Cross-view Discriminative Feature Learning for Person Re-Identification


Published in:
IEEE Trans. on Image Processing

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2018 IEEE
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
Cross-view Discriminative Feature Learning for Person Re-Identification
Alessandro Borgia, Yang Hua, Elyor Kodirov, Neil M. Robertson

Abstract—The viewpoint variability across a network of non-overlapping cameras is a challenging problem affecting person re-identification performance. In this paper, we investigate how to mitigate the cross-view ambiguity by learning highly discriminative deep features under the supervision of a novel loss function. The proposed objective is made up of two terms, the Steering Meta Center (SMC) term and the Enhancing Centres Dispersion (ECD) term that steer the training process to mining effective intra-class and inter-class relationships in the feature domain of the entities. The effect of our loss supervision is to generate a more expanded feature space of compact classes where the overall level of inter-identities interference is reduced. Compared to the existing metric learning techniques, this approach has the advantage of achieving a better optimization because it jointly learns the embedding and the metric contextually. Our technique, by dismissing side-sources of performance gain, proves to enhance the CNN invariance to viewpoint without incurring increased training complexity (like in Siamese or Triplet networks) and outperforms many related state-of-the-art techniques on Market-1501 and CUHK03.

Index Terms—Viewpoint, Loss function, Multi-camera, Person re-id, Discriminative features

I. INTRODUCTION

Performing person re-identification is a challenging task because of the presence of many sources of appearance variability like lighting, pose, viewpoint, occlusions, especially in outdoor environment [1], [2] where they are even more unrestrained. Cameras calibration or cross-camera image processing may help, but they are not an option in a surveillance context where a wide-area network of cameras with non-overlapping fields of view is deployed. In such a scenario, we investigate the changing viewpoint problem. The misleading effect of this particular factor of variability is that shots of different pedestrians taken under the same camera may quite often look more similar to each other than shots of the same identity taken under different cameras. This is illustrated in Figure 1. We support the view that learning inter-camera relationships is essential to tackle this ambiguity [3], [4] since it can contribute to build, at training time, a more discriminative feature space where all classes (pedestrian identities) are less conflated and more distant from each other (Figure 2). It can be done by properly designing a loss function, tailored to the goal, that supervises the learning process of a deep architecture.

Apart from some early techniques relying on designing hand-crafted features [5] or cross-camera transformations [3], the multi-camera context, traditionally, has embraced the deep learning (DL) paradigm because of its ability to learn complex discriminative mappings that generalise well [6]–[10]. Most of the DL approaches in person re-id focus on exploiting one of the following aspects: 1) more complex deep architecture structures, aiming either to optimize jointly more tasks in the re-id pipeline [11], [12] or to learn cross-spatial/temporal representations [13]; 2) side information extraction, [14], [15], involving pose estimation strategies and misalignment correction; 3) metric learning [16], [17] that learns a distance/similarity function in a fixed feature space; 4) more training data, either by performing cross-dataset training [18] or exploiting the transfer learning paradigm [19].

With regards to the above classification, our approach based on employing a more discriminative loss function, has the ad-
vantage of being totally complementary to the these methods: it can easily be integrated into any architecture that leverages a higher structural complexity, enhanced data exploitation or ML. Substantially, our work investigates how to make the most out of a fixed available amount of training data, without relying on the exploitation of any side information [20], [21]. A notable example of this technique in face verification is [20] that proposes the center loss function that enhances the softmax loss supervision by promoting the compactness of the extracted features around the center point of each face-class. Although the idea is quite effective in face verification, we think that its limitation in person re-id is that it does not exploit the field of view information, available in multi-camera datasets, which could be helpful to mitigate the viewpoint problem. [21] tries to fill this gap with the intra- and inter-Group-Center Losses (GCL): it incorporates the field of view information in their definition, although they still face some limitations because of their mathematical formulation (Sec.III). In our work we overcome these limitations by proposing a novel loss made up of two terms, the Steering Meta-Center term and the Enhancing Classes Dispersion term. Its goal is to exploit effectively the field of view information in order to foster the separation capability of the softmax loss and enhance further the intra-class gathering behaviour of the center loss.

In summary, the main contributions of our work are the following:

- We propose a flexible method to successfully mitigate the viewpoint problem in person re-id by tailoring a new loss function that enables a CNN to learn more discriminative features.
- Our approach adapts the ML approach to the training stage for learning an inter-class distance/similarity function.
- Our results beat the best performing loss function-based approaches (Table III) and most of the state-of-the-art methods (Table II) on two of the largest datasets for outdoor person re-id.

II. RELATED WORK

A. Deep learning for person re-id

The first work to apply DL to person re-id is [6] that relies on a Siamese CNN equipped with a cosine similarity connection function. Following this research direction, lots of other works, taking advantage of the availability of new large-scale datasets (CUHK03 [7], Market-1501 [23]), have adopted either the pure data-driven paradigm [8], [10], [18] or hybrid approaches [7], [24]–[26]. In mixed architectures, hand-crafted descriptors for pedestrians are integrated into DL frameworks and exploit the complementarity of their features. Fisher vectors and deep neural networks are combined together, for example, in [24]; [7], [25] design a Siamese network with constraints on the shape of the objective to learn by adding custom hand-crafted layers to the CNN; in [26] convolutional and hand-crafted histogram features are fused together to produce a more discriminative descriptor. In order to enhance performance in deep architectures for person re-id, two popular strategies exploit either some spatial cues of the input images [27] or the side information extracted from data, like in [14] where, within a DL framework, an effective correction of full-body images misalignment is performed by using Convolutional Pose Machines for pose estimation. A recent trend consists in addressing different re-id related tasks jointly. In this perspective, [12] in its multi-task network fuses the binary classification and ranking tasks together, while [11] integrates the pedestrian detection and searching tasks in one unified end-to-end trainable net, taking a significant step ahead in the direction of filling the gap between research-oriented re-id systems and real world deployable re-id systems. Lastly, transfer learning-based architectures provide evidence of how much performance benefits from the availability of extra data in person re-id by learning generic deep feature representations from multiple domains [18], [19]. In our work we adopt the ResNet50 architecture also used in [14] which relies on the effectiveness of the residual learning paradigm for learning deeper feature hierarchies [28].

B. Loss functions

The simple probabilistic interpretation of the softmax loss and its features separation capability make it the most widely used loss [26], [29]. In multi-label prediction, the sigmoid cross-entropy classification loss is sometimes preferred for its better performance like in [30] where it is used for jointly learning correlated complementary local and global features. Common approaches in this area involve either modifying the softmax loss or replacing/combining it with novel losses. The shared goal is to enforce intra-class compactness and inter-class separability in the feature space.

A modification of the softmax loss is proposed in [11] with the random sampling softmax loss that allows supervising the
training stage with sparse and unbalanced labels. In [31] the
generalized large-margin softmax loss generalizes the softmax
loss incorporating in its new definition the cross-entropy loss and
a fully connected layer to achieve larger angular feature
separability. The reduction of intra-class variability has been
addressed in [20] by the center loss that learns the centers of
deeper features of each id-class (in face verification), and in [21]
by the two combined GCL losses that also promote inter-class
separation.

The choice of which loss function to use may heavily
affect the way the training samples are built, having an impact
eventually on the training complexity. This is the case for the
contrastive loss [32] which is used in the Siamese network
model. It proves to be quite effective in person re-id like in
[8] and [10] that consider ways of exploiting spatial relations of
images, within a single image or between image pairs.
A boosted learning capability derives from including into a
Siamese framework more losses related to different visual
tasks (identification and verification) combined together [19].

A direct competitor of the Siamese loss is the triplet loss for
the triplet network model [33], [34] that enables insensitivity
calibration which is a problem in Siamese CNNs where the
concept of similarity/dissimilarity is tied to the specific
context [35]. In [9], an improved triplet objective is used with
an upper-threshold on the maximum distance for the intra-class
features, in order to train a multi-channel parts-based CNN.
A combination of the pairwise and triplet-wise feature learning
modality is also presented in [13], [36].

C. Learning a metric in the feature space

Distance-ML based methods learn transformations of the
original feature space in order to project embedded represen-
tations belonging to different cameras onto a common space
where the view discrepancy is mitigated [37], [38]. Typically,
the existing ML approaches may be classified in two groups:
non-DL ML methods and DL-based ML methods.

The first group is characterized by performing feature em-
bedding learning and ML in two separated subsequent stages,
so that a metric is learned only after the CNN weights are
already fixed. In this group falls [39] that applies deep transfer
ML to learn a set of hierarchical non-linear transformations for
cross-domain recognition. Intra-person variability in [40] is
handled by carrying out a similarity learning that obeys some
spatial constraints accounting for the geometrical structure of
pedestrians (to match correspondent body-parts). [41] learns a
new metric as a combination of a Mahalanobis metric and a
bilinear similarity metric and benefits from considering both
the difference and commonness of an image pair and a pair-
constrained Gaussian assumption. An original definition of
similarity metric is presented in [42] as a log likelihood ratio
between the probabilities of the two identities to be matched
(in face recognition).

In contrast, in the second group of ML methods, a metric is
learned jointly with the embedding by a DL-based architecture.
One way to achieve this is in a Siamese network fashion,
exploiting the contrastive loss (or triplet loss) that allows a
distance relation between pairs (or triples) of feature points
to be learned at training time [6], [9], [43], [44]. Another
approach among the DL-based ML methods is to integrate
the ML scheme into the feature extraction ML methods, producing a
re-id system overall trainable end-to-end by gradient descent.
In [16], [38], for example, the Mahalanobis distance matrix
is factorized as one top fully-connected layer and the matrix
of its weights is learned together with the CNN weights.
Similarly, [45] finds an end-to-end globally optimal matching
in a multi-camera network by exploiting intra- and inter-
camera consistent-aware information during the training stage
of a three branches CNN.

Both the Siamese-based and the end-to-end integration-
based approaches often need to be supported by sample mining
strategies to either improve the training effectiveness [8], [25]
or reduce the over-fitting problem [16]. A hard negatives
mining strategy is applied for example in [25] and in [7] where,
after being retrieved, the hard negatives are combined with the
positive samples to further train the network. Moderate pos-
tive samples selection is required instead in [16] to mitigate
the increased over-fitting to bad positives. The loss we propose
allows us to avoid dealing with any sample mining strategy
while, at the same time, benefiting from the embedding-metric
joint learning.

III. PROPOSED METHOD

Aiming to reduce the viewpoint ambiguity connected to the
multi-camera scenario by producing more discriminative deep
features, we assert that learning a metric only over a fixed
CNN is limiting compared to conveniently shaping the feature
space itself while it gets built, at training time. Therefore an
approach is required that address the two tasks of learning an embedding and of learning a metric jointly, avoiding the
main drawbacks of the Siamese architectures, but still retaining
their inter-camera relationships learning capability, critical for
enabling the viewpoint invariance.

When ML is performed separately by the feature extraction
task, besides reaching a suboptimal solution, usually, either di-
mensinality reduction or regularization are required to avoid
singularity in the intra-class scatter matrix due to the limited
number of training samples for a single identity compared
to their feature dimensionality (small sample size problem,
[46]). On the other hand, the joint learning performed by a
Siamese network-based approach may result in an increased
training complexity for several reasons. Firstly, the explosion
of the number of samples due to the need to build the training
samples by selecting pair/triplet of input images. Secondly, the
need of sample mining strategies to improve the effectiveness
of the training [8], [25]. Thirdly, the contrastive loss only relies
on weak re-id labels (same id or different id), as pointed out in
[47], and does not exploit fully the entire information carried
by the re-ID labels on class membership. Lastly, because of
the unbalanced training data problem, by learning a network
by binary classification, their predictions are usually biased
towards negatives. Countermeasures to this, usually require to
increase the training complexity even further as in [48] where
a unified deep learning-to-rank framework is proposed.

Our approach replicates the capability of Siamese networks
to carry out a joint features-metric learning process while
at the same time keeps the training complexity low, since one training sample corresponds to a single input image, like in the non-DL metric learning methods. We design a novel loss function that has the nice properties of a) being additive with regards to the softmax loss; b) being suitable to be easily integrated in a simple one branch shaped CNN, being trainable by gradient descent; c) being suitable for fast search requirements since it scales well to large datasets; d) producing embeddings discriminative enough that simple metrics such as the normalized Euclidean distance can be used for comparing the multi-dimensional feature points representing the identities instances.

The loss function we design is made up of two additive terms: the Steering Meta-Center term (SMC) and the Enhancing Classes Dispersion term (ECD). Used in linear combination with the softmax loss, they promote two desirable properties of the deep features distribution: the properties of intra-class compactness and of inter-class separation. In particular we test them in two combinations: SMC+ECD as in Equation 1 and SMC only, without the ECD term contribution. Beyond producing a more discriminative feature space under the SMC+ECD loss supervision, we investigate also the relationship between our method and the traditional ML approach, in order to understand whether combining them together a further improvement can be gained (Sec. IV-C).

\[
L = L_{\text{softmax}} + \lambda_{\text{SMC}} \cdot L_{\text{SMC}} + \lambda_{\text{ECD}} \cdot L_{\text{ECD}}
\]

A. Steering Meta-Center (SMC) Loss Term

The SMC loss definition exploit the camera information of all the dataset images aiming at two goals: a) Improving the center loss [20] compactness in person re-id; accounting for the camera information helps, indeed, balance the contributions of the different views to defining a more discriminative deep representation of the overall identity. b) Learning to some extent inter-camera relationships, which allows to outdistancing different identities. This task is usually deferred to after the training stage and performed by metric learning schemes.

In the following, we will illustrate how these two aspects are dealt with jointly by the SMC loss term, by introducing a new virtual point in the feature space, referred to as meta-center which steers the learning and shapes the feature space. With regards to an identity, a meta-center point is defined simply as the vector sum of its sub-centers, as clear from Equation 2, defining the SMC loss,

\[
L_{\text{SMC}} = \frac{1}{2} \sum_{i=1}^{m} \left\| \mathbf{x}_i^{(y_i)} - \sum_{j=1}^{s_i} c_{y_i}^{(j)} \right\|^2 + \frac{1}{2} \sum_{i=1}^{m} \left\| \mathbf{x}_i^{(y_i)} - c_{y_i}^{(\text{SMC})} \right\|^2
\]

(2)

where \(m\) is the training mini-batch size; \(y_i\) denotes the identity ground-truth label of the \(i^{th}\) mini-batch image; \(s_i\) represents the number of cameras that capture the identity \(y_i\); \(g_i\) is the camera ground-truth label of the \(i^{th}\) mini-batch image (\(1 \leq g_i \leq s_i\)); \(\mathbf{x}_i^{(g_i)}\) denotes the feature vector of the \(i^{th}\) input image viewed under camera \(g_i\); \(c_{y_i}^{(j)}\) represents a sub-center point of the identity \(y_i\), calculated by averaging the points of \(y_i\) that are viewed under camera \(j\); \(c_{y_i}^{(\text{SMC})}\) is the sum of all the \(s_i\) camera-related sub-centers \(c_{y_i}^{(j)}\) of identity \(y_i\).

It should be noted that Equation 2 does not represent a global objective to be minimized by the overall training process; instead, its minimization is carried out locally, with scope limited to each individual iteration (similarly to what happens for the triplet loss), with regards to the current value only of the meta-center.

At training time, each new training image being processed is pulled toward the meta-center of its class according to Equation 2. The subsequent step consists in updating the position of the correspondent sub-center that, in turn, moves
away from the reference system origin, following the meta-center (Figure 3). The repetition of this training cycle, on the one hand, makes the feature points of all identities drift away from the system origin; on the other hand, it makes each class tend to approach more and more tightly the steering direction of the meta-center. The overall effect of these two potentials is to simultaneously:

- produce a more scattered space of classes because of the identities progressive movement behind their own meta-centers.

- achieve a high intra-class compactness and a less sub-clustered space structure (more inter-mixed points regardless of their camera view), as a result of the increased insensitivity to the camera viewpoint (Figure 4).

The training process is concurrently supervised by the softmax loss and lasts until the produced features start losing generalization power due to data over-fitting.

It is interesting to analyze the mathematical relation between the meta-center point \( c^{(SMC)} \) defined by the SMC loss and the center point \( c^{(center)} \) defined by the center loss, in order to gain a deeper insight into their difference. In Equation 3 we reformulate the center loss as a function of the sub-centers variables, for a fixed identity.

\[
\begin{align*}
  c^{(center)} &= \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{1}{N} \sum_{j=1}^{s_i} \sum_{i=1}^{N_j} x_i^{(j)} = \frac{1}{N} \sum_{j=1}^{s_i} N_j c^{(j)} \\
\end{align*}
\]  

(3)

where \( N \) is the number of images of the identity of interest \( y_i \) shot by \( s_i \) cameras; \( N_j \) is the number of images of the considered identity belonging to the camera view \( j \); \( c^{(j)} \) is the sub-center associated to camera view \( j \). This formulation shows that \( c^{(center)} \) is equivalent to the weighted mean of the sub-centers of the class, with weights given by the cardinality of the population of each sub-class. In contrast, the meta-center, defined in Equation 2 as \( c^{(SMC)} = \frac{1}{N} \sum_{j=1}^{s_i} c^{(j)} \) is a scaled version, by a factor \( N \), of the unweighted mean of the class sub-centers \( \frac{1}{N} \sum_{j=1}^{s_i} c^{(j)} \), as clear from Figure 5. Therefore, the SMC loss has the nice property of referring to the unweighted mean instead of the weighted mean which allows to account for an equal contribution from all the sub-classes in defining an identity representation, regardless of the number of points they have. Since we want to address the invariance of an identity representation from the camera views, we give all the sub-classes the same relevance in determining the point that summarizes the class.

The problem of accounting for the camera information as a mean to improve the center loss class compactness has already been attempted by the intra-GCL loss [21]. By the way, its formulation shows some limitations due to the fact that it does not constrain the sub-centers of one identity (which condense the camera information) to converge to each other and the effect is to get compact sub-classes in a still wide overall class with consequently reduced performance.

B. Enhancing Classes Dispersion (ECD) Loss Term

The ECD loss is designed to enforce explicitly higher inter-class dispersion by imposing a relative constraint between
The inter-GCL expression flatten to zero when even only one sub-center of one class is badly initialized, affecting negatively the right-hand term of the product of Equation 4. Under this circumstance, indeed, it is interesting to analyze how better the behaviour of the right-hand term of the product accounts for how much extended the reference class is in the feature space (class range); the right-hand term, instead, expresses a measure of the isolation of the reference identity from each of the other identities in the current mini-batch (Figure 6).

\[
L_{ECD} = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \| x_i^{(g_i)} - c_{y_i}^{(k_i)} \|_2^2 \cdot \sum_{t=1}^{m} \sum_{k=1}^{s_t} \frac{1}{\| x_t^{(y_t)} - c_{y_t}^{(k_t)} \|_2^2} 
\]

(4)

Our starting point in the definition of Equation 4, is noting that the inter-GCL loss minimization [21] may not be effective in case the centers initialization values were subject to a large spread with respect to the data (unbalanced centers): namely, \( \| x_i^{(g_i)} - c_{y_i}^{(k_i)} \|_2^2 \) very large for some \( y_t \) and \( k \). Under this circumstance, indeed, it is interesting to analyze how better the effect of our constraint is expected to be, how confirmed from the best results achieved on Market-1501 dataset than on CUHK03. As to the viewpoint problem, learning a CNN under the joint supervision of the SMC and ECS loss terms produces a positive impact, as illustrated in Figure 7 for a small subset of accounting at once for all the dataset identities.

Differently, our solution bypasses this problem by constraining the class range to be small with regards to each individual sub-center distance at a time. In the ECD loss formulation (Equation 4), even if several terms of the summation flatten to zero due to some unbalanced centers, the other addends finite contribution will not be affected. Furthermore, another difference is that we limit the summation of the inter-class distances only to the identities in the current training mini-batch instead of accounting at once for all the dataset identities.

The larger the number of the sub-classes is, the stronger the effect of our constraint is expected to be, how confirmed from the best results achieved on Market-1501 dataset than on CUHK03. As to the viewpoint problem, learning a CNN under the joint supervision of the SMC and ECS loss terms produces a positive impact, as illustrated in Figure 7 for a small subset.
of identities and in Figure 8 for the whole dataset (Market-1501). The ECD loss reproduces at training time what non-DL ML methods do on top of a CNN already learned, learning a distance relation between inter-class pairs.

IV. EXPERIMENT

A. Settings

Database. We evaluate our approach against two of the largest person re-id dataset: CUHK03 [7] and Market-1501 [23]. With regards to CUHK03, all results refer to its labeled subset (Table II). In CUHK03, each identity is shot by one pair of cameras out of the three pairs available \((s_i = 2\) in Equation 2) and counts maximum 10 images: the first 5 images are viewed under a different camera than the remaining ones. In Market-1501, each identity is seen under up to 6 different views \((3 \leq s_i \leq 6)\) for up to 70 images. The 12936 images of the train set correspond to 751 identities completely disjoint from the 750 identities of the test set, having 13115 instances. 2798 more images representing heavily misaligned detections with ID identifier ‘0000’ (distractors) are added to the test set in order to make the re-identification task more challenging. The database includes as well a query set of 3368 images (probes) which are picked from the test set so that, for each id, only one image per camera view is selected (mislabeled examples reported in Appendix Table VII). The images of both the datasets are generated by the DPM detector [49].

Evaluation Metric and Protocols. Our ranking-based evaluation method is conducted by matching (by cosine similarity) the feature vectors of all the test images against a probe representation and then sorting the correspondency similarity scores in decreasing order (Figure 7-c). The common evaluation metric we use for measuring performance against both the datasets is the Cumulative Matching Curve (CMC, Table III). For Market-1501 we also employ the mean-Average Precision (mAP, Table III) since it has multiple ground-truths for each query and both precision and recall need to be taken into account.

Market-1501 comes in the form of three directories corresponding to train set, test set and query set. We follow its original evaluation protocol [23] for which all the query images have to be tested against their own gallery set. Each gallery set excludes the test images that have a filename starting with '-1' and that belong to the probe junk set (made up of all the test images sharing the probe same identity and field of view). Our experiments are performed both in single query testing mode (results in Table II and Table III) and in multiple query mode (Table IV). As to the former, only one query image is selected for each camera view of a given id. In the multiple query mode, the presence of multiple query images in a single camera for each identity allows to achieve superior performance in re-identification [50]. We apply the max pooling strategy (which performs better than the average pooling) to merge them into a single query for speeding up the process, as done in [23]. For CUHK03, we reproduce the evaluation protocol in [7], according to which the first 5 images of each identity represent view A, the remaining 5 represent view B. View A includes camera 1, 3, 5 while view B camera 2, 4, 6. All images belonging to view A form the probe set. To each probe corresponds a gallery set of 100 images selected randomly from view B, such that one image is picked for each of the 100 identities in the test set. This selection mode of the gallery set is repeated 50 times in order to calculate the mean CMC curve. Each of the 20 test sets counts 100 images and the validation set includes 100 identities.

Besides the rank 1 accuracy and the mAP, we introduce the figure of merit \(F_{\text{rank}1}^{(L)}\) in order to quantify which fraction of the overall improvement achieved by a new loss \(L\) translates, specifically, in an improvement of the viewpoint effects. \(F_{\text{rank}1}^{(L)}\) is associated with the rank 1 accuracy metric (likewise, we can define \(F_{\text{mAP}}^{(L)}\)) and its definition comes

### TABLE I

<table>
<thead>
<tr>
<th>layer name</th>
<th>output size</th>
<th>kernel size</th>
<th>stride</th>
<th>pad</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>112x112</td>
<td>[3x7, 64]</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>conv2</td>
<td>56x56</td>
<td>[1x1, 64]</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>conv3</td>
<td>3x3x128</td>
<td>128x1024</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>conv4</td>
<td>14x14</td>
<td>[3x3, 256]</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>conv5</td>
<td>7x7</td>
<td>[1x1, 128]</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>avg pool5</td>
<td>1x1</td>
<td>[1x1, 64]</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>fc8</td>
<td>1280x751</td>
<td>1280x751</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>softmax</td>
<td>1</td>
<td>2048</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SMC, ECD</td>
<td>1</td>
<td>2048</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
from the analysis of the negatives of interest produced by the ranking process. If we choose the softmax loss as our baseline, $F_{\text{rank}}^L$ can be defined as: $F_{\text{rank}}^L = \frac{\text{Neg}^{\text{soft}}(\text{Neg}^L)}{\text{rank}^L} = \frac{\text{Neg}^{\text{soft}}(\text{Neg}^L)}{\text{rank}^L}$, where $\text{Neg}^{\text{soft}}(\text{Neg}^L)$ is the percentage of negatives of interest (Figure 1) produced under the supervision of the softmax ($\text{Neg}^L$) loss; $\text{rank}^L$ is the ranking score ($\text{rank}^L$) represents the performance achieved by the softmax ($\text{rank}^L$) loss (Table V).

Also we build the mAP camera-pairs confusion matrix (Figure 10) for the SMC+ECD loss in Market-1501 against the baseline represented by the single softmax loss function, in order to measure the relative improvement of re-id between camera pairs, that is of the viewpoint problem. With regards to a pair of cameras $(X, Y)$, representing respectively the field-of-view of the probe and a test camera, the corresponding mAP value for the considered probe is calculated by limiting its positive samples to only those ones viewed under camera $Y$. This process is repeated for all probes and values corresponding to the same camera pair are averaged together to produce a single value for the related cell in the matrix. Camera pairs sharing more similar characteristics are characterized by higher values.

**Implementation Details.** The SMC and ECD losses are implemented in C++ within the Caffe framework, as separated layers and their output added to each other and to the softmax loss. Differently from the softmax layer that is connected to the fully connected layer $f_{c8}$, the other two losses are fed by the $pool5$ layer. The derivatives of the overall loss function (ECD contribution in Appendix A) are back-propagated by Stochastic Gradient Descend using mini-batches of 16 images and, at each iteration, the centers of each class (identity) and the sub-class (field of view) are updated accordingly. In our experiments, we integrate the losses in ResNet50, a state-of-the-art residual learning-based CNN formed by 53 convolutional layers [28], stacked in 16 residual blocks (Table I). Each residual block counts three convolutional layers except the first one of each $convX$ aggregate that count one more as shown in Figure 9. The ResNet50 input are RGB images preprocessed by channel mean subtraction (calculated against the entire dataset) and resized to 224x224 pixels. The output of the $pool5$ layer with dimension 2048, according to [14], is selected as deep representation of the input data. The network pretrained on the ImageNet dataset [29] is fine-tuned on the re-id datasets for the identity classification task, setting 1260 classes for CUHK03 and 751 for Market-1501, with stop point in all simulations set at 15000 iterations.

The convergence of the optimization process is regulated by a momentum $\mu = 0.9$ and a basic learning rate $\gamma = 10^{-3}$ except for the last three loss layers and the fully connected layer $f_8$ where we apply a learning rate multiplier of 10 in order to speed up their learning without getting to far from the original optimal point reached by pre-training. A stepwise decay policy for the learning rate is also used with dropping factor $\gamma = 0.1$ and associated step interval of 9000 iterations to progressively slow down the learning. A weight decay factor equal to $\lambda = 5 \cdot 10^{-4}$ limits the learned network weights size. All our deep learning experiments are performed on a single machine equipped with one NVIDIA GeForce GTX Titan X GPU and an Intel Core i7-5960X CPU @3.00GHz, 64.0 GB RAM.

**B. Experimented Results**

**State-of-the-art Methods.** Our technique outperforms most of the best state-of-the-art methods in person re-identification (Table II). On CUHK03 we reach the best performance for the rank 1 accuracy (SMC+ECD, 69.55%) of all the methods listed in Table II. Also in Market-1501, SMC+ECD reports the best rank 1 accuracy (80.31%) and the best mAP (59.68%) achieving an improvement respectively of 2.1% and 10.8% of the correspondent values of the second best performing method [14]. Compared to the latter, our approach is more efficient in terms of network structure since [14] employs an architecture with two ResNet50-based branches non-sharing their weights and also performs pose estimation for generating a pose-invariant embedding and producing a confidence store to incorporate into the final data combined representation. Furthermore, we outperform the triplet loss based methods in
and +25% of their mAP. It is noteworthy that the additional ECD loss term supervision in CUHK03 (where all identities are shot under 2 cameras only) does not provide any further improvement compared to the SMC loss term, because the ECD enforced constraint (the sum of solid lines distances in Figure 6) becomes looser and the SMC reaches by itself the maximum achievable degree of inter-class separation.

On Market-1501 SMC outperforms the softmax, GCL and center loss rank 1 accuracy respectively of +34.8%, +9.3%, +5.1% of their value. It is noteworthy that the additional ECD loss term supervision in CUHK03 (where all identities are shot under 2 cameras only) does not provide any further improvement compared to the SMC loss term, because the ECD enforced constraint (the sum of solid lines distances in Figure 6) becomes looser and the SMC reaches by itself the maximum achievable degree of inter-class separation.

We ascertain the effectiveness of the proposed losses in the perspective of the viewpoint problem also building the mAP confusion matrix (Figure 10): it shows a significant improvement of the cross-camera re-id performance as a consequence of mitigating the effects of the viewpoint problem. The highest relative improvements are about the camera pairs (1, 6) (66.7% and +75.0%), (2, 6) (+40.0% and +50.0%), (5, 6):

Ablation Analysis. Our tests performed with ResNet50 on both the single SMC loss term and the combined SMC+GCD formulation show that they outperform the single softmax loss [14], the GCL losses [21] and the center loss [20]. Our analysis has been carried out investigating the space generated by the 1D parameter $\lambda_{SMC}$ for SMC and the space spanned by the 2D parameter $(\gamma_{SMC}, \lambda_{ECD})$ for SMC+GCD. Figure 11,13,14 report the rank 1 accuracy and mAP curves of SMC, SMC+GCD and also of the competing losses.

Fig. 10. Re-id performance between camera pairs on Market-1501. (a) mAP confusion matrix for the softmax baseline. (b) mAP Incremental confusion matrix (%) for SMC+ECD compared to (a).

Fig. 11. Rank 1 accuracy curves to changing $\lambda$ for Market-1501. $\lambda$ indicates different hyper-parameters depending on the loss referred. With regards to Equation 1, when referring to the: a) SMC loss, $\lambda$ denotes the $\lambda_{SMC}$ parameter, with $\lambda_{ECD} = 0$; b) SMC-ECD loss, $\lambda$ denotes $\lambda_{ECD}$, with $\lambda_{SMC} = 10^{-3}$. The corresponding full 2D rank 1 accuracy surface of the SMC+ECD loss, for both $\lambda_{SMC}$ and $\lambda_{ECD}$ changing, in Market-1501, is plotted in Figure 12.

Fig. 12. Rank 1 accuracy surface of the SMC+ECD loss to changing $(\gamma_{SMC}, \lambda_{ECD})$ hyper-parameters defined in Equation 1, for Market-1501.
Improvement (%) of the "viewpoint problem" in terms of rank 1 and mAP related "figures of merit" in Market-1501 (not applicable to CUHK03).

### Table V

<table>
<thead>
<tr>
<th></th>
<th>GCL</th>
<th>Center</th>
<th>SMC</th>
<th>SMC+ECD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_{\text{rank1}} )</td>
<td>15.5</td>
<td>23.4</td>
<td>24.3</td>
<td>26.3</td>
</tr>
<tr>
<td>( F_{\text{mAP}} )</td>
<td>33.4</td>
<td>35.7</td>
<td>43.6</td>
<td>50.7</td>
</tr>
</tbody>
</table>

Performance improvement achieved by the joint-Bayesian (JB) ML method. Applying JB to the softmax baseline performs worse than performing the training under a better loss function supervision and using the simple cosine similarity.

### Table VI

<table>
<thead>
<tr>
<th></th>
<th>Softmax</th>
<th>SMC</th>
<th>SMC+ECD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market-1501</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( F_{\text{rank1}} )</td>
<td>(+5.5)</td>
<td>(+0.5)</td>
<td>(+0.1)</td>
</tr>
<tr>
<td>( F_{\text{mAP}} )</td>
<td>53.77</td>
<td>58.40</td>
<td>59.73</td>
</tr>
<tr>
<td>CUHK03</td>
<td>(+26.0)</td>
<td>(+3.5)</td>
<td>(+3.2)</td>
</tr>
</tbody>
</table>

### C. Further Results

#### Negatives Analysis

With regards to the analysis of the negatives of interest, Table V reports the figure of merit of all the considered losses. On Market, \( F_{\text{rank1}}^{\text{SMC+ECD}} = 26.3\% \) and \( F_{\text{mAP}}^{\text{SMC+ECD}} = 50.7\% \), respectively, +8.2\% and +6.5\% higher for the correspondent figures for the center loss. This proves the effectiveness of our loss with regards to the viewpoint problem. On CUHK03, \( F_{\text{rank1}} \) cannot be calculated because of the way its evaluation protocol [7] is defined.

#### Training Losses vs ML

The application of the joint-Bayesian learning method to the softmax baseline shows that its performance (rank1 77.06\% and mAP 53.76\% in Market; rank1 65.03\% in CUHK03) is lower than that achieved by SMC+ECD without learning any metric, using the simple cosine similarity instead. This confirms that learning a similarity function in the feature space when the network weights are already fixed is sub-optimal to doing that jointly with learning the network itself, under the supervision of a more discriminative objective. Furthermore, by applying the joint-Bayesian technique to SMC and SMC+ECD there is still space for further improvement, in a measure depending on the depth and the cardinality of the dataset used (+0.1\% for Market-1501 and +3.2\% for CUHK03 as from Table VI).

### V. Conclusion

In the context of a network of disjoint cameras, we have proposed a new loss function for supervising a CNN that were less prone to the effects of the viewpoint problem. The SMC+ECD loss represents a re-interpretation in person re-id of the center loss introduced in face verification. Furthermore, by applying the joint-Bayesian technique to SMC and SMC+ECD there is still space for further improvement, in a measure depending on the depth and the cardinality of the dataset used (+0.1\% for Market-1501 and +3.2\% for CUHK03 as from Table VI).

### ACKNOWLEDGEMENT

This work was supported in part by the UDRC consortium (University Defence Research Collaboration in Signal Processing) and Roke, Part of the Chemring Group.

### APPENDIX A

#### Loss Derivatives

The following equations hold true:

\[
\frac{\delta L_{\text{ECD}}}{\delta x_i^{(k)}(y)} = \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{s_i} x_i^{(j)}(y) - c(y)
\]

\[
- \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2 / \sum_{t=1}^{m} \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2
\]

\[
\frac{\delta L_{\text{ECD}}}{\delta x_i^{(k)}(y)} = \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2 \sum_{t=1}^{m} \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2
\]

\[
\frac{\delta L_{\text{ECD}}}{\delta x_i^{(k)}(y)} = \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2 \sum_{t=1}^{m} \sum_{j=1}^{s_i} \left| x_i^{(j)}(y) - c(y) \right|^2
\]


Alessandro Borgia is currently a PhD student in computer vision at Heriot-Watt University and University of Edinburgh, UK. His main research interests cover machine learning and data analytics and his expertise also includes mobile network technologies and information security. He received his BSc and MSc degree in Telecommunications Engineering (2009) from La Sapienza University of Rome, Italy. He was awarded a research prize by HWU (2015) and a research contract by DSTL following an academic competition on signal processing.

Yang Hua is presently a lecturer at the Queen’s University of Belfast, UK. He received his Ph.D. degree from Universit Grenoble Alpes / Inria Grenoble Rhône-Alpes, France, funded by Microsoft Research Inria Joint Center. He won PASCAL Visual Object Classes (VOC) Challenge Classification Competition in 2010, 2011 and 2012, respectively and the Thermal Imagery Visual Object Tracking (VOT-TIR) Competition in 2015. His research interests include machine learning methods for image and video understanding. He holds three US patents and one China patent.

Elyor Kodirov is currently a senior researcher at AnyVision, Belfast, UK. He received his Ph.D. degree in the School of Electronic Engineering and Computer Science, Queen Mary University of London, 2017 and the Master’s degree in computer science from Chonnam National University, Korea, in 2014. His research interests include computer vision and machine learning.

Neil M. Robertson is Professor and Director of Research for Image and Vision Systems in the Centre for Data Sciences and Scalable Computing, at the Queens University of Belfast, UK. He researches underpinning machine learning methods for visual analytics. His principal research focus is face and activity recognition in video. He started his career in the UK Scientific Civil Service with DERA (2000-2002) and QinetiQ (2002-2007). Neil was the 1851 Royal Commission Fellow at Oxford University (2003-2006) in the Robotics Research Group. His autonomous systems, defence and security research is extensive including UK major research programmes and doctoral training centres.