2018

Video analytics system for surveillance videos

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Thesis

VIDEO ANALYTICS SYSTEM FOR SURVEILLANCE VIDEOS

by

YANNAN BAI

B.S., Shanghai Jiao Tong University, 2016
M.S., Boston University, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2018

Approved by

First Reader

Venkatesh Saligrama, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Computer Science

Second Reader

David A. Castañón, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Third Reader

Wenchao Li, Ph.D.
Assistant Professor of Electrical and Computer Engineering
Assistant Professor of Systems Engineering

Acknowledgments

Thank you Prof. Saligrama for his support and help in my thesis. I am proud to work with Prof. Saligrama and his great research team.

Thank you Hanxiao Wang and Fred Feng for providing expertise and help with the research project.

Thank you Prof. Castanon and Prof. Li for being on my Master's Thesis Committee. Thank you to all the professors I met at Boston University.

Thank you my Mother and Father for the love and care. The support of my family allows me an opportunity to pursue a Master's in Electrical and Computer Engineering.

Yannan Bai

Student

ECE Department

VIDEO ANALYTICS SYSTEM FOR SURVEILLANCE VIDEOS

YANNAN BAI

ABSTRACT

Developing an intelligent inspection system that can enhance the public safety is challenging. An efficient video analytics system can help monitor unusual events and mitigate possible damage or loss. This thesis aims to analyze surveillance video data, report abnormal activities and retrieve corresponding video clips. The surveillance video dataset used in this thesis is derived from ALERT Dataset, a collection of surveillance videos at airport security checkpoints.

The video analytics system in this thesis can be thought as a pipelined process. The system takes the surveillance video as input, and passes it through a series of processing such as object detection, multi-object tracking, person-bin association and re-identification. In the end, we can obtain trajectories of passengers and baggage in the surveillance videos. Abnormal events like taking away other's belongings will be detected and trigger the alarm automatically. The system could also retrieve the corresponding video clips based on user-defined query.

Contents

List of Tables

List of Figures

List of Abbreviations

ALERT	.....	Awareness and Localization of Explosives Related Threats
CLASP	.....	Correlating Luggage and Specific Passengers
TSA	.....	Transportation Security Administration
CNN	.....	Convolutional Neural Networks
SSD	.....	Single Shot Multibox Detector
YOLO	.....	You Only Look Once
RPN	.....	Region Proposal Network
mAP	.....	mean Average Precision
RoI	.....	Region of Interest
IoU	.....	Intersection over Union

Chapter 1

Introduction

1.1 Motivation

As cameras have been installed in many public places like street corners, stations and airports, the amount of surveillance video keeps increasing. Currently these surveillance videos are mainly used for forensic evidence, however, they can be used to monitor unusual events and mitigate possible damage or loss. While more than four billion hours of surveillance are captured every week in the U.S., which makes it impossible for humans to inspect all the screens. Thus the intelligent surveillance system could filter the large amount of videos and narrow it down to a few video clips, so that the security personnel could focus on a couple of people and the abnormal events.

The goal of this thesis is to develop a video analytics system. Given the surveillance videos as input, the system will trigger alert for suspicious activities and retrieve video clips based on user-defined query. The surveillance videos in this thesis come from ALERT Airport Re-Identification Dataset. As part of the ALERT video analytics effort, the dataset was constructed using video data from the surveillance cameras installed post central security checkpoint at an active commercial airport within the United States.

The system is composed of detection module, tracking module, person-bin association module, re-identification and video retrieval module. Though these problems

have been studied and great progress has been achieved, they are hardly considered in a pipeline for the surveillance applications. In this thesis, the video analytics system is accomplished by first preprocessing the video data via object detection and tracking. The output bounding boxes of person and bin are associated with each other based on spatio-temporal and appearance information. Integrated with cross-camera re-identification, abnormal activity such as a passenger takes baggages that are not his own belongings will trigger the alert for a possible theft. The system could also retrieve the trajectory of a specific passenger based on an input query.

1.2 Related Work

Many approaches have been proposed for video analysis and retrieval. Traditionally, video-based retrieval are implemented based on summarization and scene understanding (Castanon et al., 2012). Summarization methods search for a set of viewpoint invariant region descriptors. Scene-understanding methods focus on classifying observed activities in terms of activities learned from a training video. But these approaches don't work well on surveillance videos because there are no scene transitions or a given knowledge of activity classes contained in the video.

Recent advances in deep learning models makes it a game changer in area of computer vision. Deep learning based methods of object detection and re-identification achieve better efficiency and accuracy. We want to improve the performance of smart surveillance system by integrating these methods. In this thesis, the proposed video analytics system employs the state-of-art object detection networks Faster R-CNN for extracting objects as in (Ren et al., 2017) and Simple Online and Realtime Tracking (SORT) for multiple object tracking as in (Wojke et al., 2017). Such methods along with re-identification allow the system to track people and bins in the cross-camera

surveillance videos. Once the activity of one picking up others' belongings is recognized, the system can automatically trigger the alert and avoid a possible theft.The features of the objects are also extracted and stored in database so that a specific object can be indexed based on user-defined descriptions.

Chapter 2

System Design

2.1 System Block Diagram

The overall goal of the video analytics system is to automatically trigger the alerts of possible theft and retrieve the video clips of user-defined query. Figure 2·1 shows the block diagram of the surveillance video analytics system.

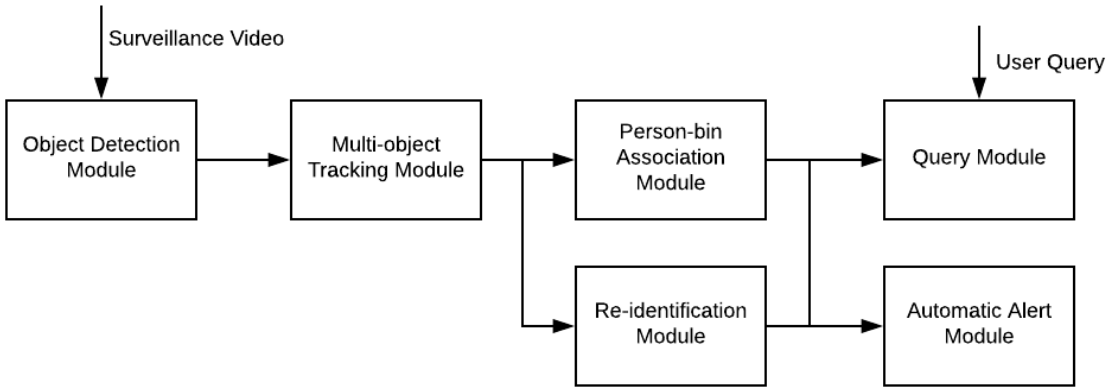


Figure 2·1: Block diagram of Video Analytics System

The input videos are collected from surveillance cameras at airport security checkpoints. The videos are processed to detect moving objects of interest, in our case are people and bins. These objects are tracked as they move around in the checkpoints. The cross-camera tracking is solved by re-identification algorithm. The ownership

between person and his or her baggages is determined by the spatio-temporal relationship and hand movements. After we associate the person with the plastic bins, re-identification results help detect suspicious activity, such as picking up other's belongings, and trigger the alert automatically. To help locate the target passenger, the system could also retrieve the respective video clips with the appearance of user-input query object.

Compared with the traditional smart surveillance system, our system employs the state-of-art deep learning algorithms and improve the accuracy at each step of the system.

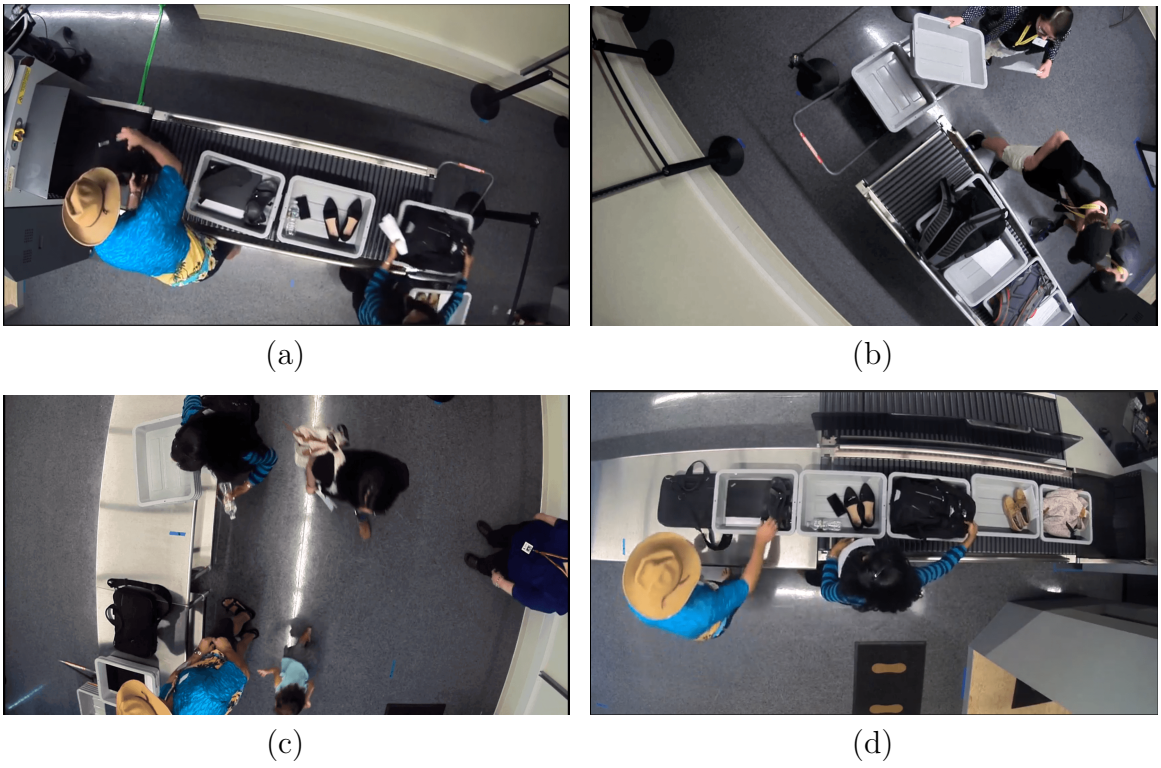
2.2 Dataset Introduction

The ALERT dataset is video data collected from indoor surveillance cameras installed post central security checkpoint at an active commercial airport within the United States(Karanam et al., 2016).

The dataset comes from CLASP (Correlating Luggage and Specific Passengers) project. CLASP primarily focus on using video technologies to assist the Transportation Security Administration (TSA) in effectively identifying security incidents like theft of items, or bags left behind at the checkpoint. The dataset contains video clips collected from 13 cameras installed at a mock airport security checkpoint that simulates real-world conditions. Each camera has a frame size of 1920×1080 pixels and the video is captured at 30 frames per second.

Up to now, 18 experiments have been conducted on this scenario, including activities like passengers dropping off baggages, walking through the metal detector and picking up items. Sample images of the dataset is shown in Figure 2·3. The first row shows activities of passenger divesting their items in a plastic bin and placing it di-

rectly on conveyor. The second row shows activities of passenger picking up divested items at collection table. Currently we only focus on the cameras that have a clear view of the conveyor as shown in Figure 2·3 (a) and (d) so that we can associate person-bin relationship and detect possible theft of items. The rest of cameras could be used later to keep track of a certain passenger.



(a) (b)

(c) (d)

Figure 2·2: Sample images in the ALERT dataset

2.3 Object Detection

The first step of our video analytics system is to detect the moving objects in surveillance videos. Conditions such as pose variation, occlusions and illumination changes make it a challenging problem for object detection. Traditionally, the object detection methods can be categorized into four forms: Background Subtraction,

Frame Differencing, Temporal Differencing and Optical Flow (Kulchandani and Dangarwala, 2015). We employed the background subtraction method on our dataset and it works badly. In the surveillance video, some passengers tend to stand in front of the conveyor without any movement for a long term, which makes it difficult to separate the passengers from the background. Fig 2·3 shows a mis-detection of this case. Besides, the background subtraction method classifies the objects into two groups of foreground and background. Since some of the passengers in foreground are close to each other, blob detection cannot recognize each identity and segment the foreground into bounding boxes of objects.

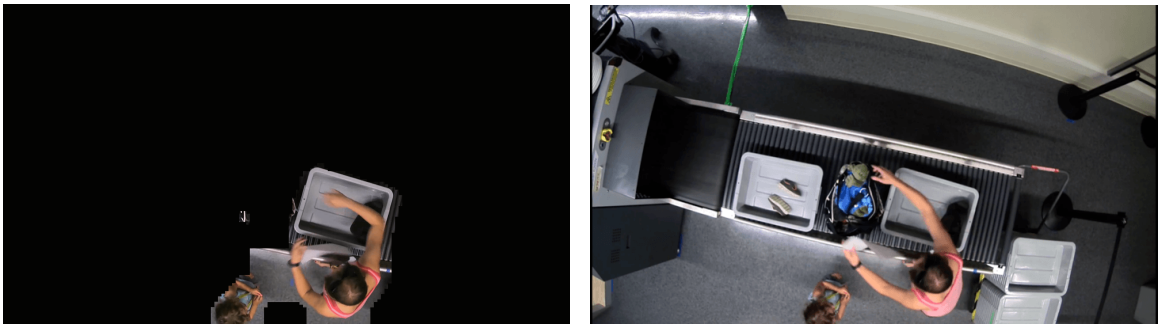


Figure 2·3: Bad results of Background Subtraction method

Recent progress of deep learning models leads to the great success in modern object detectors based on convolutional neural networks. The deep learning based detection approaches can be distinguished to two parts: two-stage detection approach like the family of R-CNN and single-shot detection approach like Single Shot Multibox Detector(SSD) (Liu et al., 2016), You Only Look Once(YOLO) Detector (Redmon and Farhadi, 2016).

Faster-RCNN (Ren et al., 2017) is the most popular method from the first part, where region proposals are generated first, then these bounding boxes will be classified

and regressed in the second stage. In comparison, SSD skips the step of region proposal and uses multiple convolutional layers to get bounding boxes of various scales in every location. The prediction and classification of the bounding box are completed in one step and the Non-Maximum Suppression technique is used to merge all detections of the same object. YOLO is similar to SSD except that the image is divided into a grid and the bounding boxes are fixed.

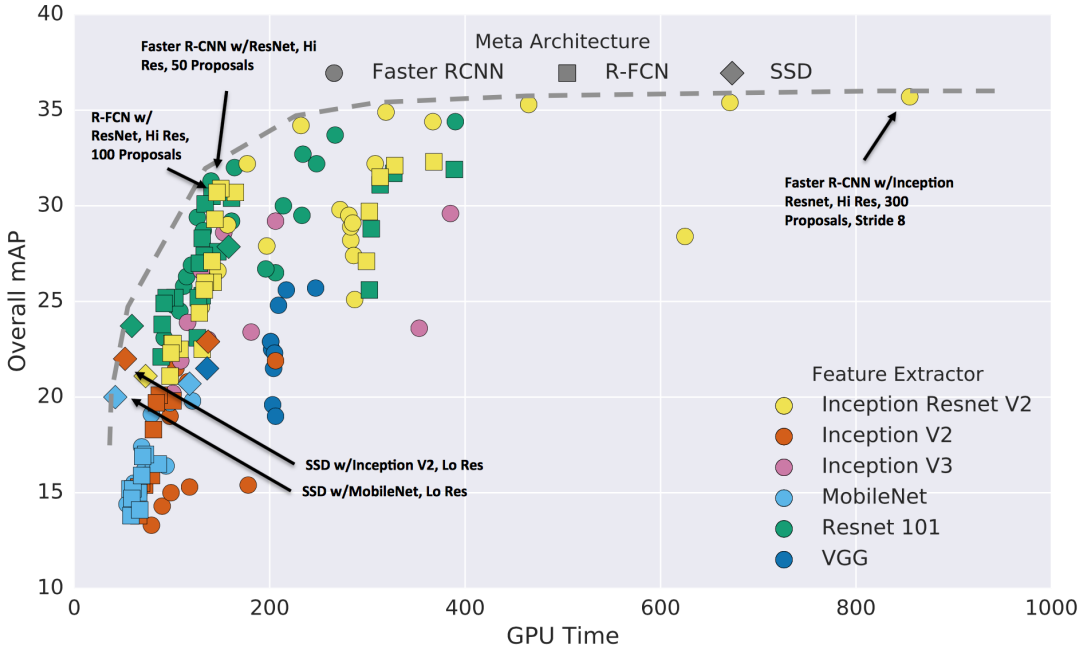


Figure 2·4: Accuracy vs time, with marker shapes indicating meta-architecture and colors indicating feature extractor. (Huang et al., 2017)

Figure 2·4 is a comprehensive visualization of the trade-off between accuracy and speed. Generally SSD and YOLO models are faster on average while Faster R-CNN achieves more accurate results at the cost of slower speed. For the same detector, the complexity of Convolutional Neural Networks(Feature Extractor) also affects the detector's speed and accuracy. Table 2.1 shows some detailed statistics of the speed and

mAP(mean Average Precision) performance on COCO dataset for different models. The architecture of ResNet-101 achieves a better balance between speed and accuracy.

Table 2.1: Speed and mAP of different models for detection task on COCO dataset. Fig.2·4 is a visualization of the performance.

Model name	*Speed(ms)*	*COCO mAP*[∧1]
SSD (Inception-v2)	42	24
Faster R-CNN (Inception-v2)	58	28
Faster R-CNN (ResNet-101)	106	32
Faster R-CNN (Inception-ResNet-v2)	620	37
Faster R-CNN (nas)	1833	43

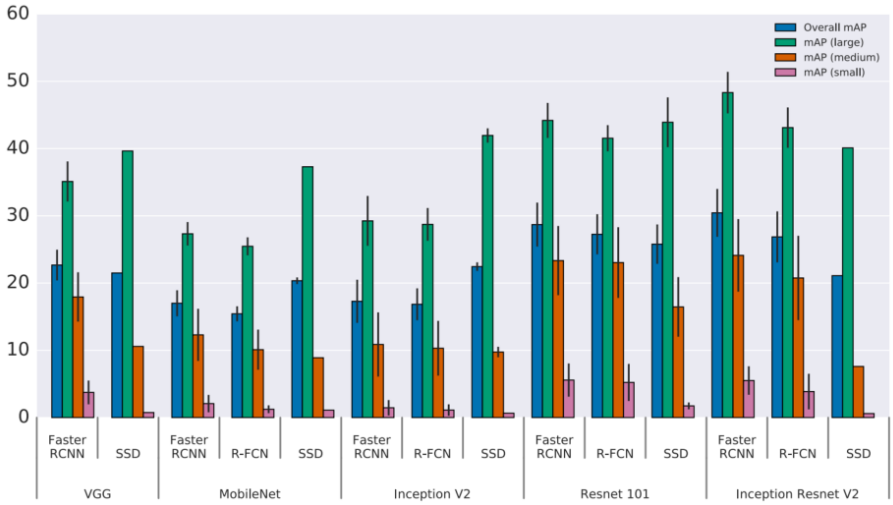


Figure 2·5: Accuracy stratified by object size, meta-architecture and feature extractor. (Huang et al., 2017)

Figure 2·5 shows the performance of each model on various sizes of objects. Faster R-CNN works better on small and medium size objects because the small sized object won't cover many anchors and is hard for SSD to detect it. Besides, the structure of the feature extractor(convolutional neural networks) can affect the performance. The more complicated and deeper the feature extractor is, the more accurate the

detection results is, while the more time it may cost.

In our dataset, passengers and bins occupy small part of the image. And object detection is the first step of the system where high accuracy is required. After comparing these three popular object detection models and its feature extractors, we choose Faster R-CNN with ResNet-101.

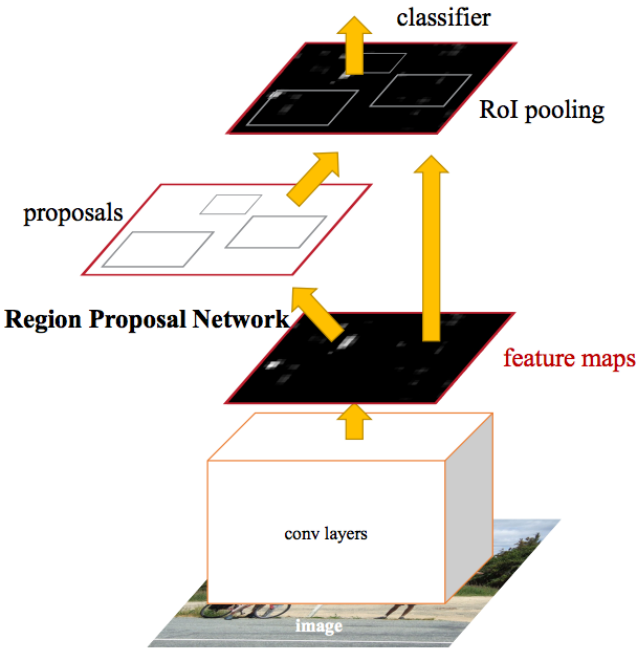
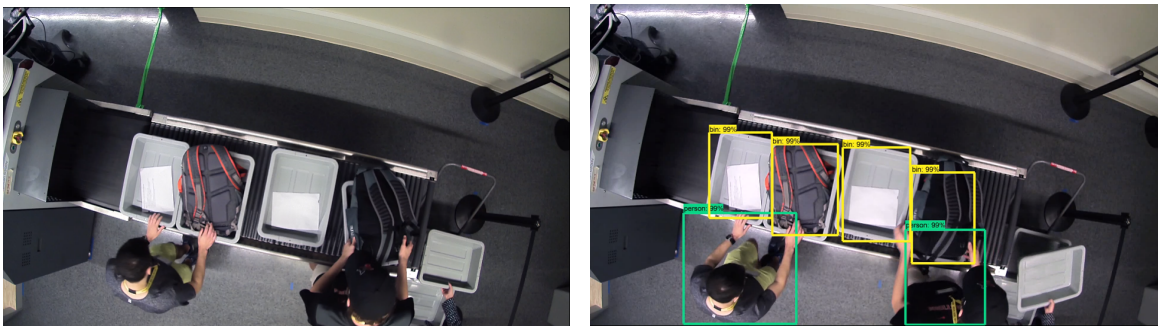


Figure 2·6: Faster R-CNN is a single, unified network for object detection. (Ren et al., 2017)

As shown in Figure 2·6, an input image passes through a pre-trained CNN and computes its feature maps. A Region Proposal Network(RPN) uses the features to generate a number of bounding boxes(region proposals) that may contain objects and corresponding objectness scores for its possibility. After that, Region of Interest (RoI) Pooling is applied to extract the features of these bouding boxes. Finally, the R-CNN module will output classifications and tightened bouding box coordinates.

In this thesis, we use ResNet-101 network pretrained on MS COCO dataset(Lin

et al., 2014) as convolutional layer to extract the feature map. MS COCO dataset contains 83K training images for 80 object classes with labels of loacation and class annotated by humans. Considering the MS COCO is a web-based image dataset different from our airport domain, the ALERT dataset contains a new class of plastic bins, thus we manually annotate videos shot by one camera and fine-tune ResNet-101 on these labeled samples. Figure 2·7 is an example image of detected passengers and bins from our airport surveillance dataset.



(a) (b)

Figure 2·7: An example of detected objects in ALERT dataset

Table 2.2: Faster R-CNN detector performance

	Person				Bin			
Video Sequence	*TP*	*FP*	*Precision*	*Recall*	*TP*	*FP*	*Precision*	*Recall*
A9_C9	189	34	84.7	83.3	124	52	70.4	68.9
A9_C11	235	26	90.0	88.2	342	51	87.0	86.0

2.4 Multi-Object Tracking

Given the bounding boxes produced by object detector, we use multi-object tracking algorithm to associate objects across video frames and yield tracklets for each identity. Generally, there are two types of tracking method depending on the way

to associate obervations: Bayesian theory based and data association based. In (Fan et al., 2016), the author briefly reviewed the Bayesian theory-based methods like Kalman filter and Particle filter, the data association-based methods like Hungarian algorithm and network flow. Since the number of objects in our dataset is relatively small, we use Kalman filter(Kalman, 1960) as tracking method for its efficiency.

In our dataset, the state of each object is defined as an eight-dimension vector $[x, y, w, h, v_x, v_y, v_w, v_h]$, which contains centroid coordinate (x, y), width and height of the bounding box (w, h) and their respective velocities. We use a standard Kalman filter to predict the object's new location and gain its trajectory. The cost function to assign the predicted Kalman state to newly arrived detections is defined as their Mahalanobis distance. When the minimum value of cost function is found, we use the newly arrived detected bounding box to update parameters of Kalman filter motion model, and use it as the input in the next frame. Repeatedly doing this to finish the model update until the moving objects disappeared(Li et al., 2010).

To better handle the problem of occlusion and rapid displacements, appearance information is introduced into the cost function(Wojke et al., 2017). We compute the cosine distance between the predicted Kalman state and newly arrived detection in appearance space using the feature map extracted by the convolutional layer of ResNet-101. Both metrics are combined into the cost function so that the tracker could take into consideration both location information and appearance information.

Table 2.3: Multi-object tracking performance

	Person				Bin			
Video Sequence	*TP*	*FP*	*MOTP*	*Recall*	*TP*	*FP*	*MOTP*	*Recall*
A9_C9	177	40	81.6	83.3	111	57	66.07	68.9
A9_C11	231	29	88.8	88.2	338	53	86.4	86.0

2.5 Person-bin Association and Re-identification

Now we have the tracklets for every identity. To trigger the alert for a possible theft, the person-baggage relationship should be matched. Observing the surveillance videos, it turns out passengers are usually close to their own belongings at checkpoints. The distance between bounding boxes of person and bin over the same period is calculated and ranked by the score. If a passenger and a baggage appear and disappear almost over the same time slot during which they are close to each other, it is more likely that the baggage belongs to the passenger. The plastic bin is matched to the person based on the temporal-spatial distance.

Location matching works for most of the person-bin association. However, the constant posture change and occlusions lead to miss detection, which increases the error rate of the person-bin association. A hand detector is trained to complement the location matching. An object detector(Faster R-CNN) pretrained on ImageNet is finetuned on a hand dataset (Mittal et al., 2011). The dataset contains a total of 13050 annotated hand instances and Figure 2·8 shows a couple of sample images from the hand dataset.

Figure 2·8: Sample images of the hand dataset. (Mittal et al., 2011)

The hand detector will generate bounding boxes of hands in the videos. Combined with the bounding boxes of passengers and bins obtained in previous step, two score matrices for hand-person and hand-bin relationship are calculated based on the intersection-over-union or spatial distance if not overlapped. For each bounding box

of the hands, if the highest ranking scores in the two matrices are above a certain threshold, it indicates an association between the person and the plastic bin with a mediation of the same hand. Such association is given the first priority for videos shot at the security entrance since passengers need to put their baggage on the conveyor belt.

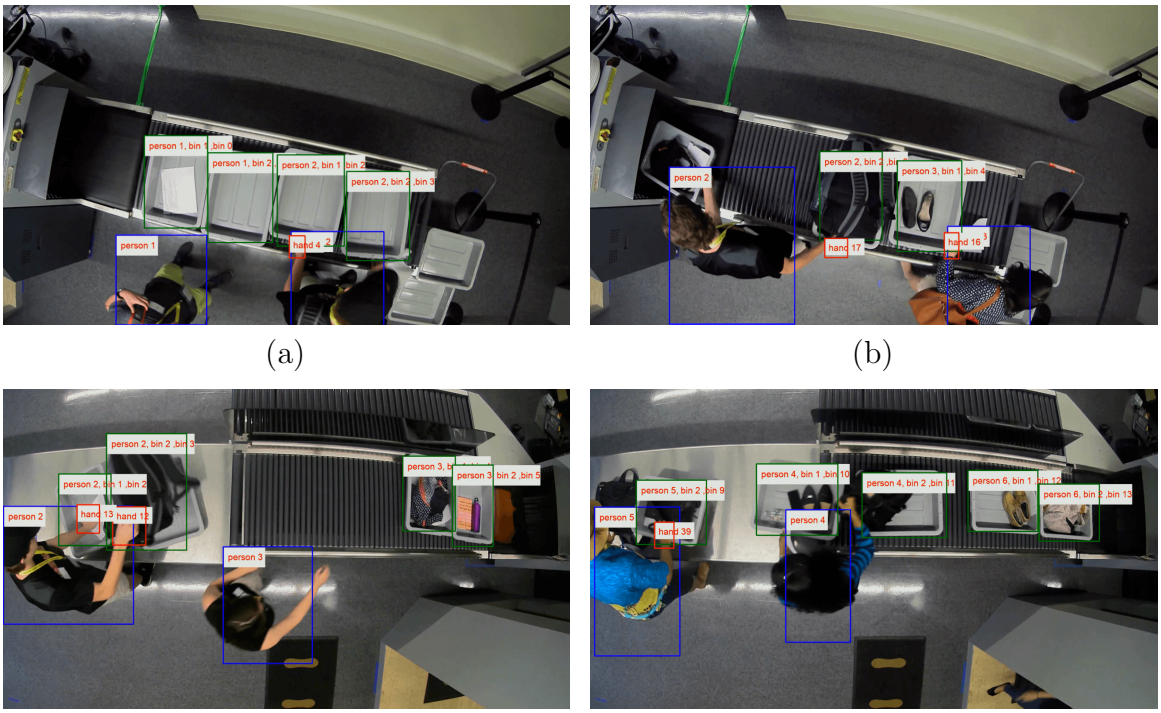


Figure 2·9: Examples of the tracking and association

2.6 Re-identification

Since the video data is collected from multiple cameras installed at the security checkpoint, before retrieving the video clip of a passenger or detect the possible theft, we need to recognize the same person from disjoint camera views, which is know as person re-identification.

Person re-identification across multiple cameras is a challenging problem because of the pose changes, illumination variations and image scaling. Its importance in helping obtain complete trajectories from surveillance camera network in public places draws many attention and a lot works have been done in this area.

The person re-identification problem mainly focuses on two parts: feature extraction and similarity measure. In the first part, color, texture and shape are the appearance features commonly used in the state-of-the-art methods for person re-identification (Mazzon et al., 2012). Many effective features of these three types have been proposed, color features like LAB, HSV and ELF, texture features like LBP and Gabor. Once a suitable feature representation is obtained, metric learning algorithm will be chosen to compare the similarity distance and associate the candidate objects across cameras. Euclidean distance, Mahalanobis distance and L1-Norm are often used to measure similarity between the features directly. An alternative approach is to train a classifier to learn the cross-camera feature difference. Methods like Large Margin Nearest Neighbor(LMNN) (Weinberger and Saul, 2009) and Probabilistic Relative Distance Comparison(PRDC) (Zheng et al., 2011) are more robust to appearance changes and less sensitive to feature selection.

Instead of the low-level features, deep neural networks is used in this thesis to extract deep features, which make full use of the large-scale dataset. As discussed in (Wang et al., 2017), Convolutional Neural Networks present an effective way of feature extraction for person re-identification.

(Wang et al., 2017) use CNN structure to extract pedestrian feature and apply it to the Cosine Distance method directly. Figure 2·10 shows the CMC curve and rank-1 identity rate of the proposed Feature-Net architecture. They beat most of the traditional methods on three challenging and common person re-identification

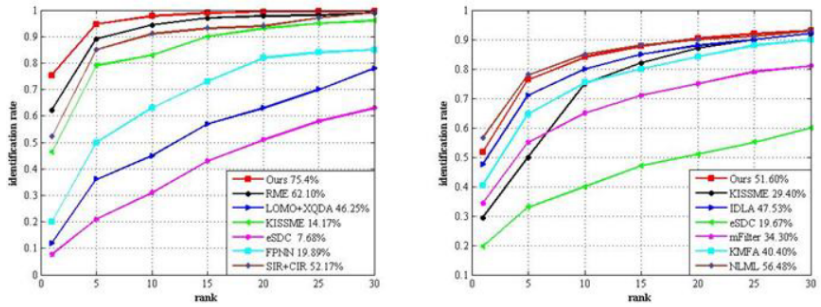


Figure 2·10: CMC curves and rank-1 identification rates on CUHK03 dataset (left) and CUHK01 (right) dataset. (Wang et al., 2017)

datasets (CUHK03, CUHK01, VIPeR), especially on CUHK03 they obtained the state-of-the-art result with rank-1 accuracy of 75.4%.

Considering the effectiveness of deep learning algorithms in re-identification problem, we use Inception-v3 network (Szegedy et al., 2016) pretrained on ImageNet dataset in the stage of feature extraction. Compared with the VGG model used in (Wang et al., 2017), the Inception-v3 model achieves better performance in classification with lower computational cost. Given the bounding boxes provided by the Faster R-CNN detector, the Inception-v3 network extracts features of the cropped images and store the features with the bounding boxes information. Considering deep convolutional neural networks usually learn a common feature that can be transfered to a target domain different from training dataset, which is also known as transfer learning, we employ the same Inception-v3 network to extract features of the bins. The deep neural network is pretrained on ImageNet dataset, thus it could extract a common feature that is eligible for the re-identification of the plastic bins.

In the stage of metric learning, the Cosine distance is calculated between every two sets of features extracted from two cropped object images. Since the whole trajectory of each object has been obtained through multi-object tracking, the mean value of

Cosine distance between every pair of objects is computed. Smaller value of Cosine distance indicates a higher similarity between the two objects. Thus, each person is matched with its highest ranking identity from a different camera.

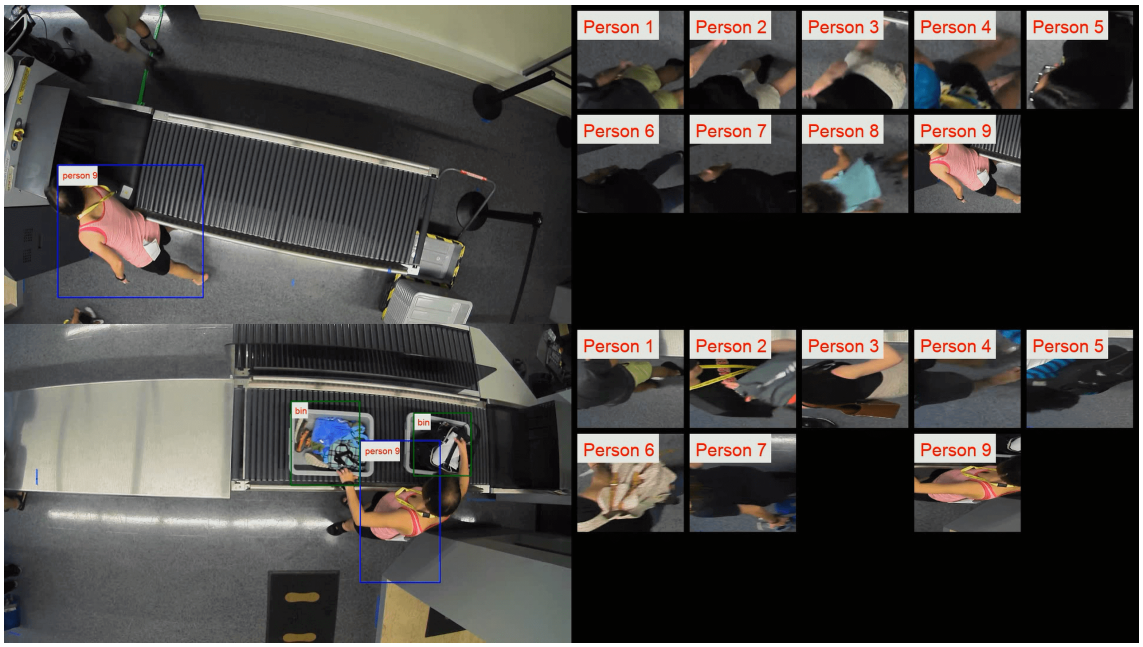


Figure 2·11: Results example of person re-identification between camera 9 and 11 in ALERT surveillance dataset

Figure 2·11 shows a sample result of the person re-identification for two surveillance videos from different viewpoint in the airport security checkpoint. As we can see, the videos are shot from top. Person 4 and 5 both wear a blue shirt and can be separated clearly. Since Person 8 is not detected in the bottom video, there is no matching identity for it.

We employ the same Inception-v3 network to extract features of the plastic bins. In the stage of metric learning, the time factor is also taken into account for the bin re-identification because the bins are transfered by the conveyor and always remain the same order. Figure 2·12 shows a sample result of the bin re-identification.

Figure 2·12: Results example of bin re-identification between camera 9 and 11 in ALERT surveillance dataset

As we can see from Figure 2·12, bins in two videos from different viewpoint have been associated with each other. The last bin in the bottom left video is associated with Bin 12 by mistake because the item has been taken away immediately and left with an empty plastic bin for a long time. Another problem happens during the re-identification is caused by the mis-detections in camera 9. The black computer bag are missed in camera 9 due to variations of viewpoint, thus it is associated with the highest ranking object Bin 9. Similarly, a broken trajectory in camera 11 also leads to a wrong re-identification. Table 2.4 shows the performance of our re-identification algorithm. Here accuracy is the ratio of correct matches to the total matches.

Table 2.4: Re-identification performance(rank-1)

Class Name	*# ID*	*# matches*	# misses	# mismatches	*Accuracy*
Person	10	8	1	0	100
Bin	16	18	1	3	83.3

2.7 Image query retrieval and automatic alert

The image query retrieval problem is similar to person re-identification in the sense that it will search through images captured by the cameras and generate a ranking list of the matching object IDs.

Given the input image of the query target, we search through all the objects in the video dataset and retrieve the object ID with the highest matching score. The trajectory of the query target is displayed in the result. Figure 2·13 shows the search results of the image query. Given the input of the surveillance video and an image query of the target, we can retrieve a ranking list of the matched object. Here we only show the object with highest similarity.

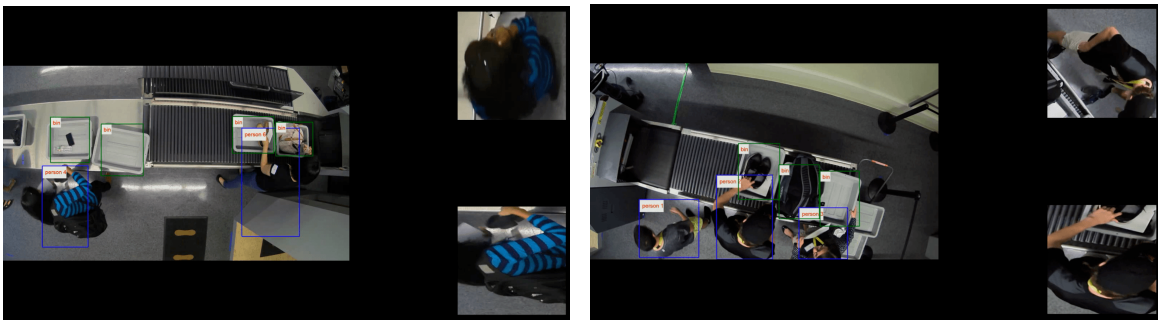
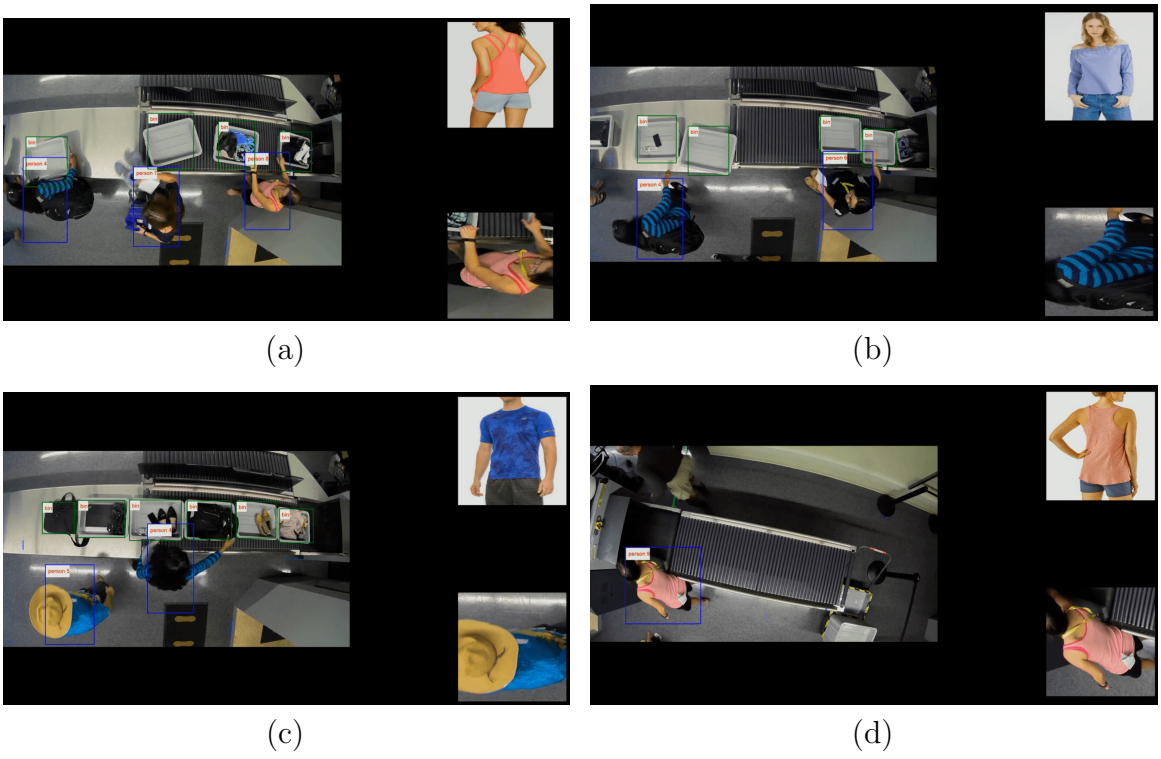


Figure 2·13: Results of the image query search

Figure 2·13 shows video retrieval of the image query shot from different camera viewpoints in the same dataset. The person can be correctly retrieved even if there exits another person with very similar appearance. In comparison, Figure 2·14 also

shows video retrieval results of some random images. The query images are downloaded from the Internet, which share the distinct attribute with a specific passenger from the surveillance dataset. Since it happens in real life that there is no clear image of the target we are looking for and the algorithm could capture the distinct features in the query image, such as a pink tank or a blue shirt, and retrieve the detected object with the highest similarity.



(a) (b)

(c) (d)

Figure 2·14: Results of the image query with similar appearance

The video analytics system could also generate behavioral alerts. These alerts are generated based on certain movement patterns and context information. (Hampapur et al., 2003) In our case, detecting suspicious behavior of picking up baggage that doesn't belong to their own could trigger alerts. Figure 2·16 shows the alert of a possible theft. Person 3 pick up the wrong divested item in Bin 8 which should

belong to Person 2.

Now our system is able to retrieve video clips based on user input image query. I am working on the attribute recognition module so that the system could process text query such as 'a woman in pink top'. An Inception-v3 model has been modified and trained on Market-1501 dataset (Zheng et al., 2015). Each identity is annotated with 27 labels of attributes, such as gender, age and clothing. However, the model failed on our ALERT dataset because the passengers in our dataset are shot from the top. Difference in viewpoint leads to the failure in attribute recognition. Further work will be done on the color extraction of the segmented passenger image so that we can obtain clothing color information. Such detailed attribute information will help match a target with user's text description and retrieve the corresponding video clips.

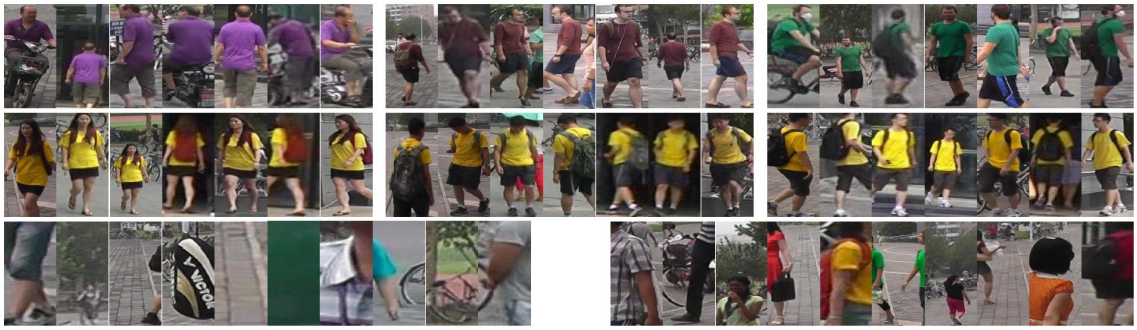


Figure 2·15: Sample images in Market-1501 dataset(Zheng et al., 2015)

To better illustrate the event of a possible theft, Figure 2·17 shows the screenshots from both cameras. As shown in the left picture, the man in black(Person 2) divested his phone in the plastic bin, while in the right picture, the woman in black(Person 3) took the phone out of the man's hand. In this event, the woman takes the phone which doesn't belong to her, which triggers the alert.

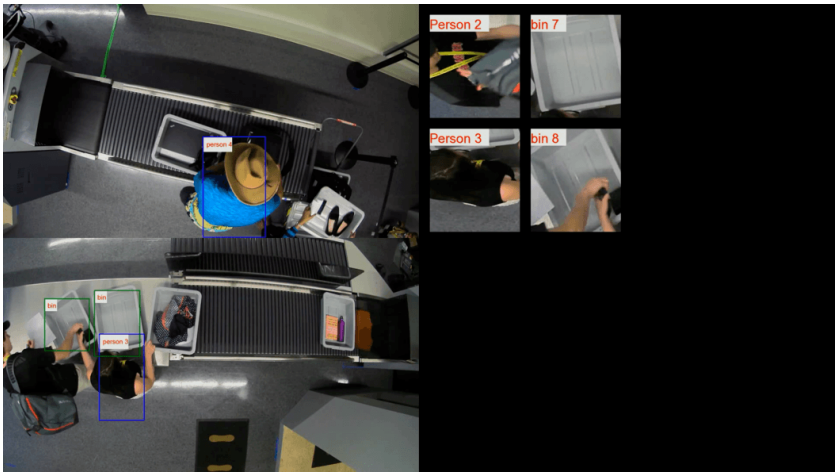


Figure 2·16: Results example of correct alerts in ALERT surveillance dataset

Figure 2·17: Screenshots from two cameras to illustrate the theft that triggers the alert

Figure 2·18 shows the final results of the triggered alerts. There happens some false alarm except for the correct alert of a possible theft shown in Figure 2·16. The objects in Bin 9 is missed in the object detection of the top left video. Person 8 is also missed in the object detection of the top left video. The miss rate of the detector leads to the false alarm. Thus our system has a high sensitivity of 100% and a low precision. Further work will be done in detector training to improve the precision.

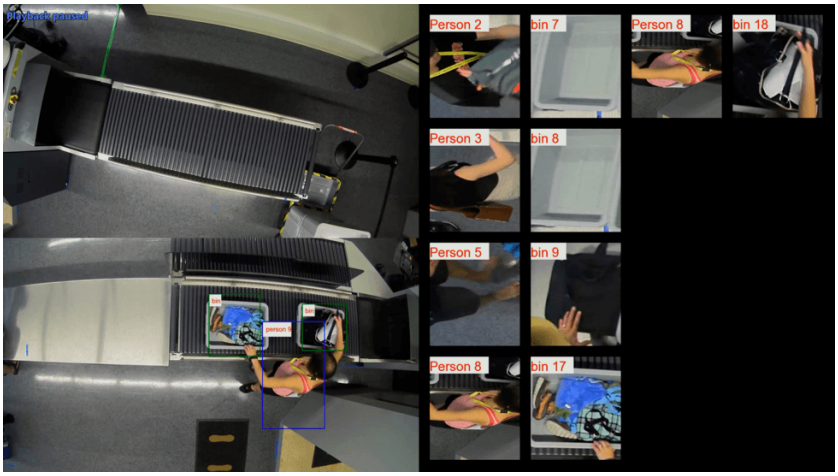


Figure 2·18: Results example of all triggered alerts in ALERT surveillance dataset

Chapter 3

Conclusions

3.1 Conclusion

In this MS thesis, we analyze the surveillance video dataset using deep learning algorithms. The input video is processed through a pipeline of detection, tracking, association and re-identification. In the end, the proposed system could generate alerts of suspicious activity and retrieve video clips that match user input queries.

As the explosive growth of video data and widely spread of surveillance cameras, our video analytics system will significantly reduce the labor work and eliminate the possible loss caused by suspicious activity. The video analytics system proposed in this thesis could effectively locate the possible thefts in the surveillance videos of checkpoints and retrieve the whole trajectory of a user-defined target. Such system could help the Transportation Security Administration(TSA) staff provide passengers with better security services and make full use of the security cameras.

Further work will be done on the attribute recognition part, we plan to segment bounding boxes of passengers so that attribute information can be extracted and generated as labels of clothing color, clothing style, gender and other information. The attribute recognition module could perfectly supplement the query module when users only have some general text description about the interested target(possibly a suspect) and retrieve the corresponding trajectory.

References

Castanon, G. D., Caron, A. L., Saligrama, V., and Jodoin, P.-m. (2012). Exploratory search of long surveillance videos. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 309–318. ACM.

Fan, L., Wang, Z., Cail, B., Tao, C., Zhang, Z., Wang, Y., Li, S., Huang, F., Fu, S., and Zhang, F. (2016). A survey on multiple object tracking algorithm. In *2016 IEEE International Conference on Information and Automation (ICIA)*, pages 1855–1862. IEEE.

Hampapur, A., Brown, L., Connell, J., Pankanti, S., Senior, A., and Tian, Y. (2003). Smart surveillance: applications, technologies and implications. In *Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia*, volume 2, pages 1133–1138. IEEE.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE Conference on Computer vision and pattern recognition (CVPR)*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., and Radke, R. J. (2016). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*.

Kulchandani, J. S. and Dangarwala, K. J. (2015). Moving object detection: Review of recent research trends. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–5. IEEE.

Li, X., Wang, K., Wang, W., and Li, Y. (2010). A multiple object tracking method using kalman filter. In *2010 IEEE International Conference on Information and Automation (ICIA)*, pages 1862–1866. IEEE.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single Shot Multibox Detector. In *European conference on computer vision*, pages 21–37. Springer.

Mazzon, R., Tahir, S. F., and Cavallaro, A. (2012). Person re-identification in crowd. *Pattern Recognition Letters*, 33(14):1828–1837.

Mittal, A., Zisserman, A., and Torr, P. H. S. (2011). Hand detection using multiple proposals. In *British Machine Vision Conference*.

Redmon, J. and Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger. *arXiv preprint arXiv:1612.08242*.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Wang, S., Wu, S., Duan, L., Yu, C., Sun, Y., and Dong, J. (2017). Person re-identification with deep features and transfer learning. In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC)*, volume 1, pages 704–707. IEEE.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244.

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *arXiv preprint arXiv:1703.07402*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*.

Zheng, W.-S., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *2011 IEEE conference on Computer vision and pattern recognition (CVPR)*, pages 649–656. IEEE.

CURRICULUM VITAE