https://doi.org/10.2352/ISSN.2470-1173.2020.10.IPAS-095

Introducing Scene Understanding to Person Re-Identification using a Spatio-Temporal Multi-Camera Model

Xin Liu, Herman G.J. Groot*†, Egor Bondarev, and Peter H.N. de With*
Eindhoven University of Technology, Dep. Electrical Eng., Video Coding and Architectures Res. Group, Eindhoven, the Netherlands
**equal contribution, †corresponding author: h.g.j.groot@tue.nl*

Abstract

In this paper, we investigate person re-identification (re-ID) in a multi-camera network for surveillance applications. To this end, we create a Spatio-Temporal Multi-Camera model (ST-MC model), which exploits statistical data on a person's entry/exit points in the multi-camera network, to predict in which camera view a person will re-appear. The created ST-MC model is used as a novel extension to the Multiple Granularity Network (MGN) [1], which is the current state of the art in person re-ID. Compared to existing approaches that are solely based on Convolutional Neural Networks (CNNs), our approach helps to improve the re-ID performance by considering not only appearance-based features of a person from a CNN, but also contextual information. The latter serves as scene understanding information complimentary to person re-ID. Experimental results show that for the DukeMTMC-reID dataset [2][3], introduction of our ST-MC model substantially increases the mean Average Precision (mAP) and Rank-1 score from 77.2% to 84.1%, and from 88.6% to 96.2%, respectively.

Index Terms— Scene understanding, person re-identification, spatial constraints, temporal constraints, context information DukeMTMC, DukeMTMC-reID, CNN.

Introduction

Nowadays, surveillance cameras are installed everywhere in cities, especially in public areas, such as shopping and city centers, bus and railway stations, campuses and airports. These cameras constitute networks offering wide spatial coverage. As an example, Fig. 1 shows a multi-camera network of a part of the Duke University campus. Multi-camera networks generally provide a plurality of videos that are used for surveillance or security purposes.

When a surveillance operator would like to track a specific person across multiple camera field-of-views (FOVs), the operator needs to observe which camera FOV the person came from and which camera FOV he is going to. The accuracy of the observed result directly determines the success of the multi-camera tracking. Furthermore, this manual observation is extremely time-consuming. Hence, automated person re-identification (re-ID) is vital and preferred over manual search.

Person re-ID is defined as a task in computer vision that aims to determine whether a specific person that enters the FOV of one camera has been previously detected in other cameras. In current research practice, plenty of labeled pedestrian images are cropped from the surveillance videos. These images are divided into a 'Query' set and a 'Gallery' set, as shown in Fig. 2. These two sets are utilized as a validation dataset for the re-ID process. For each query sample, re-ID aims to find correspondences with the large gallery set, such that the query person can be identified.

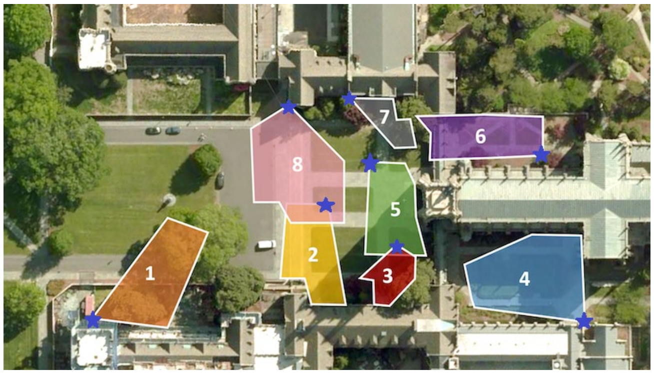


Figure 1. *Top view of the multi-camera network setup of the Duke University campus, as used in the DukeMTMC dataset [2]. The blue stars indicate the position of the cameras. The colored regions illustrate the fields of view for every camera*

Person re-ID is an essential, yet still very challenging task. One of the main challenges is the variation in the appearance of a person over different cameras, since cameras may operate at different exposure settings, as well as under different lighting, illumination, weather, and daytime conditions. Moreover, the same person may wear completely different clothing on the next day, which restricts the applicability of re-ID. A second serious challenge is a difference in the person pose towards different cameras, since a person may be oriented to a camera frontally, sideways or even backwards. Thirdly, the captured pose of the same person may also be different. Since a person may run, walk, or move slowly, his pose also differs over time. Finally, low camera density in the surveilled area, as imposed by typical person re-ID use cases, leads to non-overlapping FOVs between these cameras and thus causes ambiguity in spatio-temporal relations.

Face information is hard to be utilized in re-ID, since faces are not always visible in the camera view. Additionally, since the FOV of a typical surveillance camera is rather wide, the imaging quality of a visible face is insufficient in terms of resolution. Therefore, it is virtually impossible to extract discriminative features of a face.

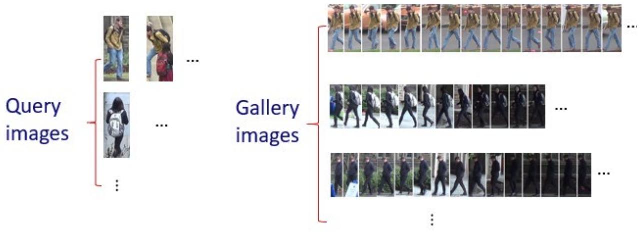


Figure 2. *Example of Query images and Gallery images for the DukeMTMC-reID [3] [2] dataset. For each person, the query images are randomly selected from each camera FOV. The remaining images are stored in the gallery set.*

Existing work on person re-ID can be generally divided into non-contextual and contextual approaches [4]. From the deep learning concept which has emerged recently, non-contextual approaches have achieved success by applying convolutional neural networks (CNNs) on feature descriptor learning [5][6] and metric learning [7]. Consequently, non-contextual approaches solely rely on features that are learned from the sample images alone. On the other hand, the contextual approaches aim to also exploit the external contextual information, such as inter-camera relationships [8], and then implement re-ID based on that information.

When considering re-ID algorithms in general, ideally a proposed match from the gallery should not only have high similarity in appearance to the query, but also satisfy the contextual requirements, such as spatio-temporal constraints. To illustrate this, consider the camera network, as shown in Fig. 1, and assume we have a query image from Camera 1 and a gallery image from Camera 7, where the time difference between these two images is e.g. 5 seconds and both images contain similarly dressed persons. Since we know it takes on average around 1 minute to travel from Camera 1 directly to Camera 7, we conclude that this image pair cannot satisfy the spatio-temporal relationship. As a result, it is possible that a feature-based approach returns an incorrect match, but if we involve the contextual (timing) constraints, we can correct for this error. Therefore, a hybrid deployment of both contextual and non-contextual approaches forms an attractive proposition.

In this project, the re-ID performance is improved by increasing the matching score for correct candidates from neighboring cameras of the query camera. To achieve this, we focus on a hybrid deployment of (1) a contextual re-ID verification approach with a spatio-temporal multi-camera model, and (2) a non-contextual CNN-based re-ID approach that adopts the multiple granularity network (MGN). First, we propose and experiment with several spatio-temporal model generation techniques. Second, we investigate the optimal ways of integrating the spatio-temporal multi-camera model into the re-ID process, which leads to improved re-identification performance.

Related Work

As described in the introduction, person re-ID research can be generally classified into non-contextual and contextual approaches.

Non-contextual approaches mainly focus on finding powerful and discriminative feature descriptors using metric embedding. Originally, CNNs were trained to learn single global features [9][10][11][12][13][14]. Instead of learning a global feature from the entire image of a person sample, the works in [6][15][16][17] divide an image into a few local branches, to generate local features and then aggregates a global feature. During the training process, metric embeddings [7] are learned to ensure that the distances between the feature vectors of images of the same person are small, while those of different persons are large.

Although these non-contextual approaches achieve high person re-ID performance [5][6][9], there is a potential improvement by also involving contextual constraints. Several research works have proven that the pre-known contextual constraints such as camera topology, are valuable for consideration [18][19].

Contextual approaches are employed to extract the contextual constraints and then implemented person re-ID based on those constraints. A few papers involve the human pose in re-ID, the contribution of [20] utilized change of poses to obtain metrics that match the appearances of a person in different poses. Furthermore, many researchers aim to find the spatio-temporal relationship between cameras [21][22]. Several works observed the entering and exiting events of people, and thereafter design the camera topology based on the measured event correlation [23][24]. Cho and Yoon exploited the feature that people walk at different speeds and proposed a distance-based camera topology network to implement re-ID [21].

Although existing contextual approaches already report improving person re-ID performance, a proper spatio-temporal camera model allows to verify the re-ID candidates by employing the spatio-temporal statistics on the scene exit and entry events. Furthermore, the optimal ways of integrating the generated contextual model into the re-ID process are investigated.

Our Approach

In this section, we first introduce the generic re-ID training setup and datasets in Subsection A. Afterwards, we explain several frequently mentioned definitions in Subsection B. Then, the two main parts, such as the generation of the spatio-temporal multi-camera model and the integration of the camera model into the re-ID process (see in Fig. 3) are explained in the remaining subsections.

A. Training setup

A set of labeled pedestrian images, which are cropped from surveillance videos obtained from multiple cameras, is defined as the re-ID dataset. As shown in Fig. 4a, the re-ID dataset is typically divided into three sets, 'Train', 'Query' and 'Gallery'. The 'Train'

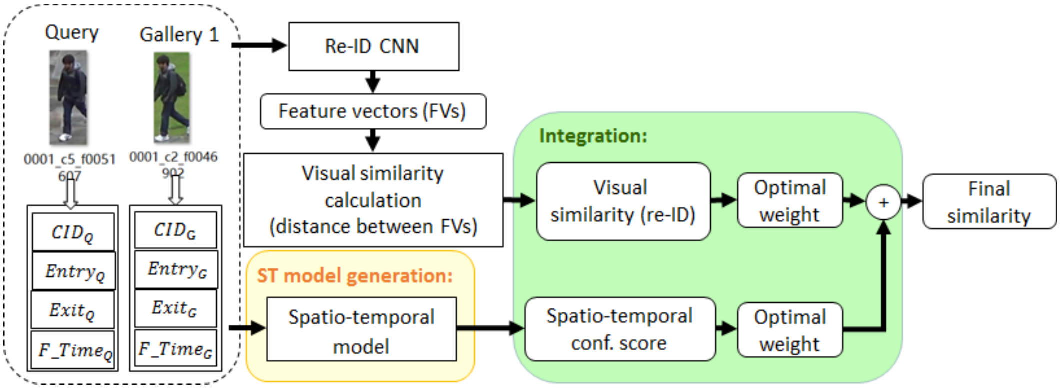


Figure 3. *Structure of our approach. The query and gallery image will be sent to the CNN to calculate the visual similarity and their spatio-temporal (ST) information, such as entry/exit points. This information is supplied into our ST model for obtaining the ST score. The computed re-ID and ST scores are optimally combined in the integration phase, which returns the final similarity score for the query and gallery image. The two colored blocks indicate our additional modeling.*

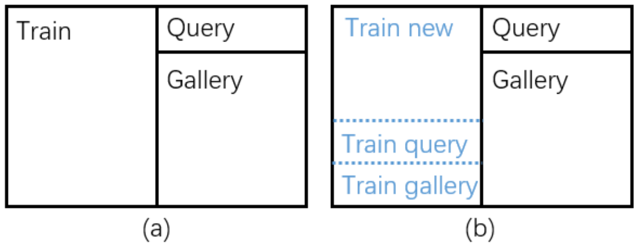


Figure 4. *Illustration of re-ID data distributions. (a) Typical re-ID data distribution. (b) New re-ID distribution used in our hybrid approach. 90% of the samples in 'Train' set are assigned in 'Train new', while 'Train query' and 'Train gallery' contain the remaining 10% of samples.*

set is used for re-ID model training, while the 'Query' and 'Gallery' sets are employed only for validation. The 'Gallery' set contains several images for each query person at different cameras. In other words, each query image has multiple candidates in the 'Gallery' set. To find the candidate ranking list of one query image, we need to calculate a score for each gallery image. A candidate with a higher score means that the candidate is more likely to be identical to the person in the query image.

Since we have added an extra spatio-temporal model, the learning of the features of this supplementary model needs to be incorporated in our training procedure. Therefore, a new data distribution is necessary (Fig. 4b) for our integration phase. That is, in our first phase, a spatio-temporal model is trained to explore the spatio-temporal constraints and the Multiple Granularity Network (MGN) [1] is trained to obtain the (visual) discriminative features of people. Thereafter, in our integration phase, these parts are combined to refine the most likely matches that the CNN would have returned alone. Nevertheless, we have observed that if we utilize the 'Train' set for both the training and integration step, learning the weights for optimal integration failed because the CNN was already 100% effective on the 'Train' set, leaving insufficient data to tune the integration. Furthermore, we cannot use the 'Query' and 'Gallery' sets to investigate the weights, since these two sets can only be used for validation. Consequently, the 'Train' set is divided into a 'Train new', a 'Train query' and a 'Train gallery' set (Fig. 4b), where 'Train new' is used to train both the multi-camera model and the CNN, while the latter two sets are used to tune the integration.

B. Frequently mentioned definitions

Prior to proceeding to the spatio-temporal model, we explain important definitions.

1) QG pair

In our approach, we regard one query image and one gallery image as a query-gallery image pair (QG pair), see e.g. Fig. 5.

2) QG path

The camera IDs (CIDs) of the two images indicate a 'QG path' for the QG pair. In the example in Fig. 5, the CIDs of the query and gallery image are five and two, respectively. Thus, the QG path would be 'Camera 2 to Camera 5 (c2→c5)' or 'c5→c2'. The choice of the direction is based on the next definition 'Past (Future) QG pair'.

3) Past (Future) QG pair

A QG pair can be either a past pair or a future pair, as illustrated in Fig. 6. We define a QG pair as a 'past pair' when the gallery image is captured before the query image and vice versa for the 'future pair'. We determine the direction of the indicated path based on the capturing times of the two images, as indicated by the frame time. The direction is important when choosing the correct algorithm to calculate the spatial confidence score, more details on this follow in Subsection C.

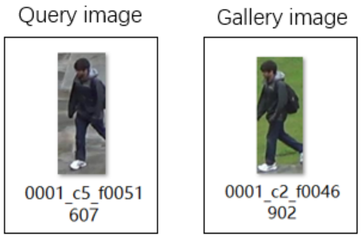


Figure 5. *Example of a QG pair. Naming rules for the bottom label are "person ID_camera ID_current frame time".*

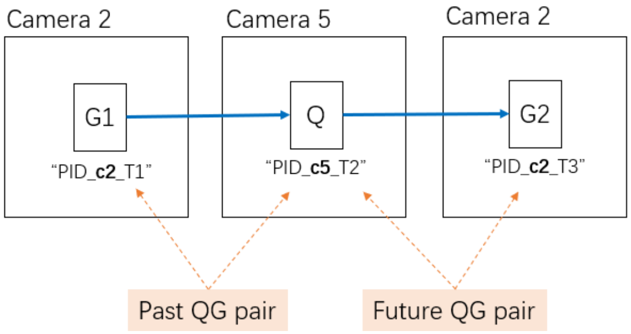


Figure 6. *Illustration of a 'Past QG pair' and a 'Future QG pair'. In this example T1<T2<T3, thus QG_1 is a past pair and the indicated QG_1path is c2→c5. Future pair QG_2 indicates a path c5→c2.*

4) Exit/Entry point of a person

Considering that a person passed several camera FOVs, the image coordinates of the starting and ending points of his tracklet in each FOV are the entry and exit points for that person (see Fig. 9). We explore the entry/exit points of all people in both the query and gallery sets, to determine the spatio-temporal constraints.

5) Spatial (Spatio-temporal) confidence scores

Essentially, our spatio-temporal multi-camera model provides a proper source for computing a spatial confidence score and a spatio-temporal confidence score for each candidate QG pair in the 'Train new' dataset. The spatial confidence score indicates the likelihood that the query person has walked via this specific path (in other words has been seen or will be seen in Camera X) as suggested by the currently considered QG pair. The spatio-temporal score indicates the likelihood of the camera-to-camera transition time of this QG path.

Using these definitions, we proceed to the spatio-temporal multi-camera model.

C. Generation and application of the spatio-temporal multi-camera model

As shown in the overview in Fig. 7, the spatio-temporal model is divided into a spatial part and a temporal part. In the spatial part, we first introduce the model generation approach (Approach C.1) and then its usage during inference (Approach C.2). Similarly for the temporal part, the modeling and inference is covered by Approach C.3 and Approach C.4, respectively.

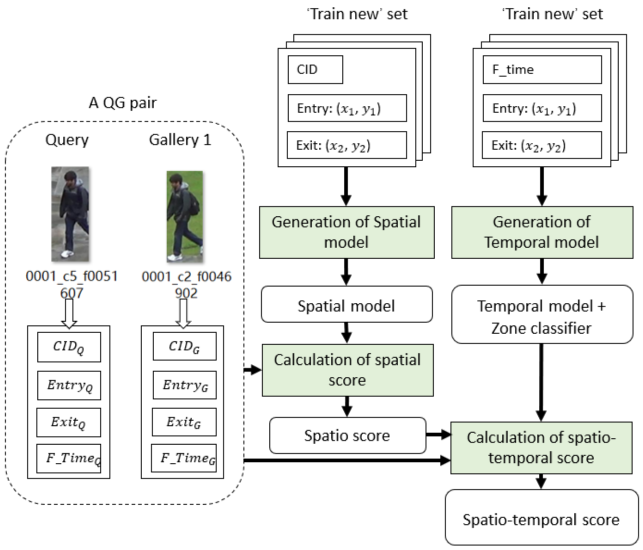


Figure 7. *Overview of the generation and application of the spatio-temporal model. 'Train new' set is used to generate the spatio-temporal model, QG pairs from the 'Train query' and 'Train gallery' set are the inputs of the spatio-temporal score calculation.*

1) Generation of Spatial Model

For each camera FOV, we plot the entry and exit points of all pedestrians (see the example of Fig. 9). To augment the training data, we assume that if a person travels from camera FOV A to B, then there is an additional travel path from camera FOV B to A with the interchanged entry/exit points. In other words, each point in Fig. 9 indicates a bi-directional travel path. For example, a green point in Fig. 9 can be the entry point of a c2→c5 travel path, but also the exit point of path c5→c2. Observing all entry/exit points in each camera, we have found that the points with the same color are clustered around the same image area. This clustering indicates that people always cross a specific area when traveling between two specific cameras. Therefore, since we can obtain statistical data on the entry/exit point clustering by considering all persons in the 'Train new' set, we are able to design the spatial model for that camera.

In Fig. 8, the images 'Query' and 'Gallery image 1' are forming a QG pair example. For this past QG pair, QG_1, we need a spatial model that can use the image coordinates of the query's entry point ('Entry$_Q$' in Fig. 8) to provide the confidence score that the query person has been seen earlier by the camera of the gallery image, in this case Camera 2, i.e. P(Query last appeared in Camera 2 |$Entry_Q$).

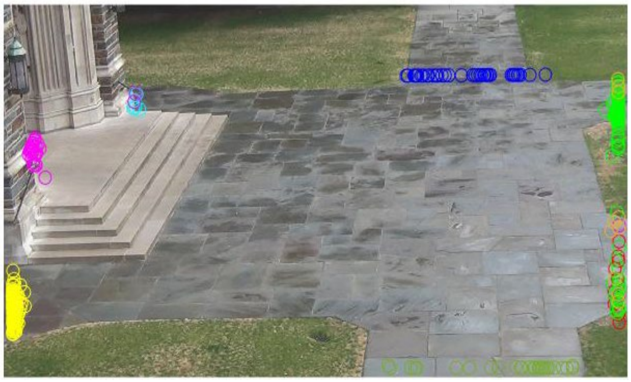


Figure 9. *Example of the entry/exit points in Camera 5. Coloring is used to label the entry/exit points for different paths. Red: travel path c1↔c5. Green: travel path c2↔c5. Cyan: travel path c3↔c5. Blue: travel path c4↔c5. Magenta: travel path c5↔c5. Yellow: travel path c6↔c5. Olive green: travel path c7↔c5. Orange: travel path c8→c5.*

Given the clustering nature of our entry/exit points, the function of our spatial model is comparable to a soft classifier. The soft classifier first estimates the class conditional probabilities and then uses the estimated probabilities to implement classification [25]. In our case, what we need are the intermediate results, which are the probabilities. Therefore, we have decided to adopt a soft classifier as our spatial model. We have tested several widely used classifiers, such as support vector machine (SVM) and random forest (RF), and finally selected the RF, since it proved to perform more accurately in our application.

To train the spatial model, we first determine the entry/exit points of all persons in the 'Train new' set, in each camera FOV. In practice, technologies of person tracking in single-camera FOV are mature, so we assume that we utilize a perfect tracker, which means the tracker returns the starting and ending points of a single tracklet without error. Hence, the entry/exit points are extracted from the ground-truth table of the dataset. These points are used as labels that additionally annotate from which camera that person comes from (or goes to) directly. Next, due to the random nature of RF models, we use 90% of these labeled points to train several RF models [26] for each camera and the remaining 10% to select the best RF model for each camera. Consequently, our final spatial camera model is composed of eight different RF models, one for each camera in the dataset.

Finally, each trained RF model contains T trees, as illustrated by Fig. 10 [27], where each tree will give different estimated class

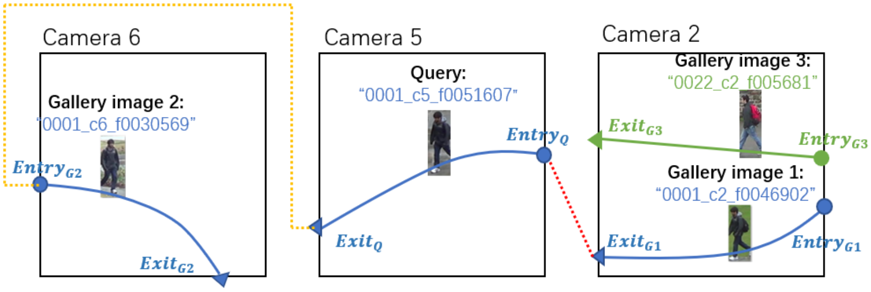


Figure 8. *Example of the trajectory of person ID 0001 (c2→c5→c6) and 0022 (in c2). In this example, QG_1 is a past pair (linked by the red dashed line), QG_2 is a future pair (linked by the yellow dashed line).*

conditional probabilities. The output of each tree is combined by taking the average of these conditional probabilities, as specified by:

$$\mathrm{p}(c|v) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|v)\,, \qquad (1)$$

where parameter T is the number of trees and $p_t(c|v)$ indicates the estimated conditional probability of the t^{th} tree that the entry/exit point (v) belongs to a specific camera ID (CID c). The next subsection will introduce the method to obtain the spatial confidence score as determined during inference, based on the trained spatial models.

2) Calculation of Spatial Score

The trained spatial camera model can now be used to obtain the spatial confidence score that a query person is entering to (or exiting from) a specific neighboring camera. The procedure to obtain the spatial confidence score is illustrated in Fig. 11. In this subsection, we will still consider QG_1 from Fig. 8 (i.e. 'Query' and 'Gallery image 1') as a QG pair example.

From the ground-truth data of the DukeMTMC-reID dataset, all spatial and temporal information of the images can be determined, such as camera ID (CID), image coordinates of entry/exit points and frame times. The latter can be used to obtain an overall synchronized time for all images. However, the start time of each camera is different, so we need to add the corresponding start time to the frame time to obtain the synchronized time.

To calculate the spatial confidence score, we first need to judge if the selected QG pair is a past pair or a future pair, as this affects the choice of feeding the entry or exit point to the spatial model. As shown in Fig. 8, a past pair (linked by the red dashed line) will utilize the entry point of the query image ($Entry_Q$) and the exit point of the gallery image ($Exit_{G_X}$), while a future pair (linked by the yellow dashed line) will use $Exit_Q$and $Entry_{G_X}$.

Because of our QG_1 is a past pair, we feed $Entry_Q$ and $Exit_{G_1}$ to the trained spatial models to calculate two scores, as shown in Fig. 11. The first score is the probability that the query person comes from Camera 2 (the CID of G_1), thus the likelihood value of $Entry_Q$ leads to Camera 2. The second is the probability that the person of the gallery image goes to Camera 5 (the CID of Q), so the likelihood value of $Exit_{G_1}$ leads to Camera 5. The exact values of these scores are determined by applying Eqn. (1). If one of the two scores is zero, it means that the selected QG pair does not satisfy the spatial constraint, so we set the spatial confidence score to zero. This bidirectional check helps to filter out QG pairs, like QG_3 in Fig. 8, which have high 'score1' but 'score2' equal to zero. If both 'score1' and 'score2' are non-zero, we take the average and use that as the spatial confidence score of the considered QG pair. The averaging helps to compromise the result for those QG pairs having unbalanced scores (e.g., score1=0.5 and score2=0.99). This phenomenon happens in the mixture area (see bottom-right part of Fig. 9) where the same entry/exit area can lead to several other cameras. Our spatial model is a soft classifier and provides probabilities like 0.5 in the areas with high uncertainty.

We calculate the spatial confidence score for every QG pair in the dataset. Only QG pairs that have a non-zero spatial confidence score will be used in computation of the spatio-temporal score.

3) Generation of Temporal Model

The temporal model is obtained by determining the distribution of camera-to-camera transition times for all travel paths, as introduced in Approach C.1. Fig. 12 shows two typical time distributions of two possible travel paths, where the distribution of Fig. 12.a can be represented by a single Gaussian distribution that is directly learned from the data. However, for the distribution in Fig. 12.b, there are two possible single pedestrian roads between Camera 2 and Camera 5 (see Fig. 13). In this case, we can obtain a more accurate travel time prediction if we separate between the two roads, as the walking distance of each road may be different. Hence, we now fit two Gaussian distributions for modeling the probabilities.

To determine the number of spatial clusters of entry/exit points for each travel path, we deploy the density-based spatial clustering

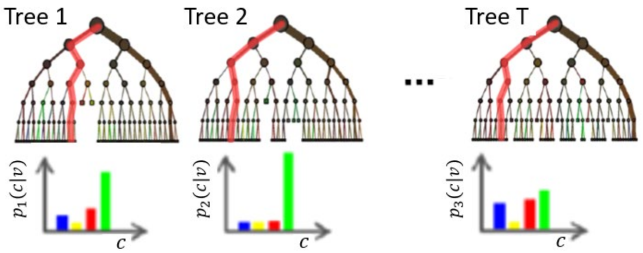


Figure 10. Example of the structure of random forest (RF). In this RF example, there are T trees in the forest, and each tree returns different class conditional probabilities.

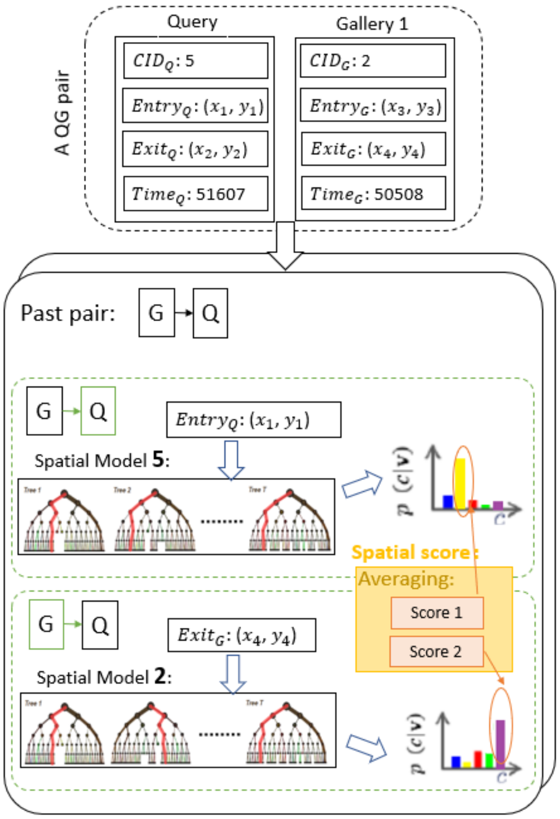


Figure 11. Example of obtaining the spatial score. By comparing $Time_Q$ and $Time_G$, we select the algorithm for the past pair. $Entry_Q$ and $Exit_G$ are supplied into spatial Model 5 and 2 separately, to calculate the probability that $Entry_Q$ can lead to Camera 2 (Score 1) and the probability that $Exit_{G1}$ can lead to Camera 5 (Score 2). The spatial score is the average of Score 1 and Score 2.

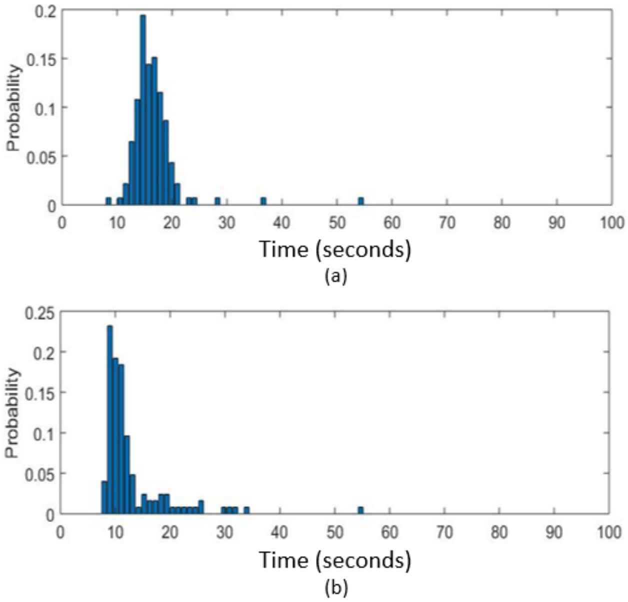


Figure 12. Examples of time distributions. (a) Time distribution for path 'Camera 3 to Camera 4' (c3→c4). (b) Time distribution for path c2→c5.

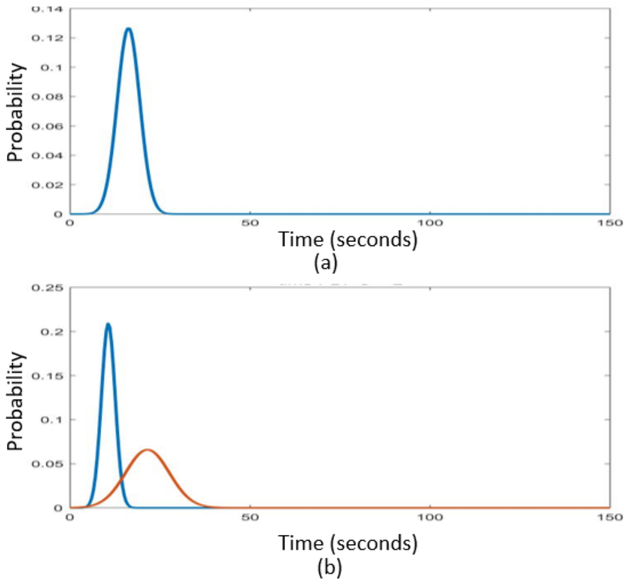


***Figure 14.** Example of obtained temporal models. (a) Temporal model for path c3→c4. (b) Temporal models for path c2→c5.*

algorithm (DBSCAN) [28]. So, for the mentioned example of Fig. 13, DBSCAN allows us to detect that there are two roads between Camera 2 and Camera 5. In the remainder of the paper, we refer to these spatial clusters as 'zones'. Thus, each QG path may have one or more query zones.

Once we have fitted a Gaussian probability distribution to each zone, we have obtained the temporal model, which is a set of Gaussian parameters, for each possible pedestrian road. Fig. 14 shows the obtained probabilistic temporal models of the examples in Fig. 12. In the multiple zone situation, a zone classifier is required to choose the correct time distribution. We investigated both RF and SVM algorithms to classify which zone the entry point of the query sample belongs. More details about this comparison will be explained in the Experiments and Results section. The next subsection explains application of the trained temporal models and calculation of the spatio-temporal confidence score during inference.

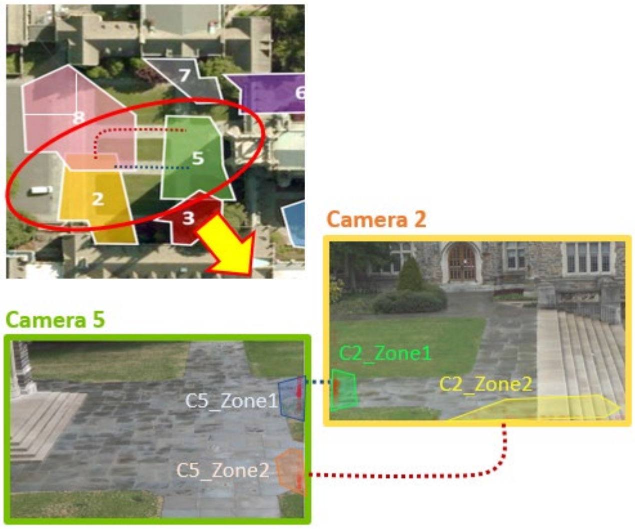


***Figure 13.** Illustration of two possible pedestrian roads for a single travel path. Red points in each camera FOV indicates the entry/exit points of these two roads. DBSCAN helps to determine that there are two spatial clusters (zones) of entry/exit points for path c2↔c5 in both Camera 5 and Camera 2.*

4) Calculation of Spatio-temporal Score

We utilize both the generated temporal model and the trained zone classifiers to obtain the temporal confidence score for a QG pair. Fig. 15 illustrates the procedure to obtain the spatio-temporal confidence score when evaluating the example QG pair. Essentially, the spatio-temporal score employs spatial condition information to select the useful temporal information. For this reason, the output of a temporal model is called a spatio-temporal score.

The timestamp and entry/exit coordinates from the two images of the selected QG pair and its corresponding spatial score are used as input for the temporal part. We only consider QG pairs having a non-zero spatial confidence score, since only then the travel path as implied by the QG pair, is possible. Subsequently, if the currently considered travel path has multiple zones, we supply $Entry_Q$ (or $Exit_Q$) into the trained zone classifier to determine which zone and corresponding temporal model applies (e.g, the blue line of Fig. 15 for the example QG pair). Once we have selected the correct temporal model for a QG pair, the spatio-temporal confidence score is simply the probability that corresponds to the time difference of the QG pair. In other words, we determine how likely the actual time difference between the entry/exit points of the currently evaluated QG pair is.

Finally, if the query and the gallery image are from the same camera, we enforce the temporal confidence score to zero because the transition time under this situation is very diverse and there is thus no suitable time distribution. The above procedure is repeated to calculate the spatio-temporal confidence score for every QG pair, and the obtained spatio-temporal confidence scores are used as one of the inputs in the integration part.

D. Integrating spatio-temporal model with CNN

At this point we have obtained a spatio-temporal confidence score for each QG pair, so that we can integrate our spatio-temporal model and our CNN for visual similarity features. As such, a weight optimization procedure is required to combine the following three parameters: the re-ID confidence score from our fully trained MGN network (Parameter *P1*), the spatial confidence score (*P2*) and the spatio-temporal confidence score (*P3*). We investigate the following three different combination approaches to obtain the final similarity score for a single QG pair:

$$Final\ score\ v1 = P1 * w1 + P3 * w3, \qquad (2)$$

$$Final\ score\ v2 = P1 * w1 + P2 * w2 + P3 * w3, \qquad (3)$$

$$Final\ score\ v3 = P1 * w1 + (P2 * P3 * w4), \qquad (4)$$

with weights *w1-w4.* Eqn. (2) is specifically considered since the optimal value for *w2* proved to be very small. As shown in Approach A, we apply this weight optimization procedure on the ‘Train query’ and ‘Train gallery’ dataset.

After ranking all gallery images (for one query at a time) using these final similarity scores, we evaluate the performance of each equation using three objectives. The first objective is to maximize the overall mean average precision (mAP). The second objective is inspired by the fact that for a few query samples, the spatio-temporal model reduces the AP score. Therefore, we determine the weights that maximize the sum of positive and negative mean changes of the AP. The final objective is that the optimal weights should also maximize the Rank-1 score. The optimal weights are thus obtained by solving:

$$W_{opt} = \underset{w}{^{argmax}}(mAP + |avg(AP \uparrow) + avg(AP \downarrow)| + Rank1), \quad (5)$$

which is a challenging nonlinear optimization problem. Parameters $avg(AP \uparrow)$ and $avg(AP \downarrow)$ indicate the mean value of net positive and net negative change in AP (Obj. 2). Value W_{opt} is the obtained optimal weight list.

We tried to find the weights automatically by using the trust-region-reflective algorithm [29], but the algorithm returns local minima for all non-zero weights. This is likely caused by the gradient of the provided cost surface being too small to start the algorithm. Therefore, we defined the optimal weights instead by creating a weights table between zero and unity, evaluated the performance for each setting and then manually identified the weights that satisfy Eqn. (5). Once the optimal set of weights is found, our approach will be validated using the ‘Query’ and ‘Gallery’ dataset.

Experiments and Results

In this section, we describe our experimental results and validate the performance of our approach. The considered imagery originates from the DukeMTMC-reID dataset, since this is currently the only broadly used re-ID dataset that publishes the original camera output and separate ground truth of the sample data including pedestrians locations in the camera frames.

The DukeMTMC-reID dataset originates from eight cameras in a network. After counting the number of transition events in the ‘Train new’ set (see Fig. 16) we found that these eight cameras are not fully connected. In detail, there are some camera pairs that seldomly have transition events, which means that few people travel directly between these cameras without appearing in other cameras. Therefore, we decided to filter out the amount of considered pairs.

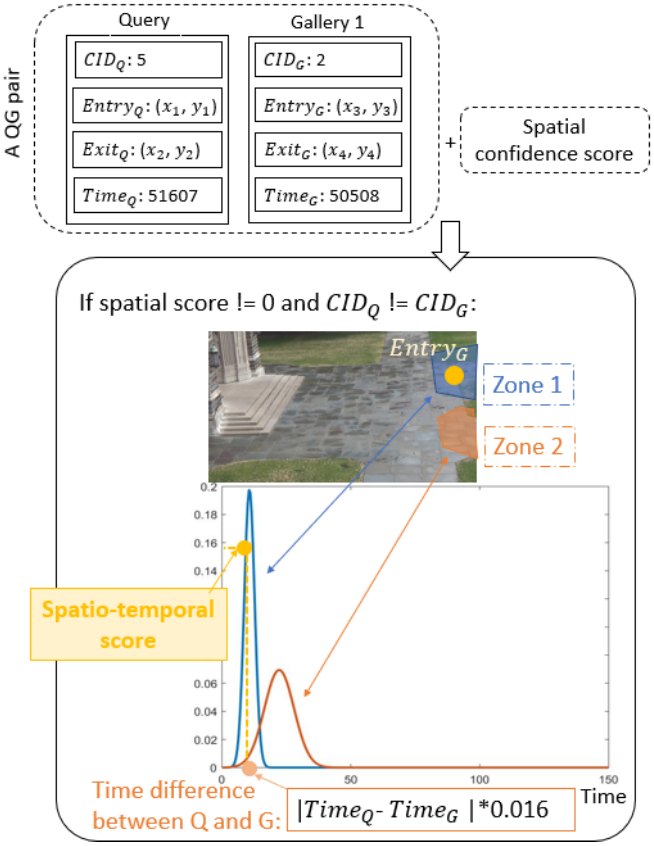


Figure 15. Example of obtaining the spatio-temporal confidence score. The temporal calculation is based on the computed spatial score. $Entry_Q$ *is fed into the zone classifier to choose the correct temporal model (blue or red time distribution), the spatio-temporal score is calculated by finding the probability corresponding to the time difference between the QG pair.*

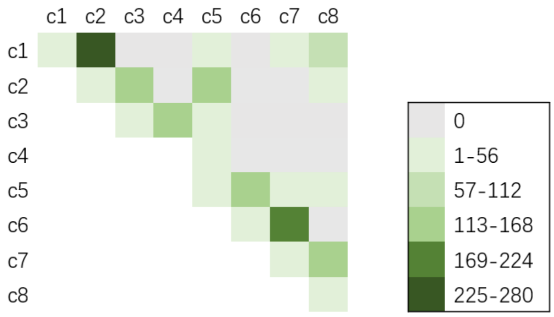


Figure 16. The number of people that travel directly between two cameras. For example, the grid with the darkest green (first row, second column) indicates that 225-280 people walked via travel path c1↔c2.

We defined that two cameras are only linked if there are more than two transition events between these cameras. After filtering, we obtain 15 pairs of linked cameras, which simplifies the camera topology network, as shown in Fig. 17.

We utilize the simplified camera topology network to generate our spatio-temporal multi-camera model based on the ‘Train new’ dataset. We start with training the spatial model. Since there are eight cameras in the DukeMTMC-reID dataset, we train eight different classifier models considering both RF and SVM (RF1 to RF8, and SVM1 to SVM8) to calculate the spatial confidence scores. After the spatial training phase, we proceed to the evaluation of these trained models. The accuracies of the trained classifier models are listed in Table 1. Overall, RF always outperforms SVM or obtains equal accuracy. Additionally, all RF models except RF1 and RF5 show a near-optimal accuracy. The reason for these exceptions is that there are mixture areas in the corresponding Cameras 1 and 5, from which people go to multiple cameras. As an example, consider the bottom right part of Fig. 9. People that walk by following paths c5↔c2 or c5↔c1 will pass through that same area. Consequently, the uncertainty in these mixture areas is high, which increases the difficulty of providing proper classification probabilities.

After the spatial model is computed, we can start calculating the temporal model. As given in Approach C, our temporal model is created by fitting multiple transition time distributions for the linked camera pairs, with example paths 'c3→c4' and 'c2→c5' shown in Fig. 14. To classify which 'zone' to use for a QG pair, in case there are multiple transition time distributions, we consider both RF and SVM zone classifiers. The results in Table 2 show that the RF and SVM zone classifier provide identical Rank-1 and mAP results on 'Train Query' and 'Train Gallery'. Therefore, we have added an experiment with a larger dataset, where training and test data were mixed. We are aware that this can make the results less reliable, but the experiment is interesting to consider some of the influence of larger datasets. Based on this, Table 3 shows that the RF and SVM zone classifiers perform differently on the official evaluation set. The reason is that a broader evaluation dataset offers higher diversity of data, which helps to explore the ignored difference. Besides this discussion, Table 1 shows that the RF classifier also outperforms the SVM classifier. Hence, in the remainder of the paper, we focus on the RF zone classifier based spatio-temporal model.

Once the spatio-temporal multi-camera model is generated, we use it to obtain the spatio-temporal confidence score for each QG pair in 'Train query' and 'Train gallery' set. Fig. 18 visualizes the visual similarity- (CNN re-ID), the spatial-, and the spatio-temporal confidence scores. By this illustration, we notice that there are some imperfections in the visual similarity score (the grey parts in Fig. 18a) and spatial score (see Fig. 18b), but not in the spatio-temporal scores (see Fig. 18c). This finding confirms that our spatio-temporal confidence scores can help to emphasize the more-likely correct matches.

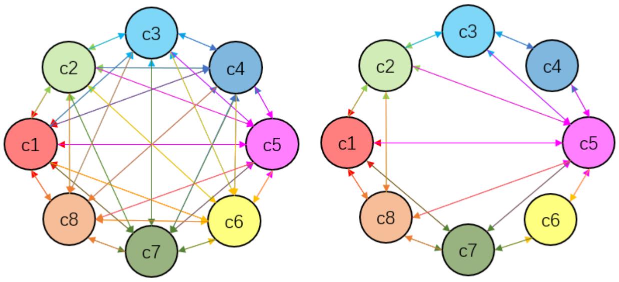


Figure 17. *After analyzing the number of transition events, the camera topology network has been simplified from a fully connected graph (left) to a graph with only 15 connections (right).*

Table 1: Accuracy (acc.) of the trained spatial models (RF and SVM classifiers) on the spatial model validation set (10% of the 'Train new' dataset), where accuracy=$\frac{\textit{number of correct predictions}}{\textit{number of validation samples}}$

Model	Acc. (RF)	Acc. (SVM)	Model	Acc. (RF)	Acc. (SVM)
Cam.1	84.2%	78.5%	Cam.5	85.4%	80.5%
Cam.2	98.3%	98.3%	Cam.6	94.3%	94.3%
Cam.3	100%	91.2%	Cam.7	94.7%	89.5%
Cam.4	93.3%	93.3%	Cam.8	92.3%	92.3%

To combine the obtained re-ID confidence scores and the calculated spatio-temporal confidence scores properly, we follow the procedure explained in the Approach D section. Table 4 lists the computed optimal weights. Given these optimal weights, it is evident that integrating our spatio-temporal multi-camera model into the re-ID process is beneficial. However, we noticed that the optimal *w2* for Eqn. (3) is very small, which means the spatial score (*P2*) has a low contribution to the total score. Furthermore, the result of an ablation study (Table 5) proves that the spatial score in the integration phase is redundant. That is, after abandoning the spatial score, the mAP and Rank-1 do not change, but we receive fewer samples where AP becomes lower, and there is a larger disparity

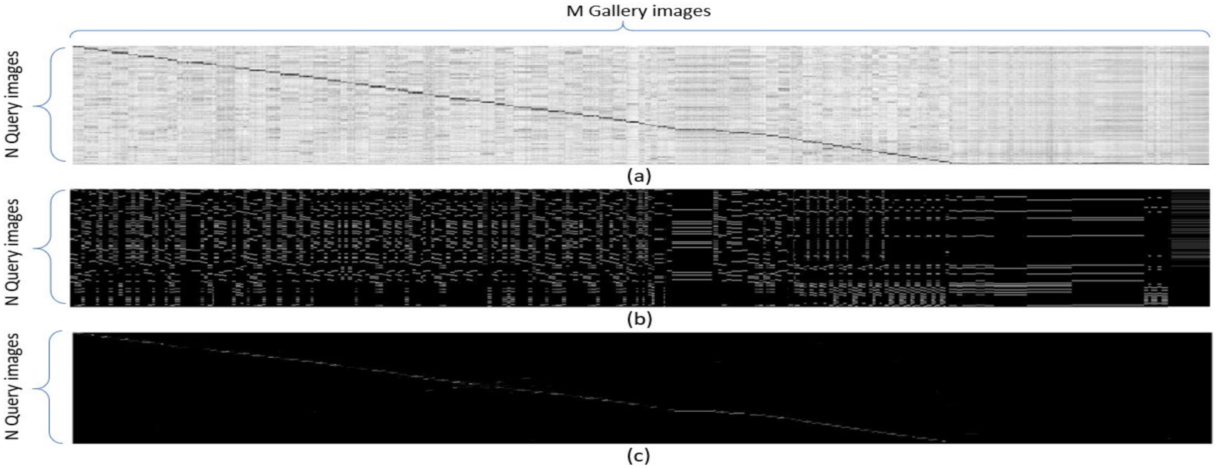


Figure 18. *Illustration of three kinds of confidence scores for every QG pair in the 'Train query' & 'Train gallery' set. (a): "visual similarity score of the re-ID CNN", where a point is dark if the current query (row index) and the current gallery image (column index) has high similarity. (the feature vector distance is small, hence it's dark); (b): "Spatial confidence scores", where white means the current query and the current gallery image satisfies the spatial constraints. (c): "Spatio-temporal confidence scores", where white means the current query and the current gallery image satisfies the spatio-temporal constraints.*

Table 2: Comparison of re-ID performance on 'Train query' and 'Train gallery' dataset with spatial models (st) using the RF and SVM zone classifier.

	Rank-1	mAP
CNN + ST model (RF)	100%	92.7%
CNN + ST model (SVM)	100%	92.7%

Table 3: Experiment of Table 2 repeated on 'query' and 'gallery'

	Rank-1	mAP
CNN + ST model (RF)	95.1%	83.0%
CNN + ST model (SVM)	94.9%	82.9%

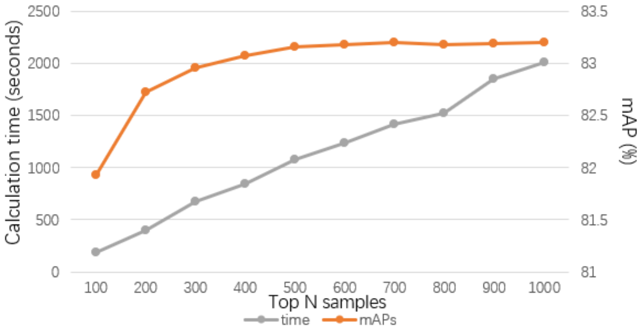


Figure 20. *Influence on mAP and calculation time when calculating the spatio-temporal confidence scores only for the N top-performing samples, according to the CNN.*

between the mean value of the positive and negative changes of AP. Therefore, Eqn. (3) is not considered in the remainder of the paper.

A. Re-ID Performance on the official evaluation set

Previously, only the 3 newly created training sets were used, since these steps involve training of algorithms. With the model training completed, we validate our approach based on the 'Query' and 'Gallery' datasets. The mAP and Rank-1 scores on the DukeMTMC-reID dataset with and without applying our model are shown in Table 6. From the results, the score aggregation with Eqn. (2) performs better than the other weight aggregation approach. The reason is that Eqn. (4) involves the spatial score (*P2*), which actually lower the performance. Therefore, we decide to apply Eqn. (2) as our integration approach in calculating the final similarity score for a QG pair. The CMC curve is shown in Fig. 19. Consequently, our spatio-temporal multi-camera model improves the mAP and Rank-1 score by 7.4% and 7.8% respectively.

B. Computational performance considerations

Currently, the above algorithm takes around 6 hours with a Threadripper 1920X processor (12 cores, 24 threads, 3.5 GHz, 4.0 GHz max clock frequency) to finish calculating the spatial-temporal confidence scores. To improve computational efficiency, the algorithm can be modified. Instead of calculating the spatio-temporal score for every QG pair, we first rank the gallery images based on the CNN re-ID confidence scores and calculate the spatio-temporal confidence score only for the top-N samples. As shown in Fig. 20 and Fig. 21, considering the top-400 ranked samples already lead to a similar re-ID performance as when all samples were included (see Table 6), but save 97.3% of the calculation time.

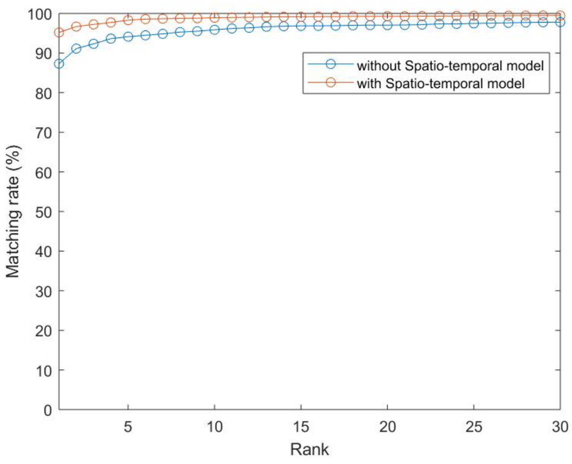


Figure 19. *Effectiveness of our novel spatio-temporal model on DukeMTMC-reID dataset. Blueline: CMC with only CNN. Redline: CMC with CNN + spatio-temporal model.*

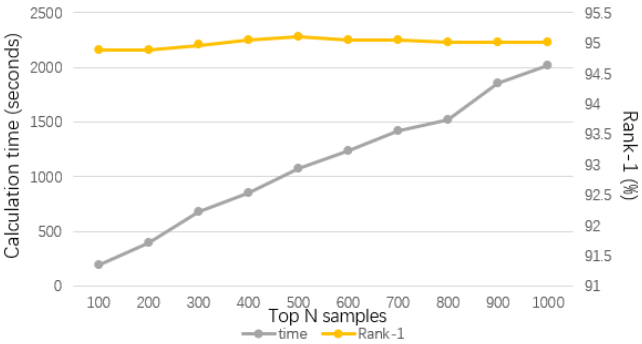


Figure 21. *Influence on Rank-1 score and calculation time, when calculating the spatio-temporal confidence scores only for the N top-performing samples, according to the CNN.*

C. Final Re-ID Performance on the official evaluation set

In Subsection A, the utilized re-ID CNN is trained on the smaller 'Train new' set. To improve the re-ID performance further, we keep the obtained optimal weights, as shown in Table 4, but train a new CNN model on the full size 'Train' set. After obtaining its new visual similarity score, we utilize Eqn. (2) to combine the CNN model with the spatio-temporal model. As shown in Table 7, the mAP and Rank-1 score significantly improved, by 6.9% and 7.6%, respectively. Note that the gain is different due to the changed baseline.

D. Comparison with State-of-the-art

We compare our approach with the top-performing state-of-the-art methods in Table 8. To ensure a fair comparison, we omit post-processing such as re-ranking. When comparing the performances in Table 8, several interesting trends are visible. First, when comparing our performance with that of MGN [1], the large increase of performance by adding our novel spatio-temporal model becomes directly visible. Second, the performance of Parameter-

Table 4: Optimal weights for the three integration approaches.

	$w1$	$w2$	$w3$	$w4$
Eqn. (2)	0.37	-	0.64	-
Eqn. (3)	0.37	-0.0005	0.64	-
Eqn. (4)	0.4	-	-	1

Table 5: Ablation study results of the spatial confidence score

	CNN+ST+S	CNN+ST
mAP	92.7%	92.7%
Rank-1	100%	100%
N samples with AP↑	92	85
N samples with AP↓	13	9
Mean positive change of AP↑	+12.7%	+13.7%
Mean negative change of AP↓	-2.6%	-3.7%

Table 6: Comparison of the re-ID performance on DukeMTMC-reID dataset with different spatio-temporal (ST) integration approaches. CNN is trained on 'Train new' dataset.

	Rank-1	mAP
CNN	87.3%	75.6%
CNN + ST model (with Eqn. (2))	95.1%	83.0%
CNN + ST model (with Eqn. (4))	95.0%	82.1%

Table 7: Re-ID performance after integrating our spatio-temporal model with the CNN trained on the 'Train' dataset.

	Rank-1	mAP
CNN	88.6%	77.2%
CNN + ST model (Eqn. (2))	96.2%	84.1%

Table 8: Comparison to the state-of-the-art methods on the DukeMTMC-reID dataset. (Red and blue indicate best and second-best results respectively)

Method	Rank-1	mAP
St-ReID [22]	94.0%	82.8%
BoT Baseline [11]	86.4%	76.4%
DG-Net [30]	86.6%	74.8%
Parameter-Free Spatial Attention [12]	89.0%	85.9%
ABD-Net [13]	89.0%	78.6%
HPM [15]	86.6%	74.3%
OSNet [16]	88.6%	73.5%
PCB (RPP) [17]	83.3%	69.2%
Incremental Learning [14]	80.0%	60.2%
MGN [1]	88.7%	78.4%
Ours	96.2%	84.1%

Free Spatial Attention algorithm [12] actually shows the performance including re-ranking (without re-ranking is not clearly reported), which probably explains its high performance. However, even if its non-re-ranking score is higher than MGN, our model would benefit from its increased performance if we would change the adopted CNN. Finally, St-ReID [22] is also very interesting in this list, as this approach also studies spatio-temporal effects, but in a different way. Furthermore, in their approach, time differences between different roads when traveling between the same set of cameras was not considered.

Summarizing, it is fair to conclude that our approach outperforms all state-of-the-art approaches by a large margin. This conclusion is mostly motivated by the Rank-1 performance, since we consider the Rank-1 score as most important if person re-ID is applied in practice (open set considerations, etc.).

Discussion

To design a spatio-temporal model, we initially planned to construct a full-fledged 3D multi-camera model by implementing camera calibration based on frames from the selected re-ID dataset (DukeMTMC-reID). This 3D multi-camera model would provide 3D coordinates for all persons in the scenes. Based on these 3D coordinates of the query person and the candidates in the gallery, the spatio-temporal constraints can be precisely identified. However, we have found that constructing such a 3D multi-camera model for the DukeMTMC-reID dataset is not feasible due to the following reasons. First, not every camera FOV contains enough structural scenes, such as buildings, to provide straight lines in three orthogonal directions (as shown in Fig. 22). Second, it is difficult to estimate the height of pedestrians with a single camera per view. This information is required to find vanishing points and achieve camera calibration [31] [32]. Therefore, instead of defining a 3D multi-camera model, we have decided to construct the spatio-temporal multi-camera topology model, as in Approach C.

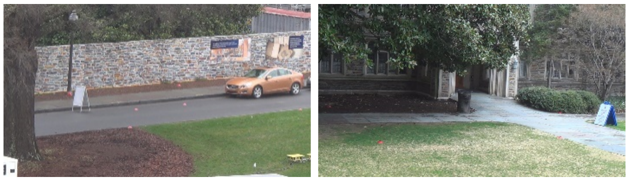


Figure 22. Examples of camera FOV which contains not enough structured scenes to explore vanishing points in three orthogonal directions.

Conclusion

In this paper, a spatio-temporal multi-camera model has been generated, mainly based on the entry/exit points (and timestamps) of pedestrians in the camera views. Our model determines how likely a query image and its candidates from the gallery satisfy the identified spatial and temporal constraints. We deploy Random Forest (RF) to train the spatial model on the DukeMTMC-reID dataset, while the fitted Gaussian probability distributions represent the temporal constraints. By empirical experiments, we have found the optimal way of integrating our novel spatio-temporal multi-camera model with the MGN network. Experimental results show that for the DukeMTMC-reID dataset, after integrating the multi-camera model, the mean average precision increases from 77.2% to 84.1%, while the Rank-1 score augments from 88.6% to 96.2%. This result outperforms all current state of the art by a large margin.

Acknowledgment

The work in this paper is funded by the European PS-Crimson project, in the framework of the ITEA research program.

References

[1] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou, “Learning Discriminative Features with Multiple Granularities for Person Re-Identification,” arXiv:1804.01438 [cs.CV].

[2] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking,” in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[3] Z. Zheng, L. Zheng and Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[4] A. Bedagkar-Gala and S. K.Shah, “A survey of approaches and trends in person re-identification,” *Image and Vision Computing,* pp. 270-286, 2014.

[5] T. Xiao, H. Li, W. Ouyang and X. Wang, “Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] X. Zhang , H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang and J. Sun, “AlignedReID: Surpassing Human-Level Performance in Person Re-Identification,” arXiv:1711.08184, 2018.

[7] A. Hermans, L. Beyer and B. Leibe, “In Defense of the Triplet Loss for Person Re-Identification,” arXiv:1703.07737v4, 2017.

[8] N. Martinel, G. Luca Foresti and C. Micheloni, “Person Reidentification in a Distributed Camera Network Framework,” *IEEE Transactions on Cybernetics,* pp. 3530-3541, 2017.

[9] J. Almazan, B. Gajic, N. Murray and D. Larlus, “Re-ID done right: towards good practices for person re-identification,” arXiv: 1801.05339, 2018.

[10] Y. Sun, L. Zheng, W. Deng and S. Wang, “SVDNet for Pedestrian Retrieval,” arXiv:1703.05693, 2017.

[11] H. Luo, Y. Gu, X. Liao, S. Lai and W. Jiang, “Bag of Tricks and A Strong Baseline for Deep Person Re-identification,” arXiv:1903.07071, 2019.

[12] H. Wang, Y. Fan, Z. Wang, L. Jiao and B. Schiele, “Parameter-Free Spatial Attention Network for Person Re-Identification,” arXiv:1811.12150, 2018.

[13] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren and Z. Wang, “ABD-Net: Attentive but Diverse Person Re-Identification,” arXiv:1908.01114, 2019.

[14] P. Bhargava, “Incremental Learning in Person Re-Identification,” arXiv:1808.06281, 2018.

[15] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao and T. Huang, “Horizontal Pyramid Matching for Person Re-identification,” arXiv:1804.05275, 2018.

[16] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, “Omni-Scale Feature Learning for Person Re-Identification,” arXiv:1905.00953, 2019.

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang, “Beyond Part Models: Person Retrieval with Refined Part Pooling,” arXiv:1711.09349, 2018.

[18] Y. Cai and G. Medioni, “Exploring context information for inter-camera multiple target tracking,” in *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[19] A. Rahimi, B. Dunagan and T. Darrell, “Simultaneous calibration and tracking with a network of non-overlapping sensors,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[20] S. Bak, F. Martins and F. Bremond, “Person re-identification by pose priors,” in *IST/SPIE Electronic Imaging*, 2015.

[21] Y. Cho and K. Yoon, “Distance-based Camera Network Topology Inference for Person Re-identification,” arXiv: 1712.00158, 2017.

[22] G. Wang, J. Lai, P. Huang and X. Xie, “Spatial-Temporal Person Re-identification,” arXiv:1812.03282, 2018.

[23] X. Li, W. Dong, F. Chang and P. Qu, “Topology Learning of Non-overlapping Multi-camera Network,” *International Journal of Signal Processing, Image Processing and Pattern Recognition,* pp. 243-254, 2015.

[24] Y. Nam, J. Ryu, Y.-J. Choi and W.-D. Cho, “Learning Spatio-Temporal Topology of a Multi-Camera Network by Tracking Multiple People,” *World Academy of Science, Engineering and Technology International Journal of Computer and information Engineering,* pp. 1549-1554, 2007.

[25] Y. Liu, H. Zhang and Y. Wu, “Hard or soft classification? large-margin unified machines,” *American Statistical Association,* pp. 166-177, 2011.

[26] L. Breiman, “Random Forest,” in *Machine Learning*, 2001, pp. 5-32.

[27] A. Criminisi, Decision Forests for Computer Vision and Medical Image Analysis, 2013.

[28] MATLAB, “dbscan,” 2019. [Online]. Available: http://www.matworks.com/help/stats/dbscan.html.

[29] MATLAB, “MathWorks,” [Online]. Available: https://nl.mathworks.com/help/optim/ug/lsqnonlin.html#buul744-1. [Accessed 6 2019].

[30] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang and J. Kautz, “Joint Discriminative and Generative Learning for Person Re-identification,” arXiv:1904.07223, 2019.

[31] F. Lv, T. Zhao and R. Nevatia, "Camera calibration from Video of a Walking Human," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* pp. 1513-1518, 2006.

[32] S. Li, V. Nguyen, M. Ma, C. Jin, T. Do and H. Kim, “A simplified nonlinear regression method for human height estimation in video surveillance,” *EURASIP Journal on Image and Video Processing,* 2015.

[33] Z. Zheng, L. Zheng and Y. Yang, “Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in vitro,” *arXiv preprint arXiv:1701.07717,* 2017.

Author Biographies

Xin Liu received her MSc degree in Electrical Engineering in 2019 from the Eindhoven University of Technology, in research on person re-identification. She is currently a PhD candidate at the same university. Her research interests include computer vision and simultaneous localization and mapping.

Herman Groot is a PhD at the Electrical Engineering faculty of Eindhoven University of Technology (TU/e, the Netherlands). His PhD study currently focuses on person re-identification, but halfway through his PhD, the focus will shift more towards robotics, since – ultimately – he strives to be involved in future space-exploration missions. To this end, he would eagerly want to broaden his image processing skills in order to become an expert in space-related image processing techniques. Fittingly, he finalized several MSc elective courses at the Aerospace Engineering faculty of Delft University of Technology (TU Delft, the Netherlands) and did his MSc internship at the Netherlands Aerospace Centre in Amsterdam (NLR, the Netherlands).

Egor Bondarev obtained his PhD degree in the Computer Science Department at TU/e, in research on performance predictions of real-time component-based systems on multiprocessor architectures. He is an Assistant Professor at the Video Coding and Architectures group, TU/e, focusing on sensor fusion, smart surveillance and 3D reconstruction. He has written and co-authored over 50 publications on real-time computer vision and image/3D processing algorithms. He is involved in large international surveillance projects like APPS and PS-CRIMSON.

Peter H.N. de With is Full Professor of the Video Coding and Architectures group in the Department of Electrical Engineering at Eindhoven University of Technology. He worked at various companies and was active as senior system architect, VP video technology, and business consultant. He is an IEEE Fellow and member of the Royal Holland Society of Sciences, has (co-)authored over 400 papers on video coding, analysis, architectures, and 3D processing and has received multiple papers awards. He is a program committee member of the IEEE CES and ICIP and holds some 30 patents.

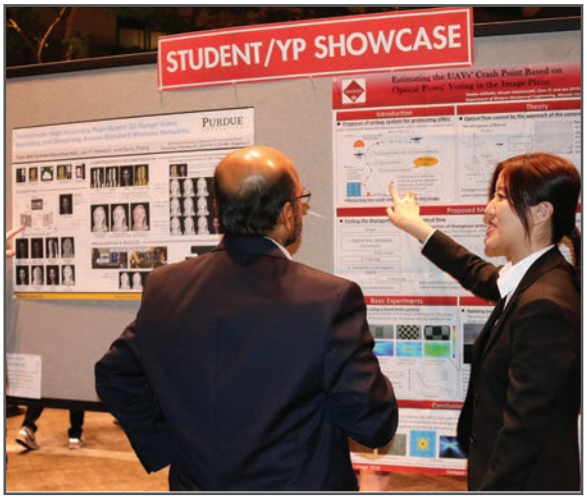
JOIN US AT THE NEXT EI!

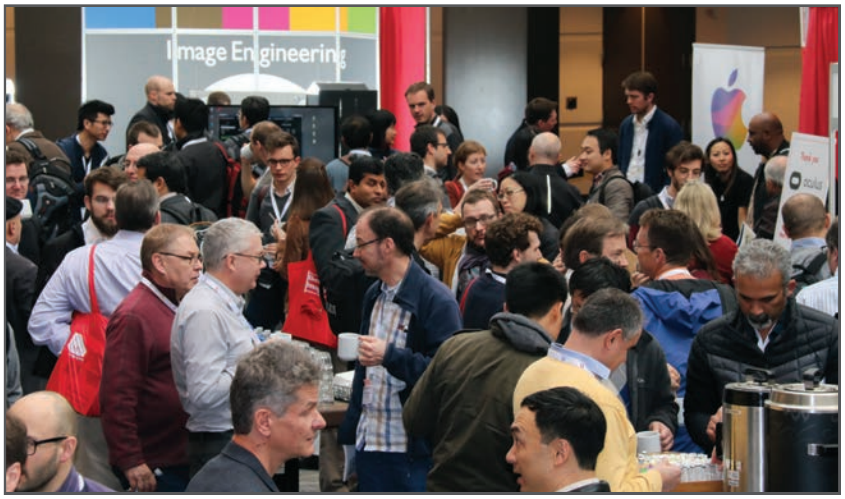
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org