World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:7, 2024

Person Re-Identification Using Siamese Convolutional Neural Network

Sello Mokwena, Monyepao Thabang

***Abstract*—**In this study, we propose a comprehensive approach to address the challenges in person re-identification models. By combining a centroid tracking algorithm with a Siamese convolutional neural network model, our method excels in detecting, tracking, and capturing robust person features across non-overlapping camera views. The algorithm efficiently identifies individuals in the camera network, while the neural network extracts fine-grained global features for precise cross-image comparisons. The approach's effectiveness is further accentuated by leveraging the camera network topology for guidance. Our empirical analysis of benchmark datasets highlights its competitive performance, particularly evident when background subtraction techniques are selectively applied, underscoring its potential in advancing person re-identification techniques.

***Keywords*—**Camera network, convolutional neural network topology, person tracking, person re-identification, Siamese.

I. Introduction

PERSON re-identification serves as a technological solution aimed at extracting specific images of individuals from cameras positioned across different, non-overlapping areas. This technology holds significant importance in the realm of security, particularly in tasks like target tracking and retrieving individuals. The resolution of the person's image, in such cases, often falls below the threshold required for accurate identification using facial recognition. Adding to the complexity, these images frequently feature intricate backgrounds, further compounded by obstructions and variations in the individual's pose. The challenge is compounded by the fact that cameras with diverse orientations naturally provide distinct viewing angles, thereby amplifying the difficulty of recognizing individuals [1]. As a result, person re-identification remains a consistently demanding task.

Moreover, within a camera network, the detection of individuals and subsequent image capture often leads to lower image resolutions, thereby complicating the extraction of distinct features necessary for person re-identification. Given these aforementioned obstacles, a successful person re-identification system should comprehensively address these challenges. In our research, we present an approach that tackles the detection, tracking, image capture, and person re-identification processes. We accomplish this by methodically investigating the impact of robust differentiating factors within an image, such as clothing colour. Public areas such as railway stations, airports, and Universities are using surveillance cameras to prevent, track or identify wrongdoers. Studies have been conducted to secure the public place as much as possible [3]. Person re-identification is the process of identifying a target person in the field of view (FOV) of a single camera and associating the images of that person with images in another camera's FOV. Over the last decade, numerous studies on person re-identification have been undertaken [2], [3].

However, multi-camera networks remain a difficult task due to the uncertainty of appearance features in multiple cameras for a huge group of people in non-overlapping FOVs of cameras. Re-identification of a person should be done by exhaustively distinguishing the target individual from every other person appearing in the camera network.

II. Person Tracking

Person tracking is an active area in computer vision research. Single-camera tracking associates the targeted person with his/her trajectory as he/she moves in a single camera's FOV. In non-overlapping cameras' FOVs, the association of images of the persons appearing in those different FOVs becomes difficult. This problem can be approached by first associating the persons in a single camera before associating the images with another camera in a camera network. A method that handles the similarities of tracking in single or multiple cameras differently and acquires the optimisation in a global graph model was suggested by [4]. However, the system does not handle the integration of tracklets very well which leads to a large number of false negatives. The proposed algorithm in this study makes use of the centroid tracking algorithm to solve this problem.

Reference [5] proposed a method that utilises the Kalman filter to address the problem of tracking feature points along with image sequences efficiently. The system automatically detects every single individual that appears in the cameras' FOV, and use the information obtained by detection to track those individuals. However, the Kalman filter predicts future states of a system by using past estimates and produces hidden variable estimates based on erroneous and imprecise data, which leads to tracking mistakes [6].

Our proposed method makes use of the centroid tracking algorithm which is based on the Euclidean distance between the centroid of the old detected person, that is, the person that the centroid tracker has observed previously, and the centroid of the new person between subsequent video frames.

Sello Mokwena is with the University of Limpopo, Socvenga RSA (corresponding author, phone: +27828830534; e-mail: sello.mokwena@gmail.com).

Monyepao Thabang was as student at University of Limpopo but now works industry RSA (e-mail: monyepaosamuel@gmail.com).

III. Camera Network Topology

Within the realm of person re-identification, as an individual moves through a network of cameras, the task of tracking and subsequently re-establishing their identity across multiple camera points becomes progressively challenging. Essentially, the larger the camera network grows, the greater the complexity in effectively tracking and re-identifying the same person. The utilization of a camera network topology serves the purpose of alleviating the necessity to predict the potential presence of the target person within the network of cameras. In essence, when the topology of the network is understood, the process of person re-identification is simplified. This is because cameras where the target person is highly unlikely to appear can be excluded from the search for images of the target person. In other words, the knowledge of the network topology streamlines the re-identification process by eliminating cameras where the target person is not expected to be captured.

A distributed re-identification approach that exploits the distance vector routing algorithm was proposed by [7]. The method makes use of multiple cameras to measure the cost of reidentification performance between pairs of cameras, using the costs as distances between nodes in the distance vector algorithm which allows prioritisation and limits the set of inquired cameras. To track people across cameras, the approach used temporal information and imposed temporal restrictions.

They do not, however, directly infer the topologies of the camera network, instead relying on people's imprecise temporal information. Furthermore, when re-identification is performed, they do not continuously update or improve the initially estimated topology. As a result, these approaches may not fully utilise topology data.

A unified approach for solving both person re-identification and inference of the topology of the camera network challenges with limited previous knowledge of the environment was presented by [8]. The same approach will be used in this study since it proves to improve re-identification of individuals by utilising the topology of the camera network effectively. This approach will allow us to focus on the appearance features of a person across multiple cameras with the knowledge of possible appearance of that person in a camera network.

IV. Siamese Convolutional Neural Network

Matching images that depict the same individual poses a challenge due to factors like changes in camera viewpoints and obstructions. To address these complexities, current methods concentrate on either creating robust feature representations, as demonstrated by [9], or on acquiring optimal matching metrics, as explored by [10]. Many deep re-identification models presently focus on learning feature representations at a single scale, which may not adequately capture compact and style-invariant characteristics. In this study, we introduce a Siamese Convolutional Neural Network (CNN) as a solution to the aforementioned problem. This approach aims to overcome these obstacles by enabling better image matching through the utilization of multi-scale and style-invariant feature representations.

Our proposed approach employs deep learning methodologies, specifically supervised learning. For training purposes, we utilize a dataset comprising over 300 images captured from diverse cameras. The objective is to effectively re-identify the target person within a camera network. The adoption of deep learning techniques has exhibited its advantages over the last decade. Notably, they obviate the need for manually engineered features and exhibit strong performance when ample training data are available, as evidenced by the work of [11].

Over the past decade, deep neural networks (DNN) have been successfully applied in many areas of computer vision, such as face recognition [12] and person re-identification [13], and in these areas, they have largely replaced traditional computer vision pipelines based on hand-crafted features. DNNs have also been used to learn classifying images based on image pairs [14], or triplets of images for person re-identification. Certain techniques, which use network architectures like the Siamese network [15], develop a direct mapping of the pixel data from a raw image to a feature space in which images of the same individual are grouped together and images of different persons are set far apart from each other.

While some researches focus on a single image [14] for person re-identification, our method made use of a video to capture multiple images of the same person to allow for the improvement of the process of person re-identification by using the multiple different images of the same person in the training classifiers that vary in posture to associate the best possible pair of images. Our method also included a colour-based feature to represent the appearance feature of the person in the camera network.

We present a DNN technique for video-based person re-identification in this work. By utilising a Siamese CNN to create an invariant representation for each person's image from multiple cameras' FOVs, our DNN-based method uses appearance data with representation learning. Our proposed method differs from existing methods that rely on hand-crafted features in the sense that it automates the process of learning to extract strong appearance traits that are relevant for person re-identification in a unified framework.

V. Proposed Method

The Camera Network (CamNet) dataset was downloaded from an online open-source data repository called Video Computing Group (VCG). The dataset contained eight cameras covering both indoor and outdoor scenes. At first, the target person was identified and tracked through the camera network's FOV using the centroid tracking algorithm. This algorithm continuously tracked the coordinates of the target person by calculating the Euclidean distance between the current coordinates and the previous coordinates of the target person and if the value was between a certain threshold value, that is 90, refer to (1) for the Euclidean distance equation, then it continued to track the target person else, it considered the target person as another person and assigned that new person a new ID number.

$$ED = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (1)$$

where X_1-Y_1 are the coordinates of one point; X_2-Y_2 are the coordinates of the other point; ED is the distance between X_1-Y_1 and X_2-Y_2.

Then the image of the target person as he was about to exit the camera's FOV was taken. The reason for tracking the target person up to the exit point of the camera's FOV was that we would not compare pairs of images of the target individual from dissimilar FOVs if we were not 100% sure that the target person has exited the first camera's FOV or not.

Fig. 1 represents the topology of the camera infrastructure. We suppose the target person was tracked through the FOV of camera 29. It showed that when the target person exited the FOV of camera 1, he could only appear in the FOV of either camera 22 or 23, or not appear in either of the cameras' FOV, which meant that the target person would have exited the camera network.

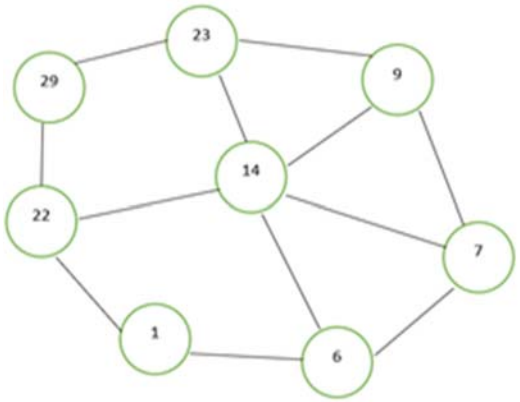

Fig. 1 The structure of the camera network topology

The trajectory of the target person with ID55 from camera PRG29's FOV was analysed using the k-means clustering algorithm. We first specified the number of clusters K. Therefore, to get the number of clusters we applied the elbow method. Secondly, we initialised the centroids by randomly selecting a set of K points from the dataset without replacement. Thirdly, we repeated the assignment of data points to clusters until there was no change in the distribution of data points. Fourth, we calculated the sum of the squared distance between data points and all centroids. Refer to (2) for the mathematical representation of the sum of squared errors.

$$SSE = \sum_i^n (X_i - \overline{X})^2 \quad (2)$$

where SSE = Sum of squared errors; X_i = Observation in the same cluster; $\overline{X}$ = mean of observations in the same cluster; n = number of observations X_i.

Fifthly, we assigned each data point to the closest cluster (centroid) using the Euclidean distance formula. Refer to (1) for the mathematical representation of the formula. Lastly, we calculated the centroids for the clusters by averaging all of the data points that belonged to each cluster. Refer to (3) and (4) for a mathematical representation of how the average was calculated.

$$\overline{X} = \frac{\sum_i^n X_i}{n} \quad (3)$$

where: X = x – coordinates of the data points in the same cluster; n = number of observations.

$$\overline{Y} = \frac{\sum_i^n Y_i}{n} \quad (4)$$

where: Y = y – coordinates of the data points in the same cluster; n = number of observations.

After obtaining the input image of the target person, the Siamese CNN was then used to check whether the persons appearing in those cameras' FOV were the same as the target person or not. Suppose two images x and y, represented as picture 1 and picture 2 in Fig. 2 respectively, were sent to the Siamese CNN as input images. Then, the Siamese CNN outputs a similarity score of whether both images are of the same person or not. Our method generated the similarity score of the two images being compared while other person re-identification applications classified the stored images in the repository based on their similarity to the detected images. Refer to Fig. 2 for the Siamese Convolutional Neural Network model.

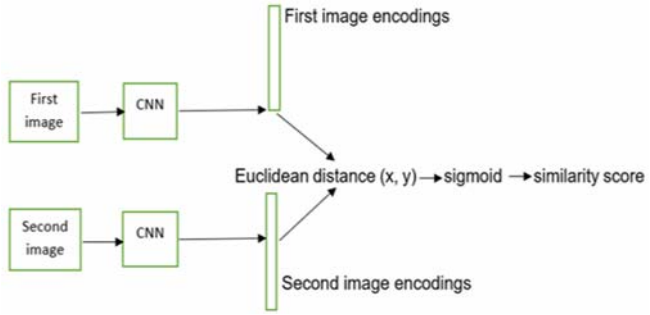

Fig. 2 The structure of the Siamese CNN which is defined by two CNN, followed by filters to produce an output

VI. Discussion

The study encompassed two distinct scenarios: person tracking and person identification. These scenarios were subjected to comprehensive evaluation to gauge the effectiveness of our proposed approach. The evaluation of the first scenario involved the application of the k-means clustering algorithm, while the second scenario was assessed through the utilization of the confusion matrix.

The test results yielded promising outcomes, indicating a notable enhancement in person re-identification facilitated by our proposed method. Particularly, our model exhibited an incremental improvement in its capacity to assimilate images from the training gallery as the dataset size expanded. It was discerned that the partitioning of data in a specific manner contributed to the augmented performance of our model.

Furthermore, the integration of background subtraction across all images yielded a significant boost in our model's overall performance. This enhancement was particularly

evident in the increased accuracy of image comparisons. By effectively removing extraneous background elements, our model was better equipped to focus on the core aspects of individuals, thus refining its accuracy in matching and identifying persons.

In summary, the simulation encompassing person tracking and identification scenarios unveiled that our proposed method yielded promising advancements in person re-identification. As our model learned from an expanding dataset, and with the strategic handling of data partitioning and background subtraction, the performance gains were discernible, emphasizing the effectiveness of our approach.

ACKNOWLEDGMENT

We thank the University of Limpopo for making available lab equipment for the data analysis.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[5] E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *World Academy of Science, Engineering and Technology Trans. Antennas Propagat.*, to be published.

[6] J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *World Academy of Science, Engineering and Technology J. Quantum Electron.*, submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces (Translation Journals style)," *World Academy of Science, Engineering and Technology Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 (*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301).

[9] M. Young, *The Technical Writers Handbook.* Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility (Periodical style)," *World Academy of Science, Engineering and Technology Trans. Electron Devices*, vol. ED-11, pp. 34–39, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *World Academy of Science, Engineering and Technology Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.

[12] R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.

[13] S. P. Bingulac, "On the compatibility of adaptive controllers (Published Conference Proceedings style)," in *Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory*, New York, 1994, pp. 8–16.

[14] G. R. Faulhaber, "Design of service systems with priority reservation," in *Conf. Rec. 1995 World Academy of Science, Engineering and Technology Int. Conf. Communications,* pp. 3–8.

[15] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.

.