MOTION-FOCUSING KEY FRAME EXTRACTION AND VIDEO SUMMARIZATION FOR LANE SURVEILLANCE SYSTEM

Congcong Li[1], *Yi-Ta Wu*[2], *Shiaw-Shian Yu*[2], *Tsuhan Chen*[3]

[1]Department of Electrical & Computer Engineering, Carnegie Mellon University
[2]Industrial Technology Research Institute
[3]School of Electrical & Computer Engineering, Cornell University

ABSTRACT

This paper proposes a motion-focusing method to extract key frames and generate summarization synchronously for surveillance videos. Within each pre-segmented video shot, the proposed method focuses on one constant-speed motion and aligns the video frames by fixing this focused motion into a static situation. According to the relative motion theory, the other objects in the video are moving relatively to the selected kind of motion. This method finally generates a summary image containing all moving objects and embedded with spatial and motional information, together with key frames to provide details corresponding to the regions of interest in the summary image. We apply this method to the lane surveillance system and the results provide us a new way to understand the video efficiently.

Index Terms— key frame extraction, video summarization, motion-focusing

1. INTRODUCTION

With the development of the digital video processing technology, video surveillance has been playing an important role for security and management. Due to the high volume of videos, manually retrieving information from these videos is very time-consuming. It is necessary and important to allow the computer to automatically extract the parts of interest form videos. Key frame extraction and video summarization [1, 2, 3, 4] are approaches towards tackling this problem, by creating a brief version to represent the original video.

In previous work, there are mainly two kinds of video summarization: dynamic video skimming [1, 2], which itself is still a video but a shorter version, and static video summary [3, 4], which is one or a set of images extracted or synthesized from the original video. Key frame extraction often serves as an important step for video summarization in previous research.

In this work, our goal is to summarize the video with a synthesized image. we integrate key frame extraction and summary-image generation into one interdependent process.

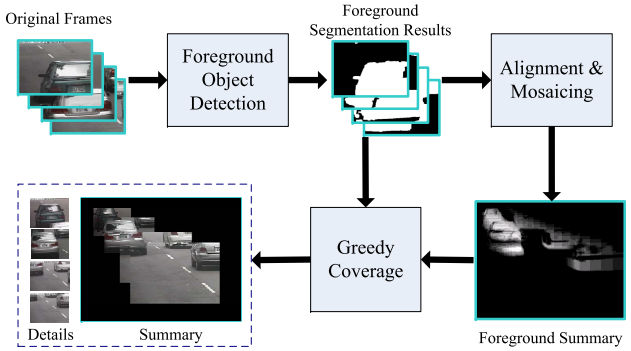


Fig. 1. Overview of the proposed method

Instead of doing frame clustering or specific feature learning, we extract key frames according to their importance on constructing the summary image. The summary image and key frames respectively provide us a general impression and details of interesting objects in the video. A motion-focusing method is proposed to achieve this goal. Unlike other mosaic-based video summarization methods[5, 6], the mosaic here is done based on the consistency within the focused motion, instead of the correspondences between the background scenes.

2. OVERVIEW

The proposed method is initially designed for lane surveillance system, but not limited to this application. This method attempts to focus on one constant-speed motion and then aligns the video frames by fixing the focused motion in a static situation. According to the relative motion theory, objects in other types of motion including the surveillance background in the video are moving relatively to the selected motion. It is reasonable to consider this motion-focusing scheme for lane surveillance video, for a continuous vehicle stream on one lane may move at very similar speed. Fig. 1 gives an overview of the proposed method. The input is a video sequence and the output contains two parts: a summary image and key frames for the video. The whole scheme works in the following steps:

Step 1: Apply background subtraction method to extract the moving foreground for each frame. Video is then cut into

shots by intervals that contain very few foregrounds. The following steps are done for each shot.

Step 2: Estimate the parameters for image alignment. Since this method focuses on a motion with almost constant speed, only the first few frames are needed for the estimation. The correspondences between these frames are constructed by tracking an object in the selected motion. The correspondences for the following frames can be deduced.

Step 3: Construct an initial foreground summary image, in which every object in the focused motion appears at a unique position. The binary segmented images are first scaled and shifted with the parameters gained from Step 2, and then mosaiced together to form the foreground summary, as shown at bottom-right of Fig. 1. The light regions indicate occurrences of objects in the video.

Step 4: Find out a local optimal solution for the problem: using as few as possible binary segmented images to cover no less than 95% foreground region in the foreground summary image. The selected frames are considered key frames and a final summary image can be mosaiced with the key frames.

The final outputs provide information in two different aspects: The summary image provides a whole impression of all moving objects present in the video, and also the spatial and motional relationships between objects that are not captured directly by the camera; the key frames complementarily show details for regions of interest on the summary image.

3. FOREGROUND SEGMENTATION

Within a video shot, the background changes very little. To segment the moving foreground we start with background construction. The naive method is to build a unitary Gaussian model for each pixel to represent the color distribution for a pixel being background. This method is simple and works well only when the background condition remains very stable. However, we may need to handle with more complex cases such as the white balance problem. The automatic white-balancing function may be turned on when some lane videos are captured. In these videos, when a vehicle passes through, the background illumination changes. The previous model may fail to give a good segmentation result. Here we combine the Gaussian background model with the min-cut method [7] to improve the foreground segmentation.

In this method, Background subtraction is combined together with min-cut to get a smooth segmentation of foreground objects. Let I be the current frame to be processed. The frame is represented in gray-color. Let V be the set of all pixels in I and N be the set of all adjacent pixel pairs in I. A labeling function f labels each pixel i as foreground $f_i = 1$ or background $f_i = 0$. The labeling problem is solved by minimizing the Gibbs energy [7], defined as below.

$$E(f) = \sum_{i \in V} E_1(f_i) + \lambda \sum_{(i,j) \in N} E_2(f_i, f_j) \quad (1)$$

$$E_1(1) = \begin{cases} 0 & d_i > k_i^1 \\ k_i^1 - d_i & k_i^2 < d_i < k_i^1 \\ \inf & d_i < k_i^2 \end{cases} \quad E_1(0) = \begin{cases} 0 & d_i < k_i^3 \\ d_i - k_i^3 & k_i^3 < d_i < k_i^1 \\ \inf & d_i > k_i^1 \end{cases} \quad (2)$$

$$E_2(f_i, f_j) = \delta(f_i - f_j) \quad (3)$$

d_i is the absolute difference between the current frame and the previous calculated Gaussian mean value for the i^th pixel, $\{k_i^t \mid \text{t =1,2,3}\}$ are thresholds set that are respectively 3, 0.5, 1.5 times of the calculated Gaussian standard deviation for the i^th pixel. This method provides us a smooth foreground segmentation result and can deal with some complex situations. Fig. 2 shows a foreground segmentation example with both methods. The improvement by the min-cut method is obvious for frames where illumination changes much, e.g, the corresponding frames in the red rectangle and green rectangle.

Fig. 2. A segmentation example comparing results from the pixel-based Gaussian model and those from the min-cut embeded model.

4. IMAGE ALIGNMENT

To synthesize a video shot into a summary image, we need to find out the correspondences between frames. Frames are aligned by fixing objects in the focused motion into unique positions. We approximately assume an affine transform between frames for objects in the focused motion and consider only scaling and translation, as defined below. $[x(t-1), y(t-1)]$ and $[x(t), y(t)]$ are the coordinates for corresponding points in two consecutive frames. $S_x(t)$, $S_y(t)$ and $D_x(t)$, $D_y(t)$ are respectively scaling and shifting parameters.

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} S_x(t) & 0 & D_x(t) \\ 0 & S_y(t) & D_y(t) \end{bmatrix} \begin{bmatrix} x(t-1) \\ y(t-1) \\ 1 \end{bmatrix} \quad (4)$$

With the assumptions of affine transform and constant speed for the focused motion, $S_x(t)$, $S_y(t)$ and $D_x(t)$, $D_y(t)$ can be expressed approximately as the forms below:

$$S_x(t) = \frac{t + a_1}{t + b_1}; S_y(t) = \frac{t + a_2}{t + b_2}; \quad (5)$$

$$D_x(t) = \frac{t}{c_1 t + e_1}; D_y(t) = \frac{t}{c_2 t + e_2}; \quad (6)$$

So we don't need to calculate the transform matrix for each frame; instead, we can figure out the parameters as below:

1) Track the first arisen object in the focused motion. The focused motion can be selected manually or automatically by

the algorithm. In the automatic case, the algorithm chooses the motion of the first-present object in the video. Then it extracts a rectangle region based on the foreground segmentation result and then uses it for template tracking [8] in the following five frames. $S_x(t)$, $S_y(t)$ and $D_x(t)$, $D_y(t)$ are calculated in this process for $t = 1, 2, \ldots, 5$.

2) With the known $S_x(t)$, $S_y(t)$ and $D_x(t)$, $D_y(t)$, we apply least-square-error method to calculate the parameters $\{ a_1, a_2, b_1, b_2, c_1, c_2, e_1, e_2 \}$. These parameters are related to the camera parameters and the speed of the focused motion.

Finally, frames can be aligned using the inversions of the transform matrices to scale and shift the images. Though the affine assumption may fail when objects are moving fast, improvement on alignment is not our main goal. Advanced alignment methods can be introduced in the future.

5. KEY FRAME EXTRACTION AND SUMMARIZATION

The key frame extraction and summary image generation is done through two steps of mosaicing. The correspondences between the summary image and the original frames can be computed with the parameters gained in Section 4.

The initial mosaicing is done with the foreground segmentation results. The reason for doing this is to reduce the redundant occurrences of objects in the focused motion. There is only one occurrence for each of them in the mosaic foreground image. Although this image is created by mosaicing all frames, many foreground regions are covered by foregrounds from only a few frames. So we propose an optimization problem: how to select the minimum number of frames that can also cover the foreground region in the mosaic image, i.e. contain all information of interest?

In practice, we use a greedy search method to find out the frames, which works well though not always providing the optimal solution. Every iteration, we pick one frame that can increase the foreground coverage on the mosaic foreground image most. The searching iteration stops until the coverage is higher than 95%. Fig. 3 illustrates this process. The lightest region in the first row indicates how much the coverage increases by adding a new efficient frame.

The frames being picked out are then considered to be key frames since they are informative and complementary. Then a second-time mosaicing is carried on by mosaicing the key frames to generate the summarization image.

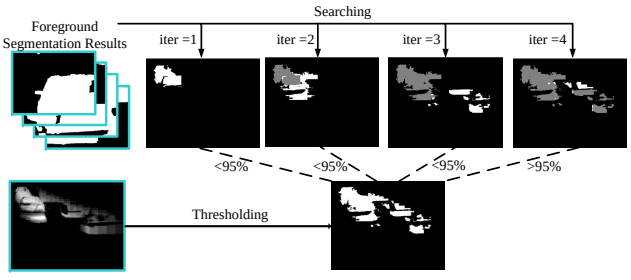


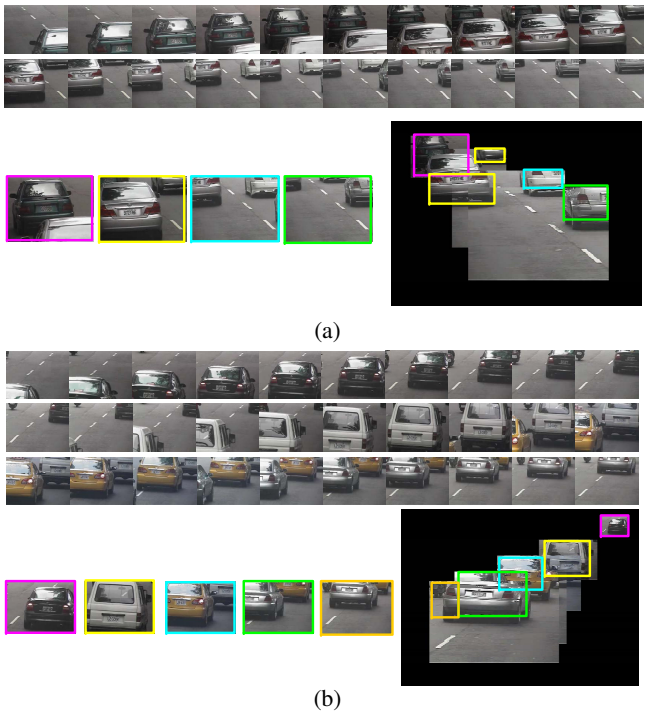
Fig. 3. Greedy search for finding key frames

6. EXPERIMENT RESULTS

We test the proposed method on 50 lane surveillance video shots, 40 of which are from the ITRI surveillance video database and the remaining are downloaded from YouTube.

Fig. 4 gives two examples of video shots and the results from the proposed method. Each of the two examples contains only one specific motion and the moving speed is almost the same for different objects. The frame rate of the original video is 30 f/s. The frames are down-sampled to 10 f/s for being shown here. The moving speed and direction in one video is different from those in the other. The method automatically adjusts the parameters used for frame scaling and shifting. We can see that every vehicle present in the video appears clearly in a unique position in the summary image. Moreover, with the correspondences between the summary image and key frames, we can easily check the details for a region of interest in the summary image. The correspondences are built from the greedy searching process, by knowing the coverage-increased region when adding a new frame.

We can see that the summary image not only represents all objects in the focused motion clearly, but also provides their relative spatial information which is not directly shown in the original video. Moreover, since it is motion-focusing, it is also obvious for us to catch the temporal relation. In previous work as [1], video summarization is generated by mosaicing the foregrounds belonging to the same object in different frames into a single image or shorter video, while keeping the background scene static as that in the reality. Fig. 5 shows the resulted frames for Fig.4 (a) with a similar method described



(a)

(b)

Fig. 4. Two examples of videos containing unique constant-speed motion and the results from the proposed method

Fig. 5. Resulted frames with method [1] for Fig. 4 (a)

Fig. 6. An example with objects moving at different speeds

Fig. 7. Another example with objects moving at different speeds and results generated by focusing on different motions

in [1]. Although the dynamic summary keeps the video dynamics, it cannot provide temporal and spatial relation as the proposed method here.

Fig. 6 and Fig. 7 give two examples with objects moving at different speeds. Fig. 6 is a video shot before an accident from a tunnel surveillance video. The vehicles on the same lane should keep moving at similar speed to avoid accidents, but the second car did not. The method focuses on the first car's constant motion. In the summary image, the first car has only one clear occurrence while the following car has multiple occurrences, showing its relative motion to the first car. Fig. 7 gives another example and shows different results by focusing on the vehicle motion and focusing on the pedestrian motion. Again, each object in the focused motion has a unique occurrence while those in other motions have multiple occurrences to indicate relative motion.

7. CONCLUSIONS

In this work, we propose a motion-focusing method to extract key frames and generate summarization for surveillance videos. The proposed method focuses on a certain constant-speed motion within a video shot and summarizing the video shot by fixing objects in the focused motion in static positions. According to the relative motion theory, the other objects in the video are moving relatively to those in the selected motion. With the proposed method, all moving objects in the video are included in the summary image: objects in the focused motion appear clearly in unique positions while the others have multiple occurrences according to their relative movement to the focused motion. The summary image reflects the spatial relationship within objects in the focused motion and also the relative motion relationship between objects in non-focused motion and those in focused motion. At the same time, the extracted key frames provide us detailed information about the regions of interest in the summary image. The summary image generation and the key frame extraction are integrated into an interplaying process.

8. ACKNOWLEDGMENT

This work is a partial result of Project 83522Q1200 conducted by Industrial Technology Research Institute under sponsorship of the Ministry of Economic Affairs, Taiwan.

9. REFERENCES

[1] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Trans. on PAMI*, vol. 30, no. 11, pp. 1971–1984, Nov. 2008.

[2] H. Kang, X. Chen, Y. Matsushita, and X. Tang, "Space-time video montage," *CVPR2006*, vol. 2, pp. 1331–1338, 2006.

[3] Y. Hadi, F. Essannouni, and R. Thami, "Video summarization by k-medoid clustering," in *Proc. of SAC 2006*, pp. 1400–1401.

[4] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. on Circ. & Sys. for Video Tech.*, vol. 9, no. 8, pp. 1280–1289, 1999.

[5] A. Aner-Wolf and J. R. Kender, "Video summaries and cross-referencing through mosaic-based representation," *CVIU*, vol. 95, no. 2, pp. 201 – 237, 2004.

[6] M. Irani, P. Anandan, and S. C. Hsu, "Mosaic based representations of video sequences and their applications," in *ICCV*, 1995, pp. 605–611.

[7] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," *IEEE Trans. on PAMI*, vol. 26, no. 9, pp. 1124–1137, Sept. 2004.

[8] L. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Trans. on PAMI*, vol. 26, no. 6, pp. 810–815, June 2004.