Video Summarization for Object Tracking in the Internet of Things

Chu Luo
Department of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
cl7e13@ecs.soton.ac.uk

***Abstract*—Object tracking in the Internet of Things (IoT) has become a hot topic over the past ten years. Currently, the integration of video and radio-frequency identification (RFID) technology plays a crucial role in item-level activity recognition. Various techniques and applications have been proposed for visual object tracking. However, identifying semantic features of item-level objects in huge size of video content is a non-trivial task, especially in supply chain management. To alleviate this problem, this paper presents a novel method that applies IoT information to facilitate video summarization. Differing from common video summarization techniques, we use IoT information to select keyframes of the video content during the background model establishment. Then we match other keyframes with the background to extract important features. Finally, a compact summarization image for queried objects is generated according to a clustering analysis. We have also performed experiments to confirm the effectiveness of the proposed work.**

Keywords*—*Internet of Things; RFID; Video Summarization; Background Modeling; Object Tracking

I. Introduction

Object tracking is an important precondition to supply chain management of smart cities. To improve the monitoring of moving objects, video surveillance systems and techniques are widely adopted in supply chain management. Video techniques are able to provide visual information of objects including appearance and motion. However, current video analysis methods are difficult to monitor behaviors and the identification of single objects without other integrated information [1], [2]. These studies point out that the poor applicability of current video tracking techniques is due to several critical factors: limited detection accuracy, insufficient object identification and lack of processing capacity. For example, video tracking may produce inaccurate results when objects are small in size or have the same appearance in supply chains. Hence, there is a need for supply chain management to physically identify objects.

The Internet of things (IoT) greatly bridges the gap between the real objects and data representations. IoT is a global infrastructure where services, devices and objects are highly interconnected. In IoT, physical objects are attached with radio-frequency identification (RFID) tags, including two main types: passive (no power source) and active RFID (self-powered). In general, passive RFID tags are inexpensive, while active RFID tags have longer operating range. Both passive and active RFID tags uniquely identify every single object with a tag serial number. Based on RFID tags, readers and other sensors, tracking single objects with time, location and other information, such as temperature and force, is achieved in IoT. With item-level traceability provided by RFID, IoT significantly minimizes the possibility of product shrinkage in supply chain management, including misplaced items and damaged products [3]. Furthermore, a significant number of researchers integrate RFID technology into video surveillance systems to improve the accuracy of object identification and positioning [4]–[8]. They suggest that the combination of RFID and video technologies is the key to successful analysis of moving objects. However, it is time-consuming to browse huge size of video data during the retrieval process. Although previous work uses diverse approaches [9], [10] to summarize video content, few studies focus on simplification of visual object tracking in IoT. It is worth noting that video summarization techniques are essential to information retrieval of IoT-based video surveillance systems.

To address the above problem, this paper proposes a novel approach for summarizing video of object tracking in IoT. By using IoT information, we select several keyframes of the video content to build a background model. This background model is used to match other keyframes and extract important features. Then connected components in extracted keyframes are clustered by the K-means algorithm. Based on the result of clustering analysis, we generate a compact summarization image which contains important features of queried objects without redundancy. Users are allowed to quickly scan the video content by jumping to a keyframe in summarization images, as well as jumping to an exact RFID serial number. Experimental results show that the proposed work creates lower noise of background models than that of the common Gaussian mixture model (GMM). More importantly, the generated summarization image is effective for simplifying visual object tracking in IoT. To summarize, the main contributions of this paper are:

- Providing a video summarization technique for IoT-based video surveillance systems. Compared to conventional video summarization approaches, our scheme satisfies the identification of objects and reduces the noise and computational cost. To the best of our knowledge, it is the first method applying IoT information to the process of video summarization.
- Implementing a prototype system and performing an experiment to assess its effectiveness. In a sample s-

978-1-4799-5073-7/14 $31.00 © 2014 IEEE
DOI 10.1109/NGMAST.2014.20

cenario video of IoT-based supply chain, our approach achieves a lower noise rate than GMM during the establishment of background models. Furthermore, it returns satisfactory summarization results for queried objects after the clustering and image stitching.

The other parts of this paper are organized as follows. In Section II, an overview of RFID and video technologies is presented. Section III describes our method, including the technological details of video summarization for object tracking in IoT. Section IV presents the experimental results. Finally, Section V concludes this paper.

II. Related Work

A. Visual Object Tracking in IoT

Based on RFID and video technologies, much work is proposed for object tracking. Wang and Cheng [4] present an indoor positioning and identification system consisting of a camera, a RFID reader and a group of fixed RFID tags. By analyzing signal strength of RFID tags and background of video images, the system locates objects and extracts foreground images of objects in the scene. Hasanuzzaman et al. [6] describe a framework for medicine monitoring in IoT. Their system builds background models from camera images when RFID tags of medicine bottles are out of range of the antenna. Once the antenna detects tags, the system starts to extract foreground images. In [11], the integration of RFID and video techniques is used to detect RFID-tagged objects manipulated by users. Even if RFID tags are missing in the new video, the learned models can be used for fuzzy visual pattern recognition of activities and objects. The work in [12] applies multiple cameras for tracking people. RFID antennas are deployed to identify authorized people who carry active RFID tags.

Using the integration of RFID and video techniques, object recognition and tracking systems have good performance in previous research. However, a major challenge is information retrieval in a large scale of recorded video. Previous work has not addressed the solutions for the explosive growth of visual object tracking in IoT.

B. Video Summarization

Most existing video summarization methods concentrate on identifying similarities of different video frames, and then clustering a number of frames with the most dissimilar features. For instance, Gong and Liu [9] present a method for video summarization based on singular value decomposition. They select a group of frames with the most different features to represent the major video shots. Ueda et al. [13] use a three-dimensional icon to represent a video shot with its duration. With edge detection, their system generates icons from visual objects in the video. Goldman et al. [14] propose a schematic storyboard for video summarization. Through a subject motion arrow, the main features of a sequence of keyframes are displayed in the storyboard.

Besides keyframe selection, some recent studies enhance the user interface to reduce the browsing complexity. Barnes et al. [15] present a method called multi-scale video tapestries, which contain continuous and zoomable visualizations for interactive video navigation. Unlike a timeline bar, video tapestries are a group of summary frames in the video and help users to understand the content. The work in [16] uses a video cube in a 3D volume to display keyframes in the video. Users are allowed to navigate to a part of the video content through the 3D volume. In [10], video is summarized into three levels: scenes represented with different colors, hand-drawn sketch representation, and compact frames. Each level has operations for interactive video browsing. Users are able to select, zoom, or drag the summarization navigators by using different gestures.

Although these approaches try to extract and present important features by exclusively analyzing video frames, they still require users to pick the right results of queried objects. In our work, this limitation can be overcome by the IoT infrastructure. RFID, together with timestamps, enables automatic identification for queried objects in the process of video summarization. On this basis, we propose a video summarization method for object tracking in IoT.

III. Proposed Methodology

In this section, we first give an overview of IoT-Based video summarization. Then we present video and IoT data acquisition of objects in the automated environment. Finally, we illustrate the internal details of video summarization, including foreground extraction and clustering.

A. Overview

Our method allows users to skim the video content of supply chain in two ways: image interface and RFID interface for every single item. The intention is to enhance the interaction and usability of video tracking. Image interface is generated from keyframe segmentations which focus on objects features. These keyframe segmentations are organized together for the same objects. While RFID interface shows the exact RFID serial numbers, providing a fast hop to the shot of every single item in the video content. Fig. 1 shows a screenshot of the main features in an example interface.

1) Image Interface: For every item, a sequence of keyframe segmentations is displayed as a timeline slider. Those segmentations consist of video shots showing object features. This is a popular way to locate a specific video shot in the whole video. Rather than existing approaches, IoT data is used to improve the precision and categorization of segmentation. According to the timestamp of every item in IOT, several keyframes are selected as background images. By using the Gaussian mixture model, objects images are segmented from backgrounds in the other keyframes. Finally, segmentations of the same objects are integrated together. Users can both preview a video shot for an object, or watch the video starting from the exact frame by clicking an area of the image interface. Fig. 2 gives a possible situation of image interface.

2) RFID Interface: Unique identifiers of RFID tags are shown in RFID interface. According to timestamps in IoT, RFID serial numbers are mapped to corresponding keyframes of the video content. By selecting a target RFID serial number, users can quickly seek a video shot of a target item. The other item information is optional. Applications can specify concrete types of information for different purposes. A possible situation of RFID interface is given in Fig. 3.

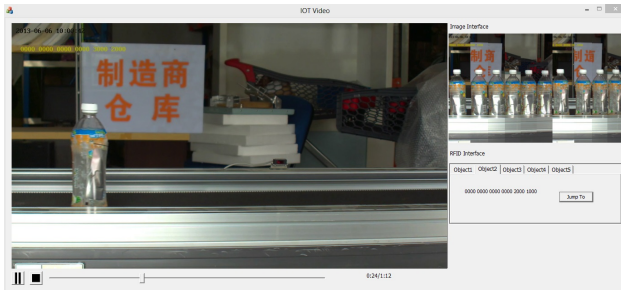

Fig. 1. User interface

Fig. 2. Image interface

B. Video and IoT Data Acquisition

When an item passes a logistics sector (typically a conveyor belt), the system both collects the RFID and video data. RFID readers send RFID data to the control server. And the system database stores the processed data. Through the control server, RFID serial numbers are sent to IP cameras. Then, IP cameras display these numbers and send video images to the control server. The architecture of the process is shown in Fig. 4. After a video file is completely created, we extract its keyframes. Timestamps of RFID tags and keyframes will be used for video summarization.

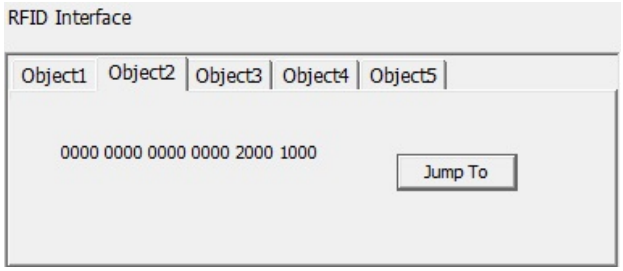

Fig. 3. RFID interface

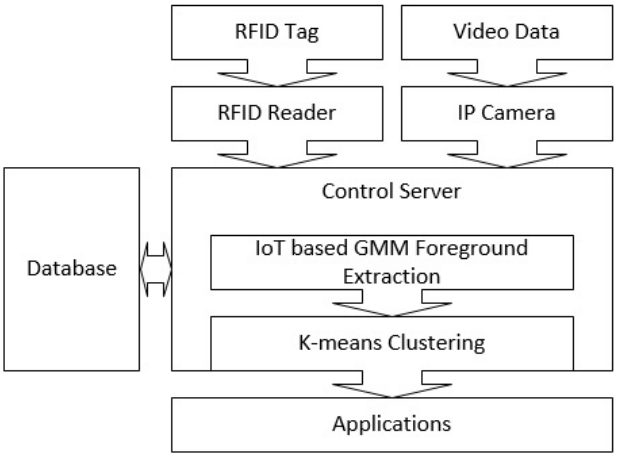

Fig. 4. Architecture of the process

C. Video Summarization

At the start of video summarization, the system queries for all the timestamps of RFID tags and selects several keyframes in the video. These timestamps are timings between two adjacent passing objects. In other words, selected keyframes contain no items in the screen. Based on the timestamps, selected keyframes have approximately regular time gaps so that all period of a video clip is covered.

1) Foreground Extraction: Our approach of foreground extraction is similar to GMM in [17], where the values of each pixel in a given time are defined as a "pixel process." With the timestamps in IoT, our approach selects fewer and preciser keyframes to establish a fast evaluation of background pixels.

In GMM, given a video clip, pixel $\{x_0, y_0\}$ has its history at a stated time t

$$\{X_1, ..., X_t\} = \{I(x_0, y_0, i) : 1 \leq i \leq t\} \tag{1}$$

where I denotes the sequence. In our method, a set of keyframes $\{F_{b_1}, ..., F_{b_m}\}$ is selected to predict the state of pixel $\{x_0, y_0\}$ at a given time t. They maintain

$$1 \leq b_1 \leq t \leq b_m \tag{2}$$

With the help of IoT information, the amount of selected frames is less than GMM, so that less computational power is needed. Moreover, according to the timestamps in IoT, several keyframes later than t are selected, unlike only concerning previous frames in other approaches. After the set of keyframes is selected, the probability of each pixel value is

$$P(X_t) = \sum_{i=1}^{K} \omega_{b_i,b_m} \times \eta(X_t, \mu_{b_i,b_m}, \Sigma_{b_i,b_m}) \tag{3}$$

The probability formula is similar to GMM, but the weight ω_{b_i,b_m} and Gaussian probability density function

$\eta(X_t, \mu_{b_i,b_m}, \Sigma_{b_i,b_m})$ are no longer sensitive to the time t. The weight ω_{b_i,b_m} , the mean value μ_{b_i,b_m} and covariance matrix Σ_{b_i,b_m} result from$\{F_{b_1}, ..., F_{b_m}\}$. Then the Gaussian probability density function is defined as follows

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_{b_i,b_m})^T \Sigma^{-1}(X_t - \mu_{b_i,b_m})} \tag{4}$$

Thus, the value prediction of each pixel is characterized. The rest of the process is similar to [17]. Since selected keyframes cover frames at any time in a video clip, this model will not be updated for new frames over time.

In order to detect foreground area of each keyframe, we use this model to go through the whole video clip. Any pixel which is outside the match of this model is labelled as the foreground area. Then the overall foreground area in a keyframe is obtained. All of the connected components in foreground areas are extracted by the two-pass, connected components algorithm [18]. Finally, the keyframes with connected components are clustered for summarization.

2) Clustering: The K-means algorithm is widely used to partition data into a specific number of clusters. The processing speed of K-means is generally faster than hierarchical clustering. As there are only two kinds of keyframes (with objects and without objects), we use K-means to cluster keyframes with K value of 2. For each keyframe, the width and length of its largest connected component in foreground area are two dimensions of a K-means vector.

Based on the above definition, the clustering is performed. Because K is stable in this context, the clustering involves no diagnostic checks for determining a better K. This improves the performance of the clustering.

After clustering, vectors are divided into two groups. We consider that the group with objects have greater widths and lengths. More importantly, the exact coordinates of keyframes in this group are used to segment images for summarization. Upon a retrieval quest, we stitch segments for a specific object or a group of objects. Applications can specify how the process of image stitching is performed. Finally, we map each region of the summarization image to the jump action of a time point in the video player. The way of interaction is also open-ended for different applications and devices, such as mouse clicks and touching screen clicks.

IV. Experimental Results

This section presents experiments carried on an example video clip. Five items on a conveyor belt are recorded in the video content. Based on IoT information, we first extract foreground areas from keyframes and compare the proposed approach to GMM. Then we cluster the keyframes with foreground areas by using the K-means algorithm. Finally, we segment and stitch keyframes for summarization.

Fig. 5(a) shows one of input frames. The extracted foreground areas based on GMM and our method are shown in Fig. 5(b) and Fig. 5(c), respectively. As we can see, the objects are successfully extracted by both methods. Also, the shadow of the object affects two methods due to the lack of shadow

(a) Original image (b) foreground area detected by GMM (c) foreground area detected by our method

(d) item region detected by GMM

(e) item region detected by our method

Fig. 5. Foreground detection

elimination [19]. However, there is difference shown in Fig. 5(d) and Fig. 5(e), which are parts of Fig. 5(b) and Fig. 5(c). We observe that the conveyor belt creates significant noise in GMM, because of poor prediction for irregular movement of the conveyor belt. Comparatively, the proposed approach reduces most of noise of the conveyor belt by considering several keyframes later than the extracted keyframes. This indicates that applying IoT information is a feasible approach to improve video tracking of objects in supply chains.

We analyze the extraction results of GMM and our method from five keyframes, where five different objects in the video clip are included. As shown in Fig. 6, about 84% noise of

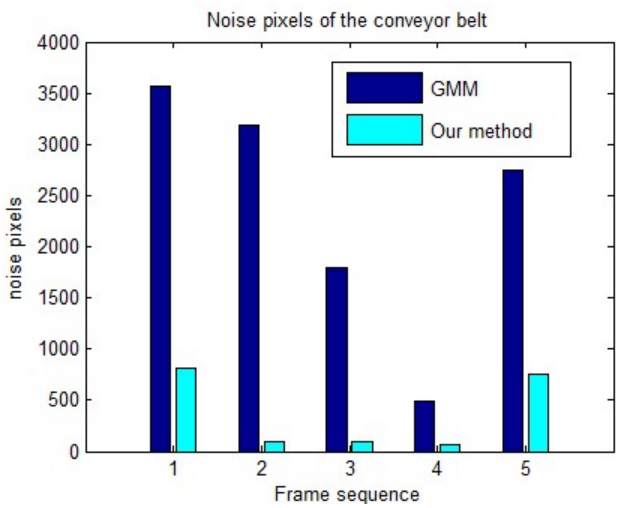

Fig. 6. A comparison of background noise pixels from the conveyor belt

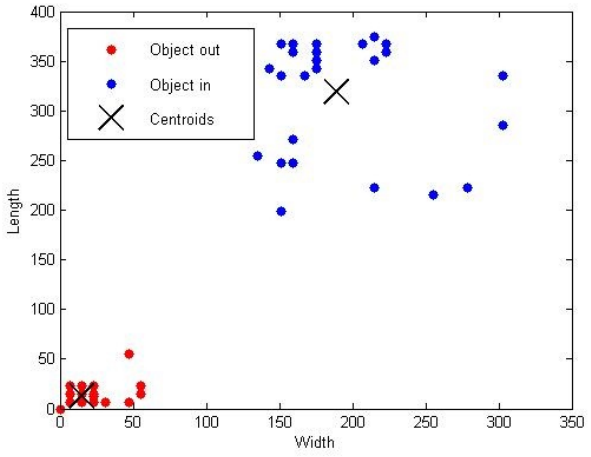

Fig. 7. Clustering result

Fig. 8. Video summarization result

the conveyor belt is eliminated by applying IoT information to GMM. At a lower level, the proposed work has less noise than GMM in every testing set. In spite of dynamic estimation for new frames, GMM lacks the recognition of the fast moving conveyor belt. In some circumstances, part of conveyor belt and objects may be mixed. However, this result shows that the accuracy of GMM can be improved. By applying IoT information to select keyframes for background models, our method extremely mitigates the effect of the fast moving conveyor belt. In addition, the memory and computational power requirements of the proposed approach are relatively lower.

After the value of K is set to 2, we execute the K-means algorithm to cluster keyframes with the extracted foreground areas. Fig. 7 shows the distribution of the clustering results. There is a clear contrast between the foreground areas in frames with objects and frames without objects. Hence, the K-means algorithm achieves desirable convergence. Based on the results, it can be seen that the K-means algorithm is feasible to cluster keyframes for our method.

Then we segment and stitch the foreground areas in keyframes according to the clustering results. As expected, foreground areas are integrated as a summarization image showing the features of an object. Fig. 8 shows an example of the summarization result. Without the redundant areas, the summarization image gives users a compact overview of an object in the supply chain. For different applications, the summarization image may contain one object or a group of objects.

From all the above figures, these experiments indicate that the combination of RFID technology and GMM is effective for foreground area extraction and clustering. More importantly, this means that the presented work enables users to quickly browse large video content for tracking objects in the supply chain.

V. Conclusion and Future Work

Object tracking in IoT provides critical capabilities for supply chain management in smart cities. With the widespread use of object tracking techniques in IoT, the integration of RFID and video technology now becomes a crucial part of object tracking applications. As a complement to this integration, in this paper we present a novel method for video summarization of object tracking in IoT. We apply IoT information to keyframe selection during the construction of background models. Using the K-means algorithm, we cluster and stitch extracted foreground areas in keyframes. Users are allowed to scan the video content through a compact

summarization image and the RFID serial number of an object. The experimental results show that the proposed method is suitable for visual object tracking in IoT-based supply chain management.

In principle, our work could also extend conventional video summarization approaches with environmental information from other sensors (temperature, sound and pressure). More useful features outside video images might improve the quality of video summarization, as long as the semantic relevance between sensor data and video images is determined.

Based on the infrastructure of IoT, future work will include video summarization techniques for object tracking with other kinds of sensors. We also plan to design multi-scale interactive user interface for object tracking.

ACKNOWLEDGMENT

This work was partially supported by NSF of Shanghai under Grant 13ZR1422500.

REFERENCES

[1] H. Hsu, Z. Cheng, T. Huang and Q. Han, "Behavior analysis with combined RFID and video information," *Proc. 3rd International Conference on Ubiquitous Intelligence and Computing*, Wuhan, China, 2006.

[2] D. Zhang, J. Zhou, M. Guo, J. Cao and T. Li, "TASA: Tag-free activity sensing using RFID tag arrays," *IEEE Trans. Parallel and Distributed Systems*, vol. 22, no. 4, pp. 558-570, April, 2011.

[3] N. Huber and K. Michael, "Minimizing Product Shrinkage across the Supply Chain using Radio Frequency Identification: a Case Study on a Major Australian Retailer," *Proc. 6th International Conference on Mobile Business*, Toronto, Canada, 2007.

[4] C. Wang and L. Cheng, "RFID & vision based indoor positioning and identification system," *Proc. IEEE 3rd International Conference on Communication Software and Networks (ICCSN)*, Xi'an, China, 2011.

[5] N. Krahnstoever, J. Rittscher, P. Tu, C. Kevin and T. Tomlinson, "Activity recognition using visual tracking and RFID," *Proc. IEEE 7th Workshops on Application of Computer Vision(WACV/MOTIONS05)*, Washington, DC, USA, 2005.

[6] F. M. Hasanuzzaman, Y. Tian and Q. Liu, "Identifying medicine bottles by incorporating RFID and video analysis," *Proc. IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Atlanta, Georgia, USA, 2011.

[7] P. Huang, R. Sawhney, D. Walker, K. Wallen, A. Bobick, S. Qin and T. Balch, "Learning a projective mapping to locate animals in video using RFID," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Algarve, Portugal, 2012.

[8] T. Germa, F. Lerasle, N. Ouadah and V. Cadenat, "Vision and RFID data fusion for tracking people in crowds by a mobile robot," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp.641-651, June, 2010.

[9] Y. Gong and X. Liu, "Video summarization using singular value decomposition," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, SC, USA, 2000.

[10] H. Wang and C. Ma. "Interactive multi-scale structures for summarizing video content." *Science China Information Sciences*, vol. 56, no. 5, pp. 1-12, May, 2013.

[11] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose and J. Rehg, "A scalable approach to activity recognition based on object use," *Proc. IEEE 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007.

[12] R. Cucchiara, M. Fornaciari, A. Prati and P. Santinelli, "Mutual calibration of camera motes and RFIDs for people localization and identification," *Proc. 4th ACM/IEEE International Conference on Distributed Smart Cameras*, Atlanta, GA, USA, 2010.

[13] H. Ueda, T. Miyatake, S. Sumino and A. Nagasaka, "Automatic structure visualization for video editing." *Proc. INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*, New York, NY, USA, 1993.

[14] D. B. Goldman, B. Curless, D. Salesin and S. M. Seitz, "Schematic storyboarding for video visualization and editing," *ACM Trans. on Graphics (TOG)*, vol. 25, no. 3, pp. 862-871, July, 2006.

[15] C. Barnes, D. B. Goldman, E. Shechtman and A. Finkelstein, "Video tapestries with continuous temporal zoom," *ACM Trans. on Graphics (TOG)*, vol. 29, no. 4, pp. 89:189:9, July, 2010.

[16] C. Nguyen, Y. Niu and F. Liu, "Video summagator: an interface for video summarization and navigation." *Proc. SIGCHI Conference on Human Factors in Computing Systems*, Austin, TX, USA, May, 2012.

[17] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747-757, August, 2000.

[18] B. K. P. Horn, *Robot vision*. Cambridge, MA: MIT press, 1986, pp. 66-69, 299-333.

[19] Z. Tang and Z. Miao, "Fast background subtraction and shadow elimination using improved gaussian mixture model," *Proc. IEEE International Workshop on Haptic, Audio and Visual Environments and Games*, Ottawa, Ontario, Canada, 2007.