on Image Processing.*, 1995, pp. 528–531.

[9] C. Zhang and P. Wang, "A new method of color image segmentation based on intensity and hue clustering," in *Proc. 15th International Conference on Pattern Recognition*, 2000, pp. 613–616.

[10] X. Haisong and Y. Hirohisa, "Visual evaluation at scale of threshold to suprathreshold color difference," *Color Research and Application*, vol. 30, no. 3, pp. 198–208, 2005.