Automatic Palette Identification of Colored Graphics

Vinciane Lacroix*

CISS Department, Royal Military Academy, 1000 Brussels, Belgium

Abstract. The median-shift, a new clustering algorithm, is proposed to automatically identify the palette of colored graphics, a pre-requisite for graphics vectorization. The median-shift is an iterative process which shifts each data point to the "median" point of its neighborhood defined thanks to a distance measure and a maximum radius, the only parameter of the method. The process is viewed as a graph transformation which converges to a set of clusters made of one or several connected vertices. As the palette identification depends on color perception, the clustering is performed in the L*a*b* feature space. As edgels are made of mixed colors not expected to be part of the palette, they are removed from the initial data set by an automatic pre-processing. Results are shown on scanned maps and on the Macbeth color chart and compared to well established methods.

Key words: palette extraction, clustering, mean-shift

1 Introduction

The first step of color graphic vectorization is the identification of its palette. According to a recent study [1], approximately 13.5 million images are vectorized in the United States every years, consuming more than 7 million man hours. These images are made of photos, artworks, logos, etc. Commercial vectorisation software [1] exist but do not provide satisfying results in a full automated mode [2].

Despite this demand, there has been limited research done on colored image vectorisation except from a specific application: scanned maps [3], [4], [5] for which the vectorisation is performed on each colored layer, thus after the color extraction process.

Color palette reduction — when the final number of classes is a power of two the process is also known as "color image quantization" — such as kmeans, fuzzy-kmeans, median-cut, and octrees or any clustering method (also called unsupervised classification) could be considered for palette identification. For a comparison of color image quantization methods see [6], for a list and discussion on unsupervised classification see [7].

* This study is funded by the Belgian Ministry of Defense

[1] see http://en.wikipedia.org/wiki/Comparison_of_raster_to_vector_conversion_software

In this paper we propose the "median-shift", a new clustering method, so called by analogy with the mean-shift [2] [8], an iterative procedure that shifts each data point towards the "median" of data points of its neighborhood. The method is used to identify the palette of some scanned maps and other graphics.

The aim of any clustering method — and in particular of palette extraction — is to find a small set of representative of the whole data set. Many authors have suggested that clustering methods applied to multi-variate images should make use of the spatial information as neighboring pixels are likely to belong to the same cluster (see [9]). The spatial information is introduced either before, during or after the clustering. For example, a common pre-processing is made by regularization or pixel grouping. Markov Random Fields [10] are examples of methods including spatial constraints into their process. The authors of the mean-shift, by adding the spatial coordinates to the feature space, also introduce spatial constraints into the clustering process. Finally, voting schemes are examples of post-processing methods. In this article, the spatial information is first used in pre-processing by filtering out edges of the luminance image, as they are most probably mixed pixels.

The full pre-processing is described in section 2. Section 3 presents the median-shift while section 4 explains how the algorithm is implemented for palette extraction and addresses the case of scanned maps. Section 5 shows the results of the procedure on several maps and on the Macbeth color chart in comparison to well established methods. Section 6 provides summary, discussion and conclusions.

2 Pre-processing

The aim of the pre-processing is twofold: removing potential mixed pixels and finding a space providing a better cluster separation.

Apart from the noise generating unexpected colors, problems for automatic palette identification come from the superposition of colors and from edge pixels generating disturbing colors. In order to remove part of these outliers, an automatic thresholding on the norm of the gradient of the luminance is performed, as according to Koschan [11] 90% of all edges are in the intensity image. This operation may however delete pixels located on lines; the latter are then brought back by a fully automatic ridge extraction process [12] on the luminance image.

A colored pixel is represented in a 3D space. As our aim is to identify some "target colors" such that each present color could be replaced by its closest target, the space and the distance are of prime importance. The *uniform* L*a*b* space (noted "Lab" in this paper), in which the Euclidean distance reflects the perceived distance, has thus been chosen for a better color separation. The data set is then made of filtered pixels described by their Lab coordinates.

[2] There exist many variants of the mean-shift, but according to the author's knowledge, none is using the median instead of the mean.

3 Clustering

Let the data be a set of points $\mathbf{x}$ embedded in a n-dimensional feature space: $\mathbf{x} = (x_1, ..., x_n)$ and let $d(\mathbf{x}, \mathbf{y})$ be a distance defined in this space. The neighborhood of $\mathbf{x}$, $V_R(\mathbf{x})$, is defined as the set of $\mathbf{y}$'s such that $d(\mathbf{x}, \mathbf{y}) < R$. In this framework, a point is *isolated* if the cardinal of its neighborhood is one, and *connected* otherwise. The "median point" $\overline{\mathbf{x}}$ of $V_R(\mathbf{x})$ is defined as the point $\overline{\mathbf{x}} = (\overline{x}_1, ..., \overline{x}_n)$ such that $\overline{x}_i$ is the median of the ith component of all points in $V_R(\mathbf{x})$.

The median-shift algorithm is an iterative process which shifts each data point $\mathbf{x}$ at time t to $\overline{\mathbf{x}}$; it can be seen as a graph transformation. Each vertex $v(\mathbf{x}, w)$ of the graph G is characterized by a vector $\mathbf{x} = (x_1, ..., x_n)$, corresponding to a point of the data set, and a weight w initially set to 1. The vertices of G are connected if the points are neighbors. A cluster $C_R(\mathbf{x})$ is defined as the connected component of G containing the vertex $\mathbf{x}$.

At time t the graph G is transformed into G' (initially empty) according to the following rule: for each vertex $v(\mathbf{x}, w)$ of G the median point $\overline{\mathbf{x}}$ is computed; if the corresponding vertex already exists in G', its weight is incremented by w, otherwise it is created with a weight equal to w. The edges of G' are then updated before the operation is repeated at $t = t + 1$ with $G = G'$.

At $t = 0$, the graph G is a set of one or several clusters. The convergence of G is thus related to the convergence of each connected component. Except in very few cases depending on the distance definition and the distance between clusters, disconnected clusters will remain disconnected.

These exceptions set aside, any cluster made of isolated vertex will remain stable. If a cluster is made of two or more points all connected to each other, the cluster will collapse in one point, would it be a new or an existing one, thus converging.

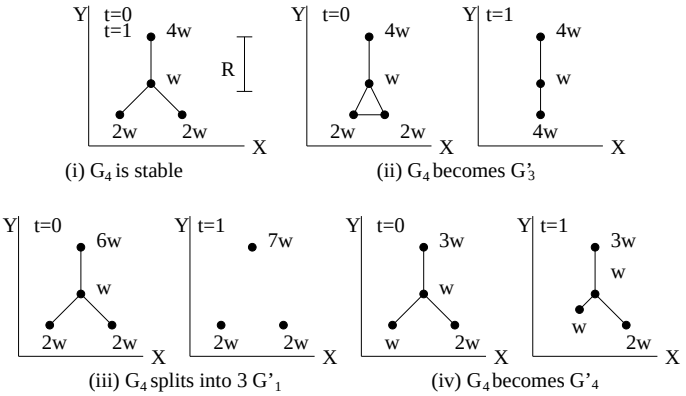

Fig. 1. Graph transformation scenarii

For a more complex cluster G_n characterized by n vertices, several scenarios may take place: (i) it is stable (ii) some vertices vanish thus leading to a graph G'_{n-i} (iii) it splits into several clusters ${G^1}_{n-i}, ..., {G^m}_{n-k}$, with a total number of vertices lower or equal to n (iv) it is transformed into G'_n. In this list of scenarios, only the last one could be problematic with respect to convergence, as in all other non stable cases, the total number of vertices is decreasing. However, though the number of vertices remains the same in (iv), the distances between them decrease, and at some distance below a threshold, the points could be considered as being at the same location, leading to G'_{n-1}. Figure 1 shows several scenarios for a 4 vertices graph in a 2D feature space. In practice convergence to complex clusters (i.e. made of several vertices) occurs when the radius is too big compared to the variations of the density.

The final result is thus a graph made of clusters containing one or several vertices, each cluster being separated from each other by at least a distance R. Note that the mean-shift may also be viewed as a graph transformation and its converging graph has the same propriety.

4 Implementation

The authors of the mean-shift [8] suggest to label a data point according to the cluster it converges to. In the palette extraction process however, this might result in assigning a point to a very different color. The following strategy is thus suggested. The median-shift algorithm is used to find the most important colors. The too small clusters ($< T$) are ignored. For each remaining cluster, the pixels are put aside if their distance to the cluster vertices is larger than R/2. If the set of all these outsiders is significant ($> PC\%$ of the initial set), the set is used again in a median-shift procedure providing additional clusters. The most important clusters are accepted until the number of ignored pixels is below the threshold.

Several distances can be used but the euclidean distance would be recommended when using the Lab color space. Several radii could also be used, but the value of 18 seems convenient for many applications.

A typical map is about 64 cm by 40 cm. Recommended scan resolution varies between 300 to 600 samples per inch, so that a scan map can be as large as 10078 pixels by 6299 pixels. A 512 by 512 sample could be too small to have the chance to get all pixel colors, while a 1024 by 1024 would seem reasonable. The following strategy is thus proposed for extracting the palette of large images.

A medium-size image (1024 by 1024) is extracted and pre-processed and divided in small images (512 by 512) the median-shifted is used on each sub-image. A new median-shift may be used for combining all clusters assigning to each cluster vertex a weight equal to the number of pixels no more distant than $R/2$.

5 Results

The procedure has been used to identify the colors of six maps and a photographic chart. The pre-processing involves a Gaussian gradient computation ($\sigma = 0.7$) used for the edge and ridge outputs. The mean of all non-zero edges m_e and non-zero bright and dark ridges, m_b and m_d respectively, are used as threshold to derive the mask of selected pixels. So far an image cut of size 512×512 has been considered in each map. Two distances have been used: Euclidean and maximum component difference ("box distance") with $R = 18$ (Euclidean distance) and $R = 15$ (box distance), $T = 100$, $PC = 4$. The computation time is highly dependent on the image content: 32, 36, 47, 64, 83 and 208 sec (box distance) and 30, 35, 43, 62, 78, and 210 sec (Euclidean distance) is needed for the median-shift computation. Results on four maps are shown on Figure 2. In order to better judge the quality of the palette extraction a labeling (i.e. assigning a color palette to all pixels) is performed. The palette is compared to the ones extracted by Vector-magic (http://vectormagic.com/home), and several color reduction algorithms provided by the VPmap-Pro software: median cut, kmeans, minimum distance and octree. For the latter, no pre-processing could be performed. Due to the bad quality of the results obtained with the octree method, these results are not reported. The first column shows the original images and in their lower right corner, a partial zoom. In order to judge the quality of the clusters extracted, the second column shows images labeled thanks to the median-shift algorithm (box distance) in which each pixel is assigned to the nearest vertex or to "undefined" if the box distance is larger than 3R/2. The last column shows all extracted palettes; from left to right: median-shift box distance $R = 15$, median-shift Euclidian distance $R = 18$, minimum distance (VPmap-Pro), Vector Magic, median-cut(VPmap-Pro), kmeans (VPmap-Pro).

All extracted palettes are displayed in the last column. For Vector-Magic the best number of classes has been chosen manually. VPmap fully automatic color reduction requires a minimum color classes of 12, value which has been chosen for all images.

The algorithm has also been tested on the photo of the Macbeth color chart used in [13] and compared to other algorithms: the implementation of the kmeans proposed in [13] and the mixture of Gaussians [14]. Figure 3 shows the Macbeth color chart in (a), the results of the median-shift in Lab space using Euclidian distance with $R = 14$ in (b), of the kmeans in RGB in (c) and in Lab space with 25 classes in (d), and finally of a mixture of Gaussians in RGB with 25 classes in (f). An "x" inside a square means that the square received a wrong label, while a "y" means the class has been correctly identified but the cluster is not a good visual representative of the class.

The median-shift parameter has been initially set to 18, as in previous experiments but at $R = 18$, two complex clusters are generated, suggesting to use a smaller R. At $R = 17$, all colors are separated but the three light grays are assigned to the same cluster and would thus have a "4x" score. A further separation occurs at $R = 14$, still keeping the two light grays in one class and the two dark grays with the background (3x); this is somehow expected as the

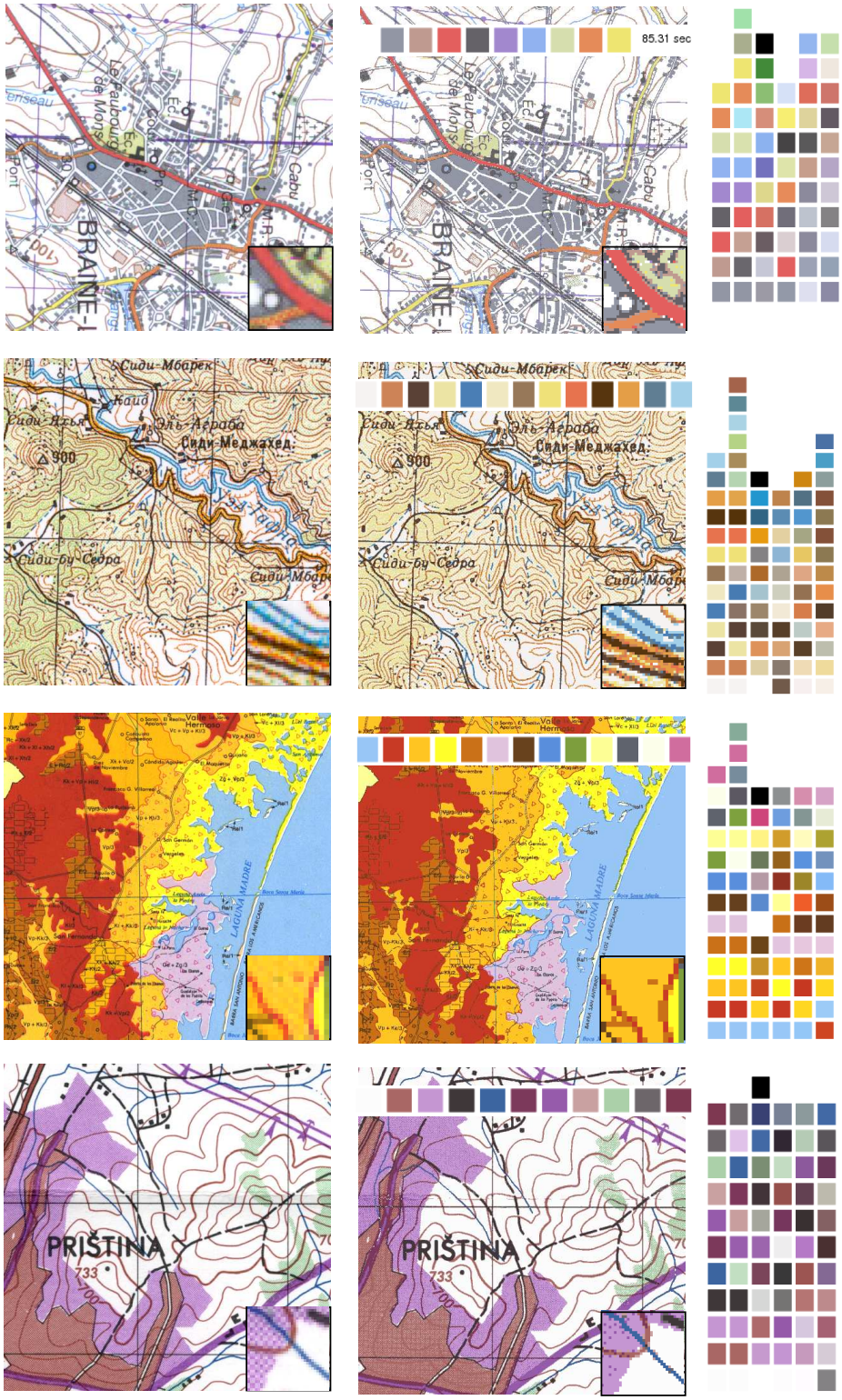

Fig. 2. Palette extraction on various maps (see text)

distance between the two light grays is about 7, and the background lies between the two darkest grays at a distance of about 9, both distances being lower than R. Note that the *original* Macbeth color chart has slightly different values; in particular each grey is separated from its neighbour by a distance of 15 (or 14 for the darkest).

The median-shift applied in Lab space is thus excellent for colors (no "x" and no "y" in colored squares); in particular, it is the only algorithm able to discriminate the deep blues (distance about 24) and the yellows (distance about 21) but is less good in discriminating shades of gray (3x). Other distances [15] like CIE19994, CIE2000, or CMC may provide better results.

The results of kmean depends on its initialization. Palus [13] proposed two initialization schemes; in this experiment both provided the same number of "x" and "y". In the RGB space three yellows and one green-yellow are merged into a unique yellow, two pairs of blues and one pair of pinks are merged, making a total of 6x. The algorithms is slightly better in the Lab space (5x, 1y).

The mixture of Gaussians gives the worse results in this experiment (8x,3y). This algorithm may also discover the best number of classes, 40 in this case, resulting in over-segmentation without resolving the difficult pairs of blues, yellows and grays.

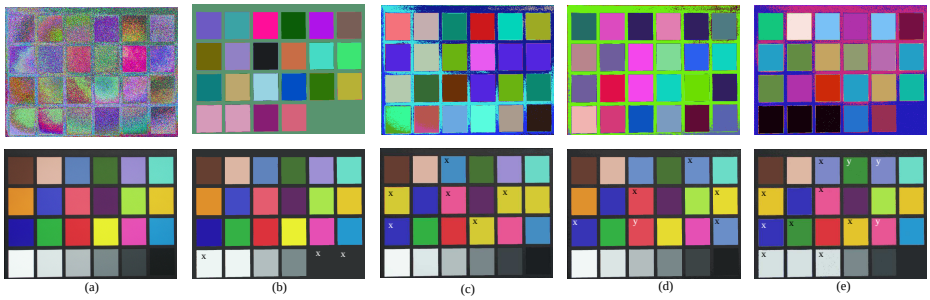

Fig. 3. Comparison of palette extraction on the Macbeth color chart. Upper row: pseudo colors; lower row: true colors. (a) Original Macbeth chart; (b) Median-shift ($R = 14$); (c) kmeans in RGB (25 classes); (d) kmeans in Lab (25 classes); (e) mixture of Gaussians in RGB space (25 classes); "x" and "y" denotes wrong class assignment, and visually not acceptable cluster value respectively.

6 Summary and conclusions

A strategy to identify the palette of scanned graphics has been proposed. A first pre-processing transforms the data into the Lab space and removes pixels with a less well defined color, for a better cluster separation.

The remaining pixels are clustered thanks to the median-shift, a new clustering algorithm which requires one parameter R related to the expected cluster

radius in the feature space associated to a distance. As far as palette extraction is concerned, the Lab feature space with Euclidian distance and R between 14 and 18 (or smaller in case of complex clusters) seems suitable for many applications. The procedure is applied on scanned graphics such as maps and a photographic chart showing improvement compared to well-established methods. In particular, the algorithm shows excellent discriminative power on saturated colors, but is slightly less efficient when dealing with low saturated ones.

7 Acknowledgments

The author wishes to thank Pasquale Nardone for the graph transformation suggestion, Henrick Palus and Dirk Borghys for the provision of results on the Macbeth color chart.

References

1. Diebel, J.R.: Bayesian Image Vectorization: The Probabilistic Inversion of Vector Image Rasterization. Phd thesis, Standford University, (2008)
2. Hilaire, X: RANVEC and the Arc Segmentation Contest: Second Evaluation. In: Wenyin Liu and Josep Lladós (eds.) Graphics Recognition—Ten Years Review and Future Pespectives. LNCS, vol. 3926, pp. 362–368. Springer Verlag, (2006)
3. Chen, Y. at al: Extracting Contour Lines From Common-Conditioned Topograpic Maps. IEEE TGARS, vol. 44, nb 4, pp. 1048–1057 (2006)
4. Deseilligny, M. P.: Lecture automatique de cartes, Phd thesis, Université René Descartes, Paris, France, (1994)
5. Robert, R.: Contribution à la lecture automatique de cartes, Phd thesis, Université de Rouen, Rouen, France (1997)
6. Braquelaire, J-P., Brun L.: Comparison and Optimization of Methods of Color Image Quantization. IEEE Trans. on Image Processing Vol 6, nb 7, pp. 1048–1052 (1997)
7. Jain, A.K., Duin R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. PAMI, vol. 22, nb. 1, pp. 4–37 (2000)
8. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans.PAMI, vol. 24, nb. 5, pp. 603–619 (2002)
9. Tran, T.N., Wehrens, H.R.M.J., Buydens, L.M.C.: Clustering multispectral images: a tutorial. Chemometrics and Int. Lab. Syst., vol. 77, pp. 3–17 (2005)
10. Price, K.: Computer Vision Biography, 8.8.3 MRF Models for Segmentation, `http://www.visionbib.com/bibliography/segment369.html` accessed 05-09
11. Koschan, A., Abidi, M.: Detection and Classification of Edges in Color Images. Sig. Proc. Mag., Spec. Issue on Color Img. Proc., vol. 22, nb. 1, pp. 64–73 (2005)
12. Lacroix, V., Acheroy, M.: Feature-Extraction Using the Constrained Gradient. ISPRS J. of Photogram. and RS, vol 53, nb. 2,pp. 85–94 (1998)
13. Palus, H.: On color image quantization by the k-means algorithm. 10. Workshop Farbbildverarbeitung, Droege, Detlev and Paulus, Dietrich (eds) Universität Koblenz-Landau, Tönning, Der Andere Verlag (2004)
14. Bouman, C. A.: Cluster: An unsupervised algorithm for modeling Gaussian mixtures `http://www.ece.purdue.edu/~bouman` (1997)
15. Ohta, N., Robertson, A. R.: Colorimetry Fundamentals and Applications. John Wiley & Sons, Ltd. (2005)