Journal of Theoretical and Applied Information Technology
20th February 2015. Vol.72 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

VIDEO OBJECT SEGMENTATION APPLYING SPECTRAL ANALYSIS AND BACKGROUND SUBTRACTION

[1]RURI SUKO BASUKI, [2]MOCH. ARIEF SOELEMAN, [3]RICARDUS ANGGI PRAMUNENDAR, [4]AURIA FARANTIKA YOGANANTI, [5]CATUR SUPRIYANTO

Faculty of Computer Science, University of Dian Nuswantoro, Semarang, Indonesia
E-mail: rurisb@research.dinus.ac.id

ABSTRACT

This study proposes an approach to segment video object semi-automatically. The issue of this study is the lack of semantic information on video object segmentation. Manual segmentation by human is not effective if the video has a large size. For initialization, we use scribble-based technique to differentiate between foreground and background. After the separation object from the background, the subtraction operation between the current and subsequent frame was performed by applying a background subtraction algorithm. Spectral analysis and background subtraction for video object segmentation becomes effective. The evaluation of this study is measured by Mean Square Error. Experiment results demonstrate the high precision of object segmentation.

Keywords: *Segmentation, Alpha Matting, Background Subtraction*

1. INTRODUCTION

The demand of video editing applications (such as video segmentation and video compositing) increases rapidly due to the advent of digital video standards such as Digital Television (DTV) in America, Digital Video Broadcasting – Terrestrial (DVB-T) in Europe and Integrated Services Digital Broadcasting-Terrestrial (ISDB-T) in Japan. It occurs since the video object segmentation in editing applications play an important role in the operation of movie production, news and advertising. Various applications such as object extraction, image recognition, augmented reality and motion understanding can be performed with the object-based technology.

The fundamental issue in video object segmentation is an ill-posed problem, namely the video object with no semantic information [1]. Therefore, the semantic information of the video object can only be identified by the human eyes by considering the video context so that the objects' withdrawal process in video editing is performed by manual segmentation. However, it is not an effective way to handle a video that has a large size. Many algorithms associated to the video object segmentation are developed to overcome this problem. The algorithms are classified into two categories, they are automatic object segmentation [2] and semi-automatic object segmentation [3] [4].

The parameters in the automatic segmentation are the specific characteristics such as color, texture and movement which are performed without human intervention. The difficulty in semantic relevant object separation is the main problem of the automatic segmentation. So there is no guarantee that the results of the automatic segmentation will be satisfactory, because the semantic object has a lot of color, texture and movement [5] [6] [7].

Several semi-automatic segmentation method which is a combination of manual and automated methods is proposed for that reason. Which in essence, the approach is a technique to withdraw the object involving human intervention on multiple frames in the segmentation process. Since semantic information can be directly made by human's assistance, the object segmentation process in the subsequent frames is performed using a tracking mechanism with temporal transformation.

In previous study, several methods related to tracking mechanism had been developed. In a region-based method, parameters of movement, texture and color were applied to keep track the related areas corresponding to the shape of the

Journal of Theoretical and Applied Information Technology
20th February 2015. Vol.72 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

object semantics. However, this method has a very complex tracking mechanism in maintaining the relationship between the areas consist of semantic objects [8]. The contour-based methods such as snake [3] would be robust when it was applied to the track on object contours, but it did not represent the whole of the object pixels, so this method might not work properly to follow the feature and the impact between edges were not connected to each other. While the model-based method applied a priori knowledge of the object shape. The shortcoming of this approach was not acceptable on the generic semantic video object segmentation since the detail required information about the object geometry [9].

Keyframe was created from one of the frames selected and considered as a still image. Matting techniques were applied to pull object of this frame. To distinguish the foreground and background object, interactive matting was applied using a scribble technique as an interface [10]. Hereinafter, the object segmentation on the subsequent frames was performed by using the background substraction algorithm.

2. KEYFRAME DESIGN

The initial step of a video segmentation process was performed by designing the selected frame of the sequences scene which became a keyframe. Since the object had no semantic information, human assistance was required to give scribble as a label to distinguish regions representing foreground and background object (white for foreground and black for background).

A. General Compositing Equation

Alpha channel [10][11][12] was applied to control the linear interpolation in the foreground and background which were depicted in matting algorithm by assuming that each pixel in the input image I_i was a linear color combination of foreground F_i and background B_i.

$$I_i = \alpha_i F_i + (1-\alpha_i) B_i, \quad \text{where } 0 \le \alpha \le 1 \tag{1}$$

Based on compositing equation Eq. (1) of each pixel, it was assumed to be a convex combination of layers K image denoted as

$$I_i = \sum_{k=1}^{K} \alpha_i^k F_i^k \tag{2}$$

the fractional contribution of each layer observed in each pixel was determined by the vector K of α^k, a component of image matting.

B. Spectral Analysis

Spectral segmentation method was associated with the affinity matrix. For example, the image A, size N x N was assumed as $A_{(i,j)} = e^{-d_{ij}/\sigma^2}$ and d_{ij}. In which d_{ij} was the space among pixels (e.g. color and geodesic space), defined as

$$L = D - A \tag{3}$$

While D was matrix degree from graph.

$$G = (V, E) \, with V = n \tag{4}$$

with diagonal matrix

$$D_{(i,j)} = \sum_j A(i,j), \quad \text{where } d_{i,j} = \begin{cases} \deg(v_i) & if\ i = j \\ 0 & Otherwise \end{cases} \tag{5}$$

$D_{(i,j)}$ was filled with degree information of each vertex (node) with D for G as rectangular matrix size $n\ x\ n$. So L was a symmetric positive semi-definite matrix with eigenvector which was able to capture a lot of image structure. Furthermore, the image was the composition of some different clusters or connected components which was captured by affinity matrix A. Subset C in image pixel was the connected component of image $A_{(i,j)} = 0$ for each (i, j) so $i \in C$ and $j \notin C$, so there was no subset C fulfilling the property. m^C was defined as indicator vector of component C ,therefore

$$m_i^C = \begin{cases} 1 & i \in C \\ 0 & i \notin C \end{cases} \tag{6}$$

with the assumption that the image consisted of connected components of K, $C_1, \ldots, C_K$ to $\{1, \ldots, N\} = \cup_{k=1}^{K} C_k$ with C_k disjoint path on the pixel, then the m^C represented 0-eigenvector (eigenvector with eigenvalue 0) from L. Since the rotation of matrix R in size $K \times K$, and vector $[m^{C_1}, \ldots, m^{C_K}]R$ was null-space based on L, then the indicator vector $m^{C_1}, \ldots, m^{C_K}$ resulted from eigenvector calculation on L was only a reaching rotation. "Spectral Rounding", a component

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction with the smallest eigenvector, was the concern in some studies [13][14][15][16][17]. K-Means algorithm was a simple approach used for clustering the image pixels [13], while the perturbation analysis algorithm was to limit errors as a function of connectivity within and across clusters.

1) Matting Laplacian

In order to evaluate the quality matte without considering colors of foreground and background, Matting Laplacian [10] was applied by using a local window w forming two different pathways in the RGB domain as denoted in Eq.(6). Furthermore, α in w is expressed as a linear combination of color channels.

$$\forall i \in w \;\; \alpha_i = a^R I_i^R + a^G I_i^G + a^B I_i^B + b \qquad (7)$$

The deviation of linear model Eq. (7) in all image window w_q was one of the findings in a matte extraction problems.

$$J(\alpha,a,b) = \sum_{q}\sum_{i \in w_q} \left(\begin{matrix} \alpha_i - a_q^R I_i^R + \\ a_q^G I_i^G + a_q^B I_i^B + b_q \end{matrix} \right)^2 + \varepsilon \left\| a_q \right\|^2 \qquad (8)$$

the requirements which should be fulfilled of the alpha was $\varepsilon \left\| a_q \right\|^2$ in which a linear model coefficients a,b that allowed elimination from Eq.(8) and the result was a quadratic cost in α

$$J(\alpha) = \alpha^T L \alpha, \qquad (9)$$

It had the ordinary minimum cost which was a constant α vector, then in framework user-assisted [12], $J(\alpha)$ was the subject minimized in user constraint. The equation L (9) was matting Laplacian Symmetric semi-definite positive matrix N x N that the matrix inserting input image function in local windows, depended on unknown foreground and background color in the coefficient of linear model. L was defined by the sum of matrix $L = \sum_q A_q$ in which each part was filled with affinity among pixels in local window w_q

$$A_q(i,j) = \begin{cases} \delta_{ij} - \frac{1}{|w_q|}\left(1 + (I_i - \mu_q)^T \left(\sum_q + \frac{\varepsilon}{|w_q|} I_{3x3}\right)^{-1} (I_j - \mu_q)\right), \\ 0 \quad Otherwise \end{cases}$$

$$where \;\; (i,j) \in w_q \qquad (10)$$

In which δ_{ij} was Kronecker delta, μ_q was the average color vector in al pixel q, $\sum_q$ is was matrix covariant size 3×3 in the same windows, $|w_q|$is was the sum of pixels in window, and I_3 was identity matrix size 3×3. By the occurrence of the smallest eigenvector, the other use of matting Laplacian property Eq. (10) was to seize information of job fuzzy cluster on image pixel, including the calculation before the limit determent by specified user [15].

2) Linear Transformation

The linear transformations track in eigenvector would produce a set of vector in which the value was adjacent to a binary. The equation denoted as $E = [e^1, \ldots, e^k]$is was converted to matrix N x K of eigenvector. Next, to locate a set of linear combination K, vector y^k minimized

$$\sum_{i,k} \left|\alpha_i^k\right|^\gamma + \left|1 - \alpha_i^k\right|^\gamma, \text{ where } \alpha^k = Ey^k$$

$$\text{subject to } \sum_k \alpha_i^k = 1 \qquad (11)$$

The robust measurement value in matting component [12] was determined by $|\alpha_i^k|^\gamma + |1 - \alpha_i^k|^\gamma$, if $0 < \gamma < 1$, thus, the value of $\gamma = 0{,}9$. Since the cost function Eq. (11) was not convex, the initialization process determine the result of Newton process. Therefore, K-means algorithm could be applied in the initialization process on the smallest eigenvector in matting Laplacian and projects indicator vector of cluster resulted from eigenvector E.

$$\alpha^k = EE^T m^{C^k} \qquad (12)$$

The matting component result Eq. (12) was then added. Thus it gave solution for Eq. (11).

3) Grouping Component

The complete results of matte extraction of the foreground object were determined by a simple summation on the foreground. For example, $\alpha^{k_1}, \ldots, \alpha^{k_n}$ is designed as a component of the foreground, so that

$$\alpha = \alpha^{k_1} + \ldots + \alpha^{k_n} \qquad (13)$$

The measurement of the results α -matte was perform by $\alpha^T L \alpha$ when the smallest eigenvector was not equal to zero, in which L was the matting Laplacian. The first calculation of correlation among matting component and L deviation in

Journal of Theoretical and Applied Information Technology
20th February 2015. Vol.72 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

matrix $\Phi K \times K$ was defined as

$$\Phi(k,l) = \alpha^{kT} L \alpha^{l} \quad (14)$$

then, matte cost was calculated as

$$J(\alpha) = b^{T} \Phi b \quad (15)$$

where *b* was the binner vector of *K-dimensional* indicating the chosen component.

3. TRACKING WORKFLOW

A. Background Subtraction

Background subtraction [18] was applied to identify differences in the intensity of the current image with the background. Frame difference was the technique applied in the background subtraction which was a non-recursive techniques. This model was assumed as BF, a binner value of foreground image.

$$BF(x,y,n) = \begin{cases} 1, if \left| I(x,y,n) - I(x,y,n) - 1 \right| \geq \alpha \\ 0, otherwise \end{cases} \quad (16)$$

The threshold (α) was applied to classify the foreground and background. Here, Otsu algorithm was applied to generate the threshold value.

B. Otsu Threshold

Otsu [19] is an adaptive threshold algorithm based on the histogram showing the value of changes in intensity of each pixel in one-dimensional image. The x-axis was applied to express the difference of intensity levels, while the y-axis was applied to declare the number of pixels that had intensity values. By applying the histogram clustering, the image pixel based on the threshold value could be performed. Optimal threshold was obtained from intensity differences of the pixels, so that it could be applied for separating groups. The information obtained from the histogram was the amount of the intensity difference (denoted by L), and the number of pixels for each intensity level was denoted by $n(k)$, with $k = 0 .. 255$). Seeking of the threshold value in Otsu algorithm was performed as follows:

1. Calculate the histogram of the normalized image denoted by p_i with $i = 0, 1, 2 \ldots L-1$

$$p_i = \frac{n_i}{MN} \quad (17)$$

where n_i was the number of the pixels at each intensity, and MN was the number of n_i starting from n_0 to $n_L - 1$

2. Compute the cumulative number of $p_1(k)$ for $k = 0, 1, 2..., L-1$.

$$P_1(k) = \sum_{i=0}^{k} p_i \quad (18)$$

3. Count the comulative average of $m(k)$ for $k = 0, 1, 2..., L-1$.

$$m(k) = \sum_{i=0}^{k} i p_i \quad (19)$$

4. Calculate the average global intensity m_G byusing ;

$$m_g = \sum_{i=0}^{L-1} i p_i \quad (20)$$

5. Compute the variance among classes, $\sigma_B^2(k)$ for $k = 0, 1, 2..., L-1$.

$$\sigma_B^2(k) = \frac{[m_G P_1(k) - m(k)]^2}{P_1(k)[1 - P_1(k)]} \quad (21)$$

6. Select a threshold value of the $k*$ if the index value of the maximum variance between classes $(\sigma_B^2 \Rightarrow \max(k))$, and if the index value was more than one value of $k*$, then the threshold value was determined from the average value of $k*$.
7. Determine the size of the separation $\eta*$ with $k = k*$

$$\eta(k) = \frac{\sigma_B^2(k)}{\sigma_B^2} \quad (22)$$

while,

$$\sigma_B^2 = \sum_{i=0}^{L-1} (1 - m_G)^2 p_i \quad (23)$$

Note: the value of *K* was obtained when $\sigma_B^2(k)$ was in maximum.

4. OBJECT SEGMENTATION PROCESS

In this section, the workflow of object segmentation process for video data was performed in the steps, as depicted in Fig. 1.

Journal of Theoretical and Applied Information Technology
20th February 2015. Vol.72 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

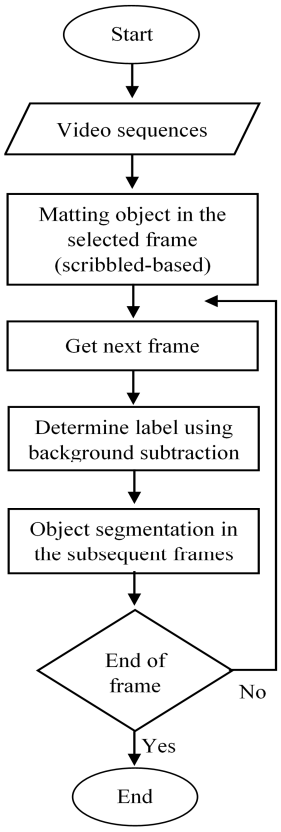
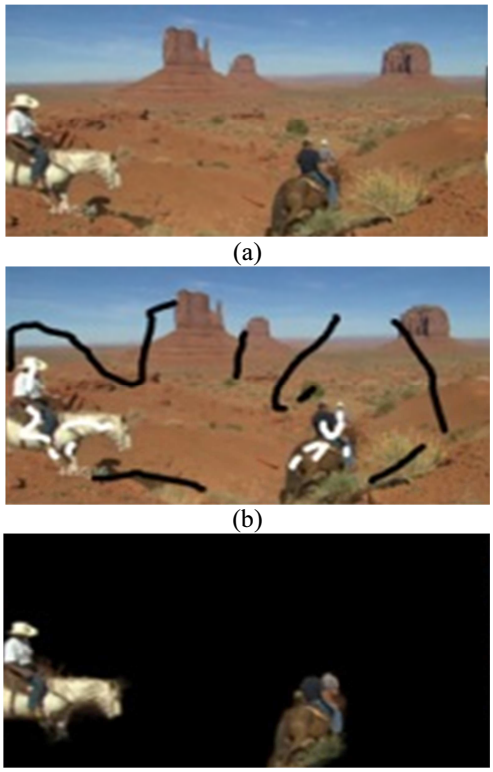


Figure 1. Flow Of Object Segmentation System

The selection frames of video sequences was considered as a still image treated as a keyframe. In order to create a keyframe, a new approach to the closed-form solution [10] with scribble-based technique was applied. After the separation object from the background, the subtraction operation between the current and subsequent frame was performed by applying a background subtraction algorithm. The value of difference subtraction process was used as a label for the object separation process in subsequent frames. This operation is performed repeatedly until the end frame of the video sequences.

5. EXPERIMENT AND EVALUATION

In this experiment, we evaluated our proposed algorithms using standard test video sequences obtained from the UCF Sports Action Data Set (i.e. riding horse, lifting, skateboarding and foreman), 30 frames respectively. Initial stages, the first frame of the video sequence was considered as a still image (shown in Fig.1 (a)). In our experiments, the selected frame considered as a keyframe was a frame which had intact object of the entire video sequence. Semi-automatic technique was performed by giving scribble (as a label) to distinguish areas of foreground and background (illustrated in Fig.1 (b)). Scribble image used background brush (black scribble in our examples) to show the background pixels ($\alpha = 0$) and foreground brush (white scribble) to show foreground pixels ($\alpha = 1$). In order to separate the foreground object from the whole image, a matting technique [10][12] was applied as depicted in (Fig. 2(c)).



(a)

(b)

(c)

Fig. 2. (A). Still Image, (B). Scribble Image, (C). Segmentation Result

Furthermore, to extract object on the subsequent frames, we applied background subtraction technique Eq. (16) to obtain difference binary value between current and previous frame. Binary value of 1 was assumed as label for foreground and 0 for the background. Later on, the value is then used to replace the role of scribble and used in the process of matting in subsequent frames. The example of segmentation results of the video data is shown in Fig. 5. To measure the accuracy of object segmentation, we evaluated using the Mean Square Error (MSE) denoted as follows:

$$MSE = \frac{\left(\sum_{i=1} \sum_{j=1} \left[Grd.Truth_{(i,j)} - Seg.Obj_{(i,j)} \right]^2 \right)}{MN} \quad (24)$$

Journal of Theoretical and Applied Information Technology
20th February 2015. Vol.72 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$Grd.Truth$ was the ground truth image resulted from manual segmentation. Whereas $Seg.Obj$ was the object produced by the segmentation process. In this experiment, the MSE calculations were performed around the frames of the video data test. The results were described in Fig. 3, and the processing time of each frame of the video data test were illustrated in Fig. 4.

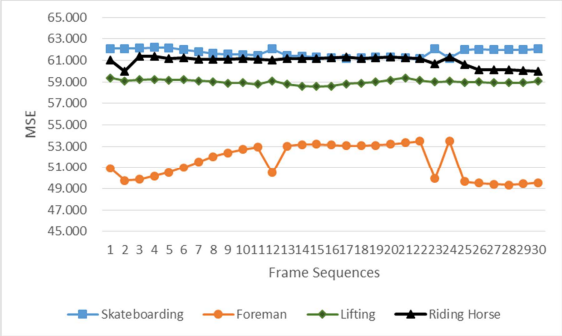


Figure 3. MSE Of Frame Sequences

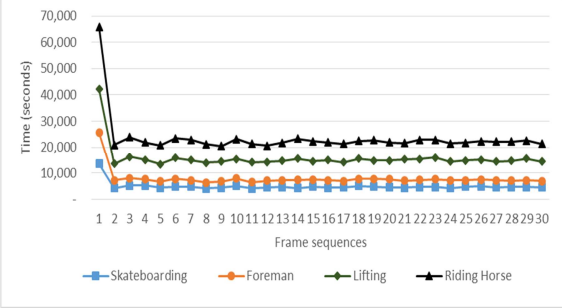


Figure 4. Processing Time Of Each Frames

6. CONCLUSION AND FUTURE PLAN

In this paper, we proposed an approach to segment video object semi-automatically. From our experiments on the 4 video datasets, 30 frame for each, the "lifting" video data indicated that segmentation accuracy of the tracking was the most stable, since it consists of most delicate object motion. While the "foreman" video data, segmentation accuracy of the tracking seemed rough on some frames, because there were objects that moved faster and all of a sudden. For future work, in order to improve the accuracy of segmentation in subsequent studies, the intensity value of video data are classified first before tracking.

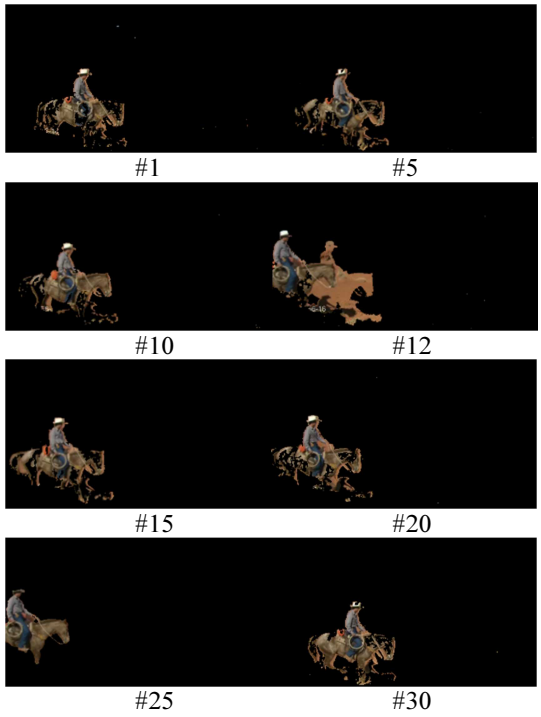


Figure 5. Object Segmented

REFERENCES

[1] A. Bovic, The Hand Book of Image and Video Processing, Academic Press, 1998.

[2] H. Xu, A. Younis and M. Kabuka, "Automatic Moving Object Extraction for Content-Based Application," *IEEE Trans. Circuits System Video Technology,* vol. 14, no. 4, pp. 796-812, 2004.

[3] S. Sun, D. Haynor and Y. Kim, "Semi-automatic Video Object Segmentation using Vsnakes," *IEEE Trans. Circuit System Video Technology,* vol. 13, no. 1, pp. 75 - 82, 2003.

[4] A. Tekalp, C. Toklu and E. A. Tanju, "Semi-automatic Video Object Segmentation in The Presence of Occlusion," *IEEE Trans. Circuit System Video Technology,* vol. 10, no. 4, pp. 624 - 629, 2000.

[5] E. Şaykol, E. Güdükbay and O. Ulusoy, "A Semi-Automatic Object Extraction Tool for Querying," in *Multimedia Databases. In Proceedings of the 7th Workshop on Multimedia Information Systems (MIS '01),*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Villa Orlandi, Capri, Italy, 2001.

[6] T. Meier and K. Ngan, "Automatic Segmentation of Moving Objects for Video Plane Generation," *IEEE Trans. Circuit System Video Technology,* vol. 8, no. 5, pp. 525 - 538, 2002.

[7] T. Tsaig and A. Averbuch, "Automatic Segmentation of Moving Objects in Video Sequences : A Region Labeling Approach," *IEEE Trans. Circuit System Video Technology,* vol. 12, no. 7, pp. 597-612, 2002.

[8] A. Cavallaro, *Semantic Video Object Segmentation Tracking and Description, Ph.D Thesis,* Ecole Polytechnique Federale de Lausanne, 2002.

[9] H. Luo and A. Eleftheriadis, "Model-based Segmentation and Trackin of Head-and-Shoulder Video Object for Real Time Multimedia Service," *IEEE Trans. Multimedia,* vol. 5, no. 3, pp. 379 - 389, 2003.

[10] A. Levin, D. Lischinski and Y. Weiss, "A Closed-Form Solution to Natural Image Matting," *IEEE Transactions on Pattern Analysis And Machine Intelligence,* vol. 30, pp. 1-15, 2008.

[11] T. Porter and T. Duff, "Compositing digital images," *Computer Graphics,* vol. 18, 1984..

[12] A. Levin, A. Rav-Acha and D. Lischinski, "Spectral matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, 2008.

[13] K. Lang, "Fixing Two Weaknesses of the Spectral Method," in *Proc. Advances in Neural Information Processing Systems*, 2005.

[14] S. Yu and J. Shi, "Multiclass Spectral Clustering," in *Proc. Int'lConf. Computer Vision*, 2003.

[15] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering," in *Proc. Advances in Neural Information Processing Systems*, 2005.

[16] A. Ng, M. Jordan and W. Y., "Spectral Clustering: Analysis and an Algorithm," in *Proc. Advances in Neural Information Processing Systems*, 2001.

[17] D. Tolliver and G. Miller, "Graph Partitioning by Spectral Rounding: Applications in Image Segmentation and Clustering," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2006.*, 2006.

[18] M. Soeleman, M. Hariadi and M. Purnomo, "Adaptive Threshold for Background Subtraction in Moving Object Detection using Fuzzy C-Means Clustering," in *Tencon Int'l Conference*, Cebu, Philippines, 2012.

[19] R. C. Gonzalez and R. E. Woods, Digital Image Processing 3rd edition, Pearson Prentice Hall, 2007.

[20] A. Levin, D. Lischinski and Y. Weiss, "A Closed-Form Solution to Natural Image Matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 1-15, 2008.