REALTIME OBJECT EXTRACTION AND TRACKING WITH AN ACTIVE CAMERA USING IMAGE MOSAICS

Chih-Ming Wang*, Chia-Wen Lin**, Yao-Jen Chang*, and Yung-Chang Chen**

Department of Electrical Engineering
National Tsing Hua University, Hsinchu 300, Taiwan
**Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan

ABSTRACT
Moving object extraction plays a key role in applications such as object-based videoconference, surveillance, and so on. The difficulties of moving object segmentation lie in the fact that physical objects are normally not homogeneous with respect to low-level features and it’s usually tough to segment them accurately and efficiently. Object segmentation based on prestored background information has proved to be effective and efficient in several applications such as videophone, video conferencing, and surveillance, etc. The previous works, however, were mainly concentrated on object segmentation 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 proposed system can segment the foreground from the background and track the moving object with an active (pan-tilt) camera such that the moving object always stays around the center of images. The processing speed is about 20 QCIF (176x144) frames per second.

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 has become a new video coding standard for multimedia communication. MPEG-4 provides many new features to cater for future multimedia applications and to enable object interactivity in video sequences. It also supports 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. Moving object extraction plays a key role in such kind of applications.

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 be also applied to home/officewon surveillance to detect and track the moving object. In this paper, our aim is to segment the foreground from the background and track the moving object with an active 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.

2. VIDEO OBJECT EXTRACTION AND TRACKING WITH AN ACTIVE CAMERA

Fig. 1 depicts the conceptual diagram of our proposed scheme, which is aimed at segmenting the foreground object from the background scene and tracking the moving object with an active camera such that the moving object always stays around the center of images. Prior to performing segmentation and tracking, a number of background images with equally spaced pan and
tile angles are captured and analyzed. A panoramic mosaic image of the background is then automatically constructed from those background images as the reference background model in the segmentation process. After the mosaic image is constructed, the live video captured from the camera is fed into the detection module. The detection module monitors scene changes and activates the segmentation module when an intruding object is detected. As the segmentation mechanism is activated, the foreground object is extracted from the background and the extracted foreground is utilized as the basis to control the active camera to track the moving object. In addition, the separated background is utilized to update the corresponding background model to improve the segmentation result. In our system, an active camera is used which enables the moving object to be extracted from the background with arbitrary pan and tilt angles, and the object can move in a wide range of background.

\begin{equation}
\alpha = \frac{w_1(x,y)}{w_0(x,y) + w_1(x,y)}
\end{equation}

where \( M_{x,y} \) is the original pixel value in the mosaic image, \( M'_{x,y} \) is the updated pixel value \( f_{x,y} \) is the pixel value of the incoming image integrated into the mosaic, \( w_0(x,y) \) is the weight at \((x,y)\) in the mosaic, and \( w_1(x,y) \) is the weight at \((x,y)\) in the incoming image.

**A. Constructing the Mosaic Reference Background**

The first step to constructing the panoramic mosaic is the alignment of images, i.e., estimate the global motion between the consecutive images. The 3-parameter translation motion model stated in Eq. (1) is used for images alignment.

\begin{align}
x' &= x + a \\
y' &= y + b
\end{align}

The global motion vector \((a^*,b^*)\) is determined by minimizing a specified error function in the overlapping region as shown in Eq. (2).

\begin{equation}
(a^*,b^*) = \arg\min_{(a,b)} \sum_{(x,y) \in S} [I_0((x+a,y+b)) - I_1(x,y)]^2
\end{equation}

where \( I_0 \) and \( I_1 \) are two consecutive background images; \( S \) stands for the overlapping region of \( I_0 \) and \( I_1 \).

Once the global motion of the consecutive frames is obtained, these frames can be integrated into a panoramic mosaic. Pixel blending is used to reduce the discontinuities in color and in luminance. In addition to constructing the panoramic mosaic, our goal is to construct an accurate background model for object extraction. We propose to use an exponential weighting function to blend the overlapping regions, as shown in Eq. (3).

\begin{equation}
w_{x,y} = e^{-\frac{(x-x_0)^2}{2C_x} - \frac{(y-y_0)^2}{2C_y}}
\end{equation}

where \( w_{x,y} \) is the weight at \((x,y)\) position in the image to be blended and \((x_0,y_0)\) represents the central position of the image. The mosaic image blended with exponential weighting function is more seamless than that with linear blending functions. 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:

\begin{equation}
M'_{x,y} = (1-\alpha) \cdot M_{x,y} + \alpha \cdot f_{x,y}
\end{equation}

**B. Object Classification Using Background Subtraction**

Fig. 2 shows an example of a sub-view in the mosaic image. The panoramic mosaic image is constructed from 15 views taken from equally spaced pan and tilt angle positions. In our method, the mosaic image is to provide an initial rough reference background model for the background subtraction method, and the background model is then refined gradually according to the segmentation result.

**Fig. 3. Procedure of the proposed object extraction and tracking with an active camera.**
For finding the corresponding background model in the mosaic image will cause delay in the system when the active camera moves.

Five steps are used to locate the sub-view in the mosaic image as the corresponding background model:
1. get the camera pan and tilt angle position from the active camera and use these parameters to roughly locate the sub-view in the mosaic image,
2. segment and remove the foreground in the current frame by the background subtraction method,
3. use the remaining background in the current frame to find the more accurate sub-view in the mosaic image,
4. update the corresponding background model,
5. iteratively repeat Steps 2 ~ 4 until the corresponding background model is stable.

C. Post-processing
Background subtraction can roughly classify pixels of background and foreground, but the resultant segmentation result may be still quite noisy due to camera noises, illumination variations, and inappropriate threshold selections. Some post-filtering operations are subsequently performed to refine the segmentation result. To mitigate the distortion of the corresponding background model, the binary segmentation result is median filtered with a 3x3 mask, then is further refined with a morphological filter.

At the final step of object discrimination, a region growing procedure, the seed point in the interested region has to be chosen first. We propose two ways to select the seed point in the system. One is to use “integral projection” proposed in [7] with the alpha plane to obtain the seed point. The other is to calculate the centroid of the skin-color region in the frame as the seed point because the human face is usually the region of interest in our system. In our method, the integral projection method is adopted to obtain the seed point in the interested region.

The separated background in the incoming frame is utilized to update the intensity mean of corresponding background model, , obtained from the mosaic image. The update mechanism is as follows:

If

else

Since the standard deviation of each pixel in the mosaic background model usually does not have significant variations during the time, it is not updated. The update mechanism is very effective in improving the segmentation result and it can also resist the slow changes in lighting conditions in the images.

D. Object Tracking with Moving Camera
The proposed segmentation method mentioned above works well when the active camera keeps stationary. However, it may fail to obtain an accurate and robust segmentation when the camera moves because the background in the incoming image is changing. In addition, the aforementioned iterative procedure for finding the corresponding background model in the mosaic image will cause delay in the system when the active camera moves.

In our method, as shown in Fig. 3, when camera moves, we get the camera pan and tilt angle position through an RS-232 port and then only use these parameters to roughly locate the sub-view in the mosaic image without iteratively finding the corresponding background model. This can save much computation. To refine the rough segmentation result, a template matching and object tracking method is adopted. Each pixel value of the current extracted object in the previous extracted object for reducing the segmentation noises in . To reduce the complexity and computation of the tracking, only the centroid of skin-color region in the extracted object is selected. In this way, the proposed strategy achieves robust tracking without any prior knowledge on the object shape.

The correspondence problem can be formulated as a back-end matching motion estimation problem, similar to that employed in predictive video compression. For fast and robust template matching, we adopted the diamond search algorithm proposed in [5]. The template matching criterion is described as follows:

\[
\begin{align*}
\text{if } & C_n(p) \in \text{foreground} \\
& \text{find } C_{1}(p), \text{ which corresponds to } C(p) \\
\text{if } & C_{1}(p) \in \text{foreground} \\
& C(p) \in \text{foreground} \\
\text{else} \\
& C(p) \in \text{background}
\end{align*}
\]

In the tracking mode, the active camera is controlled to put the moving object in the central region of the captured image according to the feature point (e.g. the centroid of the skin-color region) selected. A linear trajectory model is used to predict the feature point’s future position. Suppose the feature point position is at \( p(t) \) at time \( t \). We can predict the feature point position at time \( t+1 \), \( p(t+1) \), by assuming a constant velocity \( v \) which is obtained from the two previous frames.

\[
\begin{align*}
\dot{s}(t) &= p(t) - p(t-1) \\
p(t+1) &= p(t) + v(t)
\end{align*}
\]

Once the feature point leaves the central region, the active camera is re-adjusted. And the pan-and-tilt speed of the camera is determined by the distance between the predicted feature point, \( p(t+1) \), and the center of the central region, \( P \).

\[
\text{set Camera Speed} = \frac{S(p(t+1) - p)}{\Delta t}
\]

where \( CR \) stands for the center region of the captured image with a predefined area, \( S \) is a diagonal matrix containing scaling factors for camera tilt and pan motions, and \( \Delta t \) is the temporal interval between the consecutive frames.

3. EXPERIMENTAL RESULTS
Fig. 4 shows the simulation result of the proposed object segmentation scheme. Fig. 4(a) shows a captured image containing the foreground object. Fig. 4(b) depicts the rough segmentation results after performing the background subtraction scheme. The rough segmentation is still quite noisy. The result after applying the post-filtering is illustrated in Fig. 4(d). The small granular noises can be effectively eliminated using the post-filtering process as shown. Figs. 4(d)-(e) show the final result after region glowing.
We proposed a robust and fast segmentation algorithm and a reliable tracking strategy without the prior knowledge of object shape. The proposed method can extract the object 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.

Firstly, we introduced how to construct a mosaic image and utilize the mosaic image as the reference background image database in the background subtraction step. 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 cost by using the mosaic image.

The rough segmentation is further refined by performing post-filtering, region growing, adaptive background updating, template matching, and object tracking operations.

In the tracking mode, the centroid 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 that the tracking does not need a complex temporal filter (e.g., a Kalman filter). The proposed system can process about 20 QCIF frames per second on a Pentium-III 733 MHz PC without the need of special-purpose hardware.

The proposed scheme can be used in real-time object-based applications such as MPEG-4 video coding, home surveillance, virtual videophone and video conferencing.

5. REFERENCES


