Histogram Thresholding for Automatic Color Segmentation based on *k*-means Clustering

Adhi Prahara[1], Iwan Tri Riyadi Yanto[2], and Tutut Herawan[3]

[1]Informatics Department, Universitas Ahmad Dahlan, Indonesia
[2]Information System Department, Universitas Ahmad Dahlan, Indonesia
[3]Department of Information Systems, University of Malaya

adhi.prahara@tif.uad.ac.id, yanto.itr@is.uad.ac.id, tutut@um.edu.my

Abstract. Color segmentation method has been proposed and developed by many researchers, however it still become a challenging topic on how to automatically segment color image based on color information. This research proposes a method to estimate number of color and performs color segmentation. The method initiates cluster centers using histogram thresholding and peak selection on CIE L*a*b* chromatic channels. *k*-means is performed to find optimal cluster centers and to assign each color data into color labels using previously estimated clusters centers. Finally, initial color labels can be split or merge in order to segment black, dark, bright, or white color using luminosity histogram. The final cluster is evaluated using silhouette to measure the cluster quality and calculate the accuracy of color label prediction. The result shows that the proposed method achieves up to 85% accuracy on 20 test images and average silhouette value is 0.694 on 25 test images.

Keywords: Automatic color segmentation; Histogram thresholding; Cluster centers initialization; *k*-means clustering.

1 Introduction

One of the most difficult in image processing is segmentation step. Segmentation is a process that partitions image into segments [1]. In color image segmentation, to segment image into its appropriate color label is difficult especially when there is no provided information on how many number of color labels should be made. The problem also comes from image dimension. Many digital color images have three dimensions of color channels that each dimension is often related to each other e.g. in RGB (red-green-blue) color space or independent to each other e.g. HSV (hue-saturation-value) and CIE L*a*b* (*Commission Internationale de l'Eclairage*-Lab) color space. The purpose of color image segmentation is that each segment will have pixel that similar to pixel in its own segment and different to pixel in other segment.

There are many ways to segment color in digital images. One can segment using three-dimensional color data [2]. Two-dimensional correlation of each channel can also be used, e.g. RG (red-green), GB (green-blue), or RB (red-blue) [3]. More efficient way, color image is transformed into color histogram [3,4] or transform to other color space that has independent color channels [5,6] such as CIE L*a*b* and HSV.

Color segmentation can be done by clustering method. Many of clustering methods need to be provided with number of cluster and points to start the initial cluster centers. Some methods have been developed to estimate the number of cluster or initiate cluster centers [7,8,9] and showing good results. However, choosing the clustering methods also depend on the data that will be used. Color data can be separated well using specific distance metric e.g. Euclidean, squared Euclidean, or city block.

This research proposes a method to estimate number of color and performs color segmentation. In summary, the contribution of this work is given as follow:

a. The proposed segmentation method initiates cluster centers using histogram thresholding and peak selection on CIE L*a*b* chromatic channels.
b. *k*-means is performed to find optimal cluster centers and to assign each color data into color labels using previously estimated clusters centers.
c. Finally, initial color labels can be split or merge in order to segment black, dark, bright, or white color using luminosity histogram.
d. The final cluster is evaluated using silhouette to measure the cluster quality and calculate the accuracy of color label prediction.

The rest of this paper is organized as follow: Section 2 presents related work. Section 3 presents the proposed segmentation method. Section 4 presents the results and discussion. Finally, the conclusion of this work is described in Section 5.

2 Related Works

Many color segmentation methods have been proposed. Jassim and Altaani proposed a method for color image segmentation by combining Otsu method applied in each channel of RGB image and median filter to smoothen the distorted image caused by formulation of new color image from previous Otsu method [2]. The median filter was also used to increase the segmented regions. The result of their experiment showed that the method was fast, easy to be implemented and good for medical image processing.

Kurugollu *et al*. proposed histogram multi-thresholding on bands of color and fusion the resulting segmentation [3]. The method divided the RGB color images into subsets of pair RB (red-blue), RG (red-green), and BG (blue-green), resulting bands of two-dimensional histogram. Histogram peaks was selected using multi thresholding scheme. The result of each band-pair's segmentation would be fusion into segmentation map. Result showed that using two-dimensional histogram for segmentation was superior than using one-dimensional histogram.

Cheng and Sun proposed hierarchical approach to segment color images using homogeneity [4]. Homogeneity histogram via multi-level thresholding was used in order to find uniform regions. Peak finding algorithm was applied to select significant peak from the histogram. Hue component was used in histogram analysis to segment the image. After that, a region merging was performed to avoid over-segmentation then CIE L*a*b* color space was used to measure the color difference. The result showed effectiveness and superiority of the proposed method on tested color image.

Severino and Gonzaga used color mixture to segment color image [5]. They proposed Hue, Saturation, Mixture color space as planes that described the RGB cube. They used the method in skin classification. The proposed method was surpassed the performance of all compared methods.

Angulo and Serra proposed color segmentation method by ordered merging [6]. At the first step, they determined color space which suitable for morphological operation. Next, they compared other color segmentation methods such as pyramid of watersheds and different color gradients with the proposed method, multi-scale color segmentation using merging chromatic-achromatic partitions ordered by saturation components. The result showed that saturation component plays an important role in order to merge the chromatic and achromatic information.

Tan *et al*. proposed a color segmentation method using Fuzzy C-Means (FCM) with initialization scheme to determine the number of cluster and cluster centers called Hierarchical Approach (HA) [10]. Color image was split into multiple regions and merging technique was used to determine number of cluster. Cluster centers were obtained from FCM. Result showed the HA initialization was superior to the state-of-the-art initialization for FCM and could be applied to segment any color images.

Wang *et al*. proposed color image segmentation method using support vector machine (SVM) and fuzzy C-means (FCM) [11]. The pixel was used as input of SVM which trained by FCM with the extracted pixel-level features. They combined the advantages of local information of color images with ability of SVM classifier to segment the image. The result showed that the method was effective for decreasing computational time and increasing the quality of color image segmentation.

This research proposes automatic color segmentation method using histogram thresholding and *k*-means clustering. The method estimates number of cluster and centers using histogram thresholding and peak selection on CIE L*a*b* chromatic channels. The clustering method uses *k*-means to find optimal cluster centers and to assign each color data into color labels. The initial color labels will be split or merge in order to avoid over-segmentation and to segment black, dark, bright, or white color using luminosity histogram. The final segmentation is enhanced using median filter.

3 Proposed Method

In this section, we present the proposed method which is automatic segmentation on color images based on histogram thresholding and *k*-means clustering. The proposed method consists of four main steps, i.e. extracting color data, find cluster centers using histogram thresholding and peak selection, calculate initial segmentation, and perform final segmentation. The general steps of the method are shown in Fig. 1.

From Fig. 1, the first step consists of procedure to extract color data from input image. Input color image is enhanced using median filter to remove noise and smoothing the color image. Enhanced image is converted into CIE L*a*b* color space. Every channel of CIE L*a*b* color space is extracted and calculated its mean and deviation standard to be used as threshold.

At the second step, histogram thresholding is performed to localize cluster centers area. This step constructs a two-dimensional histogram using a* and b* chromatic channels of CIE L*a*b* color space. The histogram value is divided by its maximum value to normalize the range into 0-1. A mean threshold is applied to separate high and low-density value. In order to localize the cluster centers area, a morphological operation is performed. The morphological operation localizes the high-density value by dilating and filling the hole to find the perimeter of cluster centers area. Data inside the perimeter is likely a cluster centers area.

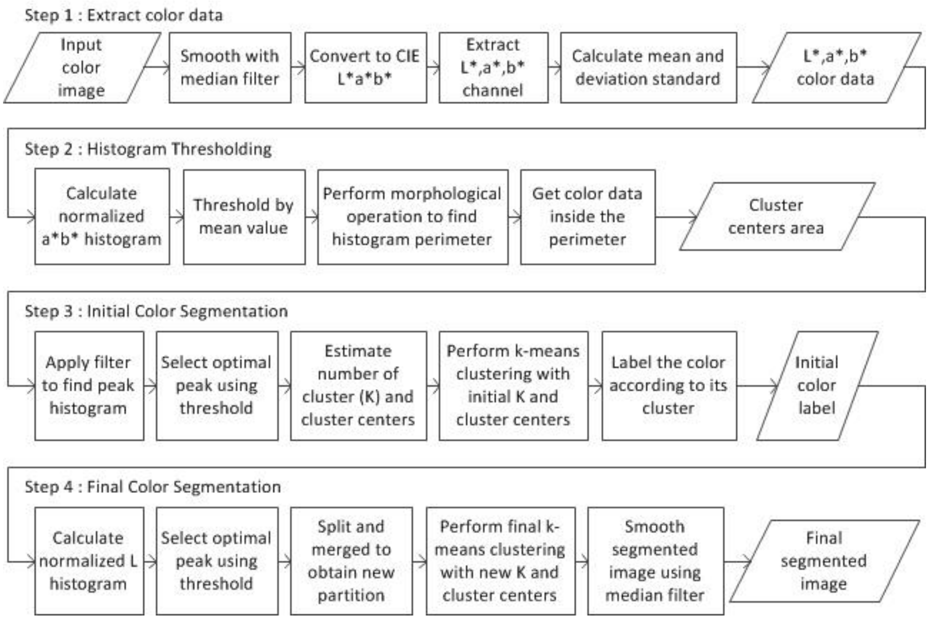


Fig. 1. The general steps of automatic color segmentation using histogram thresholding and *k*-means clustering.

At the third step, initial number of cluster and cluster centers will be estimated on cluster centers area. A peak filter is applied to find all the peaks that represent the highest density around its neighborhood. All of peaks will be sorted by descending order. An iterative procedure will compare the higher value of peak with other peak according to the order. Lower peak that has squared distance higher than variance will be kept as superior peak otherwise it will be merged with the higher peak. The final number of peaks will become initial number of cluster and its value will become the initial cluster centers. The *k*-means clustering with this defined initial number of cluster and cluster centers is performed to assign every color data into its optimal color label.

At the last step, the initial color label will be refined to differentiate the black, dark, bright, and white color using luminosity channel of CIE L*a*b* color space. One-dimensional normalized histogram is constructed using luminosity channel for each color label. To ignore low-density value, a mean threshold is applied. Optimal peak will be selected among the histogram's peak using the same iterative procedure in the step three. If there is more than one peak, the cluster might contain dark or bright color and need to be split. After all clusters are updated, the final number of cluster and initial cluster centers will be obtained. *K*-means clustering is performed with this new number of cluster and cluster centers to find the final color labels. Median filter is applied on final segmented image to remove noise and to smooth the image.

3.1 Color Histogram

The CIE L*a*b* color space is one of the most popular to quantify the visual differences of color. It is consists of luminosity layer L*, and two chromaticity layers, a* and b*. The chromaticity layer a* indicating where color falls along the red-green axis and the chromaticity layer b* indicating where the color falls along the blue-yellow axis. The color information is stored in a* and b* layers and can be measured using Euclidean distance. By using two chromaticity layers a* and b*, the histogram dimension can be reduced into two-dimensional from original three-dimensional of RGB color space.

If the size of image is $N \times M$, I_A and I_B is a* and b* intensity data, p and q is the index of histogram bins for a* and b* channel, i and j is the index of a* and b* coordinate intensity data in the image. The two-dimensional a*-b* normalized histogram (*hist_AB*) can be calculated using (1).

$$hist\_AB(p,q) = \frac{\sum_{i=0,j=0}^{M,N}(I_A(i,j)=p,\ I_B(i,j)=q)}{\max(hist\_AB)} \quad \textbf{(1)}$$

If the size of image is $N \times M$, I_L is luminosity channel intensity data, p is the index of histogram bins for luminosity channel, i and j is the index of luminosity coordinate intensity data in the image, then one-dimensional luminosity normalized histogram (*hist_L*) can be calculated using (2).

$$hist\_L(p) = \frac{\sum_{i=0,j=0}^{M,N}(I_L(i,j) = p)}{\max(hist\_L)} \quad \textbf{(2)}$$

3.2 Histogram Thresholding

Histogram thresholding uses statistical color data such as mean, deviation standard and variance value as threshold. Mean is used as threshold to ignored low-density value of histogram, deviation standard is used to determine the length of peak filter kernel, and variance is used as threshold to select optimal peak by its squared distance. Morphological operation is applied on two-dimensional normalized histogram to localize the high-density values as cluster centers area. The morphological operations are dilation and filling. Dilation is used to expand the binary image formed by normalized histogram and filling is used to fill the hole of color data distribution. The result of morphological operations is cluster centers area's perimeter. Color data that lies inside the perimeter will be kept otherwise will be ignored from the calculation of initial cluster centers. The area inside perimeter will become cluster centers area.

3.3 Number of Cluster and Cluster Centers Estimation

This method uses statistic color information to obtain the threshold, performs histogram thresholding, and peak selection to estimate number of cluster and cluster centers. Peak is a color data with value superior to other peak inside the radius of filter kernel. The length of filter kernel is determined from the value of deviation standard. A peak filter with 3×3 kernel is shown in Fig. 2.

-1	-1	-1
-1	8	-1
-1	-1	-1

Fig. 2. A peak filter with 3×3 kernel

From Fig. 2, the kernel ensures that the value will be compared to its neighborhood and if the result is higher than 0 then the value is considered as a peak. The procedure to find optimal peak after peak extraction is explained below:

a. All peak is sorted by descending order. This step ensures that higher value of peak will be prioritized than lower value of peak.
b. An iterative procedure is performed from the higher peak to compare with other lower peak below it.
c. Squared distance is used to measure the distance of current peak with other peak using variance as a threshold. If distance is higher than variance then the lower peak will be promoted to higher peak otherwise it is merged with current peak.
d. The iterative procedure is done until all peak have been evaluated.

The result of the iterative procedure is optimal number of peak and peak coordinate that will be used as initial number of cluster and cluster centers.

3.4 Color Segmentation

Color segmentation uses *k*-means clustering to assign every color data to its optimal color label. *K*-means is one of the supervised clustering methods, which usually uses Euclidean distance to assign data into its optimal cluster. The data within cluster are close as possible to other data and as far as possible from data in other cluster. This research provides *k*-means with initial number of cluster (*k*) and cluster centers. Color segmentation procedure using *k*-means will be explained in these steps:

a. Initialize number of cluster and cluster centers using statistical color information, histogram thresholding and peak selection.
b. Calculate Euclidean distance of each color data to each cluster centers.
c. Assign each color data to the nearest cluster centers using (3) where $c_{(i)}$ is the *i-th* distance data, x_i is the *i-th* color data, and μ_j is the *j-th* cluster center.

$$c_{(i)} = \arg\min_{1 \le j \le K} \left\| x_i - \mu_j \right\|^2 \tag{3}$$

d. After all color data assign to its nearest cluster, calculate new cluster centers.
e. Repeat step b to d until converge or all cluster member are not changing label.
f. Get the final number of cluster, cluster centers, and cluster label.

3.5 Evaluation

The performance of proposed method is evaluated by two methods. Segmentation accuracy is determined by comparing actual number of color in supervised color image with segmentation result. Accuracy is calculated using (4).

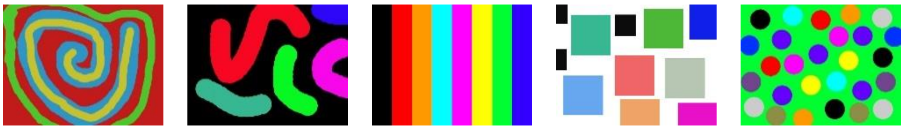
$$accuracy = \frac{number\ of\ right\ color\ prediction}{number\ of\ test\ color\ images} x100\% \tag{4}$$

Segmentation quality is evaluated using silhouette method. The silhouette measure how similar the data is to data in its own cluster when compared to data in other cluster [12]. If *a(i)* is the average distance from the *i-th* data to the other data in the same cluster as *i* and *b(i)* is the minimum average distance from the *i-th* data to data in different cluster then the silhouette value for *i-th* data *s(i)* can be calculated using (5).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad \textbf{(5)}$$

4 Result and Discussion

The proposed method is tested using Matlab 2010b and running on laptop with Intel i5 processor and 8GB of RAM. Test is conducted on 20 supervised color images and 25 natural color images. Supervised color image contain mostly 1-12 primary and secondary color as shown in Fig. 3a - Fig. 3e. Natural color image is obtained from the internet and McGill Calibrated Color Image Database [13], which consists of objects in nature. Some of natural color image are shown in Fig. 4a – Fig. 4e.



(a) (b) (c) (d) (e)

Fig. 3. Supervised color image. (a) 4 colors image (b) 6 colors image (c) 8 colors image (d) 10 colors image and (e) 12 colors image

(a) (b) (c) (d) (e)

Fig. 4. Natural color image. (a) flowers (b) sea (c) animal (d) jellyfish (e) apple

4.1 Cluster Initialization Result

Proposed method uses histogram thresholding and peak selection to estimate number of cluster and cluster centers. In initial segmentation step, the method uses mean threshold to ignore low-density of a*b* histogram and 0.8*variance as threshold to select optimal peak. In final segmentation step, the method uses mean threshold to ignore low-density of L* histogram and variance as distance threshold to select optimal peak to split the cluster. Fig. 5 shows the distribution of selected peak, initial cluster centers, and final cluster centers. From Fig. 5, the green mark is the perimeter, cyan mark is distribution of color data, and yellow mark is the high-density area. The blue cross marker is the initial peaks, red circle marker is the initial cluster centers, and the red cross marker is the final cluster centers.

The final cluster centers are likely to be inside the perimeter and lies in the high-density area. Sometimes the final cluster centers are lies outside the perimeter because of the wide distribution of color data. Notice that the number of final cluster centers is seven while the initial centers are only four. This means some of the clusters are split in the final segmentation step to accommodate dark or bright color. The distance of final cluster centers is close to each other because it is plotted on a*b* histogram while the main difference is on the L* channel that represent the dark or bright intensity.

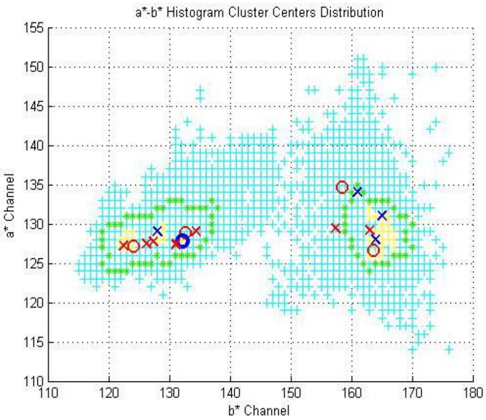


Fig. 5. Cluster centers localization

4.2 Segmentation Result

Segmentation is done by *k*-means clustering method using initial number of cluster and cluster centers obtained from histogram thresholding and peak selection. The small population of color data might be preserved if the average value of all data is low and can be merged if the average value is high. Color image are shown in Fig. 6a and Fig.6d, the initial segmentation are shown in Fig. 6b and Fig. 6e, and the final segmentation is shown in Fig. 6c and Fig. 6f.

The result between initial segmentation and final segmentation are looks different because of the different number of cluster applied in the segmentation. In Fig. 6e, the lotus flower image is segmented with *k*=4 and the final segmentation with *k*=8. In final segmentation step, initial cluster might be split to deal with dark or bright color. However, sometimes it is same as the initial cluster if the cluster is already optimal. The number of cluster becomes the number of predicted color by the method.


(a)

(b)

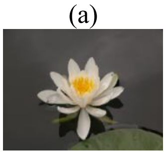
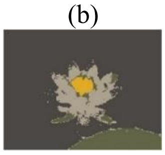
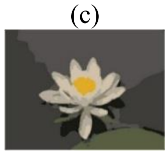
(c)

(d)

(e)

(f)

Fig. 6. Result of color segmentation. (a) lighthouse color image (b) lighthouse initial segmentation, *k=4* (c) lighthouse final segmentation, *k=5*; (d) lotus original image (e) lotus initial segmentation, *k=4* and (f) lotus final segmentation, *k=8*

Sometimes, there are color data that not assign to their right color label. Fig. 7a shows color image and Fig. 7b shows the false color label (red square marker). The number of color should be 12 but the method is predicted 10. The distance to differentiate color can become large depend on the variance, resulting the low-density peak to be merged to the higher density peak around its variance length and the color data represent by that peak is also merged to another color label. The variance threshold is enough to differentiate the color label because smaller threshold may result in over segmentation.

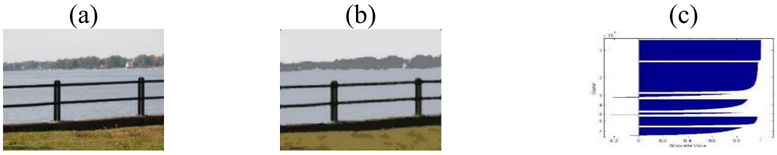
(a) (b)

Fig. 7. False color label after final segmentation of supervised color image. (a) color image (b) false segmentation result (red square marker)

4.3 Evaluation

Supervised color image is used to measure how well the proposed method segments color image by comparing the result of segmentation with actual number of color. Accuracy of color prediction with 20 test images is 85%. With that accuracy, the method can predict number of color from color image quite well. Performance of segmentation is measure using silhouette method. Silhouette value is ranged from -1 to +1. The higher value means quality of segmentation is good and the lower value means poor quality segmentation that can be caused by too many or too few cluster. The result of average silhouette value tested on 25 natural image is 0.694. Fig. 8a and Fig. 8d show the original image, Fig. 8b and Fig. 8e show the final segmentation and Fig. 8c and Fig. 8f show the silhouette visualization.

(a) (b) (c)



(d) (e) (f)

Fig. 8. Cluster performance measurement using silhouette visualization. (a) buffalo: color image (b) buffalo: final segmentation (c) buffalo: silhouette visualization (d) sea: color image (e) sea: final segmentation (f) sea: silhouette visualization

For color image in Fig. 8a, the predicted number of color is 6 and the average silhouette value is 0.758 and for color image in Fig. 8d, the predicted number of color is 7 and the average silhouette value is 0.853. As the result shows, the method produces appropriate cluster in color segmentation. Although, there are some cluster that only have small member (thin shape of the silhouette) and some part of the cluster that close to another cluster (the sharp edge of the silhouette), most of the cluster have medium-high value. The overall high value proves that this method produces appropriate number of cluster and good quality segmentation.

5 Conclusion

This paper has discussed *k*-means clustering-based method with emphasizes on histogram thresholding for automatic color segmentation. The proposed method has obtained good result in estimating number of color and segmenting the color image. This research shows that final cluster centers is likely inside the perimeter of estimated cluster centers area and lies inside the high-density of histogram value. The final cluster centers also not far from initial cluster centers obtained from peak selection. Therefore, the further development of this method is expected to localize the exact location of cluster centers without or with some minimal iterative procedure. The *k*-means clustering method is suitable for color segmentation and resulted in good quality of cluster. For future works, the method can be combined with spatial color information in order to produce good segmentation of color and even can be used to separate objects.

Acknowledgement

This research is supported by University of Malaya research grant UMRG.

References

1. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing Third Edition*. Prentice Hall (2008)
2. Jassim, F.A., Altaani, F.H.: Hybridization of Otsu Method and Median Filter for Color Image Segmentation. *International Journal of Soft Computing and Engineering* (IJSCE), vol. 3, pp. 69—74 (2013)
3. Kurugollu, F., Sankur, B., Harmanci, A.E.: Color Image Segmentation Using Histogram Multithresholding and Fusion. *Image and Vision Computing*, vol. 19, issue 13, pp. 915—928 (2001)
4. Cheng, H-D., Sun, Y.: A Hierarchical Approach to Color Image Segmentation Using Homogeneity. *IEEE Trans. on Image Processing*, vol. 9, issue 12, pp. 2071—2082 (2000)
5. Severino Jr, O., Gonzaga, A.: A New Approach for Color Image Segmentation Based on Color Mixture. *Machine Vision and Application*, vol. 24, issue 3, pp. 607—618 (2013)
6. Angulo, J., Serra, J.: Color Segmentation by Ordered Mergings. In: *Proceedings of the International Conference on Image Processing* (ICIP), vol. 2, pp. 125—128 (2003)
7. Meilă, M., Heckerman, D.: An Experimental Comparison of Model-Based Clustering Methods. *Machine Learning*, vol. 42, issue 1, pp. 9—29 (2001)
8. Khan, S.S., Ahmad, A.: Cluster Center Initialization Algorithm for K-modes Clustering. *Expert System with Applications*, vol. 40, issue 18, pp. 7444—7456 (2013)
9. Celebi, M.E., Kingravi, H.A., Vela, P.A.: A Comparative Study of Efficient Initialization Methods for the k-means Clustering Algorithm. *Expert Systems with Applications*, vol. 40, issue 1, pp. 200—210 (2013)
10. Tan, K.S., Lim, W.H., Isa, N.A.M.: Novel Initialization Scheme for Fuzzy C-Means Algorithm, on Color Image Segmentation, *Applied Soft Computing*, vol. 13, issue 4, pp. 1832—1852 (2013)
11. Wang, X-Y., Zhang, X-J., Yang, H-Y., Bu, J.: A Pixel-Based Color Image Segmentation Using Support Vector Machine and Fuzzy C-Means, *Neural Networks*, vol. 33, pp. 148—159 (2012)
12. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc. Hoboken, New Jersey (1990)
13. McGill Vision Research - *McGill Calibrated Colour Image Database*, http://tabby.vision.mcgill.ca/html/browsedownload.html