DukeMTMC4ReID: A Large-Scale Multi-Camera Person Re-Identification Dataset*

Mengran Gou[1], Srikrishna Karanam[2], Wenqian Liu[1], Octavia Camps[1], Richard J. Radke[2]
Department of Electrical and Computer Engineering, Northeastern University, Boston, MA[1]
Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY[2]
{mengran, liuwenqian, camps}@coe.neu.edu, srikrishna@ieee.org, rjradke@ecse.rpi.edu

Figure 1: Snapshots of the proposed DukeMTMC4ReID dataset.

Abstract

In the past decade, research in person re-identification (re-id) has exploded due to its broad use in security and surveillance applications. Issues such as inter-camera viewpoint, illumination and pose variations make it an extremely difficult problem. Consequently, many algorithms have been proposed to tackle these issues. To validate the efficacy of re-id algorithms, numerous benchmarking datasets have been constructed. While early datasets contained relatively few identities and images, several large-scale datasets have recently been proposed, motivated by data-driven machine learning. In this paper, we introduce a new large-scale real-world re-id dataset, DukeMTMC4ReID, using 8 disjoint surveillance camera views covering parts of the Duke University campus. The dataset was created from the recently proposed fully annotated multi-target multi-camera tracking dataset DukeMTMC [36]. A benchmark summarizing extensive experiments with many combinations of existing re-id algorithms on this dataset is also provided for an up-to-date performance analysis.

1. Introduction

Person re-identification, or re-id, is a critical component of modern surveillance systems. Consequently, this problem has drawn increasing attention from the computer vision community, evidenced by the ever-increasing number of papers published in CVPR, ICCV, and ECCV [14, 49, 51]. The fundamental problem is as follows: given some information (an image or set of images) about a person of interest in a "probe" camera view, a re-id algorithm is to rank a set of candidate persons seen in a "gallery" camera view. If the person of interest exists in this candidate set, s/he should appear near the top of the ranked list.

Existing re-id algorithms are typically evaluated on datasets that are either hand-curated or pruned with a person detector to contain sets of bounding boxes for the probes

[1]This work was supported in part by NSF grants IIS1318145 and ECCS1404163 and AFOSR grant FA9550-15-1-0392. This material is based upon work supported by the U.S. Department of Homeland Security, Science and Technology Directorate, Office of University Programs, under Grant Award 2013-ST-061-ED0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Table 1: A overview of existing widely used re-id datasets.

Dataset	Year	# people	# BBox	# FP	# distractors	# cameras	Environment	Label source	Video?	Full frame?
VIPeR [14]	2007	632	1,264	0	0	2	-	hand	N	N
ETHZ [38]	2007	148	8,580	0	0	1	-	hand	N	N
QMUL iLIDS [52]	2009	119	476	0	0	2	airport	hand	N	N
GRID [27]	2009	1,025	1,275	0	775	8	subway	hand	N	N
3DPeS [2]	2011	192	1,011	0	0	8	campus	hand	N	Y
PRID2011 [15]	2011	934	24,541	0	732	2	campus	hand	Y	Y
CAVIAR4ReID [6]	2011	72	1,220	0	22	2	mall	hand	N	Y
V47 [41]	2011	47	752	0	0	2	-	hand	N	Y
WARD [31]	2012	70	4,786	0	0	3	-	hand	Y	N
SAIVT-Softbio [4]	2012	152	64,472	0	0	8	campus	hand	Y	Y
CUHK01 [20]	2012	971	3,884	0	0	2	campus	hand	N	N
CUHK02 [19]	2013	1,816	7,264	0	0	10 (5 pairs)	campus	hand	N	N
CUHK03 [21]	2014	1,467	13,164	0	0	10 (5 pairs)	campus	hand/DPM [10]	N	N
HDA+ [11]	2014	53	2,976	2,062	20	13	office	hand/ACF [9]	N	Y
RAiD [7]	2014	43	6,920	0	0	4	campus	hand	N	N
iLIDS-VID [42]	2014	300	42,495	0	0	2	airport	hand	Y	N
Market1501 [50]	2015	1,501	32,217	2,798+500K	0	6	campus	hand/DPM [10]	N	N
Airport [16]	2015	9,651	39,902	9,659	8,269	6	airport	ACF [9]	N	N
MARS [49]	2016	1,261	1,191,003	147,744	0	6	campus	DPM [10]+GMMCP [8]	Y	N
DukeMTMC-reID [53]	2017	1,812	36,441	0	408	8	campus	hand	N	Y
DukeMTMC4ReID	**2017**	**1,852**	**46,261**	**21,551**	**439**	**8**	**campus**	**Doppia** [3]	**N**	**Y**

and the corresponding matching candidates. As noted in the recent benchmark paper by Karanam *et al*. [16], the size of a dataset, in terms of both number of identities as well as number of bounding boxes, is critical to achieve good performance. Furthermore, in real-world end-to-end surveillance systems, as noted in Camps *et al*. [5], we can use camera calibration information to predict motion patterns, potentially helping to prune out irrelevant candidates and reducing the search space. The recently proposed DukeMTMC dataset [36], while originally proposed for multi-target tracking, fulfills both of these criteria. In addition to containing 2 million full frames corresponding to 2700 identities seen in an 8-camera network, the dataset also comes with per-camera calibration information.

In this paper, we propose to adapt and re-orient the DukeMTMC dataset to address the re-id problem. To this end, we used an off-the-shelf person detector to generate candidate bounding boxes from the full frames, resulting in a re-id dataset with the largest number of unique identities to date. Figure 1 illustrates sample snapshots from the proposed dataset. We also present an up-to-date performance benchmark for this dataset, in which we adopted the evaluation protocol proposed by Karanam *et al*. [16]. Specifically, we considered hundreds of combinations of previously published feature extraction and metric learning algorithms. The goal is to systematically study how existing re-id algorithms fare on the new dataset. We provide extensive per-camera-pair evaluation results, and compare the performance on this dataset with that of existing, widely used datasets, providing useful insights for future research directions. Compared to widely used datasets such as CUHK03 [21] and Market1501 [50], the rank-1 performance on the proposed dataset is lower under similar experimental settings, suggesting future opportunities to develop better re-id algorithms, which we discuss in Sections 4.5 and 5.

2. Re-ID Datasets: An Overview

In this section, we provide a brief overview of publicly available datasets that are commonly used to evaluate re-id algorithms. Table 1 provides a statistical summary of these datasets. In the table and following content, we define an identity as a person with images in both the probe and gallery cameras, a distractor as a person only appearing in one camera, and an FP as a false alarm from the person detector.

VIPeR [14] is one of the earliest available and most widely used datasets, consisting of 632 identities from two disjoint camera views. GRID [27] has 250 paired identities across 8 cameras, in addition to 775 distractor identities to mimic a realistic scenario. 3DPeS [2] consists of 1,011 images corresponding to 192 identities, captured in an 8-camera network. PRID2011 [15] is constructed in an outdoor environment, with 200 paired identities captured in two camera views. CAVIAR4ReID [6] is constructed from two cameras placed inside a shopping mall, with 50 paired identities available. V47 [41] captures 47 identities in an indoor environment. WARD [31] captures 70 identities in a 3-camera network. SAIVT-Softbio [4] captures 152 identities in an 8-camera surveillance network installed on a campus. HDA+ [11] captures 53 identities in an indoor environment, in addition to a number of distractor identities for the gallery. RAiD [7] captures 43 identities as seen from two indoor and two outdoor cameras. iLIDS-VID [42] captures 300 identities in an indoor surveillance camera network installed in an airport. Market1501 [50] captures 1,501 identities in addition to 2,798 false positives and 500k distractors, providing for a realistic gallery. Airport [16] represents a

realistic scenario in which 1,382 identities are captured in a 6-camera indoor surveillance network in an airport. All images are automatically generated by means of an end-to-end re-id system [5, 22]. MARS [49] is a video extension of the Market1501 dataset, with long-duration image sequences captured for 1,261 identities.

As mentioned above, our proposed dataset was derived from the DukeMTMC dataset for multi-target tracking [36]. We note that Zheng *et al*. [53] also recently proposed a re-id dataset, called DukeMTMC-reID in Table 1, based on DukeMTMC. However, our proposed dataset is significantly different on several fronts. While DukeMTMC-reID uses manually labeled ground truth, the proposed dataset uses person detections from an automatic person detector. Furthermore, DukeMTMC-reID does not include any false alarms from the detector in the gallery, while the proposed dataset has over 20,000 false alarms. Therefore, the proposed dataset is more realistic in the sense that it mimics how a practical re-id system would work in the real world. Finally, we also conduct a systematic performance evaluation of existing algorithms on DukeMTMC4ReID, producing detailed per-camera performance analysis that provides useful insights for further research.

3. DukeMTMC4ReID

All frames in the DukeMTMC dataset were captured by 8 static cameras on the Duke University campus in 1080p and at 60 frames per second (Figure 2). In total, more than 2,700 people were labeled with unique IDs in eight 75-minute videos. The tight bounding boxes of each person for each frame are generated based on background subtraction and manually labeled foot positions in a few frames. Regions of interest (normal paths on the ground plane) and calibration data are also provided. The entire dataset is split into three parts: one training/validation set labeled "trainval" and two testing sets labeled "test-hard" and "test-easy". To date, only labels from the "trainval" set have been released, which contains 1,852 unique identities in eight 50-minute videos (dataset frames 49,700–227,540).

Based on this dataset, we constructed a large-scale real-world person re-id dataset: **DukeMTMC4ReID**. Following the recently proposed Market1501 [50] and CUHK03 [21] datasets, bounding boxes from an off-the-shelf person detector are used to mimic real-world systems. We used a fast state-of-the-art person detector [3] for accurate detections, which are filtered using predefined regions of interest to remove false alarms, e.g., bounding boxes on walls or in the sky. Then, following Market1501, based on the overlap ratio between the detection and ground truth (i.e., the ratio of the intersection to the union), we label the bounding box as "good" if the ratio is greater than 50%, false positive ("FP") if the ratio is smaller than 20%, and "junk" otherwise. For each identity, we uniformly sample

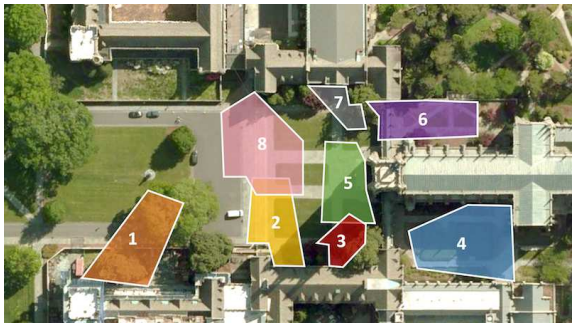

Figure 2: Layout of the cameras in the DukeMTMC dataset (from [1])

5 "good" bounding boxes in each available camera, while retaining all the "FP" bounding boxes in the corresponding frames. To summarize, the relevant statistics of the proposed DukeMTMC4ReID dataset are provided below:

- Images corresponding to 1,852 people existing across all the 8 cameras
- 1,413 unique identities with 22,515 bounding boxes that appear in more than one camera (valid identities)
- 439 distractor identities with 2,195 bounding boxes that appear in only one camera, in addition to 21,551 "FP" bounding boxes from the person detector
- The size of the bounding box varies from 72×34 pixels to 415×188 pixels

Table 2 tabulates these and other statistics of the proposed DukeMTMC4ReID dataset and Table 3 shows the number of valid identities in each camera pair.

4. Benchmark Evaluation of Re-id Algorithms

Next, we present the details of our systematic experimental evaluation of 7 existing feature extraction algorithms and 12 existing metric learning algorithms for re-id, producing an up-to-date benchmark on the proposed dataset.

4.1. Feature Extraction

Following the protocol described in [16], we evaluated 7 different feature extraction algorithms published up through CVPR 2016 (Table 4), which we briefly describe next. ELF [14] extracts color features from the RGB, YCbCr and HS color channels and texture features from the responses of multiple Schmid and Gabor filters. In HistLBP, Xiong *et al*. [46] substituted the Schmid and Gabor texture responses with LBP features, while DenseColorSIFT [48] uses dense SIFT features. gBiCov [28] uses the covariance descriptor to encode multi-scale biological-inspired features. LDFV

Table 2: Basic statistics of the proposed DukeMTMC4ReID dataset

	Total	cam1	cam2	cam3	cam4	cam5	cam6	cam7	cam8
# bboxes	46,261	10,048	4,469	5,117	2,040	2,400	10,632	4,335	7,220
# person bboxes	24,710	4,220	4,030	1,975	1,640	2,195	3,635	2,285	4,730
# "FP" bboxes	21,551	5,828	439	3,142	400	205	6,997	2,050	2,490
# persons	1,852	844	806	395	328	439	727	457	946
# valid ids	1,413	828	778	394	322	439	718	457	567
# distractors	439	16	28	1	6	0	9	0	379
# probe ids	706	403	373	200	168	209	358	243	284

Table 3: Number of valid ids in each camera pair

camera	1	2	3	4	5	6	7
2	655						
3	260	348					
4	227	292	311				
5	279	311	89	57			
6	278	261	34	9	348		
7	66	43	15	4	69	418	
8	148	42	30	27	51	374	383

Table 4: Evaluated features

Feature	Source
ELF [14]	ECCV 08
LDFV [29]	ECCVW 12
gBiCov [28]	BMVC 12
DenseColorSIFT [48]	CVPR 13
HistLBP [46]	ECCV 14
LOMO [24]	CVPR 15
GOG [32]	CVPR 16

Table 5: Evaluated metric learning methods

Metric	Source	Metric	Source
l2	-	LFDA [34]	CVPR 13
FDA [12]	AE 1936	SVMML [23]	CVPR 13
MFA [47]	PAMI 07	kMFA [46]	ECCV 14
RankSVM [35]	BMVC 10	rPCCA [46]	ECCV 14
KISSME [18]	CVPR 12	kLFDA [46]	ECCV 14
PCCA [33]	CVPR 12	XQDA [24]	CVPR 15
kPCCA [33]	CVPR 12		

[29] uses the Fisher vector representation to encode local pixel-level information. LOMO [24] extracts HSV color histogram and scale-invariant LBP features from the image in conjunction with multi-scale retinex preprocessing. In GOG, Matsukawa *et al*. [32] used hierarchical Gaussian modeling to encode local pixel-level feature descriptors.

4.2. Metric Learning

Table 5 lists all the metric learning methods that were evaluated, which we briefly describe next. FDA [12], LFDA [34], MFA [47], and XQDA [24] all solve eigenvalue problems based on general discriminant analysis to learn the distance metric. Xiong *et al*. [46] proposed kernelized variants of LFDA and MFA. RankSVM [35] formulates metric learning as a ranking problem in a soft-margin framework. KISSME [18] learns the distance metric via a maximum log-likelihood ratio test. PCCA [33] uses a hinge loss objective function, while rPCCA [46] extends it by introducing a regularization term. In SVMML [23], a locally adaptive distance metric is learned in a soft-margin SVM framework. For all the kernel-based methods, we evaluated 4 different kernels: linear (ℓ), chi-square (χ^2), chi-square-rbf (R_{χ^2}) and exponential (exp).

4.3. Implementation Details

Prior to feature extraction, all bounding boxes are normalized to 128×64 pixels and partitioned into 6 non-overlapping horizontal strips. In LDFV, the number of Gaussians for the GMM is set to 16. The number of bins in the color histogram for HistLBP and ELF is set to 16, and we use RGB as the color space in GOG. In metric learning, we set the subspace dimension to 40 and the negative-to-positive pair ratio to construct the training data to 10.

4.4. Evaluation Protocol

Out of the 1,413 valid identities, we randomly pick 707 as the training set with the rest forming the testing set. In the testing set, we follow the evaluation protocol in Market1501 [50], in which one camera is fixed as the probe and the bounding boxes in all the other cameras are combined into a large gallery set. We perform experiments across all cameras and report the average performance across all the probe identities. To further analyze per-camera performance, we perform additional experiments following the protocol in the SAIVT dataset [4], where pair-wise performance for all camera pairs is reported. In all reported results, for each probe identity, we randomly pick one bounding box from the available 5 bounding boxes.

Table 6: Rank 1 results from all feature/method combinations. The best result for each feature is marked in red and second best in blue.

Methods	Kernel	ELF	LDFV	gBiCov	SDC	HistLBP	LOMO	GOG
L2		6.26	15.28	6.93	10.86	5.36	20.6	29.04
KISSME		9.29	1.3	10.55	5.81	1.47	0.49	2.55
RankSVM		7.55	18.45	6.75	7.73	6.48	23.19	22.7
SVMML		7.55	33.29	0.31	12.11	1.74	9.61	26.9
PCCA		16.71	15.28	9.65	16.8	15.64	14.16	21.58
kPCCA	ℓ	12.47	23.15	7.28	14.52	12.69	24.44	32.84
	χ^2	19.93	-	7.46	20.55	21.13	22.7	-
	R_{χ^2}	22.74	-	10.41	25.02	22.92	26.18	-
	exp	17.25	26.94	10.55	20.73	19.62	28.64	36.64
rPCCA	ℓ	12.42	23.24	7.51	14.66	12.65	24.26	32.93
	χ^2	19.97	-	7.55	20.55	20.82	22.88	-
	R_{χ^2}	22.74	-	10.55	24.84	22.92	26.18	-
	exp	17.29	28.64	11.75	20.64	20.2	28.42	37.31
FDA		23.15	25.25	**16.53**	26.14	21.98	21	25.34
MFA		20.02	20.46	15.06	15.59	18.01	14.92	12.56
kMFA	ℓ	24.22	35.03	12.51	25.16	23.91	31.9	42.81
	χ^2	31.68	-	10.5	33.69	33.96	32.17	-
	R_{χ^2}	**34.41**	-	13.27	**34.81**	**37.44**	32.98	-
	exp	28.6	**41.06**	13.76	28.82	31.59	**38.29**	**49.46**
LFDA		23.73	26.85	**16.26**	26.41	22.07	22.34	27.35
kLFDA	ℓ	20.6	34.72	15.59	22.92	19.39	32.13	44.1
	χ^2	29.58	-	12.69	31.81	32.66	31.99	-
	R_{χ^2}	**33.82**	-	11.71	**34**	**36.86**	33.74	-
	exp	28.82	**41.02**	13.58	28.24	31.55	**38.56**	**49.55**
XQDA		19.26	23.46	0.76	23.64	8.94	27.88	34.76

4.5. Results and Discussion

In Figure 3, we report the CMC curves for all the evaluated algorithm combinations. We highlight the top 10 best-performing combinations, in terms of rank-1 accuracy, in color. The rank-1 performance for all combinations is shown in Table 6. Similar to the trends observed in [16], we found the best-performing feature extraction algorithms to be GOG, LDFV and LOMO, whereas the best performing metric learning methods were kLFDA and kMFA. GOG performs well due to its hierarchical modeling of local color and texture structures in images. kLFDA and kMFA result in the most discriminative distance metrics because they solve generalized eigenvalue problems on data scatter matrices, a framework Karanam *et al*. [16] empirically found to be most suitable for re-id.

Figure 4 shows a bar graph of the rank-1 and mean Average Precision (mAP) performance for all 8 probe cameras with the best combination of feature and leaning method. The mAP is the average precision value computed across all the queries [50]. We see that depending on the choice of the probe camera, the rank 1 accuracy varies from 30.4% to 70.8%. The hardest probe cameras are cameras 4 and 8, both of which result in less than 40% rank-1 accuracy. To further analyze the reasons for this performance variation across different cameras, we evaluated pairwise performance for all camera pairs and tabulated the rank-1 accuracy in Table 7. Because of the physical layout of the cameras, only a few people can appear in particular camera pairs (see Table 8), which leads to results with inherent statistical bias. Results from such camera pairs, typically having less than 20 probe identities, are marked in gray and ignored in our analysis. In the following, $a \rightarrow b$ denotes the use of camera a for the probe and camera b for the gallery. The $4 \rightarrow 1$ scenario gives the worst performance, with $4 \rightarrow 2$ and $4 \rightarrow 3$ scenarios also resulting in low numbers, confirming that camera 4 is the most challenging probe camera in the dataset. To qualitatively understand these numbers better, in Figure 5, we show example images from several camera pairs. The first four rows correspond to pairs involving camera 4. Examples from the $4 \rightarrow x$ scenario are depicted in the right half, where we see representative identities in camera 4 suffering from extreme appearance variations when compared to the corresponding gallery appearances of cameras 1, 2, and 3. Further complicating the

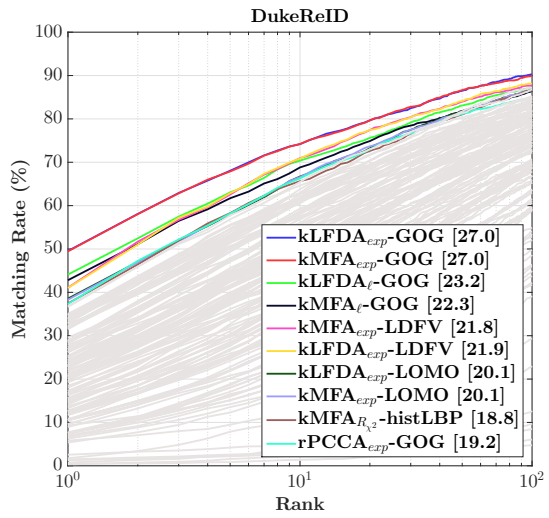

Figure 3: CMC curves for the benchmark on the DukeMTMC4ReID dataset. The top 10 performing algorithms are shown in color and the rest are shown in gray. Numbers in the brackets in the legend are the corresponding mAP values.

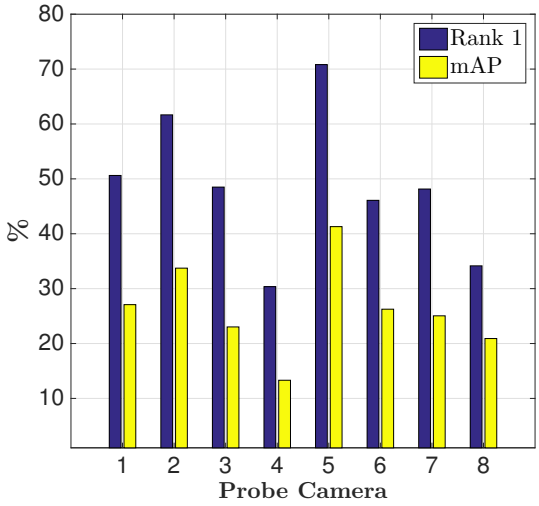

Figure 4: Results for each camera using kLFDA$_{exp}$ with GOG features.

issue is camera 4's image resolution, which is relatively lower when the person is far from the camera (see the second example in the second and third row). Similarly, the pairs $8 \rightarrow 6$ and $8 \rightarrow 7$ also have large viewpoint variations, suggesting camera 8 as the probe is quite challenging as well. On the other hand, in the $2 \rightarrow 5$ scenario, identities are captured from very similar viewpoints, resulting in more than 80% rank-1 performance accuracy.

Table 7: Rank 1 accuracy for each camera pair using kLFDA$_{exp}$ with GOG features. The first column indicates the probe camera and the first row indicates the gallery camera. Results for camera pairs with less than 20 probe persons are marked in gray. Among all the other results, the highest one is marked in red and the lowest one is marked in blue.

camera	1	2	3	4	5	6	7	8
1	-	45.9	56.9	25.2	58.9	47.2	28.1	63.0
2	52.5	-	62.6	45.8	**84.3**	59.3	31.8	66.7
3	55.4	45.4	-	27.9	40.0	47.1	22.2	36.4
4	**15.1**	36.6	30.9	-	55.6	0.0	0.0	66.7
5	69.8	82.9	44.4	59.3	-	62.2	45.0	54.2
6	42.4	49.2	64.7	80.0	58.5	-	54.1	30.9
7	25.0	36.4	33.3	33.3	27.5	55.9	-	29.1
8	57.5	58.3	0.0	33.3	50.0	26.8	32.7	-

Table 8: Number of probe instances in the pairwise evaluation protocol.

camera	1	2	3	4	5	6	7
2	316						
3	130	174					
4	119	153	165				
5	129	140	45	27			
6	125	118	17	5	164		
7	32	22	9	3	40	220	
8	73	12	11	9	24	194	199

We note that the proposed dataset also suffers from illumination variations, detection errors, occlusions, and background clutter in addition to the viewpoint variations and low-resolution images discussed above. In line with the observations in Karanam *et al*. [16], given this diversity in attributes across images, we anticipate this dataset will help further research in metric and feature learning for re-id.

Finally, to put the results for the proposed dataset in a broader context of how algorithms fare on existing datasets, we provide, in Table 9, a fairly recent benchmark on several widely used large scale re-id datasets. For a fair comparison, let us focus on the performance of the LOMO+XQDA algorithmic combination (the first row in each section). As can be noted from the table, the proposed dataset has the lowest rank-1 and mAP performance, offering opportunities for future algorithm development. In particular, such large-scale datasets provide realistic test cases containing a large number of candidates. Since real-world re-id applications typically deal with such gallery sets, the proposed dataset can be used to develop and test scalable re-id algorithms in terms of both efficiency and computability. We discuss other future research directions in Section 5.

Figure 5: Snapshots for several camera pairs. Each row gives two examples for each camera pair. The single images are the probe images.

Table 9: Comparisons with other large-scale single-shot datasets. Results of LOMO+XQDA for CUHK03 and Market1501 are directly copied from [54].

Dataset	Method	Rank 1	mAP
CUHK03 Detected	LOMO+XQDA [24]	44.6	51.5
	IDE+XQDA+re-rank [54]	58.5	64.7
	GOG+XQDA [32]	65.5	-
	Gated-SCNN [39]	68.1	-
	DTL [13]	84.1	-
Market1501	LOMO+XQDA [24]	58.6	85.7
	S-LSTM [40]	61.6	35.3
	Gated-SCNN [39]	65.9	39.6
	IDE+KISSME+re-rank [54]	77.1	63.7
	DTL [13]	83.7	65.6
	APR [25]	84.3	64.7
DukeMTMC4ReID	LOMO+XQDA [24]	27.9	13.5
	GOG+kLFDA$_{exp}$	49.6	27.0

5. Summary and Future Work

In this paper, we proposed a large-scale multi-camera re-id dataset based on the DukeMTMC [36] dataset. We conducted extensive experimental analyses to benchmark existing re-id algorithms on the proposed dataset.

While we have specifically benchmarked re-id in this paper, the problem of comparing candidate bounding boxes is only a small part of an automatic system that tracks persons of interest in a multi-camera network. As noted by Camps *et al*. [5], detection and tracking modules are key parts of such a system that would typically be operated for long periods of time. In such cases, the re-id module in the system is presented with a continuously increasing gallery set, instead of a fixed-size gallery set, resulting in several previously unaddressed temporal challenges. The DukeMTMC dataset also comes with tools to retrieve time-stamp information for every frame, potentially enabling DukeMTMC4ReID to be used to consider such temporal aspects of the re-id problem.

To fully explore the potential of the DukeMTMC dataset, we can extend DukeMTMC4ReID to construct a MARS[49]-like video-based, or multi-shot, re-id dataset. This will result in a massive dataset with full calibration information that will present unique challenges and opportunities to re-id researchers. A specific promising research direction would be the study of multi-shot ranking [26, 17, 43] in conjunction with metric learning. While these two topics are typically treated separately, we anticipate that a unified framework would lead to substantial performance gains.

Finally, a problem closely related to person re-id is multi-target multi-camera tracking (MTMCT). While re-id is generally posed as a search and retrieval problem, the goal of MTMCT is to track person(s) across multiple overlapping or non-overlapping cameras. As noted in this paper, there is substantial literature on person re-id, and much recent effort [37, 44, 8, 45, 30] has also been devoted to MTMCT. Given the obvious similarities between re-id and MTMCT, a useful future research direction would be to study how these two problems can help each other.

References

[1] DukeMTMC project. http://vision.cs.duke.edu/DukeMTMC/details.html. Accessed: 2017-03-22. 3

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *Proceedings of the 16th International Conference on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, Sept. 2011. 2

[3] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014. 2, 3

[4] A. Bialkowski, S. Denman, S. Sridharan, C. Fookes, and P. Lucey. A database for person re-identification in multi-camera surveillance networks. In *Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on*, pages 1–8. IEEE, 2012. 2, 4

[5] O. Camps, M. Gou, T. Hebble, S. Karanam, O. Lehmann, Y. Li, R. J. Radke, Z. Wu, and F. Xiong. From the lab to the real world: Re-identification in an airport camera network. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, 27(3), 2017. 2, 3, 8

[6] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 1, page 6, 2011. 2

[7] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury. Consistent re-identification in a camera network. In *European Conference on Computer Vision*, volume 8690 of *Lecture Notes in Computer Science*, pages 330–345. Springer, 2014. 2

[8] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099, 2015. 2, 8

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, 2014. 2

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 32(9):1627–1645, 2010. 2

[11] D. Figueira, M. Taiana, A. Nambiar, J. Nascimento, and A. Bernardino. The hda+ data set for research on fully automated re-identification systems. In *European Conference on Computer Vision*, pages 241–255. Springer, 2014. 2

[12] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936. 4

[13] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 8

[14] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European*

conference on computer vision, pages 262–275. Springer, 2008. 1, 2, 3, 4

[15] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 2

[16] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. I. Camps, and R. J. Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2016. 2, 3, 5, 6

[17] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with block sparse recovery. *Image and Vision Computing*, 60:75–90, 2017. 8

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012. 4

[19] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013. 2

[20] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44. Springer, 2012. 2

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 152–159. IEEE, 2014. 2, 3

[22] Y. Li, Z. Wu, S. Karanam, and R. Radke. Real-world re-identification in an airport camera network. In *Proc. Int. Conf. Distributed Smart Cameras (ICDSC)*, 2014. 3

[23] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3610–3617. IEEE, 2013. 4

[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015. 4, 8

[25] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017. 8

[26] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, 2015. 8

[27] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010. 2

[28] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machive Vision Conference*, pages 11–pages, 2012. 3, 4

[29] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 413–422. Springer, 2012. 4

[30] A. Maksai, X. Wang, and P. Fua. What players do with the ball: a physically constrained interaction modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 972–981, 2016. 8

[31] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *CVPR Workshops*, 2012. 2

[32] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4, 8

[33] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2672. IEEE, 2012. 4

[34] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325. IEEE, 2013. 4

[35] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010. 4

[36] 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. 1, 2, 3, 8

[37] E. Ristani and C. Tomasi. Tracking multiple people online and in real time. In *Asian Conference on Computer Vision*, pages 444–459. Springer, 2014. 8

[38] W. Schwartz and L. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009. 2

[39] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 8

[40] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016. 8

[41] S. Wang, M. Lewandowski, J. Annesley, and J. Orwell. Re-identification of pedestrians with variable occlusion and scale. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1876–1882. IEEE, 2011. 2

[42] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014. 2

[43] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016. 8

[44] X. Wang, V. Ablavsky, H. B. Shitrit, and P. Fua. Take your eyes off the ball: Improving ball-tracking by focusing on team play. *Computer Vision and Image Understanding*, 119:102–115, 2014. 8

[45] X. Wang, E. Türetken, F. Fleuret, and P. Fua. Tracking interacting objects using intertwined flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2312–2326, 2016. 8

[46] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, pages 1–16. 2014. 3, 4

[47] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 2007. 4

[48] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593. IEEE, 2013. 3, 4

[49] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 1, 2, 3, 8

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015. 2, 3, 4, 5

[51] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[52] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *Proc. of The 20th British Machine Vision Conference (BMVC)*, 2009. 2

[53] 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. 2, 3

[54] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017. 8