Improving Person Re-Identification with Temporal Constraints

Julia Dietlmeier*, Feiyan Hu*, Frances Ryan, Noel E. O'Connor, Kevin McGuinness
Insight SFI Research Centre for Data Analytics
Dublin City University, Glasnevin 9, Dublin, Ireland
{julia.dietlmeier,feiyan.hu,frances.ryan,noel.oconnor,kevin.mcguinness}@insight-centre.org
* All these authors contributed equally

Abstract

*In this paper we introduce an image-based person re-identification dataset collected across five non-overlapping camera views in the large and busy airport in Dublin, Ireland. Unlike all publicly available image-based datasets, our dataset contains timestamp information in addition to frame number, and camera and person IDs. Also our dataset has been fully anonymized to comply with modern data privacy regulations. We apply state-of-the-art person re-identification models to our dataset and show that by leveraging the available timestamp information we are able to achieve a significant gain of **37.43**% in mAP and a gain of **30.22**% in Rank1 accuracy. We also propose a Bayesian temporal re-ranking post-processing step, which further adds a **10.03**% gain in mAP and **9.95**% gain in Rank1 accuracy metrics. This work on combining visual and temporal information is not possible on other image-based person re-identification datasets. We believe that the proposed new dataset will enable further development of person re-identification research for challenging real-world applications.*

1. Introduction

Person re-identification (re-ID) aims at matching people across multiple non-overlapping camera views. Specifically, given a *query* image the task is to find a match in a large *gallery* of images. This task generally falls into the category of image retrieval problems. The environment in which the real-world person re-ID systems are deployed ranges from university campuses and streets to airports and train stations [4]. Currently, deep learning-based person re-ID has reached performance saturation on several public datasets. An example of this is the Market1501 dataset where some methods achieve higher Rank1 accuracy than humans [27]. On the other hand, the performance on the challenging VIPeR dataset still remains comparatively low. Furthermore, person re-ID faces several key challenges such as various camera angles, varying lighting conditions and occlusions [21]. The so-called "appearance ambiguity" [26] problem, where different individuals may have a similar appearance, makes it difficult for most person re-ID models to improve performance further by focusing solely on visual features.

Modern person re-ID research concentrates on deep learning-based approaches and over time migrated from the handcrafted features and distance metric learning. A timely and comprehensive review of the re-ID methods is provided in [27]. The size of publicly available academic person re-ID datasets has increased over time and this has facilitated the training of large deep learning-based models. Most present person re-ID methods, however, still do not scale to large datasets and ignore spatial-temporal constraints that could potentially alleviate the appearance ambiguity problem. The core of these algorithms is to learn pedestrian features and similarity functions that are view invariant and robust to camera change [19]. These methods are still far from applicable to large-scale real-world scenarios containing a large amount of candidate images (which, following the literature, we refer to as gallery images) that need to be searched to re-identify a person [26, 3]. An airport is a good example of this scenario, where there are potentially thousands of cameras, each generating tens of thousands if not hundreds of thousands of frames per minute, all of which are potential candidates for re-identification should spatial-temporal priors not be taken into account. The re-ID research on fusing visual and spatial-temporal information is, however, being hindered by the availability of suitable datasets. In this work, we make three contributions:

- First, we introduce a new dataset, called DAA[1], collected in the large and busy Dublin Airport in Ireland. The main feature of the DAA dataset is that all images contain time information (timestamps) in addition to the frame number, person, and camera IDs. Also, our dataset has been fully anonymized to comply with the strict data privacy regulations.

[1]The name DAA stands for Dublin Airport Authority

Name	Cameras	Query	Gallery	Frame #	Time	Anonymized
Airport [10, 1]	6	1,003	2,420	yes	no	no
Market1501 [29]	6	3,368	19,732	yes	no	no
DukeMTMC-reID [7]	8	2,228	17,661	yes	no	no
CUHK03 [11]	6	5,332	1,400	no	no	no
VIPeR [8]	2	316	316	no	no	no
DAA (Ours)	5	71	1,029	yes	yes	yes

Table 1: Comparison of commonly used image-based person re-ID datasets.

- Second, we show that by leveraging the available temporal information we are able to significantly increase the performance of the state-of-the-art person re-ID methods selected. Specifically, we achieve a substantial gain in mAP of **37.43**% and a Rank1 gain of **30.22**%.
- Third, we develop an efficient temporal re-ranking algorithm that further adds **10.03**% gain in mAP and **9.95**% gain in Rank1 accuracy metrics.

The remainder of the paper is organized as follows: Section 2 reviews the commonly used image-based person re-ID datasets. Section 3 describes our new dataset and its key features. Section 4 outlines the methodology and experiments. Section 5 introduces a new temporal re-ranking step and Section 6 discusses our findings.

2. Related Work

Person re-ID datasets can be mainly categorized into image- and video-based [27, 30]. In this work we focus on the first category where the query person is represented by an image (extracted from a CCTV video). Below, we review the image-based datasets commonly used in the person re-ID community in the chronological order in which they were released. Table 1 details the image-based datasets reviewed in this section.

The Viewpoint Invariant Pedestrian Recognition (**VIPeR**) [8] dataset was released in 2008 and contains 632 person images from two disjoint outdoor camera views. The dataset was collected under different viewpoints and illumination conditions and therefore is considered as one of the most challenging. Each person has one image per camera. Each image is resized to be 48 $\times$ 128 pixels. In addition, each image has a pedestrian viewpoint information tag of 0° (front), 45°, 90° (right), 135°, and 180° (back) degrees. The **CUHK03** [11] dataset was collected from six surveillance cameras in the University campus environment and released in 2014. It was the first large-scale dataset that was large enough for training deep learning-based re-ID models. Each person identity is captured by two disjoint camera views and has an average of 4.8 images in each view. Because it was collected in a more realistic setting, this dataset features such problems as misalignment, occlusions, and missing body parts. The bounding boxes were extracted using manual annotations and the Deformable Part Models (DPM) object detector [6]. The next very popular large-scale dataset, **Market1501** [29], was released in 2015. It also features images extracted using the DPM detector and thus inherits problems of background clutter and misalignment. Other problems include viewpoint variations, detection errors, and low resolution [10]. The dataset was collected from a total of six cameras (overlapping exists) placed in front of a campus supermarket in Tsinghua University. The authors provide false positive bounding boxes and also 2,793 false alarms and 500,000 distractors. There are 3.6 images on average for each of 1,501 identities and the total number of images amounts to 32,668. **DukeMTMC-reID** [7] is another large-scale dataset introduced in 2017 that features images collected outdoors on the Duke University campus. The dataset contains manually labeled bounding boxes for a total of 1,812 identities from eight cameras. There are 1,404 identities appearing in more than two cameras and 408 identities (distractors) appearing in only one camera [31]. Further, the dataset contains 2,228 query images, 17,661 gallery images, and 16,522 training images. **Airport** [1, 10] is a relatively recent large-scale image-based person re-ID dataset released in 2018 that is close to our work in both aspects: a real-world airport environment and associated image type and quality. The data were collected from six non-overlapping indoor surveillance cameras in a mid-sized airport in Cleveland, USA. Each camera captured 12-hour long videos at 30 frames per second from 8AM to 8PM. In total, there are 39,902 bounding boxes and 9,651 unique identities. The authors employed the ACF [5] framework to extract bounding boxes. Although not elaborated in [1] and [10], the images in the Airport dataset carry some relative time information. An example of an image from the query folder is `0001l004_c37_t01_0002.jpg`. From here we infer that the time is not given in seconds as t ranges from 1 to 40 in the query and gallery folders. We presume that this is the video clip number as each video was split in 40 clips (each 5 minutes long) [10]. We could not find any publication exploiting this temporal characteristic of the Airport dataset. In contrast to the previous datasets that capture data in public settings, the video data in the Airport dataset

		Market1501			Proposed DAA dataset		
		mAP	Rank1	Rank5	mAP	Rank1	Rank5
MLFN	Original	74.3	90.1	95.9	35.9	66.2	88.7
	+TR	+2.2	+1.2	+1.0	+6.8	0	−5.6
	+TLift	+1.2	+2.0	+1.0	+4.9	+1.4	−4.2
	+TLift*	+1.0	+1.5	+0.8	+7.1	−2.8	−4.2
HACNN	Original	75.3	90.6	96.2	31.8	62.0	83.1
	+TR	+2.2	+1.6	+0.5	+4.6	0	−5.6
	+TLift	+1.3	+1.7	+0.9	+3.9	−1.4	−1.4
	+TLift*	+1.2	+1.6	+0.7	+7.1	−1.4	−4.2

Table 2: Performance increase with the introduction of frame numbers on Market1501 and our proposed datasets. TR (Temporal Re-ranking) is our proposed re-ranking approach using temporal prior in Section 5. TLift, proposed in [15], models temporal probability with a Gaussian Kernel. TLift* is the same method as TLift, but uses the Gamma distribution instead.

is streamed from the inside of the secure area, beyond the security checkpoint, of an airport [10]. As highlighted by the authors, it is generally very difficult to obtain data from such a camera network, which is similar to ours.

A few papers on spatial-temporal person re-ID [14, 19, 26, 15] work with DukeMTMC-reID, Market1501, and sometimes the small-scale GRID dataset [17]. In these works a common approach is to treat the frame number as timestamp. The evaluation of the temporal performance of re-ID models is still considered to be under explored [10, 26].

In the aforementioned datasets, which contain incomplete temporal information such as frame numbers, relative temporal difference within a camera could be inferred. Temporal differences across cameras, however, are impossible to infer. To demonstrate the extra performance that knowledge of frame numbers could impart, Table 2 shows a performance increase with the introduction of frame numbers on Market1501 and the proposed DAA dataset. We benchmark two state-of-the-art re-ID models such as MLFN [2] and HACNN [12] and observe that using the provided frame numbers can boost performance, especially mAP. Even by just using time prior (Section 5) within each camera, performance is better than that achieved with more complicated ranking methods like TLift [15]. The frame number can only be used to compute the within-camera time difference; however, most applications would require cross-camera search. To compute cross-camera time differences, timestamp information is needed instead of frame numbers. In Section 5, we show that significant improvements can be made just by using the timestamp information included in the proposed DAA dataset to re-rank results. This is shown in Table 4 and Table 5.

We conclude that none of the datasets reviewed above explicitly provide timestamp information and none have been anonymized (see Table 1). It is now accepted that there are significant privacy concerns when releasing new datasets across research communities [20]. At the time of writing, some of the image-based person re-ID datasets listed in Table 1 have been taken offline due to the privacy concerns [22] and/or facing potential legal challenges. It is reasonable to envisage that anonymization of faces could lead to performance decaying. However, recent work [4] shows that face anonymization of person re-ID datasets only marginally affects the performance of state-of-the-art person re-ID methods. Based on these findings we release an anonymized version of this new person re-ID dataset.

3. New Dataset

The proposed DAA dataset was collected in one day from 14-hours long CCTV surveillance videos starting from 3AM to 5PM in one transfer level of the busy airport in Dublin, Ireland. The camera network in this transfer level consists of five digital cameras covering non-overlapping fields of view and streaming videos at variable frame rates. Varying illumination conditions, reflective floor, and reflections in windows and on the walls make up some of the challenges for this work. Figure 2 shows the camera topology of the dataset; Table 7 provides the relative walking distances between cameras.

We extract and tag the timestamp information (in seconds) when processing videos and extracting keyframes with *ffmpeg*. The inclusion of time information is the most prominent feature of our dataset. We encode metadata about the images in the filenames; an example can be given by `21_c0900_t36000_frame0002300_2.jpg`. Here the person ID is 21, and the camera ID is `c0900`. The `t36000` code means this image was sampled at 10AM so that $t = 10 \times 60 \times 60 = 36,000$ seconds. The frame number is 2300 and this is the second bounding box extracted from that frame.

We perform annotation at two levels. In the first-level annotation we extract bounding boxes containing people. The resolution of each camera is 1920×1080 pixels. For the query camera C900 we manually annotate the bounding

Figure 1: Examples of anonymized gallery images for the person identity ID=2708. It can be seen that our dataset has such challenging attributes as occlusions, viewpoint variations, pose and illumination variations, missing body parts and background clutter.

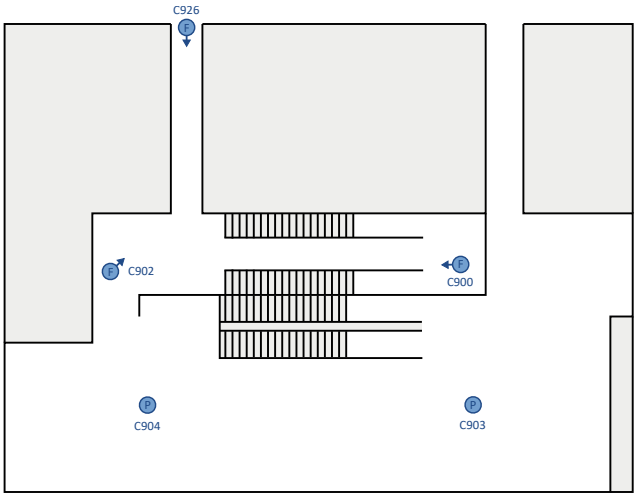

Figure 2: Camera topology of our DAA dataset

boxes using the open source *DarkLabel* tool. For all gallery cameras we use generic object detector SSD [16] (SSD512 model[2]) with the confidence threshold set to 0.25 to automatically generate candidate bounding boxes from the keyframes extracted. Therefore, the sizes of cropped bounding boxes vary in our dataset. In the second-level annotation procedure we manually match people across the four gallery cameras C902, C903, C904, and C926. We had to manually process 195,978 bounding boxes in total. This was the most time-consuming process. The resulting query folder of our dataset consists of 71 unique identities from camera C900 and the resulting gallery folder contains 1,029 images from cameras C902, C903, C904, and C926 with one or more images per identity. Figure 1 illustrates the challenging attributes of our dataset (such as occlusions, viewpoint variations, pose and illumination variations, background clutter, and missing body parts) on a sample ID from the gallery folder.

For the anonymization we adopt the procedure recommended in [4] using the pre-trained TinyFaces face detector [9] to detect the faces first and then blur the regions detected with the large-kernel Gaussian to remove all privacy-sensitive information. We manually verify the quality of anonymization. Our DAA dataset is available under this link: `https://bit.ly/3AtXTd6`.

[2]https://github.com/pierluigiferrari/ssd_keras

4. Methodology

In this section we benchmark some state-of-the-art person re-ID models on our new dataset and show that by using the available temporal constraint we are able significantly to boost performance.

4.1. Performance metrics

We evaluate the performance of re-ID models using the mean average precision (mAP) and Rank1 to Rank20 measures. These metrics rely on ranking images in the gallery dataset according to similarity with the query images. A similarity score is generated for each possible query/gallery image combination. For each query image, this similarity score is then used to rank all of the gallery images according to the estimated likelihood of the gallery image containing the individual in the query image. The top ranked image is the one that the system estimates as having the highest probability of containing the person in the query image, the second ranked image is the next most likely, and so on. Comparing the ranked results with the ground truth (images known to contain the same person as the query image) allows the computation of several metrics. The Rank1 metric evaluates the proportion of queries where the top-ranked gallery image contains the correct person (i.e. the same person appeared in both the query image and the top ranked gallery image); the Rank5 metric evaluates the proportion of queries where at least one of the top-5 ranked gallery images contains the correct person; and Rank10 and Rank20 are the proportion of queries identifying the correct person in at least one of the top 10 or 20 ranked gallery images, respectfully. Mean average precision (mAP) is calculated by taking the mean of average precision (AP) scores across all queries. It gives an estimate of how many images would contain the correct person in the top-n images for all n. In the evaluation of experiments we use a similar approach as in Market1501: gallery images, that share their person ID and camera ID

	Model	mAP	Rank1	Rank5	Rank10	Rank20
1	PCB [24]	33.3±1.6	63.8±4.5	83.5±3.3	88.8±2.2	93.7±1.8
2	MLFN [2]	**38.0±1.6**	66.7±3.3	86.9±2.8	91.7±1.7	94.5±0.9
3	HACNN [12]	31.6±0.9	65.4±2.9	82.7±2.1	88.7±2.1	93.9±1.5
4	Resnet50Mid [28]	32.9±2.0	65.7±4.7	83.5±2.9	89.2±1.6	92.7±2.3
5	MuDeep [23]	18.8±0.4	52.3±2.5	74.2±2.3	81.8±2.2	89.2±1.3
6	BoT [18]	37.1±1.8	**72.3±4.4**	**89.5±2.6**	**93.8±2.3**	**96.4±1.5**

Table 3: Benchmarks on our DAA re-ID dataset. Training for each model is repeated 20 times, and mean performance and standard deviation are reported.

with a query image, are excluded from evaluation.

4.2. Experiments

We conduct benchmark testing on the proposed DAA dataset. Table 3 details the results of this experiment. We use the Torchreid library by Kaiyang Zhou and Tao Xiang [33] to implement the state-of-the-art models 1 to 5. We adopt a cross-dataset methodology and train each model on the Airport [1, 10] dataset (3,493 images) for 60 epochs and evaluate on our DAA dataset. We train the BoT model[3] for 120 epochs on the Airport dataset. Experiments for each model are repeated 20 times. We report the mean and standard deviation of each metric in these repeated experiments. Table 3 shows that MLFN and BoT are the two best models.

One thing to note when considering the results is that the re-ID work is skewed towards individuals wearing distinctive, often brightly coloured, clothing. Annotation, even manually, of individuals wearing dark clothing can be very challenging. Although ideally we would like to have completed our performance evaluation on a more balanced dataset of individuals, in the context of Dublin Airport, being able to correctly re-identify a few individuals within a crowd and combine this with crowd density mapping has the potential to provide information about general passenger flow through a space, as timestamps at different cameras can be used to pinpoint their appearance in a certain area at a given time.

Next, we proceed eliminating irrelevant gallery images using time information. We achieve the first improvement in our re-ID performance by reducing the number of images that are considered when searching the dataset to re-identify a particular person. Our full annotated dataset contains 1,029 candidate images in total for 71 identities. Considering all of these images for every identity being searched can result in an increased potential for false positives, which can arise, for example, by two individuals appearing very similar by coincidence. By imposing time constraints on gallery images we also reduce the total number of images to be processed and reduce theoretical number of multiplications needed when computing distances between query-gallery pairs.

Clearly, there is a strong prior over the candidate images in the gallery that arises from the fact that temporally close images are more likely to contain the same people. Specifically, people entering the airport for the first time only re-appear in future frames, and after leaving one camera are more likely to appear in another camera within a short period of time. We propose to constrain the search based on the available timestamp information t. In this experiment on our DAA dataset we construct reduced galleries for each query based on $t_{\text{query}} + \Delta t$ where $\Delta t = 30$ minutes, for example. This step is based on the observation that a particular person will usually pass from the query camera C0900 through the relevant gallery cameras C0902, C0903, C0904, and C0926 in less than 30 minutes. This approach follows the procedure mentioned in [26], where authors use a spatial-temporal constraint to eliminate lots of irrelevant images in the gallery.

After constructing reduced galleries for each of the 71 query images, we run each re-ID model 71 times and compute the aggregate statistics afterwards. Table 4 shows the results for $T_{min} \leq \Delta t \leq T_{max}$ minutes, clearly demonstrating the enormous improvement in mAP and Ranks 1 to 10. Specifically, we report the average performance gain in mAP across all models for the time interval $[0, 30)$ as **37.43**% and **30.22**% gain in Rank1 accuracy.

5. A Novel Temporal Re-Ranking Approach

The previous experiments of eliminating candidate images that appear outside the Δt minute window from the query image corresponds to applying a box shaped prior distribution to the gallery images. Essentially, this sets the prior probability of an image containing the query identity to zero outside the Δt minute window, and gives equal probability to each candidate image within the Δt minute window. In the following, we describe experiments with more sophisticated priors which have been chosen to reflect the prior belief that, having left a camera, a person is likely to appear in another sooner rather than later. We then use the Bayesian update formula to integrate the evidence coming from the appearance models and the posterior probability of each candidate being relevant.

The following describes the theoretical motivation behind

[3] https://github.com/michuanhaohao/reid-strong-baseline

	Model	mAP	Rank1	Rank5	Rank10
1	PCB [24]	68.4±1.4	86.8±2.5	96.1±1.6	98.1±0.8
2	MLFN [2]	**72.4±0.9**	86.1±2.4	95.1±0.8	97.7±0.8
3	HACNN [12]	67.6±0.6	89.2±2.4	96.0±1.0	97.6±0.6
4	Resnet50Mid [28]	67.1±1.7	86.7±3.1	95.4±2.0	97.4±0.8
5	MuDeep [23]	53.8±0.6	81.1±2.0	95.8±0.9	97.2±0.0
6	BoT [18]	71.0±1.7	**90.5±2.2**	**96.9±1.1**	**98.1±0.8**

Table 4: Improved re-ID performance based on the reduced galleries with time constraints. We choose galleries with timestamp $t_{gallery}$ that satisfy $t_{query} + T_{min} \leq t_{gallery} < t_{query} + T_{max}$ to construct reduced gallery sets. $T_{max} = 30$ and $T_{min} = 0$ here use minutes as units.

	Model	mAP	Rank1	Rank5	Rank10
1	PCB [24]	76.6±0.8	94.1±1.3	98.0±0.8	98.5±0.3
2	MLFN [2]	**78.3±0.7**	90.5±1.6	97.9±0.7	**98.6±0.0**
3	HACNN [12]	75.2±0.7	93.9±0.8	97.7±0.7	**98.6±0.0**
4	Resnet50Mid [28]	73.9±1.5	91.9±1.8	97.3±0.9	98.5±0.4
5	MuDeep [23]	65.8±0.6	91.4±1.7	**98.5±0.3**	**98.6±0.0**
6	BoT [18]	77.3±1.3	**94.7±2.5**	98.0±0.7	98.5±0.3

Table 5: Improved re-ID performance by temporal re-ranking using a Gamma prior for $[0, 30)$.

the approach. Let Y be the event that a gallery image described by a feature vector x_g contains a person of the same identity as the query image described by a feature vector x_q. Also, let Δt be the time difference between the observation of x_g and x_q. Δs is the walking distance between cameras that the observation of x_g and x_q appear. By Bayes rule we have:

$$P(Y|x_q, x_g, \Delta t, \Delta s) \propto P(x_q, x_g|Y, \Delta t, \Delta s)P(Y|\Delta t, \Delta s), \tag{1}$$

where $P(x_q, x_g|Y, \Delta t)$ is the probability of observing feature vectors x_g and x_q given that they contain the same identity and $P(Y|\Delta t)$ can be interpreted as the prior probability of relevance conditioned only on time. If we further assume that the probability of observing feature vectors x_g and x_q is conditionally independent of Δt and Δs given Y, and also make the naive assumption that Δt and Δs are also independent, then this simplifies to:

$$P(Y|x_q, x_g, \Delta t, \Delta s) \propto P(x_q, x_g|Y)P(Y|\Delta t)P(Y|\Delta s). \tag{2}$$

We model the likelihood term on the right hand side of Eq 2 using a normal distribution centered on the query, and we use Laplace distribution to model spatial prior. That is:

$$P(x_q, x_g|Y) \propto \exp\left\{-\frac{\|x_q - x_g\|^2}{2\sigma^2}\right\}, \tag{3}$$

$$P(Y|\Delta s) \propto \exp\left\{-\frac{\|\Delta s\|}{\sigma_s}\right\}, \tag{4}$$

where σ is the standard deviation of the Gaussian and controls how quickly the probability decays as the gallery image moves away from the query image in feature space. Larger σ values reduce the effect of the time prior. σ needs to be re-calibrated for each model separately to account for the differing scales of the feature space. The detailed choice of σ for Table 5 is reported in Table 9.

The prior term $P(Y|\Delta t)$ in Eq 2 should be selected to reflect the fact that having left a camera, the query person is more likely to appear in another camera sooner rather than later. For example we could choose the Gamma family of distributions to allow flexible control over the prior. In this case, we model the temporal prior term using:

$$P(Y|\Delta t) \propto (\Delta t)^{\alpha-1}e^{-\Delta t/\beta}, \tag{5}$$

where α and β are hyperparameters controlling the shape of the distribution. Fig. 4 shows various probability distributions of which the parameters are estimated using empirical distribution of the DAA dataset. Table 8 gives the parameters fit for each probability distribution. Notice how the distributions place more probability that relevant images will appear earlier, and that they can be configured to account for a delay between leaving one camera and entering another.

After ranking by the updated posterior scores we obtain a significant improvement of **10.03%** in mAP and **9.95%** in Rank1 accuracy metrics compared to the results in Table 4. Again, we report the average performance gain across all models. Table 5 shows the performance improvements for the Gamma time prior distribution for the time interval $[0, 30)$.

An experiment using both spatial and temporal priors is reported in Table 6. We can see that comparing with temporal

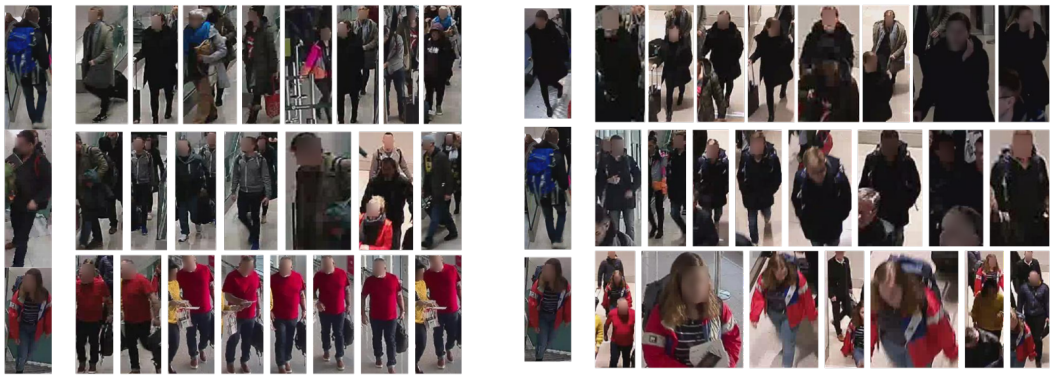

Figure 3: Examples of query images and retrieved false positive (left) and false negative (right) samples. The left columns are the query images and the remaining images are retrieved examples.

		$\sigma_s = 50$		$\sigma_s = 100$		$\sigma_s = 400$		$\Delta s*$	
	Model	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1
1	PCB [24]	−1.2	+1.1	−0.6	+0.7	-0.2	+0.3	+0.4	−0.9
2	MLFN [2]	−1.2	+1.3	−0.5	+1.0	-0.2	-0.1	+0.8	−0.4
3	HACNN [12]	−1.0	−0.4	−0.5	+0.0	-0.1	-0.1	+0.4	−0.2
4	Resnet50Mid [28]	−0.9	−0.4	−0.4	+0.0	-0.1	-0.1	+0.2	+0.1
5	MuDeep [23]	−2.2	−0.8	−1.1	−0.1	-0.3	-0.1	+0.8	+0.0

Table 6: Improved re-ID performance by temporal and spatial re-ranking using a Gamma prior for $[0, 30)$. $\Delta s*$ is using $P(Y|\Delta s) \propto \Delta s$ as spatial prior.

re-ranking alone, the result is almost identical. There are two possible explanations. First, the spatial information between subjects in the gallery and probe images are estimated using walking distances between cameras (see Table 7) and this estimation largely ignores subtle spatial differences between subjects. Second, the temporal and spatial priors are related

	C900	**C902**	**C903**	**C904**	**C926**
C900	0.0	48.5	106.0	70.0	68.0
C902	48.5	0.0	59.0	19.0	38.5
C903	106.0	59.0	0.0	45.0	97.5
C904	70.0	19.0	45.0	0.0	63.5
C926	68.0	38.5	97.5	57.0	0.0

Table 7: Estimated relative walking distances between cameras.

since the more time passed, the larger the chance a subject could have walked further. Introducing the spatial prior provides little new information to improve the ranking. Our performance gain is on a par with the results reported in [32], which are based on the commonly used but computationally expensive re-ranking approach using k-reciprocal encoding.

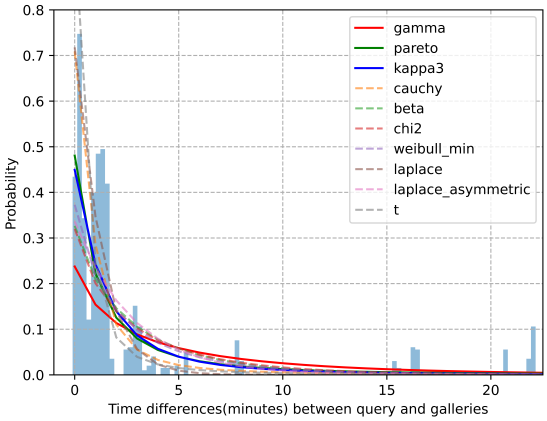

Figure 4: Empirical Distribution of $P(Y|\Delta t)$ computed from the DAA dataset. We fit various probability density functions to the empirical distribution and use them for time constrained re-ranking.

6. Discussion

The individuals annotated and re-identified in this study were typically wearing brightly colored or distinctive clothes, meaning that model performance would likely degrade with

	Gamma	Beta	Pareto	χ^2	Laplace	Laplace-asymm	t	Kappa(3)	Weibull-min	Cauchy
param	a 0.5	a, b 0.68, 98.68	b 1.2	df 1.19	N/A	κ 0.1	df 1	a 1.5	c 0.74	N/A
loc	-0.1	-0.1	-1.4	-0.1	1.1	0	0.9	-0.1	-0.1	0.9
scale	9.4	416.75	1.3	2.46	2.3	0.3	0.63	1.6	2.2	0.633

Table 8: Estimated parameters of empirical distributions.

	Gamma	Beta	Pareto	χ^2	Laplace	Laplace-asymm	t	Kappa(3)	Weibull-min	Cauchy
PCB	1.6	1.8	1.4	1.7	1.2	2	1.1	1.6	1.6	1.3
MLFN	80	80	60	80	60	80	50	60	60	60
HACNN	0.6	0.7	0.6	0.6	0.7	0.8	0.5	0.6	0.6	0.5
Resnet50mid	150	160	150	170	150	180	150	160	150	150
MuDeep	450	500	500	500	450	700	400	500	500	500
BoT	0.4	0.4	0.3	0.4	0.3	0.5	0.3	0.3	0.3	0.3

Table 9: Selected σ for different models.

more varied query individuals, as even humans find it harder to distinguish between people wearing darker clothing in CCTV images.

The proposed approach, which takes full advantage of time information, could greatly improve the number of true positive samples retrieved. In fact, by applying time constraints, it increases the number of true positives by around 200%. Applying both prior and time constraints, it increases by around 210%, while both false positives and negatives are suppressed. There are still some hard samples in the gallery that are difficult to separate in the feature space. Figure 3 illustrates some false negative and false positive examples after applying the prior and time constraints. To determine the cut-off rank that dictates the number of negative and positive examples, we use a different ranking threshold for each query. For instance, if there should be 10 matched IDs in the gallery for a certain query, the cut-off rank would be 10, meaning that the top-10 images are retrieved as positive and the remainder is marked negative. We noticed that even with the introduction of time information, pedestrian pose changes and similar clothing colors and styles could be contributing factors to the increasing false positives and negatives.

Notwithstanding, the ability to successfully re-identify at least a proportion of people from annotated images derived from CCTV footage, as demonstrated in this paper, could be of significant interest to the airport authorities. If we assume that the majority of people will move through the airport in a similar way, then the combination of two approaches: re-ID for a few individuals within a crowd and crowd density mapping, could provide key insights into passenger flow. For example, being able to evaluate whether an area of high crowd density at a given point in time is likely to dissipate or transfer to another area in the airport.

The proposed dataset with time prior re-ranking could also be of benefit to the recent development of unsupervised and/or self-supervised learning [25, 13] in re-ID research since the re-ranking process is simple, efficient, and fast. Given timestamp information, re-ranking could plug into any of the current implementations of unsupervised and/or self-supervised learning pipelines.

7. Conclusion

Most of the published person re-ID algorithms conduct supervised training and testing on single labeled datasets of small size, so directly deploying these trained models to a large-scale real-world camera network may lead to poor performance due to underfitting [19]. Also, most re-ID algorithms consider only appearance models and thus concentrate on deep learning for visual feature representation [26]. Our initial experiments with additional temporal constraints demonstrate considerable improvement in person re-ID performance on our DAA dataset, which is now publicly available. This research is not possible on other image-based datasets. Using time-reduced galleries and additional temporal re-ranking step yields the significant performance gain of **47.46**% in mAP and **40.17**% in Rank1 accuracy as compared to the results previously reported.

Acknowledgments

This publication has emanated from research supported by Science Foundation Ireland (SFI) under grant numbers SFI/20/COV/8579 SFI/12/RC/2289_2, co-funded by the European Regional Development Fund. We would like to thank Dublin Airport Authority for their contributions to this work.

References

[1] Octavia Camps, Mengran Gou, Tom Hebble, Srikrishna Karanam, Oliver Lehmann, Yang Li, Richard J. Radke, Ziyan Wu, and Fei Xiong. From the lab to the real world: Re-identification in an airport camera network. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3), 2017.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018.

[3] Yeong-Jun Cho, Su-A Kim, Jae-Han Park, Kyuewang Lee, and Kuk-Jin Yoon. Joint person re-identification and camera network topology inference in multiple cameras. *Computer Vision and Image Understanding*, 180:34–46, 2019.

[4] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E. O'Connor. How important are faces for person re-identification? In *Proceedings of the 25th International Conference on Pattern Recognition*, 2021.

[5] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014.

[6] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[7] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J. Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *CVPR Workshop*, pages 10–19, 2017.

[8] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision*, 2008.

[9] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–959, 2017.

[10] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):523–536, 2018.

[11] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[12] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.

[13] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019.

[14] Zhongyi Li, Yi Jin, Yidong Li, Congyan Lang, Songhe Feng, and Tao Wang. Learning part-alignment feature for person re-identification with spatial-temporal-based re-ranking method. *World Wide Web (Springer)*, 23:1907–1923, 2020.

[15] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 456–474. Springer, 2020.

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, 2016.

[17] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[18] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2020.

[19] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7948–7956, 2018.

[20] Bradley Malin, Kathleen Benitez, and Daniel Masys. Never too old for anonymity: a statistical standard for demographic data sharing via the hipaa privacy rule. *Journal of the American Medical Informatics Association*, 18(1):3–10, 2011.

[21] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2019.

[22] Madhumita Murgia. Who's using your face? the ugly truth about facial recognition. *Financial Times*, 2019.

[23] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5399–5408, 2017.

[24] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline. In *Proceedings of the European Conference on Computer Vision*, 2018.

[25] Haotian Tang, Yiru Zhao, and Hongtao Lu. Unsupervised person re-identification with iterative self-supervised domain adaptation. In *CVPRW*, 2019.

[26] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8933–8940, 2019.

[27] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*, 2017.

[29] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[30] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016.

[31] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv*, 2017.

[32] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[33] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv*, 2019.