Learning with Ambiguous Label Distribution for Apparent Age Estimation

Ke Chen and Joni-Kristian Kämäräinen

Department of Signal Processing
Tampere University of Technology
Tampere 33720, Finland
firstname.lastname@tut.fi

Abstract. Annotating age classes for humans' facial images according to their appearance is very challenging because of dynamic person-specific ageing pattern, and thus leads to a set of unreliable apparent age labels for each image. For utilising ambiguous label annotations, an intuitive strategy is to generate a *pseudo* age for each image, typically the average value of manually-annotated age annotations, which is thus fed into standard supervised learning frameworks designed for chronological age estimation. Alternatively, inspired by the recent success of label distribution learning, this paper introduces a novel concept of ambiguous label distribution for apparent age estimation, which is developed under the following observations that 1) soft labelling is beneficial for alleviating the suffering of inaccurate annotations and 2) more reliable annotations should contribute more. To achieve the goal, label distributions of sparse age annotations for each image are weighted according to their reliability and then combined to construct an ambiguous label distribution. In the light of this, the proposed learning framework not only inherits the advantages from conventional learning with label distribution to capture latent label correlation but also exploits annotation reliability to improve the robustness against inconsistent age annotations. Experimental evaluation on the FG-NET age estimation benchmark verifies its effectiveness and superior performance over the state-of-the-art frameworks for apparent age estimation.

1 Introduction

Chronological age estimation [1–7] is to predict persons' true age given their facial images, which is a hot yet challenging topic in computer vision. In supervised learning based frameworks for age estimation, the unique chronological age labels are provided to supervise model training. However, due to inherent ambiguities in age annotation, a large number of facial images can be readily found on the Internet, but reliable annotations of the exact age of images are usually lacking, which leads to sparsely distributed data [1] in the public benchmarks such as the FG-NET and MORPH datasets. More challengingly, apparent age estimation investigated in this paper is to estimate apparent ages of human faces (intuitively, how old the persons look like) from apparent age annotations

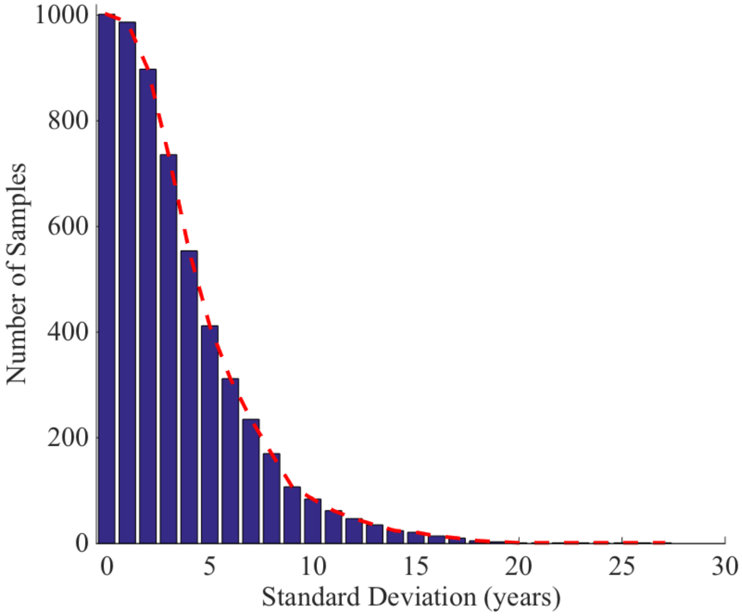

Fig. 1: Number of samples with larger standard deviation than coordinates in the horizontal axis to reflect label ambiguity on the FG-NET benchmark with manual annotations provided by Han *et al.* [8]. The maximum of standard deviation for age annotations is 27.14; standard deviation of 41.12% samples is larger than 5 years, while that of 8.38% samples is larger than 10 years.

instead of their chronological age. Apparent age estimation can be categorised into a weakly-supervised learning paradigm, as the supervised information conveyed in such a problem are implicit. Learning a mapping function between imagery feature representation and a set of age label annotations for each facial image is made even more difficult due to large variation of persons' appearance caused by both intrinsic and extrinsic factors and label ambiguity.

On one hand, low-level feature extracted from facial images is largely varied caused by intrinsic and extrinsic factors. Person-specific ageing procedure generally lies in the changes of shape (during childhood) and texture (during adulthood). In this sense, visual appearance of faces varies a lot across individuals because of different gender, hairstyle, ethnicity *etc.* In addition, changing illumination conditions and head poses of human faces also affect the extracted features, which further increases feature inconsistency and thus the difficulty in age estimation. On the other hand, label inconsistency intrinsically caused by manual annotations is the main challenge in apparent age estimation, which has been investigated in very few existing work. As shown in Fig. 1, standard deviation of apparent age annotations for each instance is first calculated and the cumulative size of samples larger than standard deviation coordinates is reported, for the purpose of visualising the uncertainty of manual annotations. Each image in apparent age estimation is associated with a number of uncertain apparent age annotations instead of a unique chronological age. The straightforward solution is to average annotated age classes to generate a pseudo age [9–13], which is readily applied to the existing supervised learning frameworks for chronological age estimation. However, the induced uncertainty has not been exploited in such a setting. We observe that the mean of apparent age classes could miss reflecting label variation across annotations. Such an observation motivates us to take incorporating annotation reliability explicitly in the label representation into account to achieve more robust performance.

We consider that *latent label correlation mining* and *reliable annotation exploiting* are two key factors for accurate and robust apparent age estimation. To this end, we propose a novel framework based on the recent label distribution learning paradigm to combine label distribution from ambiguous annotations to alleviate feature and label inconsistency. Compared to the mean and/or standard deviation over all annotated age labels in [9–13], the proposed ambiguous label distribution learning aims to construct a weighted label density space, which is then mapped from low-level feature space. The weighting strategies according to the annotations' reliability play a vital role in generating such an ambiguous label distribution. On one hand, more reliable age annotations (determined by reliability) with higher weights will contribute more to achieving robust performance from label uncertainty. On the other hand, combining a number of distribution from different annotations will have richer description degree in label distribution in comparison with single label distribution generated by mean and standard deviation of all apparent age. Specifically, the proposed ambiguous label distribution usually have more than one peak and asymmetric distribution rather than single peak and symmetric structure in original label distribution learning [4].

2 Related Work

Chronological Age Estimation – The recent frameworks for estimating persons' chronological ages given facial images can be categorised into three groups: classification based [3, 4, 14–16], regression based [1, 7, 17, 18], and ranking based [19, 20]. Considering cumulative dependent nature across age classes (*i.e.* the closer age labels of facial images are, the more visual similarity they share), the frameworks [1, 3, 4, 7, 17, 18] explicitly or implicitly mining latent label correlation are more favourable for facial age estimation problem. Chen *et al.* [1] exploited the cumulative dependency across age classes to achieve robust performance in a two-layer attribute learning framework. Geng *et al.* [4] designed a framework by learning from label distribution in the manner of multi-label learning instead of a single independent class label to capture latent age class correlation. Specifically, for each instance, a label distribution vector (whose size is equal to the age range, typically $[0, 100]$) gives description degree to each element, which reflects its describing capability. The maximum description degree is allocated to the element having the relative position of the chronological age in the label distribution vector. The label distribution vectors are then mapped from imagery feature vectors. Evidently, for any two age classes, the design of label distribution learning captures their correlation via the values of their corresponding positions in the label distribution vectors. An advanced attempt of label distribution learning with adaptive updating label distribution for each age group was introduced to mitigate the suffering of dynamic ageing procedure during different period (*i.e.* childhood and adulthood) [16].

Apparent Age Estimation – In apparent age estimation, each training facial image is associated with a number of inconsistently annotated age labels. Such a problem was cast into a partial label learning paradigm [21–23], which assumes that only one annotation is valid among a set of candidate annotations. However, the existing partial label learning frameworks were designed for classification problems without considering ordinal dependency across age classes, which are less suitable for apparent age estimation. An intuitive strategy is to construct a pseudo age label for each sample from a number of manual annotations, which can be directly applied and incorporated to the existing supervised learning frameworks originally developed for chronological age estimation. The typical strategy for pseudo age generation is to use the mean value of apparent age. Since the competition organised by ChaLearn [24] is popular, apparent age estimation has attracted wide attention in the field and a number of recent frameworks [9–13] based on Convolutional Neural Networks (CNN) [25] were proposed, which concerned mainly on training and/or fine-tuning deep CNN models for better imagery representation. Inspired by recent success of the existing frameworks such as cumulative attributes [1] and label distribution learning [4] designed for chronological age estimation, their concept have inspired to design DeepCodeAge [11] and deep label distribution learning [9] respectively. The framework proposed by Yang *et al.* [9] was one of the first attempt to handle the uncertainty of apparent age, which shares similar script as our learning with ambiguous label distribution (LALD). Nevertheless, the differences of our method lie in the utilisation of combined label distribution from independent apparent age annotations for each image instead of a single label distribution based on global statistics (*e.g.* , mean and standard deviation [9]). Moreover, the proposed method in this paper also takes the reliability of every annotation into account, which further boost the estimation performance. Consequently, owing to the introduction of label distribution combination and reliability enhancing, our ambiguous label distribution is more informative and robust than the direct utilisation of label distribution with global statistics.

Contributions – The contributions and novelties of this paper are three-fold as:

- To the best knowledge of authors, this paper is the first attempt for apparent age estimation to exploit annotation reliability to handle with label uncertainty and reduce the negative effect caused by annotation outliers.
- Compared to the mean and standard deviation over all apparent age annotations, the proposed ambiguous label distribution owing to label distribution

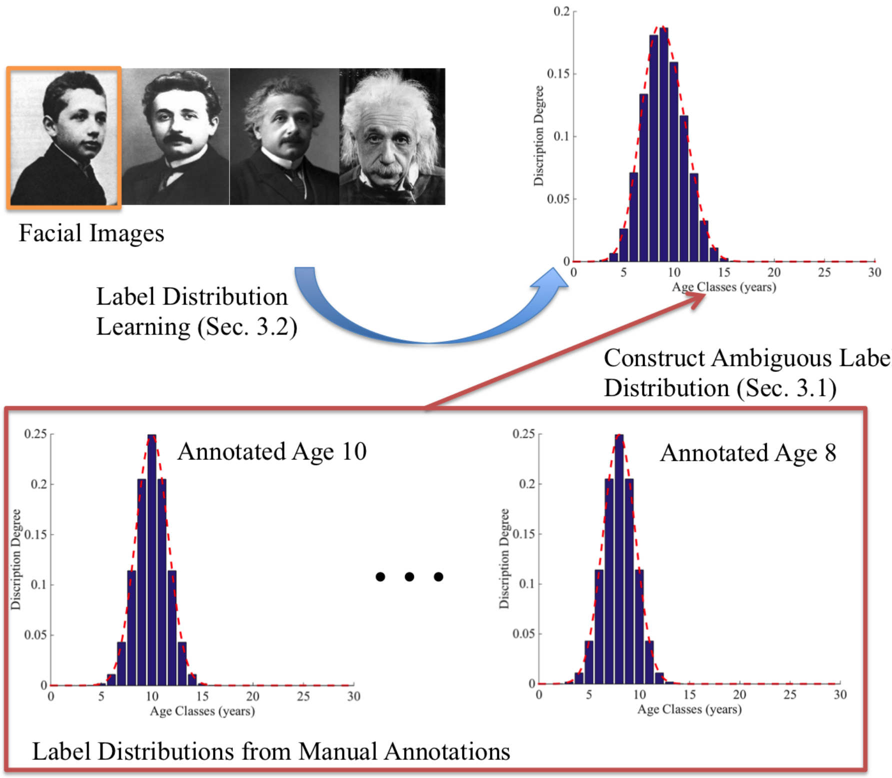

Fig. 2: The pipeline of the proposed learning with ambiguous label distribution.

combination and reliability enhancing is more informative and robust in view of using annotation density of each image.

- Extensive experiments on the public FG-NET benchmark gain notable advantage on accuracy of ambiguous label distribution learning for apparent age estimation to tackle with both feature inconsistency and label ambiguity.

3 Methodology

Given imagery feature representation $\boldsymbol{x}$ and its corresponding age annotations $\boldsymbol{y} \in \mathbb{R}^D$, training samples consist of $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}^{i=1,2,\ldots,N}$, where N denotes the number of training samples. The pipeline of the proposed algorithm illustrated in Fig. 2 is given in details as the following:

- For ith training sample, we first generate vector-formed ordinary label distribution $\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots \boldsymbol{l}_D$ of apparent age according to their relative positions $\boldsymbol{y}_i$ in a chronological age range. Label distributions are then combined together

according to their reliability, *i.e.* the distance to the pseudo age, to construct ambiguous label distribution $\boldsymbol{a}_i$ (see Sec. 3.1).
- Learning the mapping between imagery feature representation $\boldsymbol{x}$ and an ambiguous label distribution $\boldsymbol{a}$ is achieved by adopting label distribution learning [3, 4] (see Sec. 3.2).

During testing, imagery feature of an unseen image are fed into the trained model to predict the person's age, *i.e.* the age class having the maximum predicted description degree in the ambiguous label distribution.

3.1 Ambiguous Label Distribution Construction

The concept of label distribution was firstly introduced by Geng *et al.* [4] for chronological age estimation, which is investigated briefly here and named as ordinary label distribution to distinguish from the proposed ambiguous label distribution. Given a scalar-valued age label $y \in \mathbb{R}$ for each image instance, a label distribution vector $\boldsymbol{l} \in \mathbb{R}^K$ is generated, whose dimension K is equal to the size of age range. Each dimension in such a label distribution corresponds to an age class according to its relative positioning in value, which has a description degree $d_{\boldsymbol{x}} \in [0, 1]$ to indicate the capability to describe the proportion of the samples. In mathematics, description degree $d_{\boldsymbol{x}}^k$ possibly represented as conditional probability $P(k|\boldsymbol{x})$ indicates that age label $k \in [0, K-1]$ describe the proportion $d_{\boldsymbol{x}}^k$ of the sample. The sum of real-valued description degree $d_{\boldsymbol{x}}^k$ of all elements in the label distribution vector is equal to one. The assumption of label distribution is two-fold: a) true labels have the highest description degree in $\boldsymbol{l}$; and b) the farther labels are away from chronological ages, the lower description degree they have. As label distribution changing along ordinal age classes continuously and cumulatively, description degree also reflects the support of neighbouring labels contributing to the exact label y associated to instance $\boldsymbol{x}$. Consequently, all age classes k having positive values are assumed to contribute to discriminating training samples to the age class y. Typical label distributions are Gaussian and triangle distributions [4] anchored in chronological age class. In the light of Gaussian distribution consistently superior to triangle distribution, which is thus adopted in the experiments of this paper.

The setting of original label distribution designed for chronological age estimation can be readily employed for apparent age estimation by obtaining a pseudo age label for each instance from a set of unreliable annotations. Ordinary label distribution only has a single peak and the symmetric distribution. We aim to enrich the capability of label representation by proposing a novel ambiguous label distribution (ALD) $\boldsymbol{a} = [d_{\boldsymbol{x}}^0, d_{\boldsymbol{x}}^1, \cdots, d_{\boldsymbol{x}}^{K-1}] \in \mathbb{R}^K$. For each image, it combines label distribution $\boldsymbol{l}$ from each apparent age annotation, which is then normalised to satisfy that each element of description degree $d_{\boldsymbol{x}}^k \in [0, 1]$ in $\boldsymbol{a}$ and $\sum^K d_{\boldsymbol{x}}^k = 1$.

The strategies to construct the ambiguous label distribution play an important role and are sensitive to the estimation performance, which are investigated here and will be experimentally evaluated and compared. To this end, two factors need to be concerned for an informative and robust ambiguous label distribution: 1) pseudo age acquisition and 2) combination of label distribution $\boldsymbol{l}$. Evidently, averaging all annotated age labels and maximum majority voting are two types of intuitive strategies for determining pseudo ages for apparent age estimation problem. Moreover, we also introduce the third type by averaging annotated age classes, *i.e.* without counting the repeated annotations for the identical age classes. The third one consistently achieves superior performance to the rest two in our experiments with more detailed analysis given in the experimental part (Table 2). The solution to the second question is to incorporate their reliability (*e.g.* the first- or second- order statistics between apparent age and the pseudo age) as the weights for the corresponding label distribution. Adopting the weighting strategies is aimed to improve the robustness, as the less reliable annotation far away from the pseudo age should be given lower weights to reduce their negative effect. For each image, the distance measure between annotations and pseudo age is thus employed as weights of reliability, which are respectively multiplied with their corresponding label distributions and then summed up to construct the proposed ambiguous label distribution. Our framework achieves better performance than non-weighted combination in our evaluative verification (Table 3).

3.2 Mapping from Feature Input to Distribution Output

With the generated ambiguous label distribution for each image, the training set becomes $\{\boldsymbol{x}, \boldsymbol{a}\}_i, i = 1, 2, \cdots, N$. Entry $\boldsymbol{a}_j, j = 1, 2, \cdots, K$ of $\boldsymbol{a} \in \mathbb{R}^K$ denotes description degree for the jth age class, where age label for the jth age class is $k = j - 1$. The aim is to learn a conditional density function $p(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})$ to minimise the distance between the predicted $\hat{\boldsymbol{a}}$ generated by $\boldsymbol{\theta}$ and the ground truth $\boldsymbol{a}$, where $\boldsymbol{\theta}$ is the parameter vector to be optimised. Evidently, the problem is cast as a label distribution learning problem, which has been well presented in [4]. The object function for ambiguous label distribution learning can be written as:

$$\min_{\theta} \quad \sum_i P(\boldsymbol{a}_i || p(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta})), \tag{1}$$

where $P(\boldsymbol{a}^h || \boldsymbol{a}^w)$ is to measure the similarity between two distributions $\boldsymbol{a}^h$ and $\boldsymbol{a}^w$. In this paper, Kullback-Leibler divergence [26] is employed, which can be mathematically depicted as the following:

$$P(\boldsymbol{a}^h || \boldsymbol{a}^w) = \sum_j (\boldsymbol{a}_j^h \ln \frac{\boldsymbol{a}_j^h}{\boldsymbol{a}_j^w}), \tag{2}$$

where $\boldsymbol{a}_j^h$ and $\boldsymbol{a}_j^w$ denote the jth element in $\boldsymbol{a}^h$ and $\boldsymbol{a}^w$ respectively. Substituting Equation (2) into (1), object function can thus be formulated as:

$$\min_{\theta} \quad \sum_i \sum_j (d_{\boldsymbol{x}_i}^{j-1} \ln \frac{d_{\boldsymbol{x}_i}^{j-1}}{p((j-1)|\boldsymbol{x}_i; \boldsymbol{\theta})}). \tag{3}$$

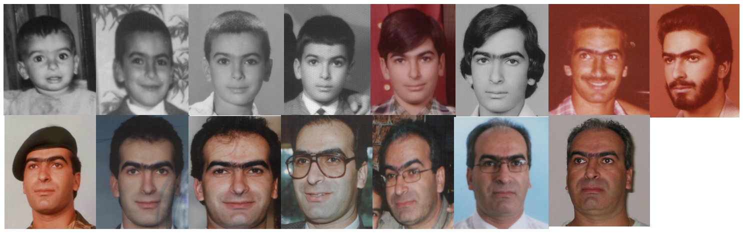

Fig. 3: Illustrative example images from the FG-NET dataset.

As a result, the optimised parameter θ^* can be determined by

$$\begin{aligned}\theta^* &= \operatorname*{argmin}_{\theta} \sum_i \sum_j (d_{\boldsymbol{x}_i}^{j-1} \ln \frac{d_{\boldsymbol{x}_i}^{j-1}}{p((j-1)|\boldsymbol{x}_i;\boldsymbol{\theta})}) \\ &= \operatorname*{argmax}_{\theta} \sum_i \sum_j d_{\boldsymbol{x}_i}^{j-1} \ln p((j-1)|\boldsymbol{x}_i;\boldsymbol{\theta}).\end{aligned} \tag{4}$$

Let us assume a maximum entropy model [27] as

$$p((j-1)|\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{\exp(\sum_r \boldsymbol{\theta}_{j-1,r}\boldsymbol{x}_i^r)}{\sum_j \exp(\sum_r \boldsymbol{\theta}_{j-1,r}\boldsymbol{x}_i^r)}, \tag{5}$$

where $\boldsymbol{x}_i^r$ denotes the rth entry of feature $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_{j-1,r}$ is the element of $\boldsymbol{\theta}$ associated to the jth label (*i.e.* age $j-1$ class in the light of starting from age 0 in the age range) and rth feature element. Substituting Equation (5) into (4) yields the object function as

$$\begin{aligned}F(\theta) &= \sum_{i,j} d_{\boldsymbol{x}_i}^{j-1} \ln p((j-1)|\boldsymbol{x}_i;\boldsymbol{\theta}) \\ &= \sum_{i,j} d_{\boldsymbol{x}_i}^{j-1} \sum_r \boldsymbol{\theta}_{j-1,r}\boldsymbol{x}_i^r - \sum_i \ln \sum_j \exp(\sum_r \boldsymbol{\theta}_{j-1,r}\boldsymbol{x}_i^r).\end{aligned} \tag{6}$$

A number of optimisation algorithms such as improved iterative scaling (IIS) [28], Conditional Probability Neural Network (CPNN) [4], quasi-Newton method BFGS [29] have been investigated and evaluated to address object function (6) in [3, 4]. In the light of stable performance of BFGS [30] as well as its high computational efficiency [3], we adopt BFGS algorithm to optimise object function (6).

4 Experiments

4.1 Datasets and Settings

Datasets – We evaluate the proposed framework on the public benchmark FG-NET [1, 4, 7, 14, 17, 19], which is the only dataset for age estimation having human annotations provided by Han *et al.* [8]. Specifically, the FG-NET dataset contains 82 persons varying from age 0 to age 69 with 1002 images in total and Fig. 3 shows all the images of the first identity. Evidently, the appearance of example faces are largely varied because of hairstyle, expression, beard style, whether wearing glasses and head poses, which makes the FG-NET dataset common and difficult for evaluating age estimation algorithm. Manual annotations for the FG-NET dataset have large uncertainty illustrated in Fig. 1, which makes apparent age estimation more challenging.

Features – Active Appearance Model (AAM) feature [31] is adopted as low-level imagery features because of its popularity in the recent works [1, 7, 14, 17, 19, 32, 33]. In details, the parameters of AAM model including visual appearance, shape, and texture cues to form a 200-dimensional feature vector.

Settings – Two experiments are conducted according to the settings of data split. In the first experiment, we followed the same leave-one-person-out setting as in [1, 7, 17, 19, 32, 33], whose testing images for each fold belong to an unseen person identity. In the second experiment, the total images of the FG-NET dataset was randomly split into 80% data for training and the remaining 20% for testing (*i.e.* 800 images for training and 202 images for testing) and we repeated the experiment 30 times.

Comparative Methods – We compare four algorithms with the proposed LALD framework, namely Instance-based PArtial Label learning (IPAL) [21], support vector regression with linear kernel (SVR) [34], two label distribute learning methods: CPNN [4] and BFGS-LLD [3]. IPAL can directly be applied to apparent age estimation with the capability of coping with multiple annotations for one instance, SVR, CPNN and BFGS-LLD employs pseudo age by averaging apparent age annotations to replace the true chronological ages. In IPAL, the number of nearest neighbours and the balancing coefficient are set to 5 and 0.45 respectively. Free parameter C in SVR to trade off the loss function and regularised term is tuned by four-fold cross-validation with $[10^{-5} : 10 : 10^{5}]$[1]. We set the size of hidden layers in CPNN to be 400.

Evaluation Metrics – Two evaluation metrics for chronological age estimation, namely *mean absolute error* (mae) and *cumulative score* (cs) [14] are not suitable for apparent age estimation because of unavailable unique chronological age labels for training and testing samples. In view of this, we adopt the evaluation metric $\in [0, 1]$ for each testing instance, introduced in [24]:

$$\epsilon = 1 - \exp\left(-\frac{(\hat{y} - \mu)^2}{2\sigma^2}\right)$$

where $\hat{y}$ denotes the predicted age having the maximum description degree in the predicted ALD, μ is the mean apparent age and σ denotes the standard deviation. For ϵ performance metric, the lower the better.

[1] Following the usage in Matlab, the notation $[x : y : z]$ represents an array starting from x to z with the step of y.

4.2 Comparative Evaluation with State-Of-The-Arts

Table 1: Comparative evaluation with state-of-the-art methods.

Methods	*Leave-One-Person-Out*	*Randomly 80% Training*
IPAL [21]	0.648±0.132	0.574±0.020
SVR [34]	0.611±0.135	0.588±0.027
CPNN [4]	0.676±0.126	0.660±0.043
BFGS-LLD [3]	0.618±0.111	0.587±0.019
LALD (ours)	**0.568±0.118**	**0.551±0.021**

In this section, we evaluate and compare the proposed LALD framework with four state-of-the-art algorithms in two data-split settings, which are shown in Table 1. Besides IPAL [21], SVR [34] CPNN [4], and BFGS-LLD [3] were designed for chronological age estimation and are applied to apparent age estimation by using the mean pseudo age generating from apparent age annotations. Evidently, our method achieves significantly better performance over the rest four comparative algorithms, in details, consistently at least 7.04% better for leave-one-person-out protocol and at least 4.01% for randomly 80% training data-split. The direct competitor of LALD is BFGS-LLD. Both methods employ the identical low-level features and optimisation method BFGS. Consequently, the performance gain achieved by LALD over BFGS-LLD can only be explained by the superiority of the proposed ambiguous label distribution. In addition, we conduct a t-test on the predictions of the state-of-the-art methods and ours. The results of both data-split protocol show consistently statistical significance (rejection on null hypothesis at the 5% significance level).

4.3 Evaluation on Pseudo Age Acquisition

Table 2: Evaluation on the strategies of pseudo age acquisition.

Methods	*Leave-One-Person-Out*	*Randomly 80% Training*
Mean Annotations	0.579±0.128	0.579±0.026
Maximum Majority Voting	0.592±0.121	0.575±0.030
Mean Classes	**0.568±0.118**	**0.551±0.021**

In this section, three types of strategies, mean apparent age annotations, maximum majority voting and mean age classes without counting the repeatability of annotations presented in Sec. 3.1 have been employed to determine pseudo age labels, which is an important factor for constructing the ambiguous label distribution. The results are illustrated in Table 2. Surprisingly, the strategy of mean classes beats the intuitive solutions of mean annotations and maximum majority voting, which are considered more robust against annotation outliers. The rational for such a phenomenon lies in the following observation. For age annotations having a roughly normal distribution (a peak in the middle), mean annotations, maximum majority voting and mean classes usually have the identical pseudo age. For tailed annotation density (a peak skewed to the boundary of annotated age range), mean classes is more robust against annotation outliers than the others, as the drift by repeated annotations leads to higher weights for outliers.

4.4 Weighted vs. Non-Weighted Combination

Table 3: Weighted vs. non-weighted combination to construct the ambiguous label distribution.

Methods	*Leave-One-Person-Out*	*Randomly 80% Training*
Non-Weighted	0.595±0.125	0.572±0.025
Weighted	**0.568±0.118**	**0.551±0.021**

This section compares the results of annotation reliability weighted and non-weighted combination for generating the ambiguous label distribution, which are given in Table 3. It is evident that weighted combination of individual label distribution from apparent age annotations can benefit to construct a more informative label representation, which can verify the motivation of the introduction of reliability weighting. It is worth mentioning here that even the results without adopting annotation reliability as weights can also beat all four state-of-the art algorithms, which further demonstrates the effectiveness of the proposed LALD framework for apparent age estimation.

4.5 Evaluation on Reliability Utilisation

This section evaluates on what kind of information is more favourable for capturing annotation reliability, with the results illustrated in Table 4. We compare the first- and second- order distance from apparent age annotations to pseudo age, and find out that the first order statistics can consistently perform better on both data-split settings. Such an observation indicates that the choice of reliability weighting is sensitive to estimation performance. Nevertheless, the performance achieved by employing second order statistics is still superior to all four comparative methods in Table 1.

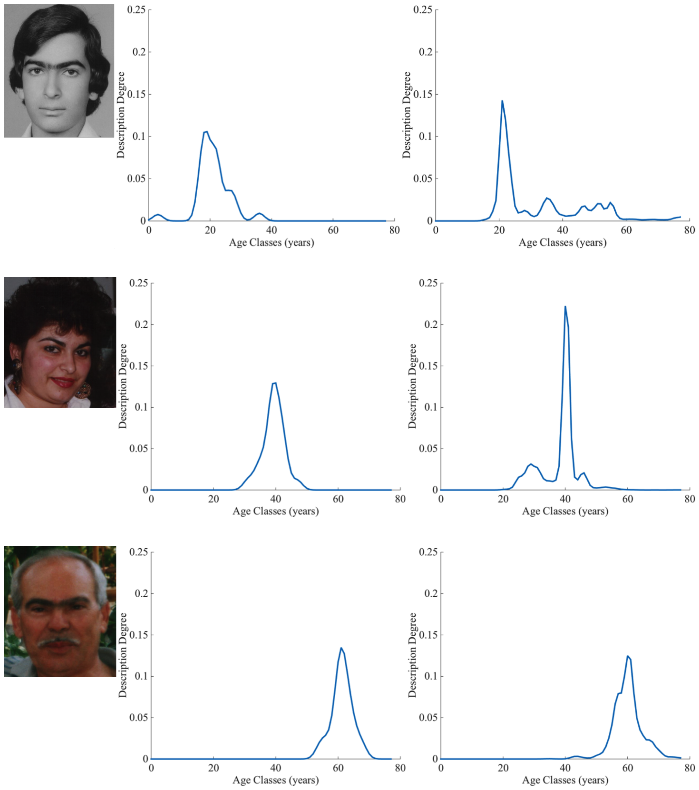

Fig. 4: Illustrative examples about the proposed LALD framework. In each subfigure, left is the original facial image, the middle is the generated ambiguous label distribution from apparent age, the right is the predicted distribution.

4.6 Illustrative Samples

In this section, we illustrate a number of successful sampled facial images belonging to young, mid-age and old age groups respectively from the FG-NET benchmark in Fig. 4, with the generated ambiguous label distribution from annotations and the predicted label distribution. Evidently, compared to single peak and symmetric distribution in original label distribution, the ambiguous label distribution could have multiple peaks and asymmetric distribution, which is more informative and robust for apparent age estimation to capture the label ambiguity.

5 Conclusion

This paper proposes a novel concept of ambiguous label distribution designed for apparent age estimation problem, whose instances have multiple unreliable age

Table 4: Evaluation on the choices of reliability utilisation.

Methods	*Leave-One-Person-Out*	*Randomly 80% Training*
First Order	**0.568±0.118**	**0.551±0.021**
Second Order	0.595±0.117	0.565±0.032

annotations. Owing to discovering latent label correlation inherited from original label distribution learning framework and utilising annotation reliability for weighting apparent age, the proposed learning with ambiguous label distribution method can achieve better and robust performance. We experimentally evaluate and analyse the variants of our algorithm, and notice that combining distributions adopted in this paper is important yet naive. In future, introducing a powerful combination method could be a promising research direction.

Acknowledgement

This work was funded by Academy of Finland under the Grant No. 267581 and 298700, and D2I SHOK project funded by Digile Oy and Nokia Technologies (Tampere, Finland). The authors wish to acknowledge CSC-IT Center for Science, Finland, for generous computational resources.

References

1. Chen, K., Gong, S., Xiang, T., Loy, C.C.: Cumulative attribute space for age and crowd density estimation. In: CVPR. (2013)
2. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: a survey. TPAMI (2010)
3. Geng, X., Ji, R.: Label distribution learning. In: ICDMW. (2013)
4. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. TPAMI (2014)
5. Luu, K., Ricanek Jr, K., Bui, T.D., Suen, C.Y.: Age estimation using Active Appearance Models and support vector machine regression. In: BTAS. (2009)
6. Pontes, J.K., Britto, A.S., Fookes, C., Koerich, A.L.: A flexible hierarchical approach for facial age estimation based on multiple features. Pattern Recognition (2015)
7. Zhang, Y., Yeung, D.: Multi-tasks warped Gaussian process for personalized age estimation. In: CVPR. (2010)
8. Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. TPAMI (2015)
9. Yang, X., Gao, B.B., Xing, C., Huo, Z.W., Wei, X.S., Zhou, Y., Wu, J., Geng, X.: Deep label distribution learning for apparent age estimation. In: CVPR Workshops. (2015)
10. Rothe, R., Timofte, R., Gool, L.: Dex: Deep expectation of apparent age from a single image. In: ICCV Workshops. (2015)

11. Kuang, Z., Huang, C., Zhang, W.: Deeply learned rich coding for cross-dataset facial age estimation. In: ICCV Workshops. (2015)
12. Zhu, Y., Li, Y., Mu, G., Guo, G.: A study on apparent age estimation. In: ICCV Workshops. (2015)
13. Antipov, G., Baccouche, M., Berrani, S.A., Dugelay, J.L.: Apparent age estimation from face images combining general and children-specialized deep learning models. In: ICCV Workshops. (2015)
14. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. TPAMI (2007)
15. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. TSMC (2004)
16. Geng, X., Wang, Q., Xia, Y.: Facial age estimation by adaptive label distribution learning. In: ICPR. (2014)
17. Guo, G., Fu, Y., Huang, T.S., Dyer, C.R.: Image-based human age estimation by manifold learning and locally adjusted robust regression. TIP (2008)
18. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR. (2009)
19. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: CVPR. (2011)
20. Wang, S., Tao, D., Yang, J.: Relative attribute SVM+ learning for age estimation. TC (2015)
21. Zhang, M.L., Yu, F.: Solving the partial label learning problem: an instance-based approach. In: IJCAI. (2015)
22. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. JMLR (2011)
23. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: CVPR. (2009)
24. Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H., Guyon, I.: Chalearn 2015 apparent age and cultural event recognition: datasets and results. In: ICCV, ChaLearn Looking at People workshop. (2015)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
26. MacKay, D.J.: Information theory, inference and learning algorithms. Cambridge university press (2003)
27. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics (1996)
28. Pietra, S.D., Pietra, V.D., Lafferty, J.: Inducing features of random fields. TPAMI (1997)
29. Nocedal, J., Wright, S.: Numerical optimization. Springer Science & Business Media (2006)
30. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: The 6th Conference on Natural Language Learning. (2002)
31. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. TPAMI (2001)
32. Yan, S., Wang, H., Huang, T.S., Yang, Q., Tang, X.: Ranking with uncertain labels. In: ICME. (2007)
33. Yan, S., Wang, H., Tang, X., Huang, T.S.: Learning auto-structured regressor from uncertain nonnegative labels. In: ICCV. (2007)
34. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing (2004)