Proceedings of the Fourth IIEEJ International Workshop
on Image Electronics and Visual Computing
Koh Samui, Thailand, October 7-10, 2014

SPECTRAL-BASED VIDEO OBJECT SEGMENTATION USING ALPHA MATTING AND BACKGROUND SUBTRACTION

[a]Ruri Suko Basuki, [b]Moch. Arief Soeleman, [c]Mochamad Hariadi, [d]Mauridhi Hery Purnomo
[e]Ricardus Anggi Pramunendar, [f]Auria Farantika Yogananti,

[a,b,c,d] Faculty of Industrial Technology, Dept. of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[a,b,e,f] Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia
E-mail: ruri.basuki10@mhs.ee.its.ac.id, rurisb@research.dinus.ac.id

ABSTRACT

The main objective of this study is to separate object from video sequences. To separate the object from the video data is performed by combine several algorithms. The first stages, the video data is split into several frames, and the initial frame is treated as a keyframe. Object extraction on the keyframe require human intervention, namely by giving scribble on the regions of foreground and background. Matting technique uses a closed-form solution method applied in this stage. Further, the results used as a reference for object extraction in subsequent frames. To get the labels on the next frames, background subtraction algorithm is applied, and the result is used as the input image on the next frames. So that the object extraction in subsequent frames can be performed repeatedly using matting techniques. The experimental results show that the object extraction at the initial frame shows good results. However, the accuracy decreases when the object moves too fast and suddenly.

Keyword : Segmentation, Alpha Matting, Background Subraction

1. INTRODUCTION

The advent of digital video standards such as Digital Video Broadcasting (DVB), Digital Video Broadcasting - Terrestrial (DVB - T) and Integrated Services Digital Broadcasting - Terrestrial (ISDB - T) is pushing the demand of the video editing applications (such as video segmentation and video compositing) and rapidly increased, since it plays an important role in the production of movies, news and advertising. The object-based technology can be used in various applications, such as object extraction, motion understanding, image recognition and augmented reality. Unfortunately, the process of object segmentation of video becomes a difficult job, since the video object has no semantic information. In other words, a video object segmentation is an ill-posed problem [1]. Therefore, the pulling of objects in video editing is performed with manual segmentation, since the semantic object can only be identified by humans vision that considers the video context. However, this method is not effective when it is applied to the video data with large volumes.

To overcome this problem, several algorithms related to video object segmentation have been developed. Generally, these algorithms are classified into two categories, namely the automatic object segmentation [2] and a semi-automatic object segmentation [3] [4]. Automatic segmentation is done without human intervention by considering specific characteristics such as color, texture and movement. The main problem of the automatic segmentation is the difficulty in objects separating which is semantically meaningful. Until today, there is no guarantee that the result of the automatic object segmentation will be satisfactory, since the semantically object has a lot of color, texture and movement [5] [6] [7].

For this reason, several semi-automatic segmentation methods are proposed as a combination of the automatic segmentation and manual segmentation. In principle, this approach is a technique to pull the object that involves user intervention at several stages of the segmentation process. Thus semantic information can be defined directly by the user. Furthermore, object segmentation in subsequent frames is performed using a tracking mechanism by temporal transformation. Some of methods used for tracking mechanisms has been applied in several previous studies. In a region-based method, the related areas are in accordance with the shape of semantic objects tracked by the motion, texture and color parameters [6] [7]. Weaknesses of the method are very complex tracking mechanism in maintaining relationships between regions composed of semantic object [8]. Meanwhile, the contour-based methods, such as snakes [3] will robust when track the object contours instead the whole of the object that comprising the pixels, so that these methods may not work well when the feature to be followed namely edges are not connected to each other. Whereas the model-based method apply a priori knowledge of the object shape. The drawback of this approach is not acceptable on the generic semantic video object segmentation due to the detail needs information about the object geometry [9].

In this paper, semi-automatic video segmentation framework is proposed to be applied to the general video data. Early stages in video segmentation is performed by creating a "keyframe" which is used as a reference for tracking process on the subsequent frames. Hereafter, the object segmentation on the subsequent frames is done by using the background substraction algorithm.

2. KEYFRAME CONSTRUCTION

The first stage of a video segmentation process is done by constructing the initial frame of the sequence scene which becomes a key frame. Since the object have no the semantic information, scribble is used as a label to distinguish areas that represent foreground and background (white color for foreground and black color for background). Next, the object is extracted with matting techniques.

A. General Compositing Equation

Alpha channel [10][11][12] is used to control the linear interpolation in the foreground and background which are depicted in matting algorithm by assuming that each pixel in the input image I_i is a linear combination of the color 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 compositing equation Eq.(1) of each pixel is assumed to be a convex combination of layers K image which 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 is determined by the vector K of α^k which is a component of image matting.

B. Spectral Analysis

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

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

while D is matrix degree from graph.

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

with diagonal matrix

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

$D_{(i,j)}$is stuffed with degree information of each vertex (node) with D for G as rectangular matrix size $n\ x\ n$ depicted. So L is a symmetric positive semi-definite matrix with eigenvector which is able to capture a lot of image structure. Furthermore, the image is composed of some different clusters or connected components which can be captured by the affinity matrix A. Subset C in image pixel is the connected component of image $A_{(i,j)} = 0$ for each (i, j) so $i \in C$ and $j \notin C$, so there is no subset C that can fulfill this property. $m^C$is 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 consists of connected components of $K, C_1, \ldots, C_K$ to $\{1, \ldots, N\} = \bigcup_{k=1}^{K} C_k$ with C_k disjoint path on the pixel, then the m^C represents 0-eigenvector (eigenvector with eigenvalue 0) from L . Since the rotation of matrix R in size $K \times K$, and vector $[m^{C_1}, \ldots, m^{C_K}]R$ is the null space base on L , then the indicator vector $m^{C_1}, \ldots, m^{C_K}$ resulted from eigenvector calculation on L is only reaching the rotation. "Spectral Rounding" which is a component extraction with the smallest eigenvector becomes a concern in some studies [13] [14] [15] [16] [17]. K-Means algorithm is a simple approach used for clustering the image pixels [13], while the perturbation analysis algorithm is 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] is 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.

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

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

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

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

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

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

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

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

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

2) Linear Transformation

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

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

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

The robust measurement value in matting component [12] is determined by $|\alpha_i^k|^\gamma + |1-\alpha_i^k|^\gamma$, If $0 < \gamma < 1$, thus, the value of $\gamma = 0{,}9$. Because the cost function Eq.(11) is not convex, the initialization process determine the result of Newton process. Therefore, K-means algorithm can be used 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) is then added. Thus it gives solution for Eq.(11).

3) Grouping Component

The complete results of matte extraction of the foreground object are 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 is perform by $\alpha^T L\alpha$ when the smallest eigenvector is not equal to zero, in which L is the matting Laplacian. The first calculation of correlation among matting component and L deviation in matrix $\Phi K \times K$ is defined as

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

then, matte cost is calculated as

$$J\left(\alpha\right) = b^T \Phi b \qquad (15)$$

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

3. TRACKING MECHANISM

A. Background Subtraction

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

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

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

B. Otsu Threshold

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

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

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

   where n_i is the number of the pixels at each intensity, and MN is the number of n_i starting from n_0 to $n_L - 1$
2. Compute the cumulative number of $p_1(k)$ for $k = 0,1,2...,L-1$.

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

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

$$m(k) = \sum_{i=0}^{k} ip_i \qquad (19)$$

4. Calculate the average global intensity m_G by using ;

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

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

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

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 more than one value of $k*$, then the threshold value is determined from the average value of $k*$.
7. Determine the size of the separation $\eta*$ with $k = k*$

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

while,

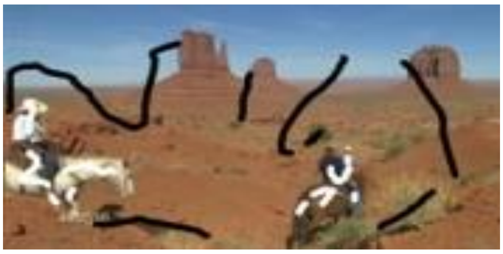
$$\sigma_B^2 = \sum_{i=0}^{L-1} (1-m_G)^2 p_i \qquad (23)$$

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

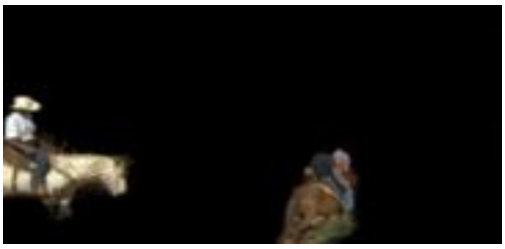
4. EXPERIMENT AND EVALUATION

In this experiment, we evaluate 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 is considered as a still image (shown in Fig.1 (a)). In our experiments, the selected frame as a keyframe is a frame that has intact object of the entire video sequence. Semi-automatic technique is performed by giving scribble (as a label) to distinguish between areas of foreground and background (illustrated in Fig.1 (b)). Scribble image uses background brush (black scribble in our examples) to show the background pixels (α = 0) and foreground brush (white scribble) to show foreground pixels (α = 1). In order to separate the foreground object from the whole image, a matting technique [10][12] is applied such as depicted in Fig. 1(c).

(a)



(b)



(c)

Fig. 1. (a). Still image, (b). Scribble image, (c). Segmentation result

Furthermore, to extract object on the subsequent frames, we used background subtraction technique Eq.(16) to get difference of the binary value between current frame and previous frame. Binary value of 1 is assumed as label for foreground and 0 for the background. This value is then used to replace the role of scribble and used in the process of matting in subsequent frames. The example of segementation results of the video data is shown in (figure 3). To measure the accuracy of object segmentation, we evaluate using the Mean Square Error (MSE) are denoted as follows:

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

$Grd.Truth$ is the ground truth image resulted from manual segmentation. Whereas $Seg.Obj$ is the object that produced by the segmentation process. In this experiment, the MSE calculations performed around the frames of the video data test. The results are described in (Fig. 2).

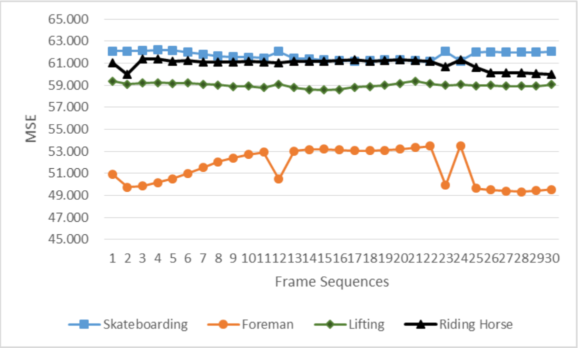


Figure 2. MSE of Frame Sequences

5. 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 each 30 frame, the “lifting” video data indicate that segmentation accuracy of the tracking is the most stable, since consist of most delicate object motion. While the "foreman" video data, segmentation accuracy of the tracking looks rough on some frames, because there are objects that move 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.

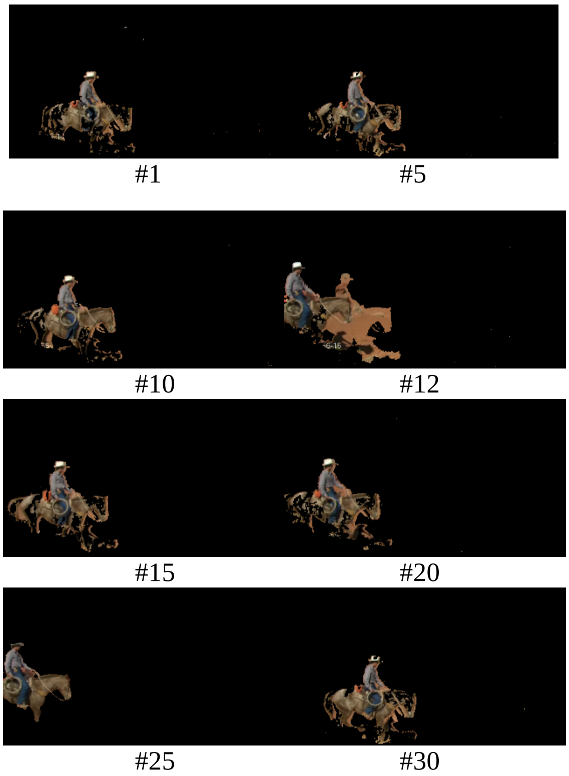


Figure 3. Object segmented

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), 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.