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

Ruri Suko Basuki, Moch. Arief Soeleman, Mochamad Hariadi, Mauridhi Hery Purnomo, Ricardus Anggi Pramunendar, Auria Farantika Yogananti

Faculty of Industrial Technology, Dept. of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
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 Subtraction

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 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 $B_i$.

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

(1)

based compositing equation Eq.(1) of each pixel is assumed to be a convex combination of layers $K$ image which denoted as

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

(2)

the fractional contribution of each layer observed in each pixel is determined by the vector $K$ of $\alpha^k$ 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 $N \times N$ is assumed as $A_{(i,j)} = e^{-d_{ij} / \alpha}$ and $d_{ij}$. In which $d_{ij}$ is the space among pixels (e.g. color and geodesic space), which is defined as

$$L = D - A$$  

(3)

while $D$ is matrix degree from graph.

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

(4)

with diagonal matrix

$$D_{(i,j)} = \sum_j A(i,j),$$

where $d_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases}$  

(5)

$D_{(i,j)}$ is stuffed with degree information of each vertex (node) with $D$ for $G$ as rectangular matrix size $n \times n$ 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

$$m^C_i = \begin{cases} 1 & i \in C \\ 0 & i \notin C \end{cases}$$  

(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 $[m^C_1, \ldots, m^C_K] \in R$ 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, $\alpha$ in $w$ is expressed as a linear combination of color channels.
\[ \forall i \in w \quad a_i = a^R I_i^R + a^G I_i^G + a^B I_i^B + b \]  
\[ J(a, b, c) = \sum_{i,j,k,l \in \Omega} \left( \frac{a_i - a_j}{\delta_{ij}} + \frac{a_j}{\delta_{ij}} + \frac{a_k - a_l}{\delta_{kl}} + \frac{a_l}{\delta_{kl}} \right)^2 + \varepsilon \| a_i \|^2 \]  
\[ J(\alpha) = \alpha^T L \alpha, \]  
\[ \alpha^k = EE^T m^c \]  
\[ J(\alpha) = \alpha^T L \alpha, \]  
\[ \alpha = \alpha^k + \ldots + \alpha^{k'} \]  
\[ BF(x, y, n) = \begin{cases} \| I(x, y, n) - I(x, y, n) \| \geq \alpha, \\ 0, \text{otherwise} \end{cases} \]  
\[ \alpha = \alpha^k + \ldots + \alpha^{k'} \]
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, with \( i = 0,1,2...L-1 \)
\[
p_i = \frac{n_i}{MN}
\]
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 \( p_i(k) \) for \( k = 0,1,2...,L-1 \).
\[
P_i(k) = \sum_{i=0}^{k} p_i \quad (18)
\]

3. Count the cumulative average of \( m(k) \) for \( k = 0,1,2...,L-1 \).
\[
m(k) = \sum_{i=0}^{k} \frac{p_i}{k} \quad (19)
\]

4. Calculate the average global intensity \( m_G \) by using :
\[
m_G = \sum_{i=0}^{L-1} \frac{p_i}{MN} \quad (20)
\]

5. Compute the variance between classes, \( \sigma^2_{g}(k) \) for \( k = 0,1,2...,L-1 \).
\[
\sigma^2_{g}(k) = \frac{[m_G P_i(k) - m(k)]^2}{P_i(k)[1 - P_i(k)]} \quad (21)
\]

6. Select a threshold value of the \( k^* \) if the index value of the maximum variance between classes (\( \sigma^2_{g} \Rightarrow \max(k) \)), 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 \( k = k^* \)
\[
\eta(k) = \frac{\sigma^2_{g}(k)}{\sigma^2_{a}} \quad (22)
\]

while,
\[
\sigma^2_{a} = \sum_{i=0}^{k} (1 - m_i)^2 p_i \quad (23)
\]

Note : the value of K is obtained when \( \sigma^2_{g}(k) \) 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 (\( a = 0 \)) and foreground brush (white scribble) to show foreground pixels (\( a = 1 \)). 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).
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 segmentation 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:

\[
MSE = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (Grd\_Truth(i,j) - Seg\_Obj(i,j))^2}{MN}
\]

(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).

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.

References


