Video Condensation for Video Forensics

Yu-Ming Liang
Department of Computer Science and Information Engineering,
Aletheia University,
Taipei, Taiwan
ymliang@mail.au.edu.tw

***Abstract*—Video forensics has become an important issue because of the setup of a large number of surveillance cameras. Video condensation can increase efficiency in quickly searching criminal events from the vast amounts of surveillance video data. In this paper, we propose a video condensation technique for video forensics to condense a long period of surveillance video into a short video clip under keeping important information. The proposed technique is comprised of the following three steps: image sequences transformation, motion oriented filter, and ribbon carving. First, the original video sequence is converted into X-T and Y-T image sequences. Then, a motion oriented filter operator based on Gaussian gradient process is applied to extract moving-object maps in a specific direction from the X-T and Y-T image sequences. Finally, we apply a ribbon carving approach to achieve the purpose of video condensation. Preliminary experiments on realistic video data demonstrate the applicability of the proposed technique.**

Keywords- video forensics; video condensation; motion oriented filter; ribbon carving

I. INTRODUCTION

With a large number of video cameras mounting on city streets and highways, airports, parking lots, and office buildings, video surveillance has become one of the most important monitoring tools. Since video records from video cameras are often treated as the important resource for crime investigation, video forensic has become an important issue in recent years. Presently, when a criminal event occurs, investigators usually collect a large number of video records from the video cameras mounted on the locations surrounding the criminal scene. If the investigators do not know the exact occurring time, they must expend much human operating cost and time to find suspicious video clips from the vast amounts of visual data. Therefore, it is very important for video forensics to remove relatively static temporal and spatial segments and preserve all the most relevant activities. In this paper, we shall focus ourselves on how to efficiently and automatically condense the vast amounts of surveillance video data for increasing efficiency in video forensics.

The objective of video condensation is to preserve all the most relevant activities by removing relatively static temporal and spatial segments so that a short summary video can be produced from a long surveillance video. A number of approaches have been proposed to achieve this aim. For example, video forwarding is an earlier approach for video condensation by skipping frames at regular time intervals [1]. This approach may exclude some relevant activities and contrarily preserve irrelevant static segments. Especially, when higher condensation ratios, defined as the ratio of video length prior to the processing to that after processing, are necessary, it may lead to serious exclusion of relevant activities due to skip numerous consecutive frames. To deal with this issue, content-adaptive skipping approaches [2-4] were proposed to skip frames that were inactive or exhibit low activity. However, since only complete frames are removed, it is limit on the realizable condensation ratios. To achieve high condensation ratios, video summarization [5-6] were proposed by extracting some key frames to represent each video shot. Video summarization is the special case of content-adaptive skipping, and thus, it also has the same disadvantage of removing only complete frames, which has the limitation on the realizable condensation ratios. In the above-mentioned approaches, all preserved frames are then presented sequentially, like a slide show. Because it may cause large temporal gaps between consecutive preserved frames, temporal continuity of events is destroyed.

Different to the above-mentioned approaches, removing complete video frames, an alternative approach extracts spatial-temporal segments from different frames, and then combines them to a new video sequence by shifting the segments in space or time. In one approach, video montage proposed by Kang et al. [7] cut off less informative spatial-temporal segments to generate a short video preserving much more compact yet highly informative. First, they extracted the visually informative space-time portions of the input videos and represented them in volumetric layers. Then, the first-fit and graph cut optimization technique was applied to pack the layers together into a small output video volume to make the total amount of visual information in the output video volume maximum. Although this approach can condense the video with higher ratios in preserving most events, it creates visible seams due to the combination of large uncorrelated segments. Pritch et al. [8] proposed video synopsis to avoid the visible seams that appear in the video montage by maintaining the spatial locations of object segments. First, moving objects were extracted and tracked in video, and the tracks of moving objects were defined as space-time tubes. Next, these tubes were realigned in time shifts by minimizing an energy function. The definition of the energy function was to preserve most of original activities in the video and to avoid collisions between different shifted tubes. Since the approach allows only temporal transformations, it can prevent the total loss of context that occurred when both the spatial and temporal locations are changing. However, this approach is complex because it requires several computational stages including object detection and tracking, object queuing, background fusion, tube selection, and tube stitching. A video condensation approach based on ribbon carving was proposed

to avoid complex computation [9]. Similar to video synopsis, this approach only allows temporal transformation to combine spatial-temporal segments from different frames into an output video. This approach uses simple background subtraction instead of object extraction and tracking to increase the computational efficiency. First, they extracted the video ribbon in space-time volume from foreground image sequences. The concept of video ribbon was an extension of that of a seam in 2-D images used for content-aware image resizing [10]. Then, similar to seam carving, they recursively carved ribbons out by minimizing an activity-aware cost function to achieve video condensation. Because this approach uses background subtraction to extract moving objects, it must face the issues of background modeling especially in complex environments. Besides, this approach cannot condense scenes with multiple objects moving at very different speeds and/or directions.

In this paper, we propose a video condensation for video forensics based on the ribbon carving approach proposed by Li [9]. However, our approach does not require complex background modeling; instead, we use Gaussian gradient process to extract moving object directly in spatial-temporal volume. Besides, we apply motion oriented filter to filter a specific direction on conditional that investigators know the moving direction of criminals, and thus we can obtains higher condensation ratios. First, the original image sequence is converted into X-T and Y-T image sequences. Next, we apply a motion oriented filter operator based on Gaussian gradient process to extract moving object maps with specific moving direction from the obtained X-T and Y-T image sequences. Finally, a ribbon carving approach is used to achieve the purpose of video condensation.

II. THE PROPOSED APPROACH

Fig. 1 shows a block diagram of the proposed video condensation process. The proposed approach is comprised three steps: image sequences transformation, motion oriented filter, and ribbon carving, which we describe in the following three subsections, respectively.

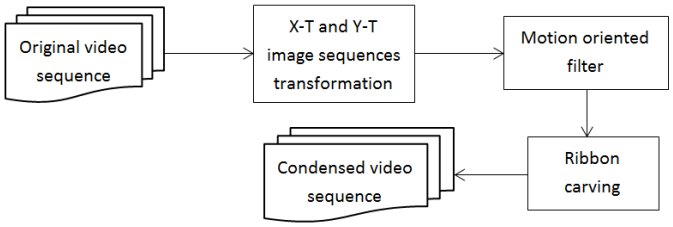


Fig. 1. Block diagram of the proposed video condensation process.

A. Image sequences transformation

Suppose that a video sequence is comprised of N video frames with resolution of $W \times H$ and it can be represented as a spatial-temporal volume, i.e. a $W \times H \times N$ cube in the X-Y-T space, as shown in Fig. 2. The purpose of this paper is to reconstruct a new video sequence comprised of N' frames with the same resolution, where N' is smaller than N. Therefore, for each time series produced by each location in the space, $(N- N')$ points will be removed. The reconstructed video sequence must preserve all the important information, e.g. moving objects and the appearing order of them. In order to reconstruct the condensed video sequence from the original one, we must keep all the moving object segments while remove relatively static spatial and temporal segments under the condition of preserving the appearing order of moving objects. If a spatial location is static, i.e. without any moving object passing, the time series of gray values produced by this location has a smooth variation. Oppositely, if a spatial location is with some moving objects passing, this location produces a non-smooth time series having an obvious variation in gray values. In order to observe the variation of the time series of gray values for each spatial location, we can transform the X-Y image sequence into X-T and Y-T image sequences, as shown in Fig. 3. In all of X-T and Y-T images, each horizontal line indicates the time series of gray values for each spatial location. In this example, it is clear that the gray value has an obvious change when a moving object passes. Furthermore, it is also clear that the movement of a moving object can be mapped into a trajectory in the X-T image when the moving object moves along the x-axis. From these trajectories, we can obtain the moving direction of each moving object and the appearing order of all moving objects.

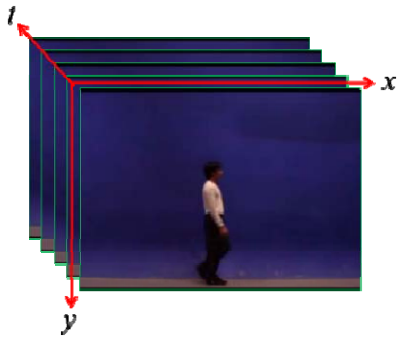


Fig. 2. A video sequence is represented as a spatial-temporal volume.

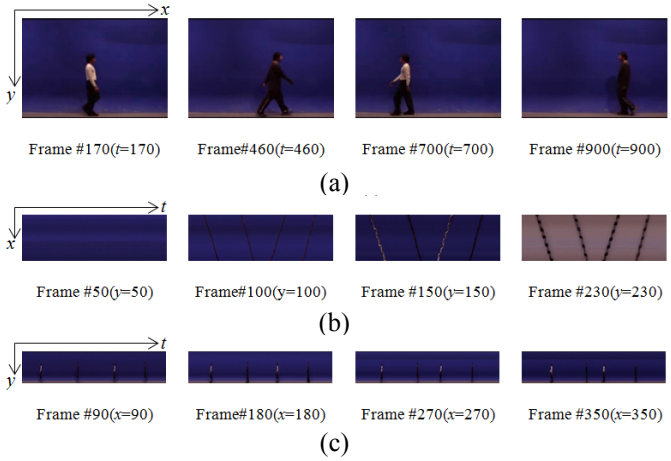


Fig. 3. (a) The original X-Y image sequence with resolution of 352×240 in total of 1100 frames. (b) X-T image sequence with resolution of 1100×352 in total of 240. (c) Y-T image sequence with resolution of 1100×240 in total of 352 frames.

B. *Motion oriented filter*

After transforming X-Y image sequence into X-T and Y-T image sequences, we must extract the important information from the image sequences. Since a non-static spatial location produces a non-smooth time series having an obvious variation in gray values, we can apply a gradient magnitude operator on the *t*-axis to detect the significant changes in gray values. Since gradient computation based on gray values of only two adjacent points is sensitive to noise, we apply derivative of Gaussian (DoG) filter [11] to calculate the gradient magnitude of *t*-axis for each image in X-T and Y-T image sequences. As shown in Fig. 4, a pixel with larger gradient magnitude, i.e. the brighter pixel, indicates a moving object passing. Besides, because the trajectories obtained from the movements of moving objects can response the moving directions of objects, the gradient vector of each pixel can be used to determine the moving directions. Therefore, we can use motion oriented filter based on the gradient vector to filter a specific direction if the moving direction of criminals is known beforehand. Figs. 5 and 6 show the gradient magnitude images after applying motion oriented filter on the left-to-right and right-to-left moving directions along the *x*-axis, respectively.

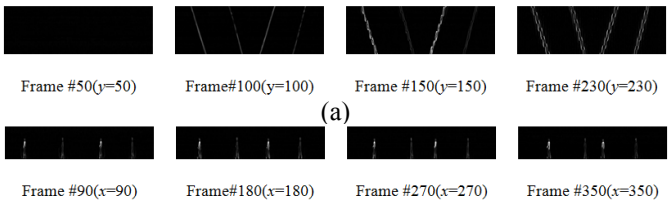


Fig. 4. Gradient magnitude images (a) in the X-T image sequence and (b) in the Y-T image sequence.

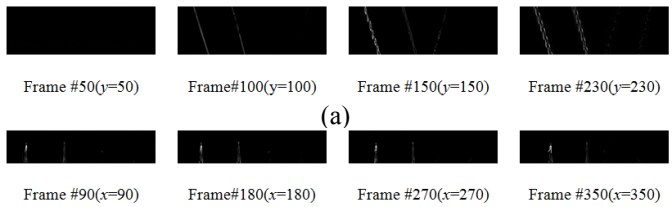


Fig. 5. The gradient magnitude images after applying motion oriented filter on the left-to-right moving direction along the *x*-axis (a) in the X-T image sequence and (b) in the Y-T image sequence.

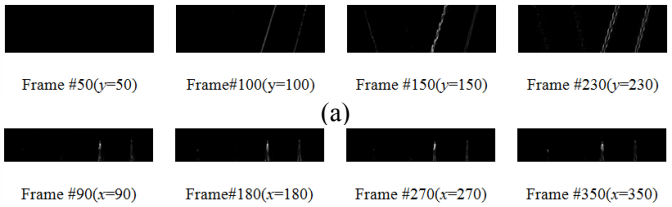


Fig. 6. The gradient magnitude images after applying motion oriented filter on the right-to-left moving direction along the *x*-axis (a) in the X-T image sequence and (b) in the Y-T image sequence.

C. *Ribbon carving*

The purpose of video condensation is to reconstruct a new video sequence by removing some static points from the time series produced by each spatial location under the condition of preserving all the important information. In order to achieve this purpose, the proposed method must satisfy two requirements. First, the number of removed points from each time series must be the same. Second, it must prevent important information distortion in either space or time. Here, we apply the ribbon carving approach proposed by Li et al. [9] for video condensation.

The ribbon carving approach is motivated by seam carving for content-aware image resizing [10]. A vertical/horizontal seam is comprised of a set of pixels, which is a path-connected with different vertical/horizontal coordinates. Therefore, the idea of image resizing is to recursively delete more unimportant seams associated with less costs. Extending the idea of seam deletion for image down-sizing, a 3-D video seam is defined as a counterpart of a 2-D image seam within the spatial-temporal volume of a video sequence. In order to satisfy the above requirements for video condensation, a video seam is defined as a connected surface within the spat-temporal volume, and no two pixels in the video seam have the same spatial location. A ribbon is defined as one of 3-D video seams, as shown in Fig. 7. A vertical ribbon is defined as follows:

$$R_v = \{(x, y, f_y(y)) \mid x = 1,...,W, y = 1,...,H\},$$
$$\text{s.t. } \left|f_y(y+1) - f_y(y)\right| \le 1 \ \ \forall y = 1,...(H-1), \qquad (1)$$

where $f_y(y)$ is a function of only *y* with range 1,...,*N*. That is, a vertical ribbon is consisted of *w* equivalent seams, each of which is located on a Y-T image. Fig. 7(a) shows an example of vertical ribbon. Similarly, a horizontal ribbon is defined as follows:

$$R_h = \{(x, y, f_x(x)) \mid x = 1,...,W, y = 1,...,H\},$$
$$\text{s.t. } \left|f_x(x+1) - f_x(x)\right| \le 1 \ \ \forall x = 1,...(W-1), \qquad (2)$$

where $f_x(x)$ is a function of only *x* with range 1,...,*N*. That is, a horizontal ribbon is consisted of *H* equivalent seams, each of which is located on a X-T image. An example of horizontal ribbon is shown in Fig. 7(b). Therefore, the ribbon carving approach is to recursively find the least-cost ribbon among the set of all horizontal and vertical ribbons to delete.

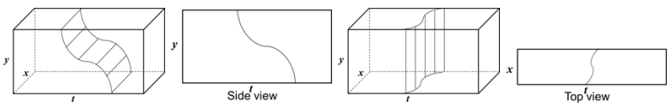


Fig. 7. The concepts of (a) a vertical ribbon and (b) a horizontal ribbon.

In our work, after applying motion oriented filter based on Gaussian gradient operator, we can obtain gradient magnitude

images, in which each pixel indicates the possibility of moving object passing. Therefore, the gradient magnitudes can be used as the costs for ribbon carving. Suppose $M_{XT}(t,x,y)$ and $M_{YT}(t,y,x)$ denote the gradient magnitudes of each point (t,x,y) in the X-T image sequence and each point (t,y,x) in the Y-T image sequence, respectively. The cost of each point (x,y,t) in the spatial-temporal volume of the original video sequence can be calculated by

$$C(x,y,t)=\sqrt{(M_{XT}(t,x,y))^2+(M_{YT}(t,y,x))^2}\ , \qquad (3)$$

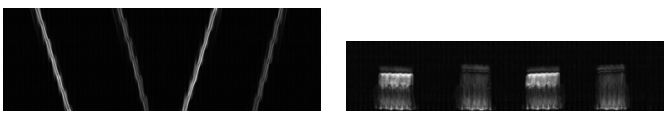
Therefore, the cost of R_v can be defined as follows:

$$\begin{aligned} C(R_v) &= \sum_{(x,y,t)\in R_v} C(x,y,t)=\sum_{y=1}^{H}(\sum_{x=1}^{W} C(x,y,f_y(y))) \\ &= \sum_{y=1}^{H} \widetilde{C}_v(f_y(y),y), \end{aligned} \qquad (4)$$

where $\widetilde{C}_v(t,y)=\sum_{x=1}^{W} C(x,y,t)$ denotes the cost of the point (t,y) in the accumulated Y-T image obtained from the Y-T image sequence. Similarly, the cost of R_h can be defined as follows:

$$\begin{aligned} C(R_v) &= \sum_{(x,y,t)\in R_v} C(x,y,t)=\sum_{y=1}^{H}(\sum_{x=1}^{W} C(x,y,f_y(y))) \\ &= \sum_{y=1}^{H} \widetilde{C}_v(f_y(y),y), \end{aligned} \qquad (5)$$

where $\widetilde{C}_h(t,x)=\sum_{y=1}^{H} C(x,y,t)$ denotes the cost of the point (t,x) in the accumulated X-T image obtained from the X-T image sequence. Fig. 8 shows accumulated X-T and Y-T images, respectively. Therefore, finding a minimum-cost vertical/horizontal ribbon can be regarded as finding a minimum-cost vertical seam in the accumulated Y-T/X-T image. The minimum-cost vertical seam can be found by the 2-D image seam carving approach using dynamic programming algorithm. After finding the minimum-cost vertical and horizontal ribbons, the ribbon with smaller cost will be selected for deletion.



(a) (b)

Fig. 8. (a) An accumulated X-T image; and (b) an accumulated X-T image.

As shown in Fig. 8, it is clear that some noise exists in the background and moving objects include some holes. To achieve higher condensation ratios and to prevent the moving object distortion in either space or time, we apply some image processing approaches to the accumulated images to obtain moving-object maps. These image processing approaches include binary operators based on automatic thresholding using Gaussian mixture models [12], morphological operators for eliminating noise and filling the holes, and connected component labeling algorithms for removing components with small pixels. Fig. 9 shows the resulting moving-object maps after applying the above processes. Therefore, the stopping criterion of ribbon craving is to delete all vertical and horizontal ribbons associated with zero cost. The resulting moving object maps after applying ribbon carving based on three different motion oriented filter operators are shown in Fig. 10. Fig. 11 shows the original video sequence in total of 1100 video frames and the three resulting video sequences in total of 341, 175, and 169 video frames, respectively.

(a) (b)

Fig. 9. Resulting moving object maps from (a) the accumulated X-T image and (b) the accumulated Y-T image.

(a) (b) (c)

Fig. 10. The resulting moving object maps after applying ribbon carving based on (a) no motion oriented filter operator, (b) motion oriented filter on the left-to-right moving direction along the *x*-axis, and (c) motion oriented filter on the right-to-left moving direction along the *x*-axis.

Fig. 11. (a) The original video sequence in total of 1100 video frames. (b) The resulting video sequence without the motion oriented filter operator in total of 341 video frames. (c) The resulting video sequence with the motion oriented filter operator on the left-to-right moving direction along the *x*-axis in total of 175 video frames. (d) The resulting video sequence with the motion oriented filter operator on the right-to-left moving direction along the *x*-axis in total of 169 video frames.

III. EXPERIMENTS

We took a video sequence to evaluate the effectiveness of the proposed method. The video sequence taken at an indoor scene was comprised of 2651 frames, and 11 moving objects passed the scene. Typical video frames are shown in Fig. 12. Fig. 13 shows the moving-object maps obtained from the X-T and Y-T image sequences transformed from the original video sequence. It is clear that the trajectories in the X-T map exactly present the moving directions and the appearing order of all objects. The resulting moving-object maps after applying ribbon carving based on the three different motion oriented filter operators mentioned in Section II.B are shown in Fig. 14. Fig. 15 shows the three condensed video sequences in total of 952, 372, and 313 frames, respectively. In the original video, only one person passed the scene at the same time, but more than one person might pass the scene at the same time and the appearing order did not be destroyed after video condensation. Therefore, we can obtain higher condensation ratios. When we apply the motion oriented filter operator on the left-to-right moving direction along the *x*-axis, all the person moving from right-to-left were removed. Similarly, when we apply motion oriented motion oriented filter operator on the right-to-left moving direction along the a-axis, all the person moving from left-to-right were removed.

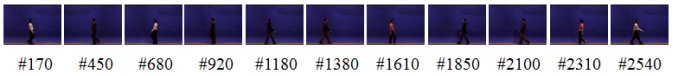


Fig. 12 A video sequence comprised of 2651 frames.

Fig. 13. The moving-object maps obtained from (a) the X-T image sequence and (b) the Y-T image sequence.

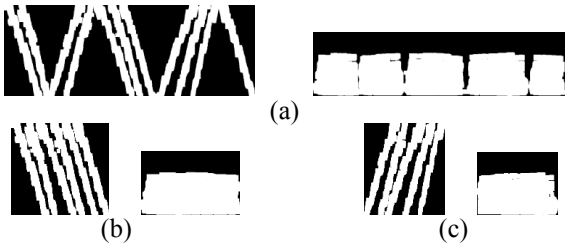


Fig. 14. The resulting moving-object maps after applying ribbon carving based on (a) no motion oriented filter operator, (b) motion oriented filter on the left-to-right moving direction along the *x*-axis, and (c) motion oriented filter on the right-to-left moving direction along the *x*-axis.

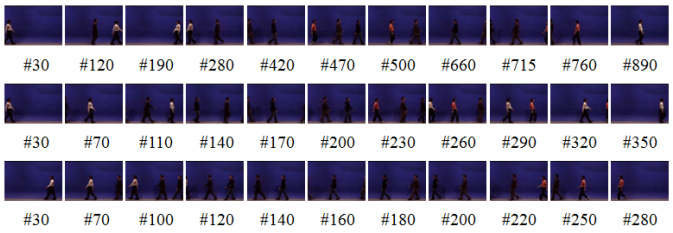


Fig. 15 The three condensed video sequences in total of 952, 372, and 313 frames, respectively.

IV. CONCLUSION

We have proposed a framework of video condensation for video forensics. The framework comprises three steps: image sequences transformation, motion oriented filter, and ribbon carving. First, we transform the original image sequence into X-T and Y-T image sequences. Next, a motion oriented filter operator based on Gaussian gradient process is applied to extract moving-object maps in a specific direction from the obtained X-T and Y-T image sequences. Finally, a ribbon carving approach is used to condense the original video sequence. The preliminary experiment results show the feasibility of the proposed framework.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council, Taiwan under Contract NSC 99-2218-E-156 -004.

REFERENCES

[1] M. Mills, J. Cohen, and Y. Y. Wong, "A Magnifier Tool for Video Data," *Proceedings of the ACM Computer Human Interface*, May 1992, pp. 93- 98.

[2] J. Nam and A. H. Tewfik, "Video Abstract of Video," *Proceedings of the IEEE Workshop Multimedia Signal Processing*, 1999, pp. 17- 122.

[3] N. Petrovic, N. Jojic, and T. S. Huang, "Adaptive Video Fast Forward," *Multimedia Tools and Applications*, vol. 26, no. 3, pp. 327- 344, 2005.

[4] M. M. Yeung and B. L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no 5, 1997.

[5] J. H. Oh, Q. Wen, S. Hwang, and J. Lee, "Video Abstraction," *in Video Data Management and information Retrieval*, S. Deb, Ed. Hershey, PA: Idea Group Inc./IRM Press, 2004, pp. 321- 346, ch. XIV.

[6] C. W. Ngo, Y. F. Ma, and H. J. Zhang, "Automatic Video Summarization by Graph Modeling," *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 104- 109.

[7] H. W. Kang, Y. Matsushita, X. Tang, and X. Q. Chen, "Space-Time Video Montage," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1331- 1338.

[8] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological Video Synopsis and Indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971- 1984, Nov. 2008.

[9] Z. Li, P. Ishwar, and J. Konrad, "Video Condensation by Ribbon Carving," *IEEE Transactions on Image Processing*, vol. 18, no. 11, Nov. 2009.

[10] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image resizing," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[11] R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision*, MIT Press and MCGraw-Hii, Inc., 1995.

[12] Z. K. Huang and K. W. Chau, "A New Image Thresholding Method Based on Gaussian Mixture Model," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 899- 907, 2008.