Gender and Age Estimation without Facial Information from Still Images

Georgia Chatzitzisi[1], Michalis Vrigkas[2], and Christophoros Nikou[1]

[1] Department of Computer Science and Engineering, University of Ioannina,
Ioannina, Greece
{gchatzizisi,cnikou}@uoi.gr
[2] Department of Communication and Digital Media, University of Western
Macedonia, Kastoria, Greece
{mvrigkas}@uowm.gr

Abstract. In this paper, the task of gender and age recognition is performed on pedestrian still images, which are usually captured in-the-wild with no near face-frontal information. Moreover, another difficulty originates from the underlying class imbalance in real examples, especially for the age estimation problem. The scope of the paper is to examine how different loss functions in convolutional neural networks (CNN) perform under the class imbalance problem. For this purpose, as a backbone, we employ the Residual Network (ResNet). On top of that, we attempt to benefit from appearance-based attributes, which are inherently present in the available data. We incorporate this knowledge in an autoencoder, which we attach to our baseline CNN for the combined model to jointly learn the features and increase the classification accuracy. Finally, all of our experiments are evaluated on two publicly available datasets.

Keywords: Gender Classification · Age Estimation · Deep Imbalanced Learning.

1 Introduction

Gender and age classification has been studied in the literature over the last decade and recently has gained much more interest due to the large availability of data [2, 7, 12]. In recent years, deep learning methods, such as CNNs, have been gradually applied to age estimation and have achieved better results than handcrafted features. Yi *et al.* [20] introduced a relatively shallow CNN architecture and a multi-scale analysis strategy to learn in an end-to-end manner the age label of a facial image. Niu *et al.* [11] formulated the age estimation problem as an ordinal regression problem using a series of binary classification tasks. Chen *et al.* [2] proposed a ranking-CNN framework, in which a series of basic CNNs were employed and their binary outputs were aggregated. A separate CNN for each ordinal age group was learned, allowing each sub-CNN to capture different patterns for different age groups.

Several hybrid methods predicting age and gender simultaneously with other facial attributes have also been reported in the literature [5]. Levi *et al.* [8]

were the among first to use a CNN architecture for the problem of age and gender classification with a relatively shallow architecture. Rodriguez *et al.* [13] introduced the visual attention mechanism to discover the most informative and reliable parts in a face image for improving age and gender classification. Dual *et al.* [4] integrated a CNN for feature extraction and an extreme learning machine (ELM) [6] for classifying the intermediate results. It is yet another popular idea to make use of body-part information and jointly utilize global CNN features with person, object and, scene attributes [18].

Visual attention mechanism has also been used in pedestrian attribute recognition. Sarfraz *et al.* [17] introduced a model with view guidance to make view-specific attribute predictions to overcome the variance of patterns from different angles. In [16], Sarafianos *et al.* extracted and aggregated visual attention masks at different scales and establish a weighted-variant of the focal loss to handle both under-represented or uncertain attributes. Although attention-based methods improve recognition accuracy, they are attribute-agnostic and fail to consider the attribute-specific information.

Other approaches are regarded as relation-based and exploit semantic relations to assist attribute recognition. Wang *et al.* [19] proposed a CNN-RNN based framework to exploit the interdependence and correlation among attributes. In [15], Sarafianos *et al.* leveraged curriculum learning, by learning first the strongly correlated attributes in a multi-task learning setup and then used transfer learning to additionally learn the weakly-correlated attributes. However, these methods require manually defined rules, e.g., prediction order and attribute groups, which are hard to determine in real-world applications.

In practice, numerous factors affect the classification performance and make the task of gender and age classification far from trivial. Datasets with gender and age annotations are usually captured in-the-wild, where often no near-frontal information is available. Also, images are taken under different illumination conditions and different camera viewing angles, providing poor visual quality. To this end, we employ CNNs and we conduct all of the experiments with the ResNet architecture as the backbone [1]. Another concern about CNNs is that they require datasets to be composed of balanced class distributions. However, datasets with gender and age labels are inherently imbalanced. To examine how a loss function affects the performance of a model, we study the performance of four different loss functions. Having the ResNet architecture as the baseline, an autoencoder is added in parallel to benefit from the appearance-based attributes, and the whole network is trained end-to-end. We consider that this combined model can learn more powerful relationships among the attributes and potentially lead in a better performance.

2 Methodology

In this work, we focus on recognizing the gender and age attributes, which are physical, adhered human characteristics belonging to the soft biometrics. Our method relies on still images of pedestrians without the presence of clear-shot face-frontal information. We opt for a three-stage strategy; (i) we first only consider the problem of gender classification, (ii) then the problem of age classification, and (iii) finally, the problem of multi-label classification, where we try to predict both attributes simultaneously. The main challenge we focused on is the class imbalanced distributions, which are inherently present in the available datasets. For all experiments, we use the ResNet50 architecture [1] as the backbone to investigate how four different loss functions perform under the class imbalance problem. Its power comes from its special architecture, which comprises of skip or shortcut connections to jump over the stacked convolutional layers. Finally, we build a model, adding an autoencoder on top of the ResNet, which we feed with appearance-based attributes. We consider that a combined model can leverage this additional information to make more accurate predictions.

2.1 Gender classification

Consider there are N pedestrian images $\mathcal{X} = \{x_i\}_{i=1}^{N}$, labeled with the gender attribute $y_i \in \{0, 1\}$. The features extracted from the ResNet are pooled and passed through a binary classifier to determine the pedestrian's gender. Our approach employs a global average pooling, which takes the average of each of the feature maps obtained from the ResNet. The output of the model is one neuron with the sigmoid activation function, representing the probability of the pedestrian being "male" or "female".

In the presence of class imbalance, the loss due to the frequent class may dominate total loss and cause instability. Hence, to see how different loss functions perform under the class imbalance problem, we explore the performance of four different loss functions. The first one is the standard binary cross-entropy, formulated as:

$$L_{bce} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \,, \tag{1}$$

where y and $\hat{y}$ are the ground truth and predicted labels, respectively. Such a loss function ignores completely the class imbalance, assigning the same weight to the two classes. Aiming to alleviate this problem, we employ a weighted-variant of the binary cross-entropy, called the binary focal loss [10], defined as:

$$L_{bfl} = -y \, (1 - \hat{y})^{\gamma} \log \hat{y} - (1 - y) \, \hat{y}^{\gamma} \, \log(1 - \hat{y}) \,, \tag{2}$$

where $\gamma \geq 0$ is a focusing parameter. Focal loss is a cross-entropy loss that weighs the contribution of each example to the loss based on the classification error. With this strategy, the loss is made to implicitly focus on the problematic cases by extending the range in which an example receives low loss. For instance, when $\gamma = 2$, an example classified with $\hat{y} = 0.9$ would have 100× lower loss and with $\hat{y} = 0.968$ it would have 1000× lower loss compared with cross-entropy. Finally, we also employ two variants of the binary cross-entropy and the binary focal loss. These two variants are the weighted binary cross-entropy and the weighted binary focal loss and are respectively defined as:

$$L_{wbce} = -w \left[y \, \log \hat{y} + (1 - y) \, \log(1 - \hat{y}) \right] , \tag{3}$$

$$L_{wbfl} = -w \Big[y\,(1-\hat{y})^{\gamma}\, \log \hat{y} + (1-y)\,\hat{y}^{\gamma}\, \log(1-\hat{y}) \Big] , \tag{4}$$

$$w = \begin{cases} \frac{1}{1-p_f} & \text{if} \quad y = 0 \\ \frac{1}{p_f} & \text{if} \quad y = 1 \end{cases} , \tag{5}$$

where w is the loss weight according to the gender label and p_f is the proportion of the females in the training set.

For the problem of gender recognition, we developed a model that can benefit from annotations already present in the available data. Specifically, instead of treating an image independently, we consider inference with help from additional attributes. We claim that introducing this kind of information into a model, gender prediction would be performed with more confidence. For instance, most datasets provide attributes related to pedestrian appearance, upper and lower body clothing styles, and accessories. We incorporate these attributes in a binary vector, hence, each pedestrian image x_i is assigned with an K-length binary vector y_i, where $y_{il} \in \{0, 1\}$ denotes the presence of the k-th attribute in x_i. Then, we employ an autoencoder to learn the "compressed" representation of the original attribute input vector. The autoencoder is a one-hidden-layer neural network, with the size of the "bottleneck" layer and the size of the output layer to be the same as the size of the input vector ($= K$). The problem that the autoencoder is trying to solve is a multi-label classification problem hence, we use the sigmoid activation function for each of the output neurons. We also employ the binary cross-entropy loss of Eq. (1) and the binary focal loss of Eq. (2) slightly modified to account for all K attributes:

$$L_{ae} = -\sum_{k=1}^{K} \Big[y_k \log \hat{y_k} + (1-y_k) \log(1-\hat{y_k}) \Big] , \tag{6}$$

$$L_{ae} = -\sum_{k=1}^{K} \Big[y\,(1-\hat{y})^{\gamma} \log \hat{y} + (1-y)\,\hat{y}^{\gamma}\, \log(1-\hat{y}) \Big] , \tag{7}$$

where K is the number of attributes and y_k, $\hat{y}_k$ are the ground truth and predicted labels for the k-th attribute.

The features from the autoencoders bottleneck layer are concatenated with the features obtained from the last fully connected layer of ResNet to form a new model. At the top, we add a binary classifier and we train this combined model, which we call ResNet+AE, with the best performing loss function from the single-ResNet architecture. The illustration of the ResNet+AE model is depicted in Fig. 1(a). This combined model is trained end-to-end and the overall loss is a combination of the autoencoders loss and the loss arising from the ResNet:

$$L_{combined} = L_{ae} + L_{ResNet} \, . \tag{8}$$

where L_{ae} is one of the Eqs. (6), (7) and L_{ResNet} is one of the Eqs. (1), (2), (3), or (4), whichever performs the best in the case of gender classification.

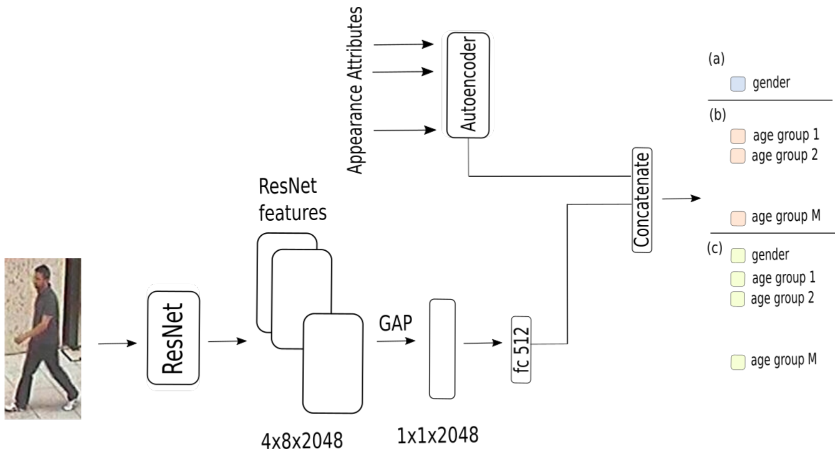


Fig. 1. The ResNet+AE model for (a) gender classification, (b) age classification and (c) multi-label classification.

2.2 Age classification

We also study the problem of age recognition, where the model should predict one of M classes corresponding to M age categories. The age label vector is a one-hot vector y, and each element of that vector is represented as $y_m \in \{0, 1\}$, with $m = 1, \cdots, M$. We now employ the ResNet architecture, with the difference that the top classifier now predicts one of M possible classes. The M output neurons use the softmax activation function, to model a probability distribution consisting of M probabilities.

For the problem of age classification, we adopt the categorical cross-entropy loss, formulated as:

$$L_{cce} = -\sum_{i=1}^{M} y_i \log \hat{y}_i \,. \tag{9}$$

where M is the number of classes, and y_i, $\hat{y}_i$ are the one-hot encoded ground truth and predicted labels for the i-th class. Since the ground-truth labels are one-hot encoded only the positive class keeps its term in the loss, discarding the elements of the summation which are zero due to zero target labels. In addition to the categorical cross-entropy loss, we also explore the performance of the categorical focal loss and their weighted variants, which can be extended to the multi-class case easily:

$$L_{cfl} = -\sum_{i=1}^{M} y_i \, (1 - \hat{y}_i)^{\gamma} \log \hat{y}_i \,, \tag{10}$$

$$L_{wcce} = -\sum_{i=1}^{M} w_i \, y_i \, \log \hat{y} \,, \tag{11}$$

$$L_{wcfl} = \sum_{i=1}^{M} -w_i \, y \, (1-\hat{y})^{\gamma} \, \log \hat{y} \, , \tag{12}$$

where the weighting factor $w_i = \frac{n_{\text{argmax}_{i \in \{1, \cdots, M\}} n_i}}{n_i}$ is the weight loss assigned to the age group i and n_i is the number of examples of the i-th age group in the training set. Finally, $n_{\text{argmax}_{i \in \{1, \cdots, M\}} n_i}$ is the number of examples of the most representative class. Besides, we conduct experiments with the combined model for the problem of age recognition, which is depicted in Fig. 1(b). The overall loss is the summation of the loss originating from the autoencoder and the loss originating from the ResNet and it is in the form of Eq. (8).

2.3 Multi-label classification

Finally, we consider the multi-label recognition problem, in which both attributes, gender and age, are predicted simultaneously. Now, each pedestrian image is labeled with a $(M+1)$-length vector, with the first element referring to the pedestrians gender and the remaining M referring to the pedestrians age range. For the multi-label recognition problem, we use the sigmoid activation function for the $M+1$ output neurons and conduct experiments with the four loss functions, which for the multi-label case are reformulated as:

$$L_{bce} = -\sum_{i=1}^{M+1} y_i \log \hat{y_i} + (1-y_i) \log(1-\hat{y_i}) \, , \tag{13}$$

$$L_{bfl} = -\sum_{i=1}^{M+1} y_i \, (1-\hat{y_i})^{\gamma} \log \hat{y_i} + (1-y_i) \, \hat{y_i}^{\gamma} \, \log(1-\hat{y_i}) \, , \tag{14}$$

$$L_{wbce} = -\sum_{i=1}^{M+1} w_i \left[y_i \, \log \hat{y_i} + (1-y_i) \, \log(1-\hat{y_i}) \right] , \tag{15}$$

$$L_{wbfl} = -\sum_{i=1}^{M+1} w_i \left[y_i \, (1-\hat{y_i})^{\gamma} \, \log \hat{y_i} + (1-y_i) \, \hat{y_i}^{\gamma} \, \log(1-\hat{y_i}) \right] , \tag{16}$$

$$w_i = \begin{cases} e^{p_i} & \text{if} \quad y=0 \\ e^{1-p_i} & \text{if} \quad y=1 \end{cases} , \tag{17}$$

where y_i, $\hat{y_i}$ are the ground truth and predicted labels for the i-th attribute, respectively, w_i is the loss weight assigned to attribute i and p_i is the proportion of the positive labels for the attribute i in the training set. The multi-label case is depicted in Fig. 1(c).

3 Experimental Results

To demonstrate the effectiveness of the proposed method, we compared with several state-of-the-art methods in two publicly available benchmark datasets. PEdesTrian Attribute (PETA) [3] dataset merges 10 consists of $19,000$ images, each annotated with 61 binary attributes. PETA dataset is randomly partitioned into three parts, of which $9,500$ for training, $1,900$ for validation, and $7,600$ for testing. Images are all captured from far view field and they exhibit large differences in terms of lighting conditions, camera viewing angles, image resolutions, background complexity, and indoor/outdoor environment. RAP v2 (Richly Annotated Pedestrian) [9] dataset has in total $84,928$ images and image resolutions range from 36×92 to 344×554. Each image is annotated with 69 binary attributes.

For the problem of gender recognition, we used five metrics, namely accuracy, precision, recall, F1 score, and mean accuracy. Accuracy quantifies the fraction of predictions the model got right. For the problem of age classification, we similarly use the accuracy, precision, recall, and F1 score, slightly modified, since age classification is a multi-class problem. In this case, accuracy quantifies how often predictions match the true labels by checking if the index of the maximal true label is equal to the index of the maximal predicted label. Finally, for the problem of multi-label recognition, accuracy, precision, recall, and F1 score are calculated per-sample.

Finally, we used the pre-trained ResNet50 architecture, which has already been trained on the ImageNet dataset [14]. All images were pre-processed and resized to 256×128 since pedestrian images are usually rectangular. To avoid overfitting, we employed some of the commonly used data augmentation techniques. As for the optimizer, we used the mini-batch stochastic gradient descent with momentum set to 0.9 and early stopping when the validation error was not improving for five consecutive epochs. The batch size is 50 samples per iteration and a dropout layer with a small dropout probability (i.e., 0.1) that may act as a regularizer is added after the feature concatenation layer. Finally, we selected $\gamma = 2$ for the focusing parameter in the focal loss and weighted focal loss.

3.1 Gender Classification

Table 1 compares the performance of the four loss functions described for the PETA dataset. Although the gender distribution is nearly balanced, it can be seen that both weighted loss functions outperform their un-weighted counterparts. Specifically, WBCE performs 0.44% better in terms of the F1 score and 0.36% better in terms of the mAcc metric compared to BCE. Similarly, WBFL is by 1.8% better in terms of the F1 score and by 1.41% better in terms of the mAcc compared to BFL. Comparing the weighted loss functions, WBCE outperforms WBFL by 3.48% in the F1 score and by 3.16% in mAcc and subsequently, it is used to train the ResNet+AE model.

The proposed ResNet+AE model leverages the appearance-based attributes in the gender classification scheme, achieving 90.71% and 91.53% in the F1 score

Table 1. Performance comparison of the four loss functions and the ResNet+AE model on the PETA dataset (in %).

	Prec	**Rec**	**F1**	**mAcc**	**Acc**
ResNet-BCE	88.80	86.07	87.42	88.57	88.81
ResNet-WBCE	**87.86**	**87.89**	**87.86**	**88.93**	**89.03**
ResNet-BFL	85.51	79.87	82.58	84.36	84.79
ResNet-WBFL	84.57	84.18	84.38	85.77	85.92
ResNet+AE	**91.67**	**89.79**	**90.71**	**91.53**	**91.70**

Table 2. Performance comparison of the four loss functions and the ResNet+AE model on the RAP v2 dataset (in %).

	Prec	**Rec**	**F1**	**mAcc**	**Acc**
ResNet-BCE	93.18	91.81	92.49	94.38	95.35
ResNet-WBCE	91.00	**94.91**	**92.91**	**95.33**	**95.49**
ResNet-BFL	92.90	91.82	92.36	94.32	95.26
ResNet-WBFL	89.46	94.17	91.76	94.57	94.73
ResNet+AE	**91.16**	92.30	91.72	94.12	94.81

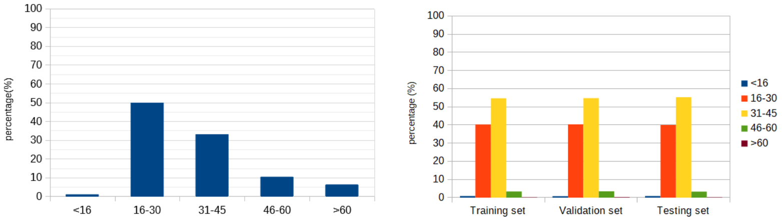
and mAcc, respectively, outperforming the single-ResNet architecture with any of the loss functions.

The gender distribution in the RAP v2 dataset is quite imbalanced given that the number of males is over twice the number of females. Table 2 compares the performance of the four loss functions for the RAP v2 dataset. WBCE performs 0.42% better in terms of the F1 score and 0.95% better in terms of the mAcc metric compared to BCE. BFL is by 0.6% better in terms of the F1 score compared to WBFL but WBFL is 0.25% better in terms of the mAcc compared to BFL. Nevertheless, WBCE outperforms WBFL by 2.87% in the F1 score and 2.47% in mAcc and this is the loss function of choice for the ResNet+AE model. The proposed ResNet+AE model performs comparably well achieving 91.72% and 94.12% in F1 score and mAcc respectively but does not outperform the single-ResNet architecture with the WBCE loss function.

The proposed ResNet+AE model demonstrates inferior performance, achieving 1.19% in the F1 score, which indicates that the performance is degraded. Therefore, the combined model cannot leverage the appearance-based attributes for the age classification, and the single-ResNet architecture with the WBCE being the best performing model.

3.2 Age Classification

The age category distribution in the PETA dataset can be seen in Fig. 2(a). There are five age classes to be predicted, < 16, $16 - 30$, $31 - 45$, $46 - 60$, and > 60, with distributions of 0.9%, 49.77%, 32.92%, 10.24%, 6.17%, respectively. Hence, it is apparent that the age attribute suffers from a severe class imbalance.



(a) (b)

Fig. 2. The distribution of the age categories in (a) the PETA dataset and (b) the RAP v2 dataset.

Table 3. Performance comparison of the four loss functions and the ResNet+AE model on the PETA dataset (in %).

	mPrec	**mRec**	**mF1**	**Acc**
ResNet-CCE	85.55	66.76	73.19	77.29
ResNet-WCCE	67.37	70.53	68.72	70.76
ResNet-CFL	**84.01**	**68.03**	**73.85**	**76.89**
ResNet-WCFL	54.23	64.64	57.80	64.04
ResNet+AE	**80.06**	**72.75**	**75.84**	**79.61**

Table 3 shows the performance of the four loss functions for the PETA dataset. Although the weighted loss functions balance each example according to the class it belongs to, giving more focus on the under-represented classes, they seem to improve none of the metrics. We consider that this behavior is caused by poor features since it is difficult for the ResNet to provide representative features given that there is no near-face information and sometimes the pedestrian is standing backward. Also, since the optimization method is performed in batches, it is not guaranteed that there are examples for each age group in each batch, hence the model is overwhelmed by the majority class and cannot ensure good discriminations among the five age categories. The categorical focal loss performs slightly better than the categorical cross-entropy by 0.66% in terms of the mF1 score and subsequently, it is used to train the ResNet+AE model.

The proposed ResNet+AE model outperforms the single ResNet architecture, achieving 75.84% in terms of the mF1 score. This means that there is some sort of interdependence among the appearance-based attributes, which helps the ResNet+AE model to yield a better age classification performance.

The age category distribution in the RAP v2 dataset can be seen in Fig. 2(b). There are five age groups to be predicted with distributions of 0.92%, 40.44%, 54.89%, 3.53%, and 0.22%, respectively. The distribution is heavily unbalanced with the second and third age categories to be more represented compared to the rest. Table 4 compares the performance of the four loss functions for the RAP v2 dataset. Similarly, with the PETA dataset, the weighted loss functions do not improve the performance compared to their un-weighted counterparts.

Table 4. Performance comparison of the four loss functions and the ResNet+AE model on the RAP v2 dataset (in %).

	mPrec	**mRec**	**mF1**	**Acc**
ResNet-CCE	41.73	29.97	31.45	65.71
ResNet-WCCE	26.27	49.92	24.51	39.36
ResNet-CFL	**48.46**	**36.82**	**39.81**	**64.82**
ResNet-WCFL	26.00	49.09	23.30	37.31
ResNet+AE	41.57	34.30	36.27	64.94

Table 5. Performance comparison of the four loss functions and the ResNet+AE model on the PETA dataset (in %).

	Prec	**Rec**	**F1**	**mAcc**	**Acc**
ResNet-BCE	79.20	77.88	78.53	81.59	91.40
ResNet-WBCE	**79.22**	**79.58**	**79.40**	**82.82**	**91.09**
ResNet-BFL	76.90	75.95	76.42	80.64	90.47
ResNet-WBFL	77.64	78.52	78.08	81.78	90.37
ResNet+AE	**80.02**	**80.80**	**80.41**	**84.49**	**91.54**

CFL is the best performing loss function, which outperforms the CCE by 8.36% in the mF1 score, and it is used to consequently train the ResNet+AE model.

The proposed ResNet+AE model demonstrates inferior performance, achieving 36.27% in the mF1 score, which indicates that the performance is degraded. Therefore, the combined model cannot leverage the appearance-based attributes for the age classification, and the single-ResNet architecture with the CFL is the best performing model.

3.3 Multilabel Classification

The performance of the four different loss functions for the PETA dataset for the task of multi-label classification, where the model classifies both the gender and the age attributes is summarized in Table 5. Since the gender attribute is nearly balanced the heavy imbalance of the age attribute (Fig. 2(a)) overwhelms the distribution to be modeled. However, the performance is not degraded even though the model now has to predict both attributes simultaneously. The weighted loss functions manage to achieve better results compared to their un-weighted counterparts. More specifically, WBCE is 0.87% and 1.23% better in F1 score and mAcc respectively compared to the plain BCE. Similarly, WBFL performs better by 1.66% in the F1 score and by 1.14% in mAcc compared to plain BFL. The best among the four loss functions is the WBCE achieving 79.4% and 82.82% in F1 score and mAcc respectively and this loss function is used to consequently train the ResNet+AE model. The proposed ResNet+AE model outperforms the single-ResNet architecture, achieving 80.41% in the F1 score and 84.49% in mAcc.

Table 6 depicts the performance of the four different loss functions for the RAP v2 dataset when the model classifies both the gender and the age attributes.

Table 6. Performance comparison of the four loss functions and the ResNet+AE model on the RAP v2 dataset (in %).

	Prec	**Rec**	**F1**	**mAcc**	**Acc**
ResNet-BCE	71.08	70.63	70.85	63.56	88.40
ResNet-WBCE	70.71	71.52	71.11	67.86	88.25
ResNet-BFL	69.31	69.53	69.42	66.37	87.97
ResNet-WBFL	**68.82**	**70.40**	**69.60**	**68.46**	**87.73**
ResNet+AE	67.63	68.30	67.96	67.76	87.29

It can be seen that the weighted loss functions perform slightly better than the unweighted counterparts. Specifically, WBCE performs 0.26% better in terms of the F1 score and 4.3% better in terms of the mAcc compared to BCE, and WBFL performs 0.18% better in terms of the F1 score and 2.09% better in terms of the mAcc compared to BFL. Overall, in the single ResNet architecture, WBCE performs 1.51% better in terms of the F1 score, but WBFL performs 0.6% better in terms of the mAcc. We chose the WBFL as the best performing loss function, as mAcc is a label-based metric and is a more important metric in the multi-label classification case. Concerning the combined ResNet+AE model, although it is quite similar in performance compared to most of the single-ResNet architectures, it does not outperform the single-ResNet case with the WBFL loss function.

4 Conclusion

In this paper, we studied the problem of gender and age classification from pedestrian images. The class imbalance which characterizes the datasets makes the task quite challenging. We focused on examining how four different loss functions. We tested our model, which concatenates the features from the ResNet backbone and the features from an autoencoder, which is trained in parallel with appearance-based attributes. Taken into consideration the experimental results, the gender classification is an easier task, as the ResNet can extract representative features to make an accurate classification. The age classification is a more challenging problem since age categories are heavily imbalanced and with no near-face information. The multi-label classification is also a challenging task, as the age category imbalance overwhelms the distribution to be modeled. The experimental results showed that high classification accuracy may be obtained when the appearance-based attributes involve some sort of relationship.

Acknowledgments: This work has been co-funded by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-04517). The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

1. Bekele, E., Lawson, W.: The deeper, the better: Analysis of person attributes recognition. In: FG. pp. 1–8 (2019)
2. Chen, S., Zhang, C., Dong, M.: Deep age estimation: From classification to ranking. IEEE TM **20**(8), 2209–2222 (2017)
3. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: ACM ICM. pp. 789–792 (2014)
4. Duan, M., Li, K., Yang, C., Li, K.: A hybrid deep learning cnn–elm for age and gender classification. Neurocomputing **275**, 448–461 (2018)
5. Gonzalez-Sosa, E., Fierrez, J., Vera-Rodriguez, R., Alonso-Fernandez, F.: Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. IEEE TIFS **13**(8), 2001–2014 (2018)
6. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing **70**(1), 489 – 501 (2006)
7. Juefei-Xu, F., Verma, E., Goel, P., Cherodian, A., Savvides, M.: Deepgender: Occlusion and low resolution robust facial gender classification via progressively trained convolutional neural networks with attention. In: CVPRW. pp. 68–77 (2016)
8. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: CVPRW. pp. 34–42 (2015)
9. Li, D., Zhang, Z., Chen, X., Huang, K.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE TIP **28**(4), 1575–1590 (2018)
10. Lin, T., Goyal, P., Girshick, R., He, K., Dollr, P.: Focal loss for dense object detection. In: ICCV. pp. 2999–3007 (2017)
11. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: CVPR. pp. 4920–4928 (2016)
12. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. TPAMI **41**(1), 121–135 (2017)
13. Rodríguez, P., Cucurull, G., Gonfaus, J.M., Roca, F.X., González, J.: Age and gender recognition in the wild with deep attention. Pattern Recognition **72**, 563–571 (2017)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Sarafianos, N., Giannakopoulos, T., Nikou, C., Kakadiaris, I.A.: Curriculum learning for multi-task classification of visual attributes. In: ICCVW. pp. 2608–2615 (2017)
16. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: ECCV. pp. 680–697 (2018)
17. Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv:1707.06089 (2017)
18. Smailis, C., Vrigkas, M., Kakadiaris, I.A.: Recaspia: Recognizing carrying actions in single images using privileged information. In: ICIP. pp. 26–30 (2019)
19. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute recognition by joint recurrent learning of context and correlation. In: ICCV, (Oct 2017)
20. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Asian Conference on Computer Vision. pp. 144–158 (2014)