

Distributed Spatial Transformer for Object Tracking in Multi-Camera

Sio-Kei Im
Macao Polytechnic University
Macao, China
marcusim@mpu.edu.mo

Ka-Hou Chan
Faculty of Applied Sciences
Macao Polytechnic University
Macao, China
chankahou@mpu.edu.mo

Abstract—Today’s video surveillance devices are almost ubiquitous and ready to collect data. Tracking targets in surveillance applications can be a challenging task. However, the work of manual surveillance is tiresome. A camera can only record information within the area it records, and that area is quite limited. Therefore, in order to monitor critical areas, it is necessary to setup many cameras in different places. As a result, personnel monitoring of all cameras remains a more difficult task. By re-identifying individuals captured by several cameras, this work presents a framework for tracking people in the presence of many cameras, while giving some potential methods to achieve fast people detection and tracking in multi-view cameras. By employing spatial transformations to achieve real-time multi-view tracking capability, the scheme is feasible and has been implemented.

Index Terms—STN, Distribution, Mult-View, Person Tracking

I. INTRODUCTION

Nowadays, CCTV has become a necessary distributed technology and is ubiquitous. For security reasons, effective monitoring of areas including colleges, industries, hospitals, markets, *etc.* has become a challenging task for the application. The requirement of multiple cameras for surveillance because these places are widely distributed [1]. In practice, it becomes difficult to monitor and track people in a multi-view form two or more cameras in different local. We have studied a framework to track some people through different camera sequences and maintain the identity of that person by re-identifying them in different cameras. According to the literature, person detection is the feature function of a tracking model, and the tracking model as a complete works better if the person detection is successful. As a results, The series of YOLO network models [2], [3], [4] are set of the state-of-the-art model for object detection among the various models available. The latest YOLOv4 had a mean average precision of accuracy increase of 10.0%. For tracking, there are two schemes for people re-identification: Centralized and Distributed:

Centralized Tracking in a structured and easily searchable single location connection is accomplished through multiple forms of detection, and collaboration around each central point.

Distributed Based on the spatial information of each camera, the tracking task can be passed from the

current camera to the other cameras by applying a spatial transformer, while enabling the detection of people and then re-identifying the subject at the angle of the other cameras.

In addition, the central point generated by the centralized part can be projected directly into other cameras through the spatial transformation matrix because the position of these CCTVs is fixed, thus allowing faster tracking fusion to be generated.

For advanced purposes, a trajectory model was introduced in [5] with the aim of learning color features, textures and metrics combined into a tracking framework where they use specified tensor states to generate relevant clusters after feature extraction and post-processing using different color channels. Furthermore, [6] proposed a symmetric structure consisting of two sub-networks connected by sinusoidal layers, where each sub-network contains two Convolutional Neural Network (CNN) layers and one Fully Connected (FC) layer as classifier for refinement of people re-identification. To deal with the significant variations in the color images in the dataset, a binomial distribution was used to evaluate the cost between labeling and similarity, robust to outliers. In [7] and [8], the authors point out that observing multiple views of a scene allows dynamic occlusion and error management, which may affect the sensor. The authors present an overview of multi-camera object tracking and summarize and classify the most common existing methods.

II. RELATED WORK

Object re-identification has long been a focus of computer vision research. Various Object detection models are used to recognize interested objects within images and video automatically [9]. While they have been utilized in a variety of situations, including pattern recognition in single image [10], and object tracking in frames sequences [11]. In this work, we present the use of object detection in multi-camera view, and compare with the Faster R-CNN [12] in the tasks of detecting objects. In practice, advanced YOLO consists of multiple CNN layers and two Fully Connected (FC) layer [13]. The network receive the original image and map it into a grid masks for further scaling and translating in a 2D domain, final produced the bounding boxes that together with a confidence weights/scores [14]. The confidence score reflects the model’s confidence that the object is within the predicted bounding

box and how accurate the predicted location is compared to the ground truth bounding box. Further, [15] introduces a unified framework that as a pre-processing aiming to learn the color feature, textures of object, and metrics simultaneously. This network is a symmetric topology composed of two sub-networks linked by a sine-wave layer [16], [17]. Each subnet is stacked of two CNN layers and FC layer for classification output. In order to deal with significant variations in portrait images in the dataset, it also makes use of the binomial distribution to evaluate the costs between labels and similarities that are resilient to outliers [18].

According to [19] and [20], a novel technical perspective is introduced to the seen scene, making possible dynamic occlusion and error control, which allows the problem of dynamic occlusion and damaged sensors to be solved. The authors present an introduction to multi-camera object tracking and describe and classify the most commonly used methods today. Currently, a popular object detection method, Faster R-CNN proposes a region suggestion layer and handles selective search with a predefined bounding box [21], which outperforms R-CNN [22] and Fast R-CNN [12] effectively. The main component of its selective search is the evaluation of regions of interest (ROIs) that is observed in the image and are faster. It aims to generate thousands of ROIs usually mapped to each image to increase the chances of perfect predictions and can be done in real time. In addition, researchers have studied the development of feature extraction on ROI. This task is widely considered as a classification problem and image recognition techniques are applied in a multilayer CNNs [23] or RNNs [24] model or complexity models for deeper learning [25]. The authors in [26] discuss feature extraction techniques that can extract features from any object in applications dealing with images or videos.

Inspired by the above model, we propose a distributed spatial transformer-related model. It corresponds to a classification task in which the underlying objects are able to identify which belong to the same object in different views. At the interface, the converter model of the proposed training model is used to extract unique embeddings of the test object pairs. We calculate the similarity of the feature embeddings. The test pair becomes an object of interest for tracking and classification in multiple view if the weight/score is above a threshold. A complete video is the result produced after tracking and re-identification.

III. DISTRIBUTED SPATIAL TRANSFORMER

The proposed distributed spatial transformer extend the standard Spatial Transformer Network (STN) model [27], [28], but the transformation parameter is obtained by considering the three sub-matrix of feature extraction, which is consisted of translation, rotation and scaling matrix can be achieved by [29]. Specifically, the extracted heterogeneous features will first be concatenated to form a joint feature, thus a standard transformation is given by:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = R(\theta) \cdot S(s_x, s_y) \cdot T(t_x, t_y) \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

where R , S and T performs rotation, scaling and translation respectively. Therefore, we propose to make use of this constraint into the training processing. Although the structure is similar to STN module, the five learning parameters in (1) would be allocated for machine training as well. Besides, it will be used to predict the transformation parameter by employing the localization network. To facilitate end-to-end training from the camera to each other, the localization network can also be optimized by two CNN layers and a regression layer, determined by a concatenate-designed classifier [30] to produce better accuracy, where each STM module is followed by a classifier and activation function layer. Another advantage is that by training these five parameters, the spatial transformer can be adapted by spatial adjustment. Since the Cartesian coordinate representation can be easily derived from the input, these methods increase the flexibility of the neural network in handling spatial transformations.

A. Features of trainable parameters

The proposed model is a combination of the standard transformation matrix and the STN model [27], but the transformation parameters are obtained by considering the three-component feature extraction in (1). Specifically, the extracted features are first flattened to form a joint feature with one-dimensional tensor projected from the input data. $H \times W \times C \in \mathbb{N}$, where H , W , C denotes the height, width and color channels of the input image. In order to fit these flattened features to (x, y) , there are two CNN (3×3) layers and one `Linear(*, 2)` is stacked before the proposed model, and each convolutional layer is connected by a Maxpooling and activation function layer. Next, to ensure spatial consistency between CNN layers and transformer, warping is sampled equally for each tensor. Therefore, the R , S and T in (1) can be represented as below:

$$R(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$S(s_x, s_y) = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

$$T(t_x, t_y) = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

To represent affine transformations with matrices, we can use homogeneous coordinates for the combination with (2), (3) and (4). They can be considered to compute its source coordinates according to the five predicted transformation parameters. Following the projective transformation as (1), the projective projection can be combined and expressed by:

$$T = \begin{bmatrix} s_x \cos \theta & -s_y \sin \theta & s_x t_x \cos \theta - s_y t_y \sin \theta \\ s_x \sin \theta & s_y \cos \theta & s_y t_x \sin \theta + s_x t_y \cos \theta \end{bmatrix} \quad (5)$$

For the target point (x', y') in the output feature, all ordinary linear transformations are contained in the set of space definitions and can be described in the form of (affine) transformation matrices. Thus, transformations in (5) can also be represented by an ordinary transformation matrix, filling the extra space with zeros, except for the lower right corner. Especially, it also ensures the gradient is available when the learning rate is a relatively small. This problem can also be solved by simply setting the gradient of the burst point to 0, and one of gradient is given by the second order partial derivative of θ :

$$\frac{\partial^2 T}{\partial \theta^2} = -T \quad (6)$$

This interesting result ensures that the use of the transformed feature representation can effectively mitigate the spatial inconsistency of the features until the training is fully converged. Thereby, the proposed model can effectively expand the field of view of the filter while reducing the number of parameters, thus excluding useless background information.

IV. OBJECT RE-IDENTIFICATION FROM MULTI-VIEW

The task of correlating frames taken of the same person at different angles in a multi-camera environment. Its main purpose is to find features that represent a person. (1) allows the re-identification of a person can be retrieved on various cameras that have some overlap. When the target person leaves the frame in an overlapping area in one camera, then it can be detected in another camera's frame after applying the spatial transformation trained by (1). This method can be processed in parallel for real-time achievement. Moreover, when the target figure leaves the frame without any overlapping area, the situation becomes a multi-view classification task. In practice, since (1) only allows an input size of 3, a linear layer is needed to project the size to the specified size. The size is projected into 3 as `Linear(*,3)`, and there is another linear used to increase the feature size by `Linear(3,*)`, for connecting the subsequently used layers. Furthermore, a Chebyshev Pooling [31] is used to optimize the predicted features. The purpose of the predicted features is to project the features onto the probability distribution used for the prediction. In addition, there is a loss function `SoftMax` that is attached to the end of our model for the generation of the output features. After that, a threshold needs to be preset to determine if the pair is the re-identified person. Otherwise, it is the other person. Once the difference is determined, an identity switch is performed. After tracking and re-identification, a complete model is thus produced.

A. Object Tracking from Multi-View

In general, a R-CNN model [32] is required to faster discover and detect all objects of interest (not only people),

and it employs a local network to generate region proposals for object detection, which is to find some text-like objects first. According to the R-CNN technique, the local of the current region can be considered as the starting point of the tracking path, and tracking can only be done by fusing many camera sequence frames for person detection. Therefore, when people are addressed in overlapping areas in the cameras, their tracking path can also be detected by applying (1) similar to the previous section. In this case, person detection is performed first, followed by tracking, re-identification and finally generating tracking fusion. From other perspectives, single-person tracking and multi-person tracking are two types of tracking. The goal of single-person tracking is to focus on one person in the frame and follow them until they leave the frame. This type of tracking is easier to use because it avoids the greater difficulty of identifying people. When doing multi-person tracking, we need to check each person in the frame. The goal in this case is to identify each person in the facility in a unique way and track them until they leave. If we provide a video sequence, the person tracker must track a specific person throughout the clip. Assuming the detector detects three people, it must see all three people in the frame. The people tracker must recognize three different detections and use a unique ID to track them in all frames.

V. CONCLUSION

We introduce a framework for person re-identification in multi-camera environments with models that track people as they approach each other. An advanced method is proposed to track projections between different multi-camera locations using a spatial transformer with Chebyshev Pooling and `SoftMax` for tracking optimization. We also give a multi-person multi-view tracking scheme where we can only track based on whether there are overlapping regions in the cameras. The model is performed to correctly assign identities to several people in the multi-camera.

ACKNOWLEDGEMENT

This work is supported by the Faculty of Applied Sciences, Macao Polytechnic University, Macau, China.

REFERENCES

- [1] K. S. Kumar, S. Prasad, P. K. Saroj, and R. C. Tripathi, "Multiple cameras using real time object tracking for surveillance and security system," in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, IEEE, nov 2010.
- [2] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jul 2017.
- [3] X. Jiang, T. Gao, Z. Zhu, and Y. Zhao, "Real-time face mask detection method based on YOLOv3," *Electronics*, vol. 10, p. 837, apr 2021.
- [4] J. Yu and W. Zhang, "Face mask wearing detection algorithm based on improved YOLO-v4," *Sensors*, vol. 21, p. 3263, may 2021.
- [5] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, mar 2018.
- [6] R. Naidoo, P. Pillay, J. Visser, R. Bansal, and N. Mbungu, "An adaptive method of symmetrical component estimation," *Electric Power Systems Research*, vol. 158, pp. 45–55, may 2018.

- [7] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: a review," *Multimedia Tools and Applications*, vol. 78, pp. 10773–10793, sep 2018.
- [8] L. Hou, W. Wan, J.-N. Hwang, R. Muhammad, M. Yang, and K. Han, "Human tracking over camera networks: a review," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, jun 2017.
- [9] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, pp. 38–49, jan 2018.
- [10] S. M. Silva and C. R. Jung, "License plate detection and recognition in unconstrained scenarios," in *Computer Vision – ECCV 2018*, pp. 593–609, Springer International Publishing, 2018.
- [11] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jun 2022.
- [12] R. Girshick, "Fast r-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, dec 2015.
- [13] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022.
- [14] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jun 2019.
- [15] A. Alzu'bi, H. Najadat, W. Eyadat, A. Al-Mohtaseb, and H. Haddad, "A new approach for detecting eosinophils in the gastrointestinal tract and diagnosing eosinophilic colitis," *The International Arab Journal of Information Technology*, vol. 18, no. 4, 2021.
- [16] V. Bharmauria, L. Bachatene, S. Cattani, J. Rouat, and S. Molotchnikoff, "Synergistic activity between primary visual neurons," *Neuroscience*, vol. 268, pp. 255–264, may 2014.
- [17] K.-H. Chan, S.-K. Im, W. Ke, and N.-L. Lei, "SinP[n]: A fast convergence activation function for convolutional neural networks," in *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*, IEEE, dec 2018.
- [18] V. H. Cene and A. Balbinot, "Resilient EMG classification to enable reliable upper-limb movement intent detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, pp. 2507–2514, nov 2020.
- [19] A. S. Rao, M. Radanovic, Y. Liu, S. Hu, Y. Fang, K. Khoshelham, M. Palaniswami, and T. Ngo, "Real-time monitoring of construction sites: Sensors, methods, and applications," *Automation in Construction*, vol. 136, p. 104099, apr 2022.
- [20] J. Guan, X. Yang, V. C. Lee, W. Liu, Y. Li, L. Ding, and B. Hui, "Full field-of-view pavement stereo reconstruction under dynamic traffic conditions: Incorporating height-adaptive vehicle detection and multi-view occlusion optimization," *Automation in Construction*, vol. 144, p. 104615, dec 2022.
- [21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, jun 2017.
- [22] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, oct 2017.
- [23] K.-H. Chan, S.-K. Im, and W. Ke, "Variable-depth convolutional neural network for text classification," in *Communications in Computer and Information Science*, pp. 685–692, Springer International Publishing, 2020.
- [24] W. Ke and K.-H. Chan, "A multilayer CARU framework to obtain probability distribution for paragraph-based sentiment analysis," *Applied Sciences*, vol. 11, p. 11344, nov 2021.
- [25] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jegou, "Going deeper with image transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, oct 2021.
- [26] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *Journal of Applied Science and Technology Trends*, vol. 1, pp. 56–70, may 2020.
- [27] Z. He, E. Liu, and Z. Xiang, "Partial fingerprint verification via spatial transformer networks," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, sep 2020.
- [28] C. Gao, X. Liu, W. Gu, B. Killeen, M. Armand, R. Taylor, and M. Unberath, "Generalizing spatial transformers to projective geometry with applications to 2d/3d registration," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 329–339, Springer International Publishing, 2020.
- [29] K.-H. Chan, S.-K. Im, V.-K. Ian, K.-M. Chan, and W. Ke, "Enhancement spatial transformer networks for text classification," in *Proceedings of the 2020 The 4th International Conference on Graphics and Signal Processing*, ACM, jun 2020.
- [30] K.-H. Chan, S.-K. Im, and W. Ke, "Multiple classifier for concatenate-designed neural network," *Neural Computing and Applications*, vol. 34, pp. 1359–1372, sep 2021.
- [31] K.-H. Chan, G. Pau, and S.-K. Im, "Chebyshev pooling: An alternative layer for the pooling of CNNs-based classifier," in *2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET)*, IEEE, aug 2021.
- [32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 386–397, feb 2020.