A Quadruplet Loss for Enforcing Semantically Coherent Embeddings in Multi-output Classification Problems

Hugo Proença, Senior Member, IEEE, Ehsan Yaghoubi and Pendar Alirezazadeh

Abstract—This paper describes one objective function for learning semantically coherent feature embeddings in multi-output classification problems, i.e., when the response variables have dimension higher than one. In particular, we consider the problems of identity retrieval and soft biometrics labelling in visual surveillance environments, which have been attracting growing interests. Inspired by the triplet loss [34] function, we propose a generalization that: 1) defines a metric that considers the number of agreeing labels between pairs of elements; and 2) disregards the notion of anchor, replacing \(d(A_1, A_2) < d(A_1, B)\) by \(d(A, B) < d(C, D)\), \(\forall A, B, C, D\) distance constraints, according to the number of agreeing labels between pairs. As the triplet loss formulation, our proposal also privileges small distances between positive pairs, but at the same time explicitly enforces that the distance between other pairs corresponds directly to their similarity in terms of agreeing labels. This yields feature embeddings with a strong correspondence between the classes centroids and their semantic descriptions, i.e., where elements are closer to others that share some of their labels than to elements with fully disjoint labels membership. As practical effect, the proposed loss can be seen as particularly suitable for performing joint coarse (soft label) + fine (ID) inference, based on simple rules as k-neighbours, which is a novelty with respect to previous related loss functions. Also, in opposition to its triplet counterpart, the proposed loss is agnostic with regard to any demanding criteria for mining learning instances (such as the semi-hard pairs). Our experiments were carried out in five different datasets (BIODI, LFW, IJB-A, Megface and PETA) and validate our assumptions, showing highly promising results.

Index Terms—Feature embedding, Soft biometrics, Identity retrieval, Convolutional neural networks, Triplet loss.

I. INTRODUCTION

Characterizing pedestrians in crowds has been attracting growing attention, with soft labels as gender, ethnicity or age being particularly important to determine the identities in a scene. This kind of labels is closely related to human perception and describes the visual appearance of subjects, with applications in identity retrieval [40][36] and person re-identification [15][27] problems.

Deep learning frameworks have been repeatedly improving the state-of-the-art in many computer vision tasks, such as object detection and classification [25][41], action recognition [19][6], semantic segmentation [24][44] and soft biometrics labelling [32]. In this context, the triplet loss [34] is an extremely popular concept, where three learning elements are considered at a time, two of them of the same class and a third one belonging to a different class. By imposing larger distances between the elements of the negative than of the positive pair, the intra-class compactness and inter-class discrepancy on the destiny space are enforced. This learning paradigm was successfully applied to various problems, upon the mining of the semi-hard negative input pairs, i.e., cases where the negative element is not closer to the anchor than the positive, but still provides a positive loss due to an imposed margin.

Based on the concept of triplet loss, this paper describes one objective function that can be regarded as a generalization of its predecessor. Instead of the binary division of the learning pairs into positive/negative, we define a metric that analyzes the similarity between any two classes (identities). In learning time, four elements of arbitrary classes are considered at a time and the soft margins between the pairwise distances yield from the number of common labels in each pair (Fig. 1). Under this formulation, different identities that are semantically close to each other (e.g., two “young, black, bald, male” subjects) are projected into adjacent regions of the destiny space. Also, as this objective function imposes different margins between
two negative pairs, it leverages the difficulties in mining appropriate learning instances, which is one on the main difficulties in the triplet loss formulation.

The proposed loss function is particularly suitable for coarse-to-fine classification problems, where some of the labels are easier to infer than others and the global problem can be decomposed into more tractable sub-components. This hierarchical paradigm is an efficient way of organizing object recognition, not only to accommodate a large number of hypotheses, but also to systematically exploit any shared attributes. The identity retrieval problem is of particular interest, with the finest labels (IDs) being seen leaves of hierarchical structures with roots such as the “gender” or “ethnicity” features.

The remainder of this paper is organized as follows: Section II summarizes the most relevant research in the scope of our work. Section III provides a detailed description of the proposed objective function. In Section IV we discuss the obtained results and the conclusions are given in Section V.

II. RELATED WORK

Deep learning methods for biometrics can be roughly divided into two major groups: 1) methods that directly learn multi-class classifiers used in identity retrieval and soft biometrics inference; and 2) methods that learn low-dimensional feature embeddings, where identification yields from nearest neighbour search.

A. Soft Biometrics and Identity Retrieval

Bekele et al. [2] proposed a residual network for multi-output inference that handles classes-imbalance directly in the cost function, without depending of data augmentation techniques. Almudhakha et al. [1] explored the concept of comparative soft biometrics and assessed the impact of automatic estimations on face retrieval performance. Guo et al. [12] studied the influence of distance in the effectiveness of body and facial soft biometrics, introducing a joint density distribution based rank-score fusion strategy [13]. Vera-Rodriguez et al. [31] used hand-crafted features extracted from the distances between key points in body silhouettes. Martinho-Corbishley et al. [29] introduced the idea of super-fine soft attributes, describing multiple concepts of one trait as multi-dimensional perceptual coordinates. Also, using joint attribute regression and a deep residual CNN, they observed substantially better ranked retrieval performance in comparison to conventional labels. Schumann and Specker used an ensemble of classifiers for robust attribute inference [35], extended to full body search by combining it with a human silhouette detector. He et al. [17] proposed a weighted multi-task CNN with a loss term that dynamically updates the weight for each task during the learning phase.

Several works relied regarded semantic segmentation as a tool to support labels inference: Gallyawala et al. [10] described a deep learning framework for person retrieval using the height, clothes’ color, and gender labels, with a semantic segmentation module used to remove clutter. Similarly, Ciparigan and Nixon [3] obtained semantically segmented regions of the body, that subsequently fed two CNN-based feature extraction and inference modules.

Finally, specifically designed for handheld devices, Samangouei and Chellappa [32] extracted facial soft biometric information from mobile phones, while Neal and Woodard [26] developed a human retrieval scheme based on thirteen demographic and behavioural attributes from mobile phones data, such as calling, SMS and application data, having authors positively concluded about the feasibility of this kind of recognition.

A comprehensive summary of the most relevant research in soft biometrics is given in [38].

B. Feature Embeddings and Loss Functions

Triplet loss functions were motivated by the concept of contrastive loss [14], where the rationale is to penalize distances between positive pairs, while favouring distances between negative pairs. Kang et al. [21] used a deep ensemble of multi-scale CNNs, each one based on triplet loss functions. Song et al. [37] learned semantic feature embeddings that lift the vector of pairwise distances within the batch to the matrix of pairwise distances, and described a structured loss on the lifted problem. Liu and Huan [28] proposed a triplet loss learning architecture composed of four CNNs, each one learning features from different body parts that are fused at the score level.

A posterior concept was the center loss [42], which finds a center for elements of each class and penalizes the distances between the projections and their corresponding class centers. Jian et al. [20] combined additive margin softmax with center loss to increase inter-class distances and avoid overconfidence on classifications. Ranjan et al. [30] described the crystal loss, that restricts the features to lie on a hypersphere of a fixed radius, by adding a constraint on the features projections such that their ℓ2-norm remains constant. Chen et al. [4] used deep representations to feed a joint Bayesian metrics learning module that maximizes the log-likelihood ratio between intra- and inter-classes distances. Based on the concept of Sphereface, Deng et al. [8] proposed an additive angular margin loss, with a clear geometric interpretation due to the correspondence to the geodesic distance on the hypersphere.

Observing that CNN-based methods tend to overfit in person re-identification tasks, Shi et al. [36] used siamese architectures to provide a joint description to a metric learning module, regularizing the learning process and improving the generalization ability. Also, to cope with large intra-class variations, they suggested the idea of moderate positive mining, again to prevent overfitting. Motivated by the difficulties in generate learning instances for triplet loss frameworks, Su et al. [39] performed adaptive CNN fine-tuning, along with an adaptive loss function that relates the maximum distance among positive pairs to the margin demanded for separate the positive from the negative pairs. Hu et al. [18] proposed an objective function that generalizes the Maximum Mean Discrepancy [33] metric, with a weighting scheme that favours...
good quality data. Duan et al. [9] proposed the uniform loss to learn deep equi-distributed representations for face recognition. Finally, observing the typical unbalance between positive and negative pairs, Wang et al. [41] described an adaptive margin list-wise loss, in which learning data are provided with a set of negative pairs divided into three classes (easy, moderate, and hard), depending of the distance rank with respect to the query.

Finally, we note the differences between the work proposed in this paper and the also (quadruplet) loss described by Chen et al. [5]. These authors attempt to augment the inter-class margins and the intra-class compactness without explicitly using any semantical constraint, in the sense that – as in the original triplet loss formulation – there is nothing that explicitly enforces to project similar classes (i.e., different identities that share most of the remaining labels) to neighbour regions of the latent space. In opposition, our method concerns essentially about this semantical coherence of assuring the projection similar classes into adjacent regions and not in obtain larger margins between the different classes. Also, even the idea behind the loss formulation is radically different in both methods, in the sense that [5] still considers the concept of anchor (as the original triplet-loss formulation), which is - again - in opposition to our proposal.

### III. PROPOSED METHOD

#### A. Quadruplet Loss: Definition

Consider a supervised classification problem, where \( t \) is the dimensionality of the response variable \( y_i \), associated to the input element \( x_i \in [0,255]^n \). Let \( f(.) \) be one embedding function that maps \( x_i \) into a d-dimensional space \( \Psi \), with \( f_i = f(x_i) \in \Psi \) being the projected vector. Let \( \{x_1,\ldots,x_b\} \) be a batch of \( b \) images from the learning set. We define \( \phi(y_i,y_j) \in \mathbb{N}, \forall i,j \in \{1,\ldots,b\} \) as the function that measures the semantic similarity between \( x_i \) and \( x_j \):

\[
\phi(y_i,y_j) = ||y_i - y_j||_0,
\]

with \( ||.||_0 \) being the \( \ell_0\)-norm operator.

In practice, \( \phi(.) \) counts the number of disagreeing labels between the \( \{x_i,x_j\} \) pair, i.e., \( \phi(y_i,y_j) = t \) when the \( i^{th} \) and \( j^{th} \) elements have fully disjoint classes membership (e.g., one "black, adult, male" and another "white, young, female" subjects), while \( \phi(y_1,y_2) = 0 \) when they have the exact same label (class) across all dimensions, i.e., when they constitute a positive pair.

Let \( \{i,j,p,q\} \) be the indices of four images in the batch. The corresponding quadruplet loss value \( \ell_{i,j,p,q} \) is given by:

\[
\ell_{i,j,p,q} = \text{sgn}(\phi(y_i,y_j) - \phi(y_p,y_q)) \left( ||f_p - f_q||^2_2^1 - ||f_i - f_j||^2_2 + \alpha \right),
\]

where \( \text{sgn}(\cdot) \) is the sign function and \( \alpha \) is the desired margin (\( \alpha = 0.1 \) was used in all our experiments). Evidently, \( \ell \) will be zero when both image pairs have the same number of agreeing labels (as \( \text{sgn}(0) = 0 \) in these cases). In all other cases, the sign function will determine the pairs for which the distances in the embedding should be minimized, i.e., if the elements of the \( \{p,q\} \) pair are semantically closer to each other than the elements of the \( \{i,j\} \) pair (i.e., \( \phi(y_p,y_q) < \phi(y_i,y_j) \)), we want to ensure that \( ||f_p - f_q||^2_2 < ||f_i - f_j||^2_2 \).

The accumulated loss in the batch is given by the truncated mean of a sample (of size \( s \)) randomly taken from the subset of the \( \binom{b}{4} \) individual loss values where \( \phi(y_i,y_j) \neq \phi(y_p,y_q) \):

\[
L = \frac{1}{s} \sum_{z=1}^{s} \ell_{i,j,p,q},
\]

where \( z \in \{1,\ldots,s\} \) denotes the \( z^{th} \) composition of four elements in the batch and \( [.]_+ \) is the \( \max(.,0) \) function. Even considering that a large fraction of the combinations in the batch will be invalid (i.e., with \( \phi() = 0 \)), large values of \( b \) will result in an intractable number of combinations at each iteration. In practical terms, after filtering out those invalid combinations, we randomly sample a subset of the remaining instances, which is designated as the mini-batch.

#### B. Quadruplet Loss: Inference

Consider four indices \( \{i,j,p,q\} \) of elements in the minibatch, with \( \phi(y_i,y_j) > \phi(y_p,y_q) \). Let \( \Delta_\phi \) denote the difference between the number of disagreeing labels of the \( \{i,j\} \) and \( \{p,q\} \) pairs:

\[
\Delta_\phi = \phi(y_i,y_j) - \phi(y_p,y_q).
\]

Also, let \( \Delta_f \) be the distance between the elements of the most alike pair minus the distance between the elements of the least alike pair in the destiny space (plus the margin):

\[
\Delta_f = ||f_p - f_q||^2_2 - ||f_i - f_j||^2_2 + \alpha.
\]

Upon basic algebraic manipulation, the gradients of \( L \) with respect to the quadruplet terms are given by:

\[
\frac{\partial L}{\partial f_i} = \sum_z \left\{ \begin{array}{ll} 2(f_j - f_i) & \text{if } \Delta_\phi > 0 \land \Delta_f \geq 0 \\ 0 & \text{otherwise} \end{array} \right.,
\]

\[
\frac{\partial L}{\partial f_j} = \sum_z \left\{ \begin{array}{ll} 2(f_i - f_j) & \text{if } \Delta_\phi > 0 \land \Delta_f \geq 0 \\ 0 & \text{otherwise} \end{array} \right.,
\]

\[
\frac{\partial L}{\partial f_p} = \sum_z \left\{ \begin{array}{ll} 2(f_p - f_q) & \text{if } \Delta_\phi > 0 \land \Delta_f \geq 0 \\ 0 & \text{otherwise} \end{array} \right.,
\]

\[
\frac{\partial L}{\partial f_q} = \sum_z \left\{ \begin{array}{ll} 2(f_q - f_p) & \text{if } \Delta_\phi > 0 \land \Delta_f \geq 0 \\ 0 & \text{otherwise} \end{array} \right..
\]

In practice terms, the model weights will be adjusted only for learning instances where pairs have a different number of agreeing labels (i.e., \( \Delta_\phi > 0 \)) and when the distance in the destiny space between the elements of the most similar pair is higher or equal than the distance between the elements of the
Fig. 2. Illustration of the key difference between the triplet loss [34] and the solution proposed in this paper. Using a loss function that analyzes the semantic similarity (in terms of soft biometrics) between the different identities, we enforce embeddings (Ψ₃) that are semantically coherent, i.e., where: 1) elements of the same class appear near each other; and also 2) elements of relatively similar classes appear closer to each other than elements with no labels in common. This is in opposition to the original formulation of the triplet loss, that relies exclusively in the elements appearance to define the geometry of the destiny space, which might result - in case of noisy image features - in semantically incoherent embeddings (e.g., in Ψ₁ and Ψ₂, classes are compact and discriminative, but the x/z centroids are too close to each other).

least similar pair (plus the margin, Δf ≥ 0). According to this idea, using (6)-(9), the deep learning frameworks supervised by the proposed quadruplet loss are trainable in a way similar to its counterpart triplet loss and can be optimized by the standard Stochastic Gradient Descend (SGD) algorithm, which was done in all our experiments.

For clarity purposes, Algorithm 1 gives a pseudocode description of the learning and batch/mini-batch creation processes during the inference step for the proposed loss function.

Algorithm 1 Pseudocode description of the inference and batch/min-batch creation processes.

Precondition: M: CNN, tₑ: Tot. epochs, s: mini-batch size, b: batch size, I: Learning set, n images
for 1 to tₑ do
  for 1 to [n/b] do
    b ← randomly sample b out of n images from I
    c ← create \binom{n}{4} quadruplet combinations from b
    c* ← filter out invalid elements from c
    s ← randomly sample s elements from c*
    M ← update weights(M, s) (eqs. (6-9))
  end for
end for
return M

**C. Quadruplet Loss: Insight and Example**

Fig. 2 illustrates our rationale in the proposed quadruplet loss. By defining a metric that analyses the similarity between two classes, we abandon the binary division of pairs into positive/negative families, and perceive which classes are more/less distinct with respect to others. This enables to explicitly enforce that the most negative pairs (e.g., with no common labels) are at the farthest possible distance from each other in the embedding. During the learning phase, we sample image pairs in a stochastic way and enforce that input elements are projected in a way that faithfully resembles the human perception of semantic similarity.

As an example, Fig. 3 compares the bidimensional embeddings resulting from the triplet and from the quadruplet losses, in the subset of the LFW identities with more than 15 images in the dataset (using \(t = 2\) : \{“ID”, “Gender”\} labels). This plot yielded from the projection of a 128-dimensional embedding down to two dimensions, according to the Neighbourhood Component Analysis (NCA) [11] algorithm.

It can be seen that the triplet loss provided an embedding where the positions of elements are exclusively determined by their appearance, with pairs of IDS with fully disjoint labels appearing close to each other (this is evident in the upper left corner of the triplet embedding, where “female” elements appear adjacent to “male tennis players”). In opposition, using the quadruplet loss we obtain a large separation between the elements of different genders, while keeping the compactness per identity. This kind of embedding will be interesting in at least two cases: 1) in identity retrieval, to guarantee that all retrieved elements have the same soft labels of the query; and 2) upon a semantic description of the query (e.g., “find adult white males similar to this image”), to guarantee that all retrieved elements meet the semantic criteria. A third application comprises the direct inference of all the fine + coarse (soft) labels in a simple k-neighbours fashion.

**IV. RESULTS AND DISCUSSION**

**A. Experimental Setting and Preprocessing**

Our empirical validation was conducted in one proprietary (BIODI) and four freely available datasets (LFW, PETA, IJB-A and Megaface) well known in the biometrics and re-identification literature.

The BIODI\(^1\) dataset is proprietary of Tomiworld\(^2\), being composed of 849,932 images from 13,876 subjects, taken

\(^{1}\)http://di.ubi.pt/~hugomcp/BIODI/

\(^{2}\)https://tomiworld.com/
from 216 indoor/outdoor video surveillance sequences. All images were manually annotated for 14 labels: gender, age, height, body volume, ethnicity, hair color and style, beard, moustache, glasses and clothing (x4). The Labeled Faces in the Wild (LFW) [16] dataset contains 13,233 images from 5,749 identities, collected from the web, with large variations in pose, expression and lighting conditions. PETA [7] is a combination of 10 pedestrian re-identification datasets, composed of 19,000 images from 8,705 subjects, each one annotated with 61 binary and 4 multi-output attributes. The IIJB-A [23] dataset contains 5,397 images plus 20,412 video frames from 500 individuals, with large variations in pose and illumination. Finally, the Megaface [22] set was released to evaluate face recognition performance at the million scale, and consists of a gallery set and a probe set. The gallery set is a subset of Flickr photos from Yahoo (more than 1,000,000 images from 690,000 subjects). The probe dataset includes FaceScrub and FGNet sets. FaceScrub has 100,000 images from 530 individuals and FGNet contains 1,002 images of 82 identities.

### B. Convolutional Neural Networks

Two CNN architectures were considered, based in the VGG and ResNet models (Fig. 4). Here, our goals were essentially to compare the performance of the quadruplet loss with respect to the baselines and also to perceive any variations in performance with respect to different learning architectures. A TensorFlow implementation of the of both architectures is

![Triplet Loss](image1.png)

![Quadruplet Loss](image2.png)

Fig. 3. Comparison between the embeddings resulting from the triplet loss [34] (top plot), and from the proposed quadruplet loss (bottom plot). Results are given for \( t = 2 \) features \{“ID”, “Gender”\} for the LFW identities with at least 15 images (89 elements).

Fig. 4. Architectures of the CNNs used in the empirical validation of our method. The yellow boxes represent convolutional layers, and the blue and green boxes represent pooling and dropout (keeping probability 0.75) layers. Finally, the red boxes denote fully connected layers. In the ResNet architecture, the dashed skip connections represent convolutions with stride 2 \( \times 2 \), yielding outputs with half of the spatial input size. The “\( \times 2 \)” symbol in the convolution layers denotes stride 2 \( \times 2 \) (the remaining layers use stride 1 \( \times 1 \)).
available at\(^3\).

All the models were initialized with random weights, from zero-mean Gaussian distributions with standard deviation 0.01 and bias 0.5. Images were resized to 256 × 256, adding lateral white bands when needed to keep constant ratios. A batch size of 64 was defined, which obviously results in too many combinations of pairs for the triplet/quadruplet losses. At each iteration, we filtered out the invalid triplets/quadruplets instances and randomly selected the mini-batch learning elements, composed of 64 instances in all cases. For all other baselines, 64 learning instances were also used as a batch. The learning rate started from 0.01, with momentum 0.9 and weight decay \(5e^{-4}\). In the learning-from-scratch paradigm, we stopped the learning process when the validation loss didn’t decrease for 10 iterations (i.e., patience=10).

Initially, we varied the dimensionality of the embedding (\(d\)) to perceive the sensitivity of the proposed loss with respect to this parameter. Considering the LFW set, the observed average AUC values with respect to \(d\) are provided in Fig. 6 (the shadowed regions denote the ± standard deviation performance, after 10 trials). As expected, higher values are directly correlated to the levels of performance, even though results stabilised for dimensions larger than 128. In this regard, we assumed that using even larger dimensions would require much more training data, having resorted from this moment to dimension \(d=128\) in all the subsequent experiments.

Interestingly, the absolute performance observed for very low \(d\) values was not too far of the obtained for much larger di-

\(^3\)https://github.com/hugomcp/quadruplets

\(^4\)http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/

\[ mAP = \frac{\sum_{q=1}^{n} \hat{P}(q)}{n}, \]   

\[ \text{where } n \text{ is the number of queries}, \hat{P}(q) = \sum_{k=1}^{n} P(k)\Delta r(k), P(k) \text{ is the precision at cut-off } k \text{ and } \Delta r(k) \text{ is the change in recall from } k-1 \text{ to } k. \]

For the LFW set experiment, the BLUFR\(^4\) evaluation protocol was chosen. In the verification (1:1) setting, the test set contained 9,708 face images of 4,249 subjects, which yielded over 47 million matching scores. For the open-set identification problem, the genuine probe set contained 4,357 face images of 3,249 subjects, and the impostor probe set had 4,350 images of 3,249 subjects, and the gallery set had 1,000 images. This evaluation protocol was the basis to design, for the other sets, as close as possible experiments, in terms of the number of matching scores, gallery and probe sets.

C. Identity Retrieval

We report the performance obtained by the quadruplet loss in the identity retrieval task, when compared to the baselines triplet loss, center loss, softmax and also to Chen et al.\(^5\)’s method. Here, we considered the LFW, Megface and IJB-A sets, and three labels: {"ID", "Gender", "Ethnicity"} (\(t = 3\)), with the annotations for the IJB-A set provided by the Face++ algorithm and subjected to human validation. In this setting, it should be noted that all the baselines classify pairs into positive/negative depending exclusively of the ID, while the proposed loss uses all the labels to determine the relative similarity between any two classes.

Considering the verification and identification scenarios, we provide the Receiver Operating Characteristic curves (ROC, Fig. 7), the Cumulative Match curves (CMC, Fig. 8) and the Detection and Identification rates at rank-1 (DIR, Fig. 9). The results are summarized in Table I, reporting the rank-1, top-10% values and the mean average precision (mAP) scores:

\[ \text{mAP} = \frac{\sum_{q=1}^{n} \hat{P}(q)}{n}, \]   

\[ \text{where } n \text{ is the number of queries}, \hat{P}(q) = \sum_{k=1}^{n} P(k)\Delta r(k), P(k) \text{ is the precision at cut-off } k \text{ and } \Delta r(k) \text{ is the change in recall from } k-1 \text{ to } k. \]
Generally, we observed that the proposed quadruplet loss slightly outperforms the other loss functions, which might also be the result of using additional information for learning. These improvements in performance were observed in most cases by a consistent margin for both the verification and identification tasks, both for the VGG and ResNet architectures.

In terms of the errors per CNN architecture, the ResNet-like error rates were roughly $0.9 \times (90\%)$ of the observed for the VGG-like networks (higher margins were observed for the softmax loss). Not surprisingly, the Chen et al. [5]' method outperformed the remaining competitors, followed by the triplet loss function, which is consistent with most of the results reported in the literature. The softmax loss got repeatedly the worst performance among the five functions considered.

Regarding the levels of performance per dataset, the values observed for the Megaface set were far worse for all objective functions than the corresponding values for the LFW and IJB-A sets. In the former dataset, we followed the protocol of the small training set, using 490,000 images from 17,189 subjects for learning (images overlapping with Facescrub dataset were discarded). Also, it is important to note that the relative performance between the loss functions was roughly the same in all sets. Degradations in performance were slight from the LFW to the IJB-A set and much more visible in case of the Megaface set. In this context, the softmax loss produced the most evident degradations, followed by the center loss.

D. Soft Biometrics Inference

As stated above, the proposed loss can also be used for learning a soft biometrical estimator. Then, in test time the position to where one element is projected to can be used to infer the soft labels, in a simple nearest neighbour fashion. In these experiments, we considered only 1-NN, i.e., the label inferred for each query was given by the closest gallery element. Better results would be possibly attained if a larger number of neighbours had been considered, even though the computational cost of classification will also increase. All experiments were conducted according to a bootstrapping-like strategy: having $n$ test images available, the bootstrap randomly selected (with replacement) $0.9 \times n$ images, obtaining samples composed of 90% of the whole data. Ten test samples were created and the experiments were conducted independently on each trial, which enabled to obtain the mean and the standard deviation at each performance value.

As baselines, we used two off-the-shelf techniques that represent the state-of-the-art: the Matlab SDK for Face++ [5] and the Microsoft Cognitive Toolkit Commercial [6]. Face++ is a commercial face recognition system, with good performance reported for the LFW face recognition competition (second best rate). Microsoft Cognitive Toolkit is a deep learning framework that provides useful information based on vision, speech and language.

5 http://www.faceplusplus.com/
6 https://www.microsoft.com/cognitive-services/
We considered exclusively the "Gender", "Ethnicity" and "Age" labels ($t = 3$), quantised respectively into two classes for Gender ("male", "female"), three classes for Age ("young", "adult", "senior"), and three classes for Ethnicity ("white", "black", "asian"). The average and standard deviation performance values are reported in Table II for the BIODI, PETA and LFW sets.

The results achieved by the quadruplet loss can be favourably compared to the COTS techniques for most labels, particularly for the BIODI and LFW datasets. Regarding the PETA set, Face++ invariably outperformed the other techniques, even if at a reduced margin in most cases. This was justified by the extreme heterogeneity of image features in this set, in result of being the concatenation of different databases. This probably had reduced the representativity of the learning data with respect the test set, being the Face++ model apparently the least sensitive to this covariate. Note that the "Ethnicity" label is only provided by the Face++ framework.

Globally, these experiments supported the possibility of using such the proposed method to estimate soft labels in a single-shot paradigm, which can be particularly interesting to reduce the computational cost of using specialized soft labelling tools.

Finally, we analysed the variations in performance with respect to the number of labels considered, i.e., the value of the $t$ parameter. At first, to perceive how the identity retrieval performance depends of the number of soft labels, we used the annotations provided by the ATVS group [38] for the LFW set, and measured the rank-1 variations for $1 \leq t \leq 4$, starting by the "ID" label alone and then adding iteratively the "Gender" $\rightarrow$ "Ethnicity" $\rightarrow$ "Age" labels. The results are shown in the left part of Fig. 10. In a complementar way, to perceive the
| Method       | mAP      | rank-1   | top-10%  |
|-------------|----------|----------|----------|
| VGG         |          |          |          |
| Quadruplet loss | 0.958 ± 3e−3 | 0.951 ± 0.020 | 0.979 ± 6e−3 |
| Softmax loss | 0.877 ± 0.011 | 0.812 ± 0.053 | 0.960 ± 9e−3 |
| Triplet loss [34] | 0.897 ± 4e−3 | 0.842 ± 0.034 | 0.953 ± 0.011 |
| Center loss [43] | 0.727 ± 0.014 | 0.615 ± 0.060 | 0.863 ± 0.017 |
| Triplet loss [34] | 0.834 ± 5e−3 | 0.939 ± 0.037 | 0.958 ± 6e−3 |
| Softmax loss | 0.849 ± 0.010 | 0.823 ± 0.039 | 0.941 ± 0.014 |
| Triplet loss [34] | 0.934 ± 4e−3 | 0.929 ± 0.033 | 0.964 ± 8e−3 |
| Softmax loss | 0.854 ± 9e−3 | 0.758 ± 0.059 | 0.946 ± 0.017 |
| Center loss [43] | 0.917 ± 5e−3 | 0.901 ± 0.040 | 0.950 ± 0.011 |
| Chen et al. [5] | 0.961 ± 2e−3 | 0.945 ± 0.022 | 0.976 ± 6e−3 |
| ResNet       |          |          |          |
| Quadruplet loss | 0.968 ± 2e−3 | 0.966 ± 0.012 | 0.981 ± 4e−3 |
| Softmax loss | 0.902 ± 9e−3 | 0.906 ± 0.048 | 0.972 ± 8e−3 |
| Triplet loss [34] | 0.912 ± 4e−3 | 0.861 ± 0.029 | 0.960 ± 8e−3 |
| Center loss [43] | 0.730 ± 0.010 | 0.745 ± 0.051 | 0.899 ± 0.011 |
| Triplet loss [34] | 0.841 ± 9e−3 | 0.860 ± 0.030 | 0.938 ± 8e−3 |
| Softmax loss | 0.947 ± 4e−3 | 0.948 ± 0.026 | 0.968 ± 9e−3 |
| Triplet loss [34] | 0.872 ± 8e−3 | 0.839 ± 0.052 | 0.957 ± 9e−3 |
| Center loss [43] | 0.919 ± 5e−3 | 0.937 ± 0.031 | 0.961 ± 0.011 |
| Triplet loss [34] | 0.939 ± 3e−3 | 0.898 ± 0.016 | 0.967 ± 9e−3 |
| Center loss [43] | 0.847 ± 9e−3 | 0.845 ± 0.048 | 0.945 ± 9e−3 |
| Triplet loss [34] | 0.877 ± 7e−3 | 0.893 ± 0.035 | 0.963 ± 9e−3 |
| Chen et al. [5] | 0.968 ± 2e−3 | 0.959 ± 0.015 | 0.983 ± 4e−3 |
| Quadruplet loss | 0.916 ± 8e−2 | 0.880 ± 0.050 | 0.975 ± 8e−3 |
| Softmax loss | 0.952 ± 4e−3 | 0.960 ± 0.022 | 0.986 ± 6e−3 |

It is interesting to observe the apparently contradictory results in both plots: at first, a positive correlation between the labelling errors and the values of \( t \) is evident, which was justified by the difficulty of inferring some of the hardest labels in the BIODI set (e.g., the type of shoes). However, the average rank-1 identification accuracy also increased when more soft labels were used, even if the results were obtained only for small values of \( t \) (i.e., not considering the particularly hard labels, in result of no available ground truth). Overall, we concluded that the proposed loss obtain acceptable performance (i.e., close to the state-of-the-art) when a small number of soft labels is available (\( \geq 2 \)), but also when a few more labels should be inferred (up to \( t \approx 8 \)). In this regard, we presume that even larger values for \( t \) (\( t \geq 8 \)) would require substantially more amounts of learning data and also larger values for \( d \) (dimension of the embedding).

### E. Semantic Identity Retrieval

Finally, we considered the problem of semantic identity retrieval, in which, along with the query image, it is also
specified some semantic criteria that filter out the retrieval elements (i.e., “Find this person” → “Find this female”), as illustrated in the top part of Fig. 11). In this setting, it is assumed that the ground truth labels of the gallery identities is known, even though the same guarantee does not exist for query elements. We used the dataset that produced the poorest identity retrieval performance (Megaface) and compared our performance to Chen et al.’s method (the most frequent runner up in previous experiments). The soft label “Gender” (automatically provided by the Microsoft Cognitive Toolkit for all the queries) was used as additional semantic data, to filter out the retrieved identities that do not match the desired gender. The bottom plot in Fig. 11 provides the results in terms of the hit/penetration rates, being particularly evident that both methods attained practically the same performance (“semantic” data series). Essentially, this means that if coarse labels are available, our method and Chen et al.’s attain embeddings of similar quality in terms of identity compactness and discriminability. However, it is noteworthy to observe that the proposed loss - the baseline version, without using auxiliary semantic information in the query - is a way to approximate the results that are attained by the state-of-the-art when using semantic information to filter out the retrieved identities. This accords the insight that have led to its proposal, as it can be seen as implicitly inferring in a joint way the coarse (soft biometrics) and fine (identity) labels.

In particular, having focused in the identity retrieval and soft biometrics problems, the proposed loss uses the “ID” and available soft labels (e.g., “Gender”, “Age” and “Ethnicity”) to produce feature embeddings that are semantically coherent, i.e., in which not only the intra-class compactness is guaranteed, but also the broad families of classes appear in adjacent regions. This enables a direct correspondence between the ID centroids and their semantic descriptions, i.e., where a “young, black, male” is surely closer to an “adult, black, male” than to a “young, white, female”. This property is in opposition to the way previous loss functions work, where elements are projected based uniquely in their appearance, being assumed that the semantic coherence naturally yields from the similarities in appearance between classes (i.e., upon the similarity of image features).

V. CONCLUSIONS AND FURTHER WORK

In this paper we proposed a loss function designed to work in multi-output classification problems, where the response variables have dimension larger than one. Our function can be seen as a generalization of the well known triplet loss, that replaces the positive/negative division of pairs and the notion of anchor, by: i) a metric that measures the relative similarity between any two classes; and ii) a quadruplet term that - based in the semantic similarity - imposes the margins between pairs of projections.

ACKNOWLEDGEMENTS

This work is funded by FCT/MCTES through national funds and when applicable co-funded by EU funds under the project UIDB/EEA/50008/2020. Also, this research was funded by the FEDER, Fundo de Coesão e Fundo Social Europeu, under the PT2020 program, in the scope of the POCI-01-0247-FEDER-033395 project. We acknowledge the support of NVIDIA Corporation®, with the donation of one Titan X GPU board.

REFERENCES

[1] N. Almudhahka, M. Nixon and J. Hare. Automatic Semantic Face Recognition. Proceedings of the IEEE 12th International Conference on Automatic Face & Gesture Recognition, doi: 10.1109/FG.2017.31, 2017. 2
[2] E. Bekele, C. Narber and W. Lawson. Multi-attribute Residual Network (MAResNet) for Soft-biometrics Recognition in Surveillance Scenarios. Proceedings of the IEEE 12th International Conference on Automatic Face & Gesture Recognition, doi: 10.1109/FG.2017.55, 2017. 2
[3] E. Cipcigan and M. Nixon. Feature Selection for Subject Ranking using Soft Biometric Queries. Proceedings of the 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance, doi: 10.1109/AVSS.2018.8639319, 2018. 2
[4] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. Patel and R. Chellappa. An End-to-End System for Unconstrained Face Verification with Deep Convolutional Neural Networks. Proceedings of the IEEE International Conference on Computer Vision Workshops, doi: 10.1109/ICCVW.2015.55, 2015. 2
[5] W. Chen, X. Chen, J. Zhang and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. Proceedings of the IEEE International Conference on Computer Vision, doi: 10.1109/CVPR.2017.145, 2017. 3, 6, 8, 9, 10
[6] V. Choutas, P. Weinzaepfel, J. Revaud and C. Schmid. PoToN: Pose MoToN Representation for Action Recognition. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pag. 7024–7033, doi: 1109/CVPR.2018.00734, 2018. 1
[7] Y. Deng, P. Luo, C. Loy and X. Tang. Pedestrian attribute recognition at far distance. Proceedings of the ACM International Conference on Multimedia, pag. 789–792, doi: 10.1145/2478686.2654966, 2014. 5
[8] J. Deng, J. Guo, N. Xue and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. https://arxiv.org/abs/1801.07869, 2019. 2
[9] Y. Duan, J. Lu and J. Zhou. UniformFace: Learning Deep Equidistributed Representation for Face Recognition. https://arxiv.org/abs/1801.07698, 2019. 2
[10] H. Galiyawala, K. Shah, V. Gajjar and M. Raval. Person Retrieval in Surveillance Video using Height, Color and Gender. https://arxiv.org/abs/1810.05080, 2018. 2
[11] J. Goldberger, G. Hinton, S. Roweis and R. Salakhutdinov. Neighbourhood Components Analysis. Proceedings to the Advances in Neural Information Processing Systems Conference, vol. 17, pag. 513–520, doi: 10.5555/2976040.2976105, 2005. 4
