2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Online Domain Adaptation for Person Re-Identification with a Human in the Loop

Rita Delussu, Lorenzo Putzu, Giorgio Fumera and Fabio Roli
Department of Electrical and Electronic Engineering, University of Cagliari – Piazza d'Armi, 09123 Cagliari, Italy
Email: {rita.delussu,lorenzo.putzu,fumera,roli}@unica.it

***Abstract*—Supervised deep learning methods have recently achieved remarkable performance in person re-identification. Unsupervised domain adaptation (UDA) approaches have also been proposed for application scenarios where only unlabelled data are available from target camera views. We consider a more challenging scenario when even collecting a suitable amount of representative, unlabelled target data for *offline* training or fine-tuning is infeasible. In this context we revisit the human-in-the-loop (HITL) approach, which exploits *online* the operator's feedback on a small amount of target data. We argue that HITL is a kind of *online* domain adaptation specifically suited to person re-identification. We then reconsider relevance feedback methods for content-based image retrieval that are computationally much cheaper than state-of-the-art HITL methods for person re-identification, and devise a specific feedback protocol for them. Experimental results show that HITL can achieve comparable or better performance than UDA, and is therefore a valid alternative when the lack of unlabelled target data makes UDA infeasible.**

I. Introduction

Person re-identification consists of matching pedestrian images across non-overlapping video surveillance camera views [1], [2]. This is a challenging task due to pose, lighting and background variations, and has attracted great interest from the computer vision community due to its relevance in security-related applications [2], [3]. Early methods relied on manually defined pedestrian descriptors (e.g., [4]), and on similarity measures defined either ad hoc [4] or by metric learning [5], [6]. More recent approaches are based on supervised learning, mainly through convolutional neural networks (CNNs) used as feature extractors, and also for similarity measure learning [2], [7]. Other supervised solutions also exist, such as one-pass online learning on streaming data [8].

Supervised methods have reached a very high recognition accuracy on benchmark data sets, such as Market-1501 and DukeMTMC-reid (see Sect. IV-B). However, in real-world application scenarios it may be not possible to collect labelled pedestrian images from the *target* camera views, i.e., the ones that will be used after system deployment. In this case a supervised model has to be used in a cross-view setting. However, although benchmark data sets contain different camera views, they are affected by a significant data set bias [9], [10], similarly to other computer vision tasks [11], which considerably affects the cross-view performance of supervised methods [10], [12]. As a solution, *unsupervised domain adaptation* (UDA) methods have been proposed [13], [14], in which a model is trained both on labelled data from a *source* domain (i.e., camera views) and on unlabelled data from the *target* domain [12], [15]–[17]. However, UDA requires a suitable amount of representative target images. This may be too demanding, or even infeasible for some end users such as law enforcement agencies (LEAs), especially in scenarios like temporary camera installations which should be operational in a short time. In this kind of scenario the accuracy of re-identification models trained *offline* only on source data can still be improved through a human-in-the-loop (HITL) approach [18]–[21], which exploits the inherent, *online* interaction with users (e.g., LEA officers in a control room or forensic investigators) by asking them a feedback on *target* images retrieved for a given query.

Whereas HITL is a complementary approach to UDA (and can also be applied to improve the accuracy of supervised methods), we revisit it as an *alternative* solution in the above scenario, when UDA cannot be applied. In particular, we argue that HITL is a specific kind of *online* domain adaptation (ODA) [22], [23], where no target data is used for offline model training. We then reconsider HITL person re-identification based on relevance feedback (RF) algorithms for content-based image retrieval (CBIR), which are less complex than state-of-the-art HITL methods [19], [21]. We also revise the feedback protocol used by the latter methods and propose a protocol more suited to RF. Experimental results show that the considered HITL implementation can attain a similar or even better performance than UDA, and is therefore a valid alternative when UDA cannot be applied.

In the rest of this paper, previous work on UDA and HITL for person re-identification is summarised in Sect. II; our revisited HITL approach is motivated and described in Sect. III; Sect. IV reports experimental results, whereas conclusions and ideas for future work are presented in Sect. V.

II. Related Work

In this section we summarise previous work on UDA and HITL for person re-identification.

A. Unsupervised domain adaptive person re-identification

Domain adaptation (DA) addresses the mismatch between the distributions of labelled samples used to train a given model at design phase (source domain) and the distribution of samples that are processed during operation (target domain), which occurs in many applications. This issue has been extensively investigated in machine learning and computer vision [13], [14]. Usually, DA methods require the availability

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412485

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

of labelled samples of the target domain at design phase for model training or fine-tuning. However this may be infeasible in many practical application scenarios, including different computer vision tasks. UDA methods have therefore been proposed to exploit *unlabelled* target samples [13], [14], for applications including person re-identification. A common approach is to **learn a shared feature space** between source and target domains where the difference between the corresponding distributions is minimised, and the discriminant capability on the source domain can be transferred to the target one. In [15] a ResNet50 CNN architecture pre-trained on ImageNet is first fine-tuned on the source domain as a feature extractor, then a further fine-tuning step is carried out on target images using pseudo-labels assigned through clustering. A variant of this method is proposed in [24], where soft pseudo-labels are generated in a refinement step to mitigate the issue of noisy, hard pseudo-labels generated by clustering. An analogous solution is proposed in [16], where the adaptation step is carried out through a module for invariance learning with unlabelled target data. A self-training scheme based on clustering is adopted in [12] to learn a feature encoder.

Some methods introduce **specific deep architectures**. A shared representation is learnt in [25] through supervised identity classification of source data, and unsupervised reconstruction of unlabelled target data. In [17] two branches based on attention modules are added to a ResNet-50 backbone to extract domain-shared and domain-specific feature maps; the former map is learnt through supervised training on the source domain, using pseudo-labels obtained by clustering on the target domain, to make the target and source distributions similar, and is then applied to the target domain. Analogously, in [26] encoder and decoder modules are used to learn a shared, domain-independent feature map and a domain-specific one; the former is learnt by using the labelled source data, and by minimising (in an unsupervised way) the reconstruction loss on both domains.

Methods based on **adversarial learning** (including generative adversarial networks, GAN) have also been proposed. In [27] the diversity in lighting conditions between the two domains is addressed; synthetic pedestrian images are collected, which are made similar to the lighting conditions of the target domain through a CycleGAN, and are then used to fine-tune a pre-trained CNN model for feature extraction. In [28] a shared feature map based on a backbone CNN is learnt through an adversarial learning approach to minimise the source and target distribution discrepancy, taking into account also within-domain camera-level discrepancy. In [29] style-transfer GAN is used to adapt images in the source domain to the style of the target one; the modified images are used to train a supervised baseline model, which is then refined through a self-supervision step through iterative pseudo-labelling of target images.

Other methods exploit different kinds of features beside visual ones, such as spatio-temporal patterns related to the transition time of pedestrian across the different camera views [30], and pedestrian attributes [31].

B. Person re-identification with a human in the loop

The HITL approach is related to previous computer vision and machine learning approaches which exploit an online interaction with the user, such as CBIR-RF, active learning, and online metric learning. Only a few works have proposed HITL methods for person re-identification systems, despite their inherent interaction with users [18], [19], [21], [32], [33].

In [32] template images are matched to the query using Euclidean distance, and the user is asked to select some positive and negative images among the top-ranked ones. A Mahalanobis metric is then learnt based on selected images, and is used to re-rank the template gallery. In [18] the user is asked whether the identity of the query is present in the top ranks; if not, a discriminative person model is learnt using the query and its worst matches, and template images are re-ranked according to it. In [19] if the top matches do not contain the query identity, the user is asked to select one "strong" negative match and optionally a few "weak" negatives. Synthetic instances of the query image are then generated, and a weighted affinity graph is constructed to model the appearance similarities among all template images, including the synthetic ones. User feedback is then propagated through the graph, a ranking function is learnt and its output is combined with the original matching scores to re-rank the gallery images. A drawback of [19] is its high processing cost, which makes it infeasible for large template galleries [21]. Differently from the above methods, in [21] an *incremental* learning formulation is adopted, aimed at better adapting a person re-identification model to the target camera views. For each query the template gallery is ranked according to the current distance metric on a given feature space, initially defined as the Euclidean distance and then updated as a Mahalanobis distance. The user is asked to select either a positive match among the top ranks, if any, otherwise a strong negative. This feedback is then used to update the current distance metric, through an online metric learning algorithm. In [33] a sequential feedback process is considered. At each iteration the user is asked to select a true match on the ranked list of a *single* camera view, if any; the selected image is used to learn an updated feature representation based on the query image and previous true matches in other camera views (if any), which is then used to rank the gallery images of the *next* camera view. A significant difference with previous methods is that only feedback on true matches is asked to the user: if no true match is found, the HITL process stops.

III. Revisiting Human-In-The-Loop Person Re-Identification

In this section we show that the HITL approach to person re-identification can be viewed as a kind of online domain adaptation, and argue the case for HITL implementations based on relevance feedback, focusing on feedback protocols.

A. HITL as a kind of online domain adaptation

UDA methods require that a suitable amount of representative, unlabelled target data are available during system design,

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

for offline training or fine-tuning of the source model. However this can be too demanding, or even infeasible for some end users or application scenarios. For instance, this is the case of LEAs, especially when a temporary camera installation has to be deployed for a specific monitoring task, to be operational in a short time. In such scenarios it may be not possible to collect beforehand a *sufficient amount* of *representative* (albeit unlabelled) data from the target camera views. Unlabelled data from other auxiliary data sets can also be leveraged to improve the robustness of the source model, although they are different from the target views. The HITL approach appears particularly appealing in the above application scenario, since it improves the accuracy of a given source model through *online* user's feedback on the target data being processed during operation. In particular, since user's feedback is a form of *supervision*, HITL can require a significantly lower amount of target data than UDA. As an *indirect* support to this fact, in [21] several HITL algorithms outperformed *supervised* methods after user's feedback on a few gallery images. On the other hand, we point out that HITL remains a *complementary* approach to UDA when the latter can be applied. A high-level view of the UDA and HITL approaches, highlighting their differences, is given in Fig. 1.

Under a methodological perspective HITL exhibits interesting analogies with *online* domain adaptation (ODA), which has been used so far for computer vision tasks such as face detection [22], object detection, categorisation and recognition [23], [34], [35], and pedestrian detection [36]. Similarly to HITL: (i) ODA exploits the target data that are processed *online* during system operation; (ii) Existing ODA methods do not reuse the source data, which are assumed to be unavailable after deployment; (iii) Although most ODA methods rely only on unlabelled target data, the object categorisation method of [23] exploits user's feedback on the classification outcome of target images; (iv) Some ODA methods *incrementally* update the source model without reusing source data [34]–[36], similarly to the HITL person re-identification method of [21], whereas other ODA methods only update the scores given by the original source model to a given testing instance [22], [23], similarly to [18], [19], [32]. It is therefore pertinent to view the HITL approach as a kind of ODA, and to consider it as an alternative to UDA in relevant application scenarios such as the one mentioned above. We finally point out that an online, *one-pass* learning method has recently been proposed for person re-identification [8], which is however a supervised one, and is nevertheless orthogonal to HITL.

B. Relevance feedback for HITL person re-identification

Existing HITL person re-identification methods are relatively complex. The corresponding user interaction is however very similar to the one occurring in RF for CBIR, for which a number of simpler algorithms exist. However RF algorithms have not been considered so far for HITL person re-identification, except for experimental comparisons [19], [21]. It is therefore interesting to investigate more thoroughly their effectiveness for this task. In this context, two relevant RF

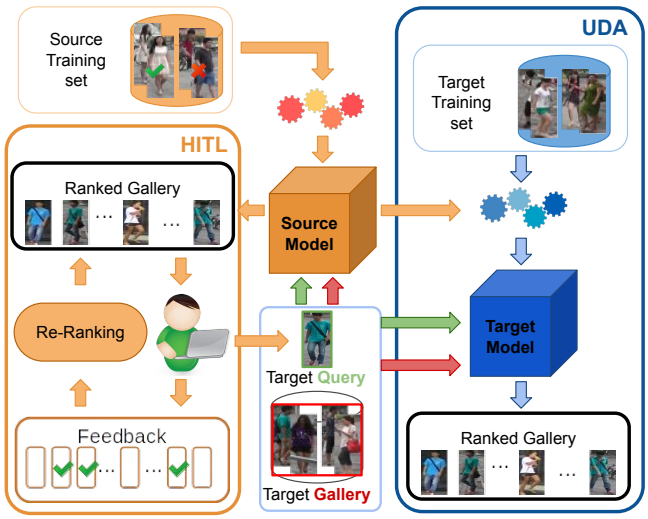

Fig. 1. High-level view of UDA and HITL approaches to person re-identification. They both start from a model trained offline on source data, and possibly fine-tuned using unlabelled data from auxiliary data sets. UDA refines it offline (before deployment) using unlabelled target data. HITL refines online (during operation) the ranked list of target gallery images provided by the source model, exploiting user's feedback (online updating of the distance metric, external to the source model, is also carried out in [21]).

algorithms are the classical Rocchio or Query Shift (QS) algorithm [37]–[39], and the Relevance Score (RS) algorithm [40]. **QS** assumes that positive images are clustered in feature space, whereas the original query $\mathbf{x}_\mathrm{q}$ could lie (relatively) far from this cluster. Accordingly, after user's feedback on N_p positive (relevant) and N_n negative (non-relevant) images, QS computes a new query and places it near the Euclidean centre of positive images in feature space and far from negative ones:

$$\mathbf{x}_\mathrm{q}^\mathrm{new} = \frac{1}{N_\mathrm{p}} \sum_{i \in D_\mathrm{p}} D_i - \frac{1}{N_\mathrm{n}} \sum_{i \in D_\mathrm{n}} D_i \qquad (1)$$

where D_p and D_n are the sets of positive and negative images, respectively, and D_i is the feature representation of the i-th image. **RS** belongs to Nearest Neighbour approaches, and is the most used and still effective RF approach to compute a score for each image [41]. For each retrieved image $\mathbf{x}$, RS computes a *relevance score* $s_\mathrm{NN}(\mathbf{x})$ based on the distances to its nearest positive and negative neighbouring images, $\mathbf{x}_\mathrm{p}^\mathrm{NN}$ and $\mathbf{x}_\mathrm{n}^\mathrm{NN}$:

$$s_\mathrm{NN}(\mathbf{x}) = \frac{\|\mathbf{x}-\mathbf{x}_\mathrm{n}^\mathrm{NN}\|}{\|\mathbf{x}-\mathbf{x}_p^\mathrm{NN}\|+\|\mathbf{x}-\mathbf{x}_\mathrm{n}^\mathrm{NN}\|} \ , \qquad (2)$$

where $\|\cdot\|$ is a given metric in feature space, typically the Euclidean distance. The relevance score increases as the distance from the nearest positive image decreases compared to the distance from the nearest negative one. Note that RF algorithms like QS and RS do not include an online learning phase, and are therefore very fast.

A relevant aspect for RF algorithms is the underlying feedback protocol. The one used in state-of-the-art HITL person re-identification methods consists in asking the user at each round to select a *single* image among the top-k ranked

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

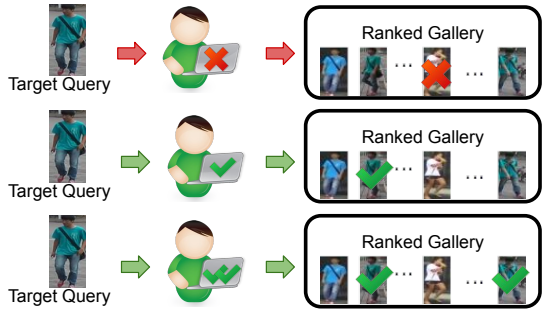

Fig. 2. Single-feedback protocol: the user selects either a positive match (if any) among the top-k gallery images (middle row) or a strong negative (top row). Multi-feedback protocol (bottom row): the user selects all positive matches (if any); the remaining images are automatically labelled as negatives.

ones ($k = 50$ in [19], [21]): either a true match, if any, or a *strong negative*, i.e., a template image very different from the query (see Fig. 2).[1] This protocol, that we call "single-feedback", may seem convenient since it limits user's effort to a single image per round. However, we point out that it appears sub-optimal for RF algorithms, that usually benefit from a larger number of feedback per round. Indeed, from Eq. (1) it can be seen that for QS a single feedback either on a true match $D_\mathrm{p} = \{\mathbf{x}_\mathrm{p}\}$ (with $N_\mathrm{p} = 1$) or on a (strong) negative $D_\mathrm{n} = \{\mathbf{x}_\mathrm{n}\}$ (with $N_\mathrm{n} = 1$) would result trivially in $\mathbf{x}_\mathrm{q}^\mathrm{new} = \mathbf{x}_\mathrm{p}$ or $\mathbf{x}_\mathrm{q}^\mathrm{new} = \mathbf{x}_\mathrm{n}$, respectively, whereas a more refined new query can be obtained from a larger set of true matches (if any) and negative examples. Similarly, the single-feedback score of RS (Eq. 2) is trivially reduces to $s_\mathrm{NN}(\mathbf{x}) = \frac{1}{1+\|\mathbf{x}-\mathbf{x}_\mathrm{p}\|}$ or $s_\mathrm{NN} = \frac{\|\mathbf{x}-\mathbf{x}_\mathrm{n}\|}{\|\mathbf{x}-\mathbf{x}_\mathrm{n}\|+\|\mathbf{x}-\mathbf{x}_\mathrm{q}\|}$ for a positive and a negative example, respectively, whereas a more refined score can be obtained from a larger amount of feedback.

Accordingly, we argue that a more suitable choice consists in asking the user to select *all* the positive matches in the top-k ranks, if any; this also allows to exploit *all* the remaining images in such ranks as negative examples. Therefore all top-k images can be used in each round of RF to update the ranked list, while asking the user to select only true matches. We call our protocol "multi-feedback" (see Fig. 2). Although our protocol may seem more demanding for the user, we point out that, as the size of template gallery increases to values typical of real application scenarios, it becomes less likely to find positive matches in the top-k ranks, for relatively small values of k such as 50. For instance, an n-fold increase of the template gallery leads to an n-fold decrease of rank-1 accuracy [21]. We also argue that asking the user to select all positive matches among the top-k images fits a scenario when the user wants to retrieve all occurrences of the query individual in the gallery to reconstruct his or her movements. On the other hand, also under the single-feedback protocol, when no positive match is present in the top-k ranks, the user has to analyse all such images, and in addition a strong negative has to be selected, which is not required by our multi-feedback protocol.

To sum up, the little attention that the HITL approach has received so far in the person re-identification literature can be justified by the effectiveness reached by state-of-the-art *supervised* methods. However, we believe that this approach remains valuable in challenging cross-view, unsupervised application scenarios characterised by the unavailability of (unlabelled) target images during system design, as well as by very large template galleries. We also believe that RF algorithms with a suitable feedback protocol can allow an effective as well as efficient implementation of the HITL approach. Accordingly, the rest of this paper is devoted to an empirical investigation of HITL person re-identification based on RF, and to a comparison with UDA.

IV. Experimental Evaluation

Based on the discussion in Sect. III, in this section we empirically compare HITL person re-identification methods based on RF algorithms with state-of-the-art UDA methods, to understand whether the former can be a valid alternative to the latter when unlabelled target data suitable to UDA are not available before operation.

A. UDA and HITL methods

For a fair comparison, UDA and HITL methods should be applied to the same source model (see Fig. 1). This requires that the source code of the considered UDA methods is available, and that it allows to train the source model only. Among the existing UDA methods only Mutual Mean Teaching (MMT) [24] and Exemplar Camera Neighbourhood invariance (ECN) [16] turned out to fulfil these requirements.[2] They both use a ResNet-50 backbone pre-trained on ImageNet. **ECN** adds a fully convolutional layer to the backbone network for feature extraction. It trains the network as an identity classifier on the source domain, together with an additional exemplar memory module for invariance learning on target data which stores the features of each target image. Invariance learning is carried out after estimating the similarities in the above feature space between a mini-batch of target samples and all such samples, to enforce exemplar invariance (related to differences in the images of the same individual), camera invariance (to account for camera style variation) and neighbourhood invariance (related to similar images in feature space). **MMT** first trains two instances of the backbone network with different weight initialisation as identity classifiers on the source domain, and generates hard pseudo-labels by clustering target images in the corresponding feature spaces. It then collaboratively trains the same networks on the target domain to predict soft pseudo-labels, under the supervision of hard labels: two temporally averaged models of such networks are updated, and the best performing one (on validation data) is finally used as a feature extractor for the inference step.

With regard to HITL methods, we considered the two RF algorithms described in Sect. III-B, namely QS and RS. We

[1] As an option, weak negatives can also be selected in [19].

[2] The code of [12], [15] is also available, but it did not work properly.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

point out that QS was used also in [19], but only as a baseline for comparison, with the single-feedback protocol. We also considered the Efficient Manifold Ranking (EMR) algorithm [42], although it is more complex than QS and RS, since it was used for comparison in [21] as well. **EMR** belongs to graph-based Manifold Ranking (MR) approaches. Instead of using K-nearest neighbours for graph construction as classical MR approaches, EMR uses K-means and a new form of adjacency matrix that optimises the ranking function by least square regression. Although MR approaches are not designed for RF, it turned out they can handle user's feedback very efficiently [42]. A more complete comparison should have included state-of-the-art HITL person re-identification methods not based on RF [19], [21], but their source code was not available.

B. Data sets

We selected two common benchmarks: Market-1501 [43] and DukeMTMC-reID [44] (for short, Market and Duke). They contain, respectively, 32,668 and 36,411 images (bounding boxes) of 1,501/1,404 identities, acquired from 6/8 cameras. Training and testing data contain *disjoint* sets of identities, 751/750 for Market and 702/702 for Duke. The number of training images is 12,936 for Market and 16,522 for Duke; the remaining ones are used for testing. The gallery contains 15,913 images for Market and 17,661 images for Duke.

C. Experimental settings

We carried out cross-data set experiments: each data set was used in turn as the source domain and the other as the target domain. For all the considered methods we adopted the authors' recommended parameter settings [16], [24], [38], [41], [42], and used as the source model a ResNet-50 pre-trained on ImageNet (see Sect. IV-A) and fine-tuned on the training partition of the source data set. For performance evaluation we used the testing partition (query set and template gallery) of the target data set. As in [21] we used a subset of 300 identities of the query set, with one image per identity, due to the time required to collect user's feedback. For HITL methods we carried out three feedback rounds as in [21]. We adopted two common performance metrics: mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) curve at rank 1, 5, 10 and 20.

D. Results

Table I shows the overall results. Note that the performance of HITL methods refers to the third feedback round. As expected, all the considered UDA and HITL methods outperformed the source model. Among the UDA methods, MMT outperformed ECN, whereas RS outperformed the other HITL methods for both feedback protocols.

Focusing now on the main comparison between UDA and HITL, it is interesting to observe that RS always outperformed either ECN or both ECN and MMT. In particular, using the multi-feedback protocol, RS was the best performing method overall, except for the re-identification accuracy at ranks 5, 10 and 20 on Market. This result is particularly interesting, taking into account that RS is a much simpler HITL algorithm than state-of-the-art ones [19], [21]. Although a comparison with such methods was not possible (see Sect. IV-A), the above results are nevertheless sufficient to show that the HITL approach is capable to achieve a performance similar to or even better than UDA.

To better analyse the performance and behaviour of the considered HITL methods, in Tables II and III we report their performance after each feedback round, respectively under the single- and multi-feedback protocol. A comparison between Tables I and II shows that RS outperforms the UDA method ECN on Duke since the second round, and even using the single-feedback protocol, in mAP, rank-1 and rank-5. Moreover, using the multi-feedback protocol, Tables I and III show that on the same target data set (Duke) RS outperforms ECN at all considered ranks and in mAP; it also performs comparably to MMT since the first feedback round.

Consider now the behaviour of the HITL methods under the two feedback protocols and over the three rounds. Tables II and III show that the proposed multi-feedback protocol considerably improved the performance of RS over single-feedback. RS also exhibits a constant and often considerable performance improvement after each round on both target data sets and for both protocols. With respect to the source model (see Table I) the improvement is already considerable at the first feedback round (about 20% in terms of mAP and 33% in terms of rank-1), up to the point that it may be not necessary to engage other rounds to further improve performance.

Tables II and III show that the simpler QS algorithm attained limited improvements with respect to RS. Nevertheless, the multi-feedback protocol was beneficial to QS as well, which exhibits remarkable improvements both in mAP and in all considered ranks with respect to single-feedback. In particular, for both target data sets the improvement was of 7% on average at all ranks, whereas mAP improved of 9% and 14% on Duke and Market, respectively.

With regard to EMR, under the single-feedback protocol its performance increased over the feedback rounds, but the third round brought a rather limited or null improvement on Duke. The multi-feedback protocol (see tab. I) provided a significant improvement only in mAP (around 11%) on both data sets, whereas the improvement in the considered ranks was small (around 1%). The fact that EMR was not capable to benefit from multiple runs of the multi-feedback protocol can be explained by the fact that MR approaches to CBIR (including EMR) were not originally designed to exploit user feedback (see Sect. IV-A).

We finally point out that the processing time required by the considered RF algorithms to re-rank the template gallery was negligible (less than one second), due to their simplicity (especially for RS and QS), including the absence of an online learning phase.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

TABLE I

RESULTS OF CROSS-DATA SET EXPERIMENTS (SOURCE → TARGET) FOR THE SOURCE MODEL, UDA METHODS (ECN, MMT), AND HITL METHODS (QS, EMR, RS) AFTER THREE ROUNDS OF SINGLE- AND MULTI-FEEDBACK PROTOCOL. BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

Method	Market ⟶ Duke					Duke ⟶ Market				
	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Source model	29.1	47.7	61.3	66.0	72.0	25.5	54.3	72.0	79.0	81.7
ECN	43.2	66.7	77.3	81.0	82.7	34.9	64.3	80.3	86.0	91.0
MMT	60.8	76.0	85.3	88.0	90.3	69.4	87.0	**95.3**	**97.0**	**97.7**
QS-single	42.71	68.67	74.33	76.33	78.0	33.9	71.0	78.0	82.0	84.0
EMR-single	36.79	72.33	72.67	73.33	73.67	30.41	70.33	72.0	73.33	75.33
RS-single	56.6	82.33	82.67	83.0	83.67	41.69	77.0	80.33	81.33	85.0
QS-multi	51.74	73.67	82.67	83.67	85.0	47.64	80.67	87.33	88.0	88.0
EMR-multi	47.23	74.0	74.33	74.33	74.33	36.13	70.67	71.67	72.0	72.0
RS-multi	**74.67**	**92.0**	**92.67**	**92.67**	**93.0**	**75.09**	**92.67**	92.67	93.33	93.67

TABLE II

RESULTS OF CROSS-DATA SET EXPERIMENTS FOR THE HITL METHODS AFTER EACH ROUND OF THE **SINGLE**-FEEDBACK PROTOCOL. BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

Method	Round	Market ⟶ Duke					Duke ⟶ Market				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
QS	1	40.49	64.33	70.0	73.0	76.0	29.59	61.33	73.67	79.33	82.67
	2	41.79	69.0	74.33	76.0	78.33	32.4	67.33	75.0	79.67	83.0
	3	42.71	68.67	74.33	76.33	78.0	33.9	71.0	78.0	**82.0**	84.0
EMR	1	33.24	61.67	64.33	67.33	70.0	23.67	50.67	60.0	63.67	68.33
	2	36.62	68.67	71.0	72.33	73.33	27.93	65.33	68.33	70.33	74.0
	3	36.79	72.33	72.67	73.33	73.67	30.41	70.33	72.0	73.33	75.33
RS	1	41.44	65.67	70.33	74.0	77.33	29.09	61.0	70.0	76.33	80.67
	2	49.51	75.67	79.0	79.0	80.33	37.11	72.67	74.67	78.33	82.0
	3	**56.6**	**82.33**	**82.67**	**83.0**	**83.67**	**41.69**	**77.0**	**80.33**	81.33	**85.0**

TABLE III

RESULTS OF CROSS-DATA SET EXPERIMENTS FOR THE HITL METHODS AFTER EACH ROUND OF THE **MULTI**-FEEDBACK PROTOCOL. BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

Method	Round	Market ⟶ Duke					Duke ⟶ Market				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
QS	1	47.68	71.0	79.0	79.67	81.33	43.23	79.67	85.67	86.67	88.0
	2	50.89	73.67	81.67	83.33	85.0	46.9	81.0	87.33	88.0	88.0
	3	51.74	73.67	82.67	83.67	85.0	47.64	80.67	87.33	88.0	88.0
EMR	1	46.27	76.0	76.33	76.33	76.33	34.86	71.0	71.33	71.33	71.33
	2	47.19	74.33	74.33	74.67	75.33	36.14	73.67	74.0	74.33	74.33
	3	47.23	74.0	74.33	74.33	74.33	36.13	70.67	71.67	72.0	72.0
RS	1	57.72	81.0	83.0	83.67	84.33	56.65	88.0	88.33	89.0	91.0
	2	68.06	87.67	88.0	89.0	90.33	68.44	91.33	92.33	92.33	92.67
	3	**74.67**	**92.0**	**92.67**	**92.67**	**93.0**	**75.09**	**92.67**	**92.67**	**93.33**	**93.67**

E. Discussion

Our results provide evidence that the HITL approach is capable to effectively address the domain gap in cross-view person re-identification scenarios by exploiting online user's feedback on target images processed during operation, instead of fine-tuning offline the source model using unlabelled target data (when available) as UDA does. Our results also confirmed that a simple RF algorithm such as RS, together with a suitable feedback protocol different from the one used by state-of-the art HITL person re-identification methods [19], [21], can be effective to this aim.

With regard to user's effort, in our experiments 18 positive matches were present on average among the top-50 ranks at the first round. The average number of feedback required to the user by the proposed multi-feedback protocol, for a single query, was therefore 18. Note that in subsequent rounds such images are likely to be still present in the top-50 ranks, and therefore the user does not have to select them again. Using the proposed protocol RS achieved similar or better performance than UDA methods, that however used *all* (unlabelled) training images of the target data set, whose number was *two* orders of magnitude larger (see Sect. IV-B). This confirms that HITL can be much more effective than UDA in terms of the required number of target images.

We further assessed the performance of HITL with our feedback protocol when asking the user a feedback on a number of gallery images lower than $k = 50$. To this aim we used $k = 10$. The results for the RS algorithm are reported

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

Fig. 3. Examples of considerable appearance changes in Market: colour (first three columns) and shape (last two columns).

Fig. 4. Examples of same individual labelled with different IDs (first two columns: Market; third and fourth column: Duke), and of different individuals labelled with the same IDs (last two columns, Duke).

in Table IV. In this case the performance of RS was lower than that of UDA methods, but it followed the same increasing trend over the RF rounds as in Table III, with a still remarkable improvement over the source model by more than 20% in mAP and 30% in rank-1 accuracy, on both data sets.

Moreover, we evaluated the performance attained by RF algorithms in further rounds after the third one. We found that the improvements were very limited, and therefore they do not justify the corresponding user's effort.

We finally highlight some issues of the data sets used in our experiments that may have penalised HITL methods. Market presents significant differences between images of the same individual from different cameras. Beside differences in lighting conditions and colours, which are inherently present in cross-view scenarios, there are also differences in the aspect ratio, most likely caused by the Deformable Part Model pedestrian detection algorithm used for this data set, which forces the bounding boxes to a fixed size (see Fig. 3). This may prevent users from recognising images of the same individual or to find differences between similar individuals; however it would not be present in a real person re-identification system, where the original pedestrian image can be shown to the user instead of its resized version, whereas only the latter is present in Market (as well as in other data sets).

Moreover, Market also contains annotation errors: different IDs can be associated to different images of the same individual, and a same ID can be associated to different individuals (see Fig. 4); this may be due to the same bounding box size issue discussed above. Also Duke presents annotation errors, which in this case appear to be mainly caused by the use of a pedestrian tracker in presence of static or dynamic occlusions. Fig. 4 shows some examples where an image is annotated with the ID of a subject, even when that subject is completely occluded. Clearly, annotation errors on target testing samples of a benchmark data set (if any) can affect the performance of HITL methods to a higher extent than UDA ones, since they are used during the RF steps to update the ranked list.

V. Conclusions

We revisited the HITL approach to person re-identification. Although it can be exploited, thanks to its inherent user interaction, to improve the accuracy of re-identification systems in any application scenario (including supervised ones), we argued that it can be particularly useful as an online domain adaptation solution in challenging *unsupervised* scenarios when even collecting representative *unlabelled* samples of the target camera view(s) is unfeasible, and therefore UDA solutions cannot be used. We showed that in such scenarios HITL can be a valid alternative to UDA, while requiring the user a feedback on a much smaller amount of target data. This result was attained using simple RF algorithms originally devised for CBIR, provided that a suitable feedback protocol is used, such as the multi-feedback protocol proposed in this paper. As an interesting research direction, our ongoing work is focused on investigating *incremental* RF algorithms capable of accumulating the operator's feedback over *subsequent* queries to better adapt the underlying re-identification model (e.g., the image similarity measure in feature space) to the target camera views, similarly to a more complex state-of-the-art HITL method not based on RF [21].

Acknowledgement

This work has been partially supported by the project LETSCROWD (Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings, https://letscrowd.eu/), funded by the European Union Horizon 2020 research and innovation programme under grant agreement No 740466.

References

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 29:1–29:37, 2013.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[3] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. I. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, 2019.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

RESULTS OF CROSS-DATA SET EXPERIMENTS FOR THE SOURCE MODEL, UDA METHODS (ECN, MMT), AND THE HITL METHOD RS AFTER EACH ROUND OF THE **MULTI**-FEEDBACK PROTOCOL ON **TOP-10** TEMPLATE IMAGES. BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN BOLD.

Method	Round	Market ⟶ Duke					Duke ⟶ Market				
		mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
Source model	–	29.1	47.7	61.3	66.0	72.0	25.5	54.3	72.0	79.0	81.7
ECN	–	43.2	66.7	77.3	81.0	82.7	34.9	64.3	80.3	86.0	91.0
MMT	–	**60.8**	76.0	**85.3**	**88.0**	**90.3**	**69.4**	**87.0**	**95.3**	**97.0**	**97.7**
RS	1	43.19	69.67	73.33	75.33	77.67	39.32	76.67	80.33	81.67	83.0
	2	48.43	75.67	78.0	79.0	81.67	44.32	81.67	83.33	86.0	86.33
	3	50.9	**79.0**	81.67	82.67	85.0	47.95	85.33	87.0	87.67	89.0

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[5] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[6] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[7] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *CoRR*, vol. abs/2001.04193, 2020.

[8] W. Li, Z. Zhong, and W. Zheng, "One-pass person re-identification by sketch online discriminant analysis," *Pattern Recognition*, vol. 93, pp. 237–250, 2019.

[9] N. McLaughlin, J. M. del Rincón, and P. C. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *AVSS*, 2015, pp. 1–6.

[10] A. Genç and H. K. Ekenel, "Cross-dataset person re-identification using deep convolutional neural networks: effects of context and domain adaptation," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5843–5861, 2019.

[11] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *ECCV*, 2012, pp. 158–171.

[12] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *CoRR*, vol. abs/1807.11334, 2018.

[13] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, 2015.

[14] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," in *Domain Adaptation in Computer Vision Applications*, 2017, pp. 1–35.

[15] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *ICCV*, 2019, pp. 6111–6120.

[16] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019, pp. 598–607.

[17] Y. Huang, P. Peng, Y. Jin, J. Xing, C. Lang, and S. Feng, "Domain adaptive attention model for unsupervised cross-domain person re-identification," *CoRR*, vol. abs/1905.10529, 2019.

[18] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011, pp. 91–102.

[19] C. Liu, C. C. Loy, S. Gong, and G. Wang, "POP: person re-identification post-rank optimisation," in *ICCV*, 2013, pp. 441–448.

[20] A. Das, R. Panda, and A. K. Roy-Chowdhury, "Active image pair selection for continuous person re-identification," in *ICIP*, 2015, pp. 4263–4267.

[21] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *ECCV*, 2016, pp. 405–422.

[22] V. Jain and E. G. Learned-Miller, "Online domain adaptation of a pre-trained cascade of classifiers," in *CVPR*, 2011, pp. 577–584.

[23] A. Royer and C. H. Lampert, "Classifier adaptation at prediction time," in *CVPR*, 2015, pp. 1401–1409.

[24] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *CoRR*, vol. abs/2001.01526, 2020.

[25] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *ECCV*, 2016, pp. 597–613.

[26] Y. Li, F. Yang, Y. Liu, Y. Yeh, X. Du, and Y. F. Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification," in *CVPR Workshops*, 2018, pp. 172–178.

[27] S. Bak, P. Carr, and J. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *ECCV*, 2018, pp. 193–209.

[28] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *ICCV*, 2019, pp. 8079–8088.

[29] H. Tang, Y. Zhao, and H. Lu, "Unsupervised person re-identification with iterative self-supervised domain adaptation," in *CVPR Workshops*, 2019.

[30] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *CVPR*, 2018, pp. 7948–7956.

[31] J. Wu, S. Liao, Z. Lei, X. Wang, Y. Yang, and S. Z. Li, "Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification," in *ICME*, 2019, pp. 886–891.

[32] S. Ali, O. Javed, N. Haering, and T. Kanade, "Interactive retrieval of targets for wide area surveillance," in *ACM*, 2010, pp. 895–898.

[33] N. K. L., R. K. Sarvadevabhatla, S. Shekhar, R. V. Babu, and A. Chakraborty, "Operator-in-the-loop deep sequential multi-camera feature fusion for person re-identification," *IEEE Trans. Information Forensics and Security*, vol. 15, pp. 2375–2385, 2020.

[34] A. Gaidon and E. Vig, "Online domain adaptation for multi-object tracking," in *BMVC*, 2015, pp. 3.1–3.13.

[35] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, and B. Caputo, "Kitting in the wild through online domain adaptation," in *IROS*, 2018, pp. 1103–1109.

[36] J. Xu, D. Vázquez, K. Mikolajczyk, and A. M. López, "Hierarchical online domain adaptation of deformable part-based models," in *ICRA*, 2016, pp. 5536–5541.

[37] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-Based image retrieval with relevance feedback in MARS," in *ICIP*, 1997, pp. 815–818.

[38] W. Lin, Z. Chen, S. Ke, C. Tsai, and W. Lin, "The effect of low-level image features on pseudo relevance feedback," *Neurocomputing*, vol. 166, pp. 26–37, 2015.

[39] W.-C. Lin, "Aggregation of multiple pseudo relevance feedbacks for image search re-ranking," *IEEE Access*, vol. 7, pp. 147 553–147 559, 2019.

[40] G. Giacinto, "A nearest-neighbor approach to relevance feedback in content based image retrieval," in *CIVR*, 2007, pp. 456–463.

[41] L. Putzu, L. Piras, and G. Giacinto, "Ten years of relevance score for content based image retrieval," in *MLDM*. Springer, 2018, pp. 117–131.

[42] B. Xu, J. Bu, C. Chen, C. Wang, D. Cai, and X. He, "EMR: A scalable graph-based ranking model for content-based image retrieval," *IEEE Trans. on Knowledge and Data Eng.*, vol. 27, no. 1, pp. 102–114, 2015.

[43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[44] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, "Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project," *CoRR*, vol. abs/1712.09531, 2017.

Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on June 19,2021 at 02:08:33 UTC from IEEE Xplore. Restrictions apply.