Abstract

Growing interest arises in segmentation for object-based video clips since the development of MPEG-4 standard [1]. The moving object extraction can also be applied to the object-based videoconference, surveillance, and so on. The difficulties of moving object segmentation are that physical objects are normally not homogeneous with respect to low-level features and it's usually tough to segment them efficiently. The previous related researches are only operated with a static camera and in a stationary background. In this paper, we propose a robust and fast segmentation algorithm and a reliable tracking strategy without knowing the shape of the object in advance. The system can segment the foreground from the background and track the moving object with an active (pan-tilt zoom) camera such that the moving object always stays around the center of images. Especially, the system can work in an unrestricted environment without the need for special purpose hardware. The proposed segmentation algorithm is based on the background subtraction, morphological operations, region growing, adaptive mechanism, template matching, and some innovative operations. The system can segment a moving object at 15 frames per second over a 176 x 144 pixel image.

Keywords: segmentation, mosaic image, active camera, tracking.

1. Introduction

Recently, there has been growing interest in segmentation for object-based video coding. This is mainly due to the development of MPEG-4 standard, which is set to become the new video coding standard for multimedia communication [1]. MPEG-4 will provide many new features to cater for future multimedia applications and to enable interactivity with objects in video sequences. It will support an object-based representation of audio-visual objects that allows the access of objects in the compressed domain, selective decoding of such objects and their manipulation. The MPEG-4 video coding standard provides an object-based bitstream to assist this functionality. Each object is coded into a separate object bitstream layer.

The moving object extraction can be applied to the popular video conferencing environments. In general multipoint video conferencing environments, each participant joins the conference together from separate places. Each participant has its own background, and the conferencing environment looks not concordant at all. This kind of video conferencing environment is quite different from traditional conference. To overcome this disadvantage, we can create a virtual environment and put the segmented objects in it, so the object-based videoconference will look more realistic. The technique can also apply to home security, which is to detect and segment the intruding objects in the house.

Researchers in the object segmentation field have proposed many methods and algorithms. All these schemes can be separated into two categories. One is only operated with a static camera and in a stationary background [2-3]. The other is operated with an active camera but is not efficient, since its strategy is to store all the background images for the background model database [4].

In the paper, our aim is to segment the foreground from the background and track the moving object with an active (pan-tilt zoom) camera such that the moving object always stays around the center of images. Hence, the system requires a robust segmentation algorithm and a reliable tracking strategy.

This paper is organized as follows. Section 2 describes the steps to construct mosaic images and how to utilize the mosaic image in our system. In section 3, we illustrate how to segment the foreground from the background and tracks the moving object with an active (pan-tilt zoom) camera such that the moving object always stays around the center of images. And the experiment results are shown in section 4. Finally, the conclusion and future work is provided in section 5.

2. Construction of Mosaic Images

In previous related works, the moving object segmentation is only operated with a static camera
and in a stationary background [2-3] and many of its image analysis algorithms would not be applied easily to images taken from an active camera. The static camera restricts the object in a limited view. And their applications are not satisfying. Some related researches also used an active camera to segment and track the moving object but their applications have some limitation, i.e. they cannot segment the moving object in arbitrary pan angle. Hence, their applications must store all views in fixed pan and tilt angles. They are hence not efficient and not flexible.

In our system, an active camera (pan-tilt zoom camera) is used and enables the moving object (foreground) to be extracted from the background in arbitrary pan angles. The object can move in a wide range of background and the moving object extraction method is based on background subtraction, which will be described in detail in section 3. One of the key benefits of a mosaic image is as a means of efficient scene presentation. Hence, the mosaic image of the non-stationary background is used as the reference of background subtraction and it would reduce the memory usage in a practical implementation.

### 2.1 Image Alignment

To construct the panoramic mosaic, the first step is the alignment of images, i.e. estimate the global motion between the consecutive images, as shown in Figure 1. 2-parameter translation motion model (1) is used to align images. The basic idea of global motion is determined by minimizing the specified error function (2) in the overlapping region.

\[
\begin{align*}
  x' &= x + a \\
  y' &= y + b 
\end{align*} \quad \ldots (1)
\]

\[
E(a,b) = \sum_{(i,j) \in S} \left[ (I_i(x+a,y+b) - I_o(x,y))^2 \right] \quad \ldots (2)
\]

\[S : \text{overlapping region}\]

### 2.2 Image Integration

Once the global motion of the consecutive frames is obtained, they can be integrated into a panoramic mosaic. In general, the pixel blending method is utilized to reduce the discontinuities in color and in luminance. But our goal is to construct more accurate background model.

We design an exponential weighting function to blend overlapping regions, as shown in equation (3) and Figure 2,

\[
 w_{x,y} = e^{-\frac{(x-x_c)^2 + (y-y_c)^2}{c_x c_y}} \quad \ldots (3)
\]

where \( w_{x,y} \) is the weight at the \((x,y)\) position in the image and the \((x_c,y_c)\) is the central position in the image. The mosaic image integrated with exponential weighting function is more seamless than that with other blending methods. Another reason is that the center region in the image is more suitable for background subtraction and the weights around the central regions should be larger than the boundary regions.

The blended pixel value of the mosaic image is computed as follows:

\[
M_{x_m,y_m} = (1-w) \cdot M_{x_m,y_m} + w \cdot f_{x_m,y_m}
\]

\[
w = \frac{w_{f}(x, y) + w_{m}(x_m, y_m)}{w_{f}(x, y) + w_{m}(x_m, y_m)}
\]

\[
w_{m}(x_m, y_m) = \left(1 - w\right) \cdot w_{m}(x_m, y_m) + w \cdot w_{f}(x, y)
\]

where \( M_{x_m,y_m} \) is the pixel value in the mosaic image, \( f_{x_m,y_m} \) is the pixel value of the image integrated into the mosaic, \( w_{m}(x_m, y_m) \) is the weight at \((x_m, y_m)\) in
the mosaic, and $w_r(x,y)$ is the weight at $(x,y)$ in the image.

3. Moving Object Extraction and Tracking

3.1 Overview of the proposed system

To overcome the inconvenience and limitation caused by a static camera, we propose a real-time moving object extraction system using mosaic technique and tracking with a active camera. Figure 3 depicts the conceptual diagram of our proposed system. Our aim is to segment the foreground object from the background scene and track the moving object with an active (pan-tilt zoom) camera such that the moving object always stays around the center of images. Hence, the system requires a robust segmentation algorithm and a reliable tracking strategy.

First of all, we automatically construct a panoramic mosaic image of the background to be used as the reference in the segmentation process. Then, the camera continuously captures color images and conveys images to the detector module. The detector module monitors any change in the scene and activates the segmentation module when an intruding object is detected. As the segmentation mechanism is activated, foreground is extracted from background and it is utilized as the basis to control the active camera to track the moving object. In addition, the extracted background is utilized to update the corresponding background model. The update mechanism is extremely effective in improving the segmentation result.

3.2 Object Segmentation Based on Background Subtraction Method

In general, background subtraction is a typical method to discriminate moving object from the static background scene [3].

We take the mosaic image as the reference background database. The panoramic mosaic image is constructed from 15 views taken from equally spaced pan and tilt angle positions. And each view needs to be analyzed over several seconds of video. Each pixel value in each view may change over a period of time due to camera noises and illumination fluctuation by light sources. Each view is modeled by representing each pixel with two parameters, mean and standard deviation, during the training period, and they are calculated as follows:

$$
\begin{align*}
\mu(\tilde{p}) &= \frac{1}{N} \sum_{i=1}^{N} R_i(\tilde{p}) = (Y_{\tilde{p}}, U_{\tilde{p}}, V_{\tilde{p}}) = (M_{\mu}(\tilde{p}), M_{\epsilon}(\tilde{p}), M_{\sigma}(\tilde{p})) : \text{mean of } Y, U, V \\
\text{std}(\tilde{p}) &= \sqrt{\frac{1}{N} \sum_{i=1}^{N} R_i^2(\tilde{p}) - \mu^2(\tilde{p})} \\
&= (\text{std}_Y(\tilde{p}), \text{std}_U(\tilde{p}), \text{std}_V(\tilde{p}))
\end{align*}
$$

where $\tilde{p}$ presents the index of pixels in the frame. $R(\tilde{p})$ is the luminance and chrominance values of the pixel $\tilde{p}$ in the current frame. $\mu(\tilde{p})$ and std$(\tilde{p})$ represent the mean and the standard deviation of the luminance and chrominance values of the pixel $\tilde{p}$ during the $N$ analyzed frames in a view. Then, fuse different views into a mosaic image. Therefore, each pixel in the mosaic image has two parameters: the mean $M(\tilde{p})$ and the standard deviation $M_{ad}(\tilde{p})$.

The criterion to classify pixels is described as follows:

$$
\begin{align*}
&\text{if } |C(\tilde{p}) - M'(\tilde{p})| > k \times M'_{ad}(\tilde{p}) \\
&\quad \tilde{p} \in \text{foreground} \\
&\text{else} \\
&\quad \tilde{p} \in \text{background}
\end{align*}
$$

where $C(\tilde{p})$ is captured from the camera at position $\tilde{p}$ and $M'(\tilde{p})$, $M'_{ad}(\tilde{p})$ are the mean and the standard deviation of the sub-view in the mosaic image, respectively.

To get the sub-view in the mosaic image as the corresponding background model, there are five steps:

1. get the camera pan and tilt angle position through an RS-232 port and use these data to interpolate roughly the sub-view in the mosaic image.
2. remove the foreground in the current frame by subtraction method.
3. use the remaining background in the current frame and find the more accurate sub-view in the mosaic image.
4. update the corresponding background model.
5. iteratively repeat Step 2 ~ 4 until the corresponding background model is stable.
After the subtraction, only non-stationary object is left, but background subtraction just roughly classifies pixels of background and foreground. There are some necessary operations to improve the resulted images. The whole procedure of the proposed object discrimination is shown in Figure 4.

To mitigate the distortion of the corresponding background model, we assume that there is one pixel displacement between the current frame and the background model. Then, each pixel $p_v$ in the current frame, $C(p_v)$, is classified as either a background or a foreground class by considering not only the corresponding pixel in the background model but also the eight neighbors of the corresponding pixel. The criterion is described as follows:

$$
\begin{align*}
\bar{p}_1, \bar{p}_2, \bar{p}_3; \\
\text{count} = 0; \\
\text{for}(i = 1; i < 9; i++) \\
\quad \text{if } (\|C(p_v) - M'(p_v)\| > k \times M^{'\infty}(p_v)) \quad \text{count}++; \\
\quad \text{if (count} = 3) \\
\quad \quad \bar{p} \in \text{foreground} \\
\quad \text{else} \\
\quad \quad \bar{p} \in \text{background}
\end{align*}
$$

A rough alpha mask is obtained (1: foreground; 0: background). Then, the median filter is followed in the procedure. It smooths the alpha mask and is thus useful in reducing noise.

Next, morphological opening operation and closing operation, eight-neighbor dilation and eight-neighbor erosion, are used for removing noises in our system. After morphological operations, two-pixel thick noises in the frame would be eliminated.

At the final step of object discrimination, a region growing method is used to reduce large noise in the alpha plane. It preserves the region of interest in the alpha plane and erases the other regions, which are considered background noises. In the region growing procedure, the seed point in the interested region must be acquired first. We propose two ways to get the seed point in the system. One way is to use “integral projection” [5] with the alpha plane to get the seed point (described in detail later). And the other way is to calculate the barycenter of the skin-color region in the frame as the seed point in the region growing procedure because the human head is the region of interest in our system. We adopt the “integral projection” method to get the seed point when the skin-color region is less than a threshold and otherwise adopt the barycenter of the skin-color region when the skin-color region exceeds a threshold.

3.2.1 Integral Projection

The procedure of region growing method which obtains a seed point from “integral projection” is summarized as follows:

(1) project the alpha plane $F(x, y)$ using horizontal integral projection and vertical integral projection as shown in Figure 5 (b), (c):

$$
\begin{align*}
\text{HistH}(y) &= \sum_x F(x, y) \\
\text{HistV}(x) &= \sum_y F(x, y)
\end{align*}
$$

where $\text{HistH}(y)$ and $\text{HistV}(x)$ are the horizontal integral projection and vertical integral projection respectively.

(2) find index of each maximum value in $\text{HistH}(y)$ and $\text{HistV}(x)$.

$$
\begin{align*}
\text{SeedPoint}_y &= \max_x \{x | \text{HistV}(x)\}; \\
\text{SeedPoint}_y &= \max_y \{y | \text{HistH}(y)\};
\end{align*}
$$

(3) preserve the region, which contains the seed point, and erase the other regions.
3.2.2 Updating Background

The extracted background model in the current frame is utilized to adapt the corresponding background model, \( M'(p) \), obtained from the mosaic image. The update mechanism is as follows:

\[
\begin{align*}
\text{If } C(p)_x \in \text{foreground} & \quad M'(p)_x = M'(p)_{x+1} \\
\text{else} & \quad M'(p)_x = (1-\eta)M'(p)_{x+1} + \eta C(p)_x
\end{align*}
\]

The update mechanism is extremely effective in improving the segmentation result and it can also resist the slow changes in lighting conditions in the images.

3.3 Tracking of Moving Object and Dynamic Segmentation

In section 3.2.2, the proposed segmentation method is robust when the active camera keeps stationary. But it is another story under the condition that camera moves. The procedure, as shown in Figure 4, can only segment a rough foreground when the camera moves, since the corresponding background model is inaccurate. And the background update mechanism fails because the background in the current frame is changing. And finding the corresponding background model in the mosaic image iteratively causes delay in the system when the active camera moves.

There are some modifications in segmentation while the camera moves, as shown in Figure 5. We get the camera pan and tilt angle position through an RS-232 port and only use these data to interpolate roughly the sub-view in the mosaic image without finding the corresponding background model iteratively. It can save a lot of computation. To refine the roughly segmented result, a template matching method is adopted. Each current segmented object \( C'(p)_x \) is matched to the previous segmented object \( C(p)_{x-1} \) for reducing the noises in \( C(p)_x \).

We also propose a reliable tracking strategy with the active camera. The active camera is controlled through an RS-232 port during the tracking process. Tracking the moving object is a challenging task because searching for features and getting the correspondence between consecutive frames is not easy. To reduce the complexity and computation of the tracking, only the barycenter of skin-color region in the segmented object is selected. In this way, the proposed strategy enables robust tracking without knowing the shape of the object in advance.

3.3.1 Template Matching

Template matching is a natural approach to pattern classification. Here, template matching assumes that the moving object is composed of moving blocks. Therefore, we determine the correspondences between the current segmented object \( C'(p)_x \) and the previous segmented object \( C(p)_{x-1} \) for reducing the noises in \( C(p)_x \). The correspondence problem can be formulated as a backward block-matching motion estimation problem, which is commonly employed in predictive video compression.

A simple, robust and efficient block-matching motion estimation algorithm, called diamond search, is adopted [6]. Template matching criterion is described as follows:

\[
\begin{align*}
\text{if } C'(p)_x \in \text{foreground} & \quad \text{find } C(p)_{x-1} \cdot \text{which corresponds to } C'(p)_x \\
\text{else} & \quad C'(p)_x \in \text{background}
\end{align*}
\]

3.3.2 Skin-Color Detection

Statistical experiments indicate that although skin-colors of different people appear to vary over a wide range, they differ less in chrominance than in brightness, specifically the skin-colors form a very compact area in the normalized \((r,g)\) space [5].

\[
\hat{z} = (r,g) = \left( \frac{R}{R+G+B} - \frac{G}{R+G+B} \right)
\]

Therefore, a Gaussian model \( N(m, \Sigma^2) \) can be utilized for representing the skin-color model,

\[
m = (r, g) = \left( \frac{1}{L} \sum_{l=0}^{L} r_{l}, \frac{1}{L} \sum_{l=0}^{L} g_{l} \right)
\]

\[
\Sigma = \begin{bmatrix}
\sigma_{rr} & \sigma_{rg} \\
\sigma_{gr} & \sigma_{gg}
\end{bmatrix}
\]
As shown in equation (4), this allows us to construct a Gaussian probability image of skin color and detect skin color regions.

\[
\hat{p} \in \text{skin-color class if } f(\hat{p}) > TH_{\text{skin}}
\]

where \( f(p) = \frac{1}{(2\pi)^d} \exp\left(-\frac{1}{2} \sum (p-m)^T \Sigma^{-1}(p-m)\right) \) \( \ldots (4) \)

In our system, the skin-color information is so useful that the barycenter of the skin-color region is utilized as the seed point in the region growing process and as the key feature for detection and tracking.

3.3.3 Tracking of Moving Object

In tracking mode, the active camera is controlled to pull the moving object inside “CR” according to the feature point. A linear trajectory model is used to predict the feature’s future position. Suppose the feature position is at \( x(t) \) at time \( t \). We can predict the feature position at time \( t+1, \bar{x}(t+1) \), by assuming constant velocity \( v \) which is obtained from previous two frames.

\[
v(t) = x(t) - x(t-1) \\
\bar{x}(t+1) = x(t) + v(t)
\]

Once the feature point leaves the “CR”, the active camera is pulled. And the pan speed of the camera is determined by the distance between predicted feature point, \( \bar{x}(t+1) \), and “O”.

\[
\text{set } \text{PanSpeed} = c \frac{\bar{x}(t+1) - O}{T}
\]

where \( c \) is a scaling factor and \( T \) is the temporal interval between consecutive frames.

4. Experimental Results

Our system has been implemented on a AMD Duron 700 Mhz system under Microsoft Windows 98 with a Winnov capture card and a Canon VCC-3 CCD pan tilt zoom camera, and can segment a moving object at 15 frames per second over a QCIF image.

4.1 Object Extraction

Figure 8 shows the simulation results of the proposed object extraction method and shows that that the proposed method can segment the foreground from the background well in real time and significantly reduce the noise around the foreground region.

4.2 Detecting the Skin Color

Figure 9 (c) illustrates the result of the skin color. Discriminating the skin color in the foreground region only can reduce computation and avoid the disturbance of skin color in the background. Therefore, the skin color information is useful for tracking and region growing in our system.

4.2 Tracking the Feature

Figure 9 (a) image captured from the camera (b) the segmented foreground (c) the skin-color region in the foreground after median filter.
The barycenter of the skin-color region is utilized as the tracking feature since it varies smoothly and it is less complex. In our system, a linear trajectory model is used to predict the feature’s future position which is obtained from previous two frames. Figure 10 illustrate that the predicted position is close to the real position, so we believe that these assumptions are a fair approximation.

4.3 Template Matching

Although the sub-view in the mosaic image does not exactly match the background in the current frame captured from the camera, we can segment a rough foreground and reduce the memory usage by the mosaic image.

The proposed segmentation algorithm is based on the background subtraction, morphological operations, region growing, adaptive mechanism, template matching, and some self-defined operations. The system can segment a moving object at 20 frames per second over a 176 x 144 pixel image.

In the tracking mode, the barycenter of the skin-color region in the foreground is utilized as the feature for detection and tracking with an active camera. The feature is reliable and varies smoothly so as to track without a complex temporal filter such as a Kalman filter.

From the experimental results, we prove that the method works well and our algorithm can be a front-end function of the object-based applications in real time, i.e. MPEG-4, very low bit rate video (VLBV), home security, videophone and video conferencing.

In our system, the mosaic image must be constructed without disturbances of the background but this restriction is inconvenient. In the future work, the reference background mosaic image could be obtained online [7] and adapt the background gradually. Besides, the foreground boundary is not smooth due to the morphological operations, so the refinement of foreground boundary may be added into our system.

References


