This is a postprint version of the following published document:

M.J. Gómez-Silva, E. Izquierdo, Escalera, A. de la and Armingol, J.M. “Transferring Learning from Multi-person Tracking to Person Re-identification”, *Integrated Computer-Aided Engineering*, vol. 26, nº 4, pp. 329-344, 1 Jan. 2019

DOI: <https://dx.doi.org/10.3233/ICA-190603>

©IOS Press

Transferring Learning from Multi-Person Tracking to Person Re-identification

María José Gómez-Silva^{a,*}, Ebroul Izquierdo^b, Arturo de la Escalera^a and José María Armingol^a

^a*Intelligent Systems Lab (LSI) Research Group, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés, Madrid, Spain.*

^b*Multimedia and Vision (MMV) Research Group, Queen Mary University of London, Mile End Road, Mile End, London, E1 4NS, UK.*

Abstract. Learning to discriminate, whether two person-images correspond to the same person or not, is a daunting challenge when only two images per person are available. This task is called single-shot person re-identification (re-id) and it assumes that each one of the two available images was captured from a different camera view entailing variations in pose, resolution, scale, illumination and background. Addressing this task through supervised training of a deep convolutional neural network is susceptible to model overfitting due to the critical lack of enough labelled data. This paper proposes to exploit the transference of learning previously acquired from a multi-object-tracking (MOT) domain. In this context, a unique deep triplet architecture has been trained on both domains. Six different levels of transfer learning have been implemented and evaluated, proving that the transference of learning from a different domain remarkably increases the re-id performance. Experimental results validate accuracy and robustness of the proposed method as comparable to other state-of-the-art techniques. These results also confirm that, despite the data problem, deep learning is also applicable to the single-shot re-id task.

Keywords: Transfer Learning, Deep Learning, Person Re-Identification, Multi-Object Tracking, Pair-wise Binary Classification

1. Introduction

Person re-identification (re-id) consists of matching people across non-overlapping camera views, at different locations and time. It has become one of the most studied tasks in video surveillance since many other applications, like tracking or behaviour analysis, rely on the person re-id performance [55].

Even though many research efforts from the computer vision field have focused on solving the re-id problem [17], it remains unsolved, [71]. This is because, in unconstrained scenarios, a person's appearance often changes dramatically across camera views due to changes in body pose, view angle, occlusion and illumination conditions. These changes cause intra-class variations, that means quite different representations of the same identity. Moreover, the presence of different people with a similar appearance introduces inter-class ambiguities, i.e. similar representations corresponding to different people. The intra-

class ambiguities tend to distance the representations of a person in the feature space. Conversely, inter-class ambiguities close the features corresponding to different people. Therefore, the clustering of the features according to the identity that they are rendering is a hard task, turning the recognition of the identities into a daunting challenge.

Recently, the use of deep re-id models has been boosted by the success of Convolutional Neural Networks (CNNs) in various vision problems, like image classification [28], objects detection [39], face identification [57], transportation [21, 42, 63], structures analysis [11, 31, 47-49, 64, 68], medical diagnosing [2, 40], or even in computational modelling [43, 61], combinatorial optimization [67], and big data management [53]. Most of the networks used in person re-id treat this task as a pairwise binary classification to discriminate between matched and mismatched pairs of images and they are trained with a supervised learning framework, like the work presented in [32].

*Corresponding author. E-mail: magomezs@ing.uc3m.es.

However, person re-identification poses a unique challenge to deep supervised learning, since a deep model with millions of parameters must be learnt from a small training set. Two types of re-id datasets can be found: multi-shot recognition, [4], where a tracklet of every individual (i.e. small sequence of images) is available for each camera view; and single shot recognition, where only one image per person and per view is used. Figure 1 shows examples of matched pairs (in each column) from the single-shot PRID2011 dataset [23]. Every pair is formed by one image captured from each one of the two different cameras views, a and b. This paper is focused on the single-shot case, where the overfitting of the model is particularly acute due to the small size of the available labelled data.

Fig. 1. Examples of matched pairs belonging to the single shot PRID2011 dataset.

The objective of the presented work is to address the re-id data problem in deep models learning. Given the requirement of a large amount of labelled data to learn a non-overfitted model, able to generalise the solution with unknown samples, and the small size of single-shot re-id datasets, the transference of learning from a different domain has been adopted.

In contrast to the use of other re-id datasets or classification datasets, this paper proposes the use of a Multi-object tracking (MOT) dataset to pre-train a re-id model, which has been subsequently fine-tuned with the re-id target dataset.

With this approach, the re-id network architecture does not require any modification to transfer learning from one task to the other, i.e. from people tracking to re-identification. Instead of that, we propose the adaptation of the MOT dataset to feed a re-id network. Therefore, a new source dataset has been created by extracting people images from MOT sequences.

The use of the proposed cross-sequence dataset provides certain benefits to the re-id model learning. First, the unbalanced nature of the re-id data, where there are only two instances of a certain individual against the

high number of different people, is attenuated by our new source dataset. In the new dataset, many detections of the same person are extracted, making easier the convergence in the model learning process.

In addition, samples from different MOT sequences have been compared in the training. This setting helps to avoid the dependence on the characteristics of a certain camera view, and consequently, to avoid the negative transfer of learning.

Therefore, this paper addresses the single-shot reidentification problem by transferring the similarity degree of appearance learnt from MOT datasets, and its main contributions are: i) the creation of a new source dataset to pre-train a person re-id model from MOT sequences, by designing and using a data generation tool, which has been publicly delivered¹; ii) the implementation and evaluation of different stages of transfer learning, according to different values for the number of layers that are initialized with the pre-trained weights, and for the number of layers that are fine-tuned on the target re-id dataset. This prior analysis provides experienced knowledge about the best configuration for transferring learning from MOT to re-id domain, which is applicable beyond the presented experiments. iii) Finally, the last contribution is the learning of a degree of appearance similarity measurement applicable to both, tracking and re-id tasks through a unique deep network architecture.

The pair-wise binary classification performance of the pre-trained model has been tested on the MOT17 dataset. Furthermore, the effects of the different stages of transfer learning on the final re-id capacity have been evaluated over two of the most challenging re-id datasets: PRID2011 [23], and VIPeR [18], proving the effectiveness of the proposed approach.

The rest of the paper is organized as follows: Section 2 presents the existing related work. The used learning model is introduced in Section 3 and Section 4 describes the different conducted stages of transfer learning. Section 5 evaluates the learning process evolution and presents the experimental results, and some concluding remarks are given in Section 6.

2. Related work

Traditionally, re-identification problem has been faced through two different types of approaches: first, those which enhance the design of distinctive feature

¹ The data generation tool has been implemented as a set of c++ functions, which are publicly available under http://github.com/magomez/dataset_factory/tree/master/data_factory_from_mot

representations, like graph representations [13] and spatial co-occurrence representations [56]. And second, the methods to optimally combine and quantify visual features, such as boosting with AdaBoost [18], Ranking Support Vector Machines (Rank-SVM), [46], Probabilistic Relative Distance Comparison (PRDC), [70], or Metric Learning algorithms, such as Linear Discriminant Analysis (LDA) [9] and Logistic Discriminant Metric Learning (LDML) [19].

These mentioned methods address the learning of a global weighting that reflects the stability of each feature component across two cameras, so they can be grouped under the paradigm of Global Feature Importance (GFI) methods. On contrary, in [33], a Prototype-Sensitive Feature Importance based method is proposed to adaptively weight features according to different clusters of population.

On the other hand, a new approach has been recently adopted. This, which has been inspired by the success of deep Convolutional Neural Networks (CNNs) in image classification [28], is based on the learning of Deep re-id models.

In general, two types of CNN models are commonly employed in computer vision. The first type is the classification model, which is used in image classification [28] and object detection [14]. The second is the Siamese model using image pairs [54] or triplets [51] as input. The first two works in re-id to use deep learning, [30, 65], employed a siamese neural network [3] to determine if a pair of images belong to the same identity. Yi et al. [65] add an additional cost function in the network, while Li et al. [30] use a finer body partitioning. Siamese Networks consist of two deep CNNs branches, sharing parameters and joined in the last layer, where the loss function performs a pairwise verification. Later, Ahmed et al. [1] improved the siamese model by computing the cross-input neighbourhood difference features, which compares the features from one input image to features in neighbouring locations of the other image.

Another variation of the Siamese network is the one known as Triplet model, presented in [51], where face re-identification is addressed. This model has been widely used to automatically find salient high-level representations from raw images, like in the work presented in [6]. The work presented in [16] demonstrated how triplet model helps to alleviate the unbalance nature of the re-id datasets, by allowing multiple triplet combinations of re-id samples.

In the Triplet model, each training input is a set of three samples, two are rendering the same person and the third one, a different identity. Therefore, this model allows the comparison between a matched and

a mismatched pair, from each triplet input, so the objective function can maximize the relative distance between them. For that reason, this has been the learning architecture used in this work to train VGG16 networks [52]. Hence, each branch of the proposed Triplet model is a VGG16 net. Other triplet approaches that chose the VGG16 model to learnt representative features are the presented in [72] and [32] to solve face and person re-identification problems, respectively. This last work also adds LSTM networks, since they adopted a multi-shot re-id approach, in contrast to the single shot re-id that is presented in this work. In [50] a combination of LSTM and VGG networks are used solve people tracking.

However, addressing the re-id problem by means of Deep learning involves an exceptional challenge, because of the lack of training data. Single-shot re-id datasets, such as PRID2011 [23] or VIPeR [18], provide only two images for each identity so currently most CNN-based re-id methods focus on the Siamese and Triplet model, treating the re-id task as a pairwise binary classification.

The requirement of a large amount of labelled data to learn a non-overfitted model, able to generalise the solution with unknown samples, severely limits its scalability in real applications. To overcome this limitation, some works employ unlabelled data from video-surveillance people detection, [8, 59]. Nevertheless, without labelled matching pairs across camera views, existing unsupervised models are not capable of learning what makes a person recognisable under severe appearance changes. Therefore, the matching performance of unsupervised approaches is generally lower than that for supervised learning methods.

Given the inefficiency of unsupervised methods in re-id and the insufficiency of labelled data to apply supervised ones, transferring the models learned from a larger auxiliary dataset becomes necessary. The literature presents transfer learning, also called domain adaptation, [44], as a widely applied tool in deep re-id model learning, where the target task is short of labelled data. The most common deep transfer learning strategy is fine-tuning, [66]: firstly, a base network is trained on a large source data, then the weights corresponding to the first n layers are copied in the target network, and the remaining layers are randomly initialised; and secondly, fine-tuning is performed over these layers or over all layers.

In theory, domain adaptation can be performed by any unsupervised deep learning method, such as, auto-encoder [22] and dictionary learning [27] which can be integrated as the later layers of a CNN network [60]. The method presented in [45] transfers a pre-trained

model to solve re-id on an unlabelled target dataset. In addition, soft-label self-training based deep unsupervised learning has become popular recently [25]. In [12], a novel co-training based unsupervised domain adaptation method is proposed, in [34] a semi-supervised approach is presented, and in [5], both supervised and unsupervised settings are proposed to perform a two-stepped finetuning strategy.

On the other hand, to keep the advantages of supervised models and to cope with the lack of labelled training data, some recent works use auxiliary source datasets from a different domain, taking a multi-task joint training approach. These methods go beyond existing re-id datasets and consider much larger sources, like classification datasets. An example of vast classification dataset is ImageNet dataset [7], which contains millions of images of thousands of object categories. This has been shown useful as source dataset for model pre-training in re-id. In [12], the ImageNet dataset is employed in a two-stepped fine-tuning strategy, to solve a re-id model.

However, transferring knowledge from a classification dataset to a re-id dataset has a number of drawbacks, due to the differences in the tasks for which they were meant, and the dissimilarities in their input data. In [12], several model architecture modifications, and different loss functions are proposed, to combine the learning on two different domains. Indeed, the inputs of a re-id model are person detection images, which have very different aspect ratios and much lower resolutions than classification images. Consequently, the architectures used to train models with them are quite different. That implies the modification and adaptation of the traditional re-id architectures to be able to receive the transferred knowledge from the classification models.

Other methods aim to minimise the discrepancy between the marginal [10, 69] or joint [35] distributions of the source and target datasets, by blurring the domain boundary [10], with a cross-domain loss. However, this method is unsuitable when the source and target domains have completely different tasks, like in [12]. A systematic study is presented in [66] which examines how transferable features of different layers are between different domains. This work concludes that the generalisation ability decreases when the discrepancy between the source and target tasks increases.

For that reason, instead of adapting the re-id model network to a completely different domain dataset, in this work, a new auxiliary source dataset is created from a different domain but keeping the necessary characteristics to feed a re-id model. A multi-object tracking dataset has been chosen, concretely, the

MOT17 dataset from the MOT challenge. As most of the datasets from the people tracking domain, the MOT17 provides the localization of the people detections and their identities at every frame of a bunch of video sequences.

In that way, the resolution and aspect ratio of the created auxiliary dataset is quite similar to those in re-id datasets, so we keep the advantages of using an auxiliary re-id dataset, but with the possibility of getting a larger training dataset. Moreover, a pairwise binary classification can be adopted in both the pre-training step, with the new dataset and the fine-tuning stage, with the re-id target dataset. Hence the same network is used as source and target model.

In order to use a unique architecture, other works use different re-id datasets as source and target datasets. The cross-dataset transfer learning has been recently adopted for re-id with the aim of providing transferable discriminative information from other re-id labelled data to a given target dataset. Among the existing cross-dataset transfer learning works, the literature presents [29], where an SVM multi-kernel learning transfer strategy is adopted, and [37, 58], where multi-task metric learning models are employed. In [62], a single model is learnt across multiple re-id datasets before the fine-tuning in each one of them.

The size of the training data, even when it is formed by a combination of re-id datasets, is relatively small, which tend to overfit a camera-pair-specific model. In consequence, that model cannot be directly transferred to solve re-id in a target re-id dataset containing people who never appeared in the source datasets, which were captured from a new camera pair.

Besides that, the use of existing re-id labelled datasets, each captured from camera pairs with different viewing conditions, to pre-train a model, allows learning the view-invariant features of a person's appearance. However, the differences between the source and target domains, due to the drastically different camera viewing conditions could result in a negative transfer [45].

Therefore, the negative transfer of the pre-learned view-to-view transitions should be avoided. For that reason, contrary to the cross-dataset approach, this work deals with the lack of data by means of creating a new vast dataset from multi-object tracking sequences. In addition, people detections from different sequences are compared in the pre-training of the model. In that way this is fed with input data captured from different cameras views, avoiding the learning, and consequently the negative transfer of specific camera-to-camera transformations.

3. Re-Identification Model

Considering single-shot re-identification task as an isolated module, its objective is to identify the person represented by an image from one view (probe image) among all the images from the other view (gallery images).

This is achieved by calculating the distances between the probe image and all the gallery images. Then, the gallery image presenting the smallest distance is selected as the correct match.

The distances are not computed directly over the raw images but over a representative feature array. Therefore, it is necessary to learn an embedding $F(I)$, from an image I into a feature space, such that the distances between samples corresponding to the same person are small and, between different people are larger.

This feature embedding has been modelled by a deep convolutional neural network, whose architecture is described in Section 3.1. To train the weights of this neural network, a Triplet model has been employed, as it is described in Section 3.2.

The objective of this model is to perform person re-identification, so re-id domain is considered as the target domain. However, learning previously acquired in a different domain, called source or auxiliary domain can be transferred to initialize the weights of the model.

The presented learning model has been used for both, its pre-training on a source dataset, and its fine-tuning on a target dataset, allowing the use of a unique architecture to transfer learning between two different data domains.

3.1. Model architecture

The transformation of every image into its corresponding point in the feature space is performed by a deep convolutional neural network, DCNN. Concretely, the 16-layered network presented as the D version of a set of Very Deep Convolutional Neural networks in [52] and hereafter called VGG16. The selection of this architecture has been inspired by its extensive use in both domains, multi-object tracking [50] and re-identification [32], with demonstrated good results. In addition, this architecture fits perfectly for the proof of concept of the proposed approach, conducted in this work, and its configuration is easily implemented with Caffe libraries [26].

VGG16 presents thirteen convolution layers and three fully connected (FC) layers. Moreover, this network presents a SoftMax layer as the final layer, which

has been removed in this work. Besides that, the input size has been modified in order to adapt it to the re-identification training samples dimensions. Therefore, the input of the proposed DCNN is a person RGB image of 128x64 pixels and its output is a point of the feature space, represented by a 1000-dimensional array. The layers specifications for the proposed VGG16-based embedding are listed in Table 1. The dimension of the inputs and outputs of every layer are determined by the size of the network input, and the successive performed operations (convolutions and pooling), whose parameters were defined in [52]. In addition, all hidden layers are provided of the rectification (ReLU) non-linearity, [28].

Table 1

Structure of the used VGG16-based model. The input and output sizes are described in #rows x #cols x #filters; the kernel, in #rows x #cols x #filters, stride, and #outputs for FC layers

Layer	Input size	Output size	kernel
Conv-1-1	128 x 64 x 3	128 x 64 x 64	3 x 3 x 3
Conv-1-2	128 x 64 x 64	128 x 64 x 64	3 x 3 x 64
Pool-1	128 x 64 x 64	64 x 32 x 64	2 x 2 x 64, 2
Conv-2-1	64 x 32 x 64	64 x 32 x 128	3 x 3 x 64
Conv-2-2	64 x 32 x 128	64 x 32 x 128	3 x 3 x 128
Pool-2	64 x 32 x 128	32 x 16 x 128	2 x 2 x 128, 2
Conv-3-1	32 x 16 x 128	32 x 16 x 256	3 x 3 x 128
Conv-3-2	32 x 16 x 256	32 x 16 x 256	3 x 3 x 256
Conv-3-3	32 x 16 x 256	32 x 16 x 256	3 x 3 x 256
Pool-3	32 x 16 x 256	16 x 8 x 256	2 x 2 x 256, 2
Conv-4-1	16 x 8 x 256	16 x 8 x 512	3 x 3 x 256
Conv-4-2	16 x 8 x 512	16 x 8 x 512	3 x 3 x 512
Conv-4-3	16 x 8 x 512	16 x 8 x 512	3 x 3 x 512
Pool-4	16 x 8 x 512	8 x 4 x 512	2 x 2 x 512, 2
Conv-5-1	8 x 4 x 512	8 x 4 x 512	3 x 3 x 512
Conv-5-2	8 x 4 x 512	8 x 4 x 512	3 x 3 x 512
Conv-5-3	8 x 4 x 512	8 x 4 x 512	3 x 3 x 512
Pool-5	8 x 4 x 512	4 x 2 x 512	2 x 2 x 512, 2
FC-6	4 x 2 x 512	1 x 1 x 4096	4096
FC-7	1 x 1 x 4096	1 x 1 x 4096	4096
FC-8	1 x 1 x 4096	1 x 1 x 1000	1000

3.2. Learning model

To learn the weights of the previously described deep neural network, a Triplet model approach is proposed since it has been demonstrated to alleviate the unbalanced nature of the re-id datasets, [16]. Indeed, the number of instances of a query identity is very limited, specifically when compared with the vast quantity of potentially available different people representations.

According to the triplet model, the network is triplicated in three branches, which are forced to share the same parameters. That means that the learnt weights are identical in the three branches, so only one feature descriptor is learnt, but it is computed for each one of the three different input images, as Figure 2 shows.

Fig. 2. Triplet learning architecture.

The first branch receives an anchor sample, I_i^a , the second one takes an image, I_i^p , rendering the same person than the anchor image, and the input of the third branch is a different person's image, I_i^n . Therefore, every sample, I_i , of the training dataset is constituted by a triplet of person images, $I_i = \langle I_i^a, I_i^p, I_i^n \rangle$. I_i^a and I_i^p present the same identification number and they form a matched or positive pair of samples. The identification number of I_i^n is different, so the pair formed by this image and the anchor image, I_i^a , is a mismatched one, also called negative pair.

The goal is to learn a transformation from the image space to the feature space, such it leads the representations for the same person near in the feature space and far away from different person representations. These constraints are imposed by the triplet loss function.

The feature representation for an image, $F_W(I_i)$ is given by the output of the deep convolutional neural network (VGG16), which maps an input image to the feature space. Such embedding depends on the learnt values for the network weights, which are represented by W .

The similarities between the input images, I_i^a, I_i^p, I_i^n , are measured by computing the squared Euclidean distance between their corresponding features arrays, $F_W(I_i^a), F_W(I_i^p), F_W(I_i^n)$. The triplet objective formulation, presented in [51], establishes a relative distance relationship. It requires that the distance of

the negative pair be larger than that of the positive one by a predefined margin, λ , as it is defined by Eq. (1).

$$\begin{aligned} \|F_W(I_i^a) - F_W(I_i^p)\|^2 + \lambda &< \\ \|F_W(I_i^a) - F_W(I_i^n)\|^2, \forall I_i \in I \end{aligned} \quad (1)$$

This equation ensures that the anchor sample is closer to all samples of the same person, in the feature space, than it is to any sample of other people since I is the set of all possible triplets in the training set.

However, in the one-shot person re-identification challenge, there is only one positive sample for every anchor image. So, the available data is not enough to adopt an individual-meant approach, which clusters the samples of the same person close from each other and distant to another person identity cluster.

On contrary, the proposed method treats all the possible positive pairs as a set rendering the condition of similarity. In the same way, negative pairs represent the dissimilarity situation. The discrimination between similarity and dissimilarity is learnt by means of comparing every positive pair with all the negative pairs generated from the same anchor, taking advantage of the triplet model architecture for that.

The huge amount of possible triplet combinations makes impossible the used of all the data in every iteration, due to the limitations of the computational memory resources. This has dealt to implement a mini-batch-based learning algorithm approach. Therefore, eventually, the loss function to minimise, f_L , is defined by Eq. (2), where B ($B=64$) is the number of elements in a batch of triplets, $I^t = \{I_i\}$.

Batch gradient descent, where the batch is formed by all the training samples, provides more accuracy in the parameters updating than Stochastic gradient descent ($B=1$). Besides, of that, with the vectorized computation performed by the employed deep learning framework (Caffe) the time taken by Batch gradient descent to perform an update has decreased until acceptable values. For that reason, in the performed training processes, the chosen batch size was that as large as the available memory resources make possible.

The value of λ parameter has been experimentally set to 0.4. Not remarkable differences have been found on the performance of the algorithm when the margin value is slightly changed. However, much higher values make the learning objective too demanding. This results in oscillating values of the loss function, and consequently, the model convergence is slower.

$$\begin{aligned} f_L(W, I^t) = \\ \sum_{i=1}^B \left[\max \left(\lambda + \|F_W(I_i^a) - F_W(I_i^p)\|^2, -\|F_W(I_i^a) - F_W(I_i^n)\|^2, 0 \right) \right] \end{aligned} \quad (2)$$

4. Transfer Learning

The re-id performance of the proposed model is measured over some re-id datasets, so these are known as target datasets. Because of the insufficiency of labelled data provided by the target re-id datasets to avoid the overfitting of the trained neural model, transferring the parameters learned on a larger auxiliary dataset becomes necessary.

This auxiliary dataset, also called source dataset, is used to pre-train the neural model. In that way, the learning acquired by the whole model or by some of its layers could be transferred to solve the re-id task with improved performance.

The employed source and target datasets, and the levels of learning transference from one to the other domain are described in the following sections.

4.1. Target and source datasets

The following two datasets have been used as target datasets: PRID2011 (Person Re-ID) dataset² [23] and VIPeR (Viewpoint-Invariant Pedestrian Re-identification) dataset³ [18]. These are two of the most widely used datasets for evaluating re-identification approaches.

Both are composed of two sets (A and B) of person images captured from two static non-overlapping camera views with remarkable differences in camera characteristics, illumination, person's poses, and background. The person images have a resolution of 64x128 pixels. The main difference between these datasets is that in PRID2011 all the images from the set A were captured from the same camera view, and all the samples from set B were acquired from a second camera view different from the first one. However, in VIPeR, the images were captured from many different views, even inside the same set, but maintaining the assumption that the representation of a person in set A was captured from different camera view than that from which its representation in set B was captured. In addition, most of the examples contain a viewpoint change of 90 degrees.

The number of images and different individuals presented by each dataset (in their single-shot version) are exposed below:

- PRID2011: set A contains 385 different images, and set B, 749 images. 200 of the captured individuals are rendered in both sets, and 90 of them

were randomly extracted to be used to generate the training set and 10 for the cross-validation set. The test set has been formed by following the procedure described in [23], i.e., the images of set A for the 100 remaining individuals with representation in both sets have been used as a probe set. The gallery set has been formed by 649 images belonging to set B (all images of set B except the 100 corresponding to the training and cross-validation individuals).

- VIPeR: it presents 632 pedestrians, each one with representation in both sets, A and B. For evaluation, the procedure described in [18] has been followed. The set of 632 pairs has been randomly split into two sets of 316 pairs each. One of them has been divided in a proportion of 90% -10% to create the training and cross-validation sets, respectively; and the other set has been used for testing the algorithm. Therefore, the gallery set is formed by 316 images from set A, and the probe set by their matching images from set B.

For both datasets, those individuals selected as training samples, have been properly combined to form a huge number of training triplets, and in the same way, for the cross-validation set. This has been performed by a data generation tool⁴, which takes the anchor samples from one of the views (A) and the positive and negative samples from the other one (B), as Figure 3 shows through some triplet examples extracted from PRID2011. In order to generate a vast number of triplets, every available positive pair has been grouped with every possible negative sample.

Fig. 3. Triplet samples from PRID2011 dataset.

² Publicly available under: <https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/PRID2011/>

³ Publicly available under: <https://vision.soe.ucsc.edu/node/178>

⁴ The data generation tool has been implemented as a set of c++ functions, which are publicly available under http://github.com/magomez/dataset_factory/tree/master/data_factory_from_reid

Fig. 4. Triplet samples generated from MOT17 dataset.

Before training and testing over the described re-id datasets, the model has been pre-trained using an auxiliary dataset. This auxiliary dataset, also called source dataset, has been created by extracting people detections from the MOT17 dataset⁵. This dataset includes seven varied and labelled real-world surveillance sequences, meant to train and test multi-person tracking algorithms. The MOT17 dataset belongs to the MOT challenge and contains the same set of sequences as MOT16 [38], but with extended ground truth.

A data generation tool⁶ has been employed to extract and combine people to create the training and cross-validation sets of triplets. From each frame, the appearing people have been extracted thanks to their bounding boxes, which are defined in the dataset ground truth, as well as, their identification number. Subsequently, triplets have been generated from the extracted people images, taking as the anchor and the positive images pairs of samples from the same individual. These samples have not why to be consecutive in time, that means that a certain time step is allowed between the frames from which they come. Then, for each pair, the negative sample is added to form a triplet. The negative sample is an image from a different individual and can be taken even from a different sequence of the MOT17 dataset, as Figure 4 shows.

The person images have been resized to the same resolution than re-id datasets samples, 64x128, in

order to be able to use the same neural architecture to train with the source and the target datasets.

4.2. Levels of transfer learning

Fine-tuning [66] has been the adopted transfer learning strategy in this work. This is composed of two training processes: firstly, the neural model (presented in Section 3) is trained on the source dataset, created from tracking sequences. Secondly, the model is fine-tuned over the target dataset. To do that, the pre-trained weights corresponding to the first l_T layers have been transferred, and the remaining weights have been initialized using Xavier method [15]. Then, the learning algorithm is conducted over the target dataset, so that, only the weights of the last l_U layers are updated thorough the second training.

Therefore, l_T is the number of pre-trained layers transferred to the final model and l_U is the number of layers whose weights are trained or updated by the target dataset. Considering a model of depth L , i.e., L is the total number of layers of the model, it can be deduced that: the weights of the first $L-l_U$ layers are learnt using the source dataset in the first training; the weights of the last $L-l_T$ layers are trained in the second training with the target dataset; and the l_t+l_U-L central layers are pre-trained in the first training process and fine-tuned in the second one. These deductions are represented in Figure 5, where the contribution of the

⁵ Publicly available under: <https://motchallenge.net/>

⁶ The data generation tool has been implemented as a set of c++ functions, which are publicly available under http://github.com/mamez/dataset_factory/tree/master/data_factory_from_mot

source dataset to the parameters learning is represented in blue, and yellow is used for the target dataset.

Depending on the values given to l_T and l_U , six different levels of transfer learning have been designed and tested in this work, as Table 2 exposes.

Fig. 5. Contribution of the source (blue) and target (yellow) datasets on the learning of the different layers weights for an example model, and for given values of l_T and l_U .

Table 2

Set-up values for l_T and l_U , for every level of transfer learning

Level	l_T	l_U	$L-l_U$	l_T+l_U-L	$L-l_T$
0	0	16	0	0	16
1	7	16	0	7	9
2	7	9	7	0	9
3	16	16	0	16	0
4	16	9	7	9	0
5	16	0	16	0	0

Level 0. No layer is pre-trained in the source dataset ($l_T=0$). So, the model is totally and uniquely learnt over the target dataset, where it is also tested. In this level, there is no transferred learning.

Level 1. Seven of the layers pre-trained on the source dataset ($0 < l_T < L$) are transferred and fine-tuned on the target dataset. So, in the second training the initial values of the first seven layers comes from the first training.

Level 2. Seven of the layers pre-trained on the source dataset ($0 < l_T < L$) are transferred to the final model, but not fine-tuned on the target dataset. So, in the final model, the final values of the first seven layers come from the first training.

Level 3. All the layers pre-trained on the source dataset ($l_T = L$) are transferred and fine-tuned on the target dataset. So, in the second training, the initial values of the layers come from the first training.

Level 4. All the layers pre-trained on the source dataset ($l_T = L$) are transferred to the final model, but only nine are fine-tuned on the target dataset. In the final model, the final values for the first seven layers come from the first training. The rest of the layers also takes the initial values from the first training but they are fine-tuned in the second training.

Level 5. All the layers are trained on the source dataset ($l_T=L$) and none of them is fine-tuned ($l_U=0$). So, the model is totally trained on the source dataset but tested on the target one. In this level the complete model is transferred from the source domain to the target one.

In general, the intuition behind fine-tuning is that low-level features are learnt from the source dataset. This dataset is formed by person images from MOT sequences, so many of the learnt features can be transferred to solve the task of re-identifying people. Then, by fine-tuning the parameters and learning the final layers of the model on the re-id target dataset, high-level descriptors are learnt. These descriptors automatically acquire the capacities of discriminating people and recognising an identity in images from different views, through their training on the target dataset. That is because, contrary to the source dataset, the re-id target dataset contains samples of images captured from different camera views.

5. Experimental Results

The transfer learning method basically consists of a first stage of pre-training on the source dataset and a second stage of fine-tuning in the target dataset. Instead of only evaluating the re-id performance of the final model, a prior test of the pre-trained model has been also conducted. Therefore, both stages of transfer learning have been evaluated.

The following subsections 5.1, and 5.2 describe the evaluation of the pre-trained and final model respectively. Both present the employed metrics and datasets, as well as the conducted experiments and the obtained results.

5.1. Pre-trained model evaluation

In this work, a pair-wise binary classification model has been pre-trained over samples that were extracted from a MOT dataset. The obtained model is meant to initialize the weights of a re-identification network. Therefore, a prior test of the performance of this model to discriminate between positive and negative pairs of person images is essential to evaluate whether it is suitable to be used as the initializer. For that reason, this model has been evaluated by measuring its performance to discriminate between positive and negative pairs of samples.

A test set has been generated by extracting pairs of samples from the MOT17 sequences [38], as it has

been explained in Section 4.1. For that, different individuals from those appearing in the training and cross-validation sets have been used. In order to provide a fair evaluation, the test set has been formed by 20000 positive pairs and the same number of negatives ones, presenting a completely balanced proportion of samples from each class.

The performance of the pre-trained model to properly classify the pairs samples of the test set has been rendered by a Relative Operating Characteristic (ROC) curve [20], shown by Figure 6. This curve plots the True Positive Rate (TPR), also called Sensitivity or Recall, against the False Positive Rate (FPR), also known as fall-out rate, at various discrimination threshold, th , settings.

The TPR and FPR metrics are defined by Eqs. (3) and (4), respectively, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. The obtained values for these metrics are listed in Table 3.

$$TPR = \frac{TP}{TP+FN} \quad (3)$$

$$FPR = \frac{FP}{FP+TN} \quad (4)$$

Fig. 6. ROC curve of the pre-trained model.

The classification test proves the high capacity of the proposed re-id architecture to learn to discriminate between positive pairs of images, belonging to the same person, and negative ones, corresponding to different people.

The distance between the features of each image of a pair measures the degree of dissimilarity between them. Therefore, the chosen threshold, th , divides the distance space in two ranges of values corresponding to each class. The objective was to learn features that make the distance between two images of the same person lower than a certain threshold, and higher for

different people images, in order to provide a high TPR and low FPR .

The results show that choosing the value 0.7 for th , the model keeps the TPR higher than 90% and the FPR lower than 10%, which are considerably good values for a challenging task as re-identification, even in a MOT sequence. In addition, these results prove the potential applicability of this model not only in re-identification task but also for a visual-based association of identities in a tracking algorithm.

Table 3

Metrics scores resulting from the binary classification performed by the model pretrained on the MOT dataset, for different values of th .

th	TN	FN	TP	FP	TPR	FPR
0	20000	20000	0	0	0	0
0,05	19997	14727	3	5273	0,26365	0,00015
0,1	19990	11599	10	8401	0,42005	0,0005
0,15	19984	9516	16	10484	0,5242	0,0008
0,2	19970	7879	30	12121	0,60605	0,0015
0,25	19936	6542	64	13458	0,6729	0,0032
0,3	19878	5428	122	14572	0,7286	0,0061
0,35	19765	4540	235	15460	0,773	0,01175
0,4	19600	3849	400	16151	0,80755	0,02
0,45	19441	3370	559	16630	0,8315	0,02795
0,5	19264	2968	736	17032	0,8516	0,0368
0,55	19030	2646	970	17354	0,8677	0,0485
0,6	18785	2340	1215	17660	0,883	0,06075
0,65	18499	2082	1501	17918	0,8959	0,07505
0,7	18166	1861	1834	18139	0,90695	0,0917
0,75	17727	1612	2273	18388	0,9194	0,11365
0,8	17093	1387	2907	18613	0,93065	0,14535
0,85	16344	1166	3656	18834	0,9417	0,1828
0,9	15356	863	4644	19137	0,95685	0,2322
0,95	13751	430	6249	19570	0,9785	0,31245
1	0	0	20000	20000	1	1

5.2. Fine-tuned model evaluation

In the second stage of transfer learning, the weights of a certain number of layers of the pre-train model are transferred to the network and then some of its layers weights are fine-tuned. In this work, six different levels of transfer learning (defined in Section 4.2) have

been conducted over two different re-id datasets, PRID2011 and VIPeR dataset. Therefore, twelve different final models have been obtained and evaluated. The procedures exposed in Section 4.1. have been applied to select test data.

The performance of the learnt models has been evaluated by computing their Cumulative Matching Characteristic (CMC) curve [41], which is a standard re-id performance measurement.

To obtain the CMC curve, first, every image from the probe set is coupled with every image from the gallery set and the distance metric (squared Euclidean distance) between them is computed. The matches presenting the lowest values of the distance metric are considered as the top matches since two images belonging to the same person should be rendered close in the feature space and further from different people representations. The distance metrics obtained by comparing a probe image with all the gallery set images are ranked. This process is repeated for each one of the probe images. Then the CMC curve renders the expectation of finding the correct match within the top r matches, for different values of r , called ranks. Therefore, the CMC curve is always increasing since the correct matches found within the top r matches will be also found within the top $r+1$ matches.

The time required to compute the proposed image descriptor using a Nvidia Titan Xp graphic process unit is 11.34 ms. This time has been measured as the average time taken to compute the output of the trained network for 65.000 different samples. The obtained computation time makes possible the use of the proposed descriptor in a real-time surveillance system.

The CMC scores obtained for the models learnt over the PRID2011 dataset by applying different transfer learning levels are listed in Table 4, and represented in Figure 7, in ease of exposition. Analogously, Table 5 and Figure 8 present the CMC scores for the models obtained with VIPeR dataset.

Table 4

CMC scores (in [%]) for different levels of transfer learning, tested over PRID2011 dataset.

	Rank, r					
	1	5	10	20	50	100
Level 0	2	7	10	20	30	52
Level 1	2	8	18	27	46	65
Level 2	2	8	11	17	34	59
Level 3	4	14	20	33	58	77
Level 4	2	12	23	29	47	75
Level 5	2	6	7	14	29	42

Fig. 7. CMC curves for different levels of transfer learning, tested over PRID2011 dataset.

Table 5

CMC scores (in [%]) for different levels of transfer learning, tested over VIPeR dataset.

	Rank, r					
	1	5	10	20	50	100
Level 0	6	16	26	39	69	88
Level 1	10	32	48	64	80	93
Level 2	11	31	48	67	83	95
Level 3	11	35	54	69	90	96
Level 4	10	33	50	70	90	95
Level 5	5	17	25	38	58	75

Fig. 8. CMC curves for different levels of transfer learning, tested over VIPeR dataset.

In general, both graphics show that the extreme levels 0 and 5 present the worst results and the intermediate levels 3 and 4, the best scores.

In level 0, the model has been trained and tested in the target dataset. Because of the small number of samples on the target dataset (especially the number of positive samples), this model suffers from overfitting and it is not able to generalise the learnt solution to the

unknown samples from the test set. Figure 9 shows the learning curve for the level 0 training on PRID2011 and VIPeR datasets.

Fig. 9. Learning curves for level 0 training with PRID2011 (a) and VIPeR dataset (b).

The curves plot the loss value, f_L , thorough the learning iterations, it . Although the proposed algorithm has been trained only using the training set, the loss value, for both the training and the cross-validation sets, has been represented, in blue and orange respectively. The training loss strongly decreases, so the model is converging, but the cross-validation loss increases since the model do not generalise with the unknown samples from the cross-validation set. These differences between both loss values indicate that the model is overfitted.

On contrary, the problem in level 5 is that the model has been trained on an auxiliary dataset, where, despite presenting a huge number of samples, they were not captured from two different camera views. In consequence, the model is not able to learn a feature robust against the intra-class variations.

On the other hand, in level 3 and 4, all the weights pre-trained in the source dataset have been transferred to the model. Then, in level 3 all the layers have been fine-tuned with the target dataset and, in level 4, only the final l_U layers. Hence, firstly the model has been trained to extract low-level features from many samples of people, provided by MOT sequences, and secondly, high-level representations have been learnt by fine-tuning with the target dataset. The target dataset allows the model to learn features able to deal with the intra-class variations and inter-class ambiguities, which is inherent to the re-id task and re-id datasets.

In addition, the model overfitting is reduced by the pre-training in a larger auxiliary dataset, as Figure 10 shows with the reduction of the cross-validation loss.

Fig. 10. Learning curves for the fine-tuning of level 4 with PRID2011 (a) and VIPeR dataset (b).

The intuition under this approach is that the most representative features of a person are automatically learnt on the MOT dataset. From the set of learnt descriptors, the low-level ones, which are learnt in the

earlier layers of the network model, are kept. These low-level features have been trained on a larger and more variate dataset, so they not overfit the re-id training data. Then, the most high-level representations, coded in the further layers, are fine-tuned on a re-id target dataset to make them more discriminative.

Furthermore, prediction accuracy have been also estimated during the training process. This metric is the percentage of well-classified triplets over the total, considering as well-classified those triplets where the negative pair distance is larger than the positive pair distance by a threshold, λ . The accuracy, Ac , for a triplet $I_i = \langle I_i^a, I_i^p, I_i^n \rangle$, is calculated with Eq. (5). Table 6 presents the final average accuracy percentage values measured over the test samples. This value has been obtained for the models learnt with levels 0 and 3 of transfer learning, over PRID2011 and VIPeR dataset. The table shows the accuracy increment produced by the transference of learning.

$$Ac(I_i^t) = \begin{cases} 1 & \text{if } \|F_W(I_i^a) - F_W(I_i^p)\|^2 + \lambda \leq \|F_W(I_i^a) - F_W(I_i^n)\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Table 6

Accuracy metric values comparison.

		Accuracy, Ac (in [%])
PRID2011 dataset	Level 0	80.34
	Level 3	85.93
VIPeR dataset	Level 0	87.84
	Level 3	92.28

The models that have been obtained by transferring the MOT dataset learning clearly outperform the models directly learnt on the re-id dataset. This result can be taken as proof of concept to verify that transfer learning can be exploited to solve the re-id data problem, which is the main objective of this work.

Although the proposed method does not directly address the inter-class variations and intra-class ambiguities problems, it implicitly alleviates them, through the pretraining of the model in a larger and more variate dataset. The higher scores of CMC demonstrate how the proposed transference of learning from MOT domain makes the re-id model capable of successfully managing a larger number of inter-class variation cases and intra-class ambiguities.

Concretely the best performance is achieved when all the network layers are pre-trained in the source dataset and fine-tuned in the target one. This is what we have call level 3 of transfer learning.

For that reason, the models obtained by applying level 3 of transfer learning have been compared with other state-of-the-art methods, and their CMC scores are listed in Tables 7 and 8, for PRID2011 and VIPeR dataset, respectively.

Table 7

Re-id methods comparison on PRID2011 dataset, '-' indicates no result was reported.

	Rank, r				
	1	10	20	50	100
Proposed method	4	20	33	58	77
PSFI+PRDC [33]	3	16	24	39	-
PRDC [70]	3	15	23	38	-
PSFI+RankSVM [33]	4	13	20	32	-
RankSVM [46]	4	13	19	32	-
LDA [9]	4	14	21	35	48
Euclidean [24]	3	10	14	28	45
GFI [36]	4	10	17	32	-
LDML [19]	2	6	11	19	32
PSFI [33]	1	4	7	14	-

Table 8

Re-id methods comparison on VIPeR dataset, '-' indicates no result was reported.

	Rank, r				
	1	10	20	50	100
Proposed method	11	54	69	90	96
PSFI+PRDC [33]	16	51	66	-	-
PRDC [70]	16	51	66	-	-
PSFI+RankSVM [33]	16	51	66	-	-
RankSVM [46]	15	50	65	-	-
LDA [9]	7	25	37	61	79
Euclidean [24]	7	24	34	55	73
GFI [36]	9	27	34	-	-
LDML [19]	6	24	35	54	72
PSFI [33]	10	31	43	-	-

Instead of automatically find high-level features by using deep learning algorithms, the compared methods are mainly based on the computation of hand-crafted low-level features and the learning of a metric distance to compare them.

Linear Discriminant Analysis (LDA), [9], and Logistic Discriminant Metric Learning (LDML), [19], are discriminant methods. They learn a metric distance over a set of previously computed features, on contrary to the proposed approach where the whole re-id model is jointly trained, resulting in a notable improvement.

In [33], a Prototype-Sensitive Feature Importance based method is proposed to adaptively weight features according to different clusters of population. On contrary, [36] present a Global Feature Importance (GFI) approach, addressing the learning of a global weighting, i.e. a vector of generic weights and invariant to the population. Other examples of GFI methods are Ranking Support Vector Machines (Rank-SVM), [46], and Probabilistic Relative Distance Comparison (PRDC), [70]. No population discrimination has been made in this work, and the general weighting of the features to create a global person descriptor have been implicitly and automatically performed by the proposed deep neural model learning. This proposed approach outperforms the previously cited methods and even the combination of some of them (PSFI+PRCD and PSFI+RankSVM).

In [24], the Euclidean distance is directly applied to compare two person representations. This paper also proposes to use the Euclidean distance, with the difference that the compared descriptors have been learnt by a deep re-id neural model, which produces a remarkable improvement in the performance.

In general, the proposed re-id model presents better results than the other compared methods, which do not use deep learning models. Therefore, this paper proves that, as long as the lack of re-id data is faced by a proper strategy like transfer learning, a deep learning approach is applicable to solve the challenging task of single-shot re-id.

6. Conclusions

The main goal of this paper was to solve a problem derived from the lack of data when addressing the single-shot re-identification task through the learning of a deep triplet neural network. With that objective, this paper proposes the transference of learning from the

Multi-Object Tracking (MOT) domain to the Re-Identification (re-id) one.

In order to avoid the overfitting of the proposed neural model, a huge amount of triplet samples have been generated from a set of MOT sequences. Therefore, the MOT domain has been adapted to feed a re-id neural model, and not the other way around.

Different levels of transfer learning have been designed and analysed, and the results have proved that the transfer of learning from a larger and different domain dataset provides a remarkable improvement in the re-identification performance of the proposed deep model.

The method has been evaluated over two of the most challenging and commonly used re-id datasets: PRID20111 [23], and VIPeR [18], proving the effectiveness of the proposed approach, which outperforms some of the most used metric learning methods.

Besides proving the potential application of deep and transfer learning to solve single-shot re-identification, this paper proposed a generalized solution to learn a degree of appearance similarity metric. The conducted evaluation presents this metric as a potential tool for comparing people images not only in the context of re-identification but also in people tracking applications. In people tracking, an identity association process based on visual appearance can be provided by our degree of appearance similarity measurement. This measurement is the result of comparing two images with our modelled deep features, and it can be used as the cost of matching a certain detection with previous captures of a certain tracked identity.

Acknowledgements

This work was supported by the Spanish Government through the CICYT projects (TRA2015-63708-R and TRA2016-78886-C3-1-R), and Ministerio de Educación, Cultura y Deporte para la Formación de Profesorado Universitario (FPU14/02143), Ayudas a la movilidad para estancias breves y traslados temporales (2018) of Programa Estatal de promoción del talento y su empleabilidad, , and Comunidad de Madrid through SEGVAUTO-TRIES (S2013/MIT- 2713). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

References

- [1] E. Ahmed, M. Jones and T. K. Marks, An improved deep learning architecture for person re-identification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3908–3916.
- [2] A. Antoniadou, L. Spyrou, D. Martin-Lopez, A. Valentin, G. Alarcon, S. Sanei, and C.C. Took (2018), “Deep neural architectures for mapping scalp to intracranial EEG,” *International Journal of Neural Systems*, 28:8, (15 template pages).
- [3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, Signature verification using a “siamese” time delay neural network. In *Advances in neural information processing systems*, 1994, 737–744.
- [4] S. Chan-Lang, Q. C. Pham, and C. Achard. Bidirectional sparse representations for multi-shot person re-identification. In *13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016, 263–270.
- [5] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian and T. Xiang. “Deep Transfer Learning for Person Re-Identification.” *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)* (2018): 1-5.
- [6] D. Cheng, Y. Gong, S. Zhou, J. Wang and N. Zheng, N, Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 1335–1344.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In *CVPR*, IEEE, 2009, 248–255.
- [8] M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [9] R. A. Fisher. The use of multiple measurements in taxonomic problems, *Annals of eugenics*, 1936, 7(2): pp. 179–188.
- [10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
- [11] Y. Gao and K.M. Mosalam, (2018), “Deep Transfer Learning for Image-based Structural Damage Recognition,” *Computer-Aided Civil and Infrastructure Engineering*, 33:9, 748–768.
- [12] M. Geng, Y. Wang, T. Xiang, and Y. Tian, Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.
- [13] N. Gheissari, T. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 580–587.
- [15] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.
- [16] M. J. Gómez-Silva, J. M Armingol, and A. de la Escalera. “Balancing People Re-Identification Data for Deep Parts Similarity Learning.” *Journal of Imaging Science and Technology* (2018).
- [17] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The re-identification challenge. *Person Re-identification*, 2014, 1–20.
- [18] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008. n *ECCV*, Springer, 262–275.
- [19] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification, *IEEE 12th international conference on Computer Vision*, 2009, 498–505.
- [20] J. A. Hanley, and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 1982, 143(1), 29–36.
- [21] H. Hashemi and K. Abdelghany, (2018), “End-to-end deep learning methodology for real-time traffic network management,” *Computer-Aided Civil and Infrastructure Engineering*, 33:10, pp. 849–863.
- [22] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006, 313(5786):504–507.
- [23] M. Hirzer, C. Belezni, P.M. Roth, and H. Bischof, Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, Springer, 2011, 91–102.
- [24] M. Hirzer, P. Roth, M. Köstinger and H. Bischof. “Relaxed pairwise learned metric for person reidentification”, *Computer Vision—ECCV*, 2012, pp. 780–793.
- [25] C. Huang, C. C. Loy, and X. Tang. Unsupervised learning of discriminative attributes and visual representations. In *CVPR*, 2016.
- [26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, R., S. Guadarrama and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 675–678). ACM. (2014, November).
- [27] K. Kenneth, M. Joseph, R. Bhaskar, E. Kjersti, L. Te-Won, and S. Terrence. Dictionary learning algorithms for sparse representation. *Neural Computing*, 15(2), 2003
- [28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, 2012, 1097–1105.
- [29] R. Layne, T. Hospedales, and S. Gong. Domain transfer for person re-identification. In *ARTEMIS*, 2013.
- [30] W. Li, R. Zhao, T. Xiao, and X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 152–159.
- [31] Y.Z. Lin, Z.H. Nie, and H.W. Ma, (2017), “Structural Damage Detection with Automatic Feature-extraction through Deep Learning,” *Computer-Aided Civil and Infrastructure Engineering*, 32:12, pp. 1025–1046.
- [32] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*, 2016.
- [33] C. Liu, S. Gong, C. C. Loy, and X. Lin. Evaluating feature importance for re-identification, In *Person Re-identification*, Springe, 2014, 203–228.
- [34] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proc. CVPR*, 2014.
- [35] M. Long, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. *CoRR*, abs/1605.06636, 2016.
- [36] C. C. Loy, T. Xiang, and S. Gong, Time-delayed correlation analysis for multi-camera activity understanding. *International Journal Computer Vision*, 2010, 90 (1), 106–129.
- [37] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 2014, 23(8):3656–3670.
- [38] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.
- [39] M. A. Molina-Cabello, R. M. Luque-Baena, E. López-Rubio, and K. Thurnhofer-Hemsi, “Vehicle Type Detection by

- Ensembles of Convolutional Neural Networks Operating on Super-resolved Images,” *Integrated Computer-Aided Engineering*, 25:4, 2018, pp. 321-333.
- [40] F.C. Morabito, M. Campolo, N. Mammone, M. Versaci, S. Franceschetti, F. Tagliavini, V. Sofia, D. Fatuzzo, A. Gambardella, A. Labate, L. Mumoli, G.G. Tripodi, S. Gasparini, V. Cianci, C. Sueri, E. Ferlazzo, U. Aguglia, “Deep Learning Representation from Electroencephalography of Early-stage Creutzfeld-Jakob Disease and Features for Differentiation from Rapidly Progressive Dementia,” *International Journal of Neural Systems*, 27:2, 2017, 1650039 (15 pages)
- [41] H. Moon and P. J. Phillips. “Computational and performance aspects of pca-based face-recognition algorithms”, *Perception*, 2001, 30(3): 303–321.
- [42] M.A. Nabian and H. Meidani, (2018), “Deep Learning for Accelerated Reliability Analysis of Transportation Networks,” *Computer-Aided Civil and Infrastructure Engineering*, 33:6, pp. 459-480.
- [43] L. Pan, G. Paun, G. Zhang, F. Neri, “Spiking Neural P Systems with Communication on Request”, *International Journal of Neural Systems*, vol. 27, no. 8, 1750042, 2017
- [44] S. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 2010, 22(10):1345–1359.
- [45] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 1306-1315.
- [46] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, volume 2, p. 6, 2010.
- [47] M.H. Rafiei, W.H. Khushefati, R. Demirboga, and H. Adeli, (2017), “Supervised Deep Restricted Boltzmann Machine for Estimation of Concrete Compressive Strength,” *ACI Materials Journal*, 114(2), 237-244.
- [48] M.H. Rafiei, and H. Adeli, (2017), “A Novel Machine Learning Based Algorithm to Detect Damage in Highrise Building Structures,” *The Structural Design of Tall and Special Buildings*, 26:18 (DOI: 10.1002/tal.1400).
- [49] M.H. Rafiei, and H. Adeli, (2018), “A Novel Unsupervised Deep Learning Model For Global and Local Health Condition Assessment Of Structures,” *Engineering Structures*, 156:1, 598-607.
- [50] A. Sadeghian, A. Alahi, and S. Savarese. “Tracking the untrackable: Learning to track multiple cues with long-term dependencies.” *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
- [51] F. Schroff, D. Kalenichenko, and J. Philbin, Facenet: A unified embedding for face recognition and clustering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 815–823.
- [52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [53] J.F Torres, A. Galicia, A. Troncoso, F. Martínez-Álvarez, “A scalable approach based on deep learning for big data time series forecasting,” *Integrated Computer-Aided Engineering*, 25:4, 2018, pp. 335-348.
- [54] R. R. Viorio, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human reidentification. In *ECCV*, Springer, 2016, 791–808.
- [55] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 2013, 46(2):29.
- [56] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE International Conference on Computer Vision*, 2007.
- [57] Wang, P. and Bai, X., “Regional Parallel Structure Based CNN for Thermal Infrared Face Identification,” *Integrated Computer-Aided Engineering*, 25:3, 2018, pp. 247-260.
- [58] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Crossscenario transfer person re-identification. 2015.
- [59] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person reidentification. In *Proc. BMVC*, 2014.
- [60] K. Wang, L. Lin, W. Zuo, S. Gu, and L. Zhang. Dictionary pair classifier driven convolutional neural networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [61] T. Wu, F. Bible, A. Paun, L. Pan, F. Neri, “Simplified and yet Turing universal spiking neural P systems with communication on request”, *International Journal of Neural Systems*, vol. 28, no. 8, 1850013, 2018
- [62] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.
- [63] Y.D. Xue and Y.C. Li, (2018), “A Fast Detection Method via Region-based Fully Convolutional Neural Networks for Shield Tunnel Lining Defects,” *Computer-Aided Civil and Infrastructure Engineering*, 33:8, pp. 638-654.
- [64] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, X. Yang and X. Yang, (2018) “Automatic pixel-level crack detection and measurement using fully convolutional network,” *Computer-Aided Civil and Infrastructure Engineering*, 33:12.
- [65] D. Yi, Z. Lei, S. Liao, S. Z. Li et al., “Deep metric learning for person re-identification.” in *ICPR*, vol. 2014, 2014, 34–39.
- [66] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014, 3320–3328.
- [67] G. Zhang, H. Rong, F. Neri, and M.J. Perez-Jimenez, “An Optimization Spiking Neural P System for Approximately Solving Combinatorial Optimization Problems”, *International Journal of Neural Systems*, Worldscientific, Volume 24, Issue 5:1440006, August 2014
- [68] A. Zhang, K.C.P. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J.Q. Li, and C. Chen, (2017), “Automated Pixel-Level Pavement Crack Detection on 3D Asphalt Surfaces Using a Deep-Learning Network,” *Computer-Aided Civil and Infrastructure Engineering*, 32:10, pp. 805-819.
- [69] X. Zhang, F. X. Yu, S. Chang, and S. Wang. Deep transfer network: Unsupervised domain adaptation. *CoRR*, abs/1503.00591, 2015.
- [70] W. S. Zheng, S. Gong and T. Xiang, Person reidentification by probabilistic relative distance comparison, *IEEE conference on Computer vision and pattern recognition (CVPR)*, 2011, 649–656.
- [71] L. Zheng, Y. Yang, and A. G. Hauptmann. Person reidentification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.
- [72] B. Zhuang, G. Lin, C. Shen, and I. Reid, Fast training of triplet-based deep binary embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5955-5964.