

# Facial Age Estimation Using Robust Label Distribution

Ke Chen and Joni-Kristian Kämäräinen  
Department of Signal Processing  
Tampere University of Technology  
Tampere, Finland  
firstname.lastname@tut.fi

Zhaoxiang Zhang  
Institute of Automation  
Chinese Academy of Sciences  
Beijing, P.R. China

## ABSTRACT

Facial age estimation, to predict the persons' exact ages given facial images, usually encounters the data sparsity problem due to the difficulties in data annotation. To mitigate the suffering from sparse data, a recent label distribution learning (LDL) algorithm attempts to embed label correlation into a classification based framework. However, the conventional label distribution learning framework only considers correlations across the neighbouring variables (ages), which omits the intrinsic complexity of age classes during different ageing periods (age groups). In the light of this, we introduce a novel concept of robust label distribution for scalar-valued labels, which is designed to encode the age scalars into label distribution matrices, i.e. two-dimensional Gaussian distributions along age classes and age groups respectively. Overcoming the limitations of conventional hard group boundaries in age grouping and capturing intrinsic inter-group dependency, our framework achieves robust and competitive performance over the conventional algorithms on two popular benchmarks for human age estimation.

## Keywords

facial age estimation; hard group boundaries; robust label distribution learning (RLDL)

## 1. INTRODUCTION

Facial age estimation [5, 9, 10, 14, 21, 22, 27] is to predict persons' age given their facial images, which is a hot yet challenging topic in computer vision. This problem and other similar problems such as crowd counting [2, 5, ?, 6, 7] are aimed to learn a mapping function from imagery feature representation to scalar-valued targets. However, due to inherent ambiguities in age annotation, data sparsity problem is encountered with the lack of sufficient samples for covering the whole data distribution. For example, a large number of facial images can be readily found on the Internet, but reliable annotation of the exact age of images is usually lacking, which leads to sparsely distributed data [5] in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '16, October 15-19, 2016, Amsterdam, Netherlands

© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2964284.2967186>

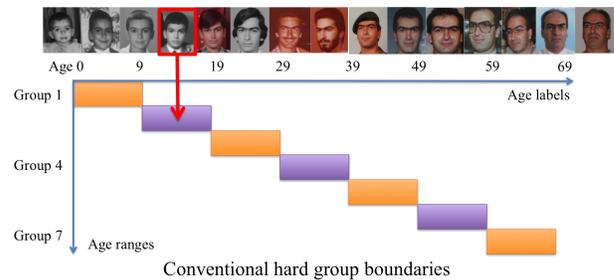


Figure 1: Conventional hard boundaries between age groups.

the public benchmarks such as the FG-NET and MORPH datasets. Sparse data distribution can increase the difficulty in learning a robust mapping function, which thus leads to poor performance of the existing algorithms especially for those age classes without sufficient training data.

Human age estimation is usually formulated into a multi-class classification problem [15, 18] or a single-variate regression problem [5, 16, 17, 27]. Compared to classification based algorithms, regression frameworks are more favourable owing to the inherently cumulative and dependent characteristics of facial age estimation, i.e. the closer age labels of two facial images are, the more visual similarity they share. Recently, label distribution learning (LDL) was proposed by Geng *et al.* [13, 14] to represent each target scalar with a vector-formed label distribution which represents label correlation in classification frameworks in the fashion of multi-labels. Compared to 1-of- $K$  target coding in the classic one-vs-all multi-class classification, label distribution vectors have positive values for neighbouring samples instead of zero, which can contribute to transferring the knowledge of neighbouring classes to those age classes without any training samples. For intuitive explanation, let us take one certain age (e.g. age 40) as an example, which can be learnt indirectly from the samples of neighbouring ages such as 39 and 41, even when age 40 has few or even no training samples, as element-wise difference of predicted and ground truth label distributions are designed to enforce to the minimal. In other words, the entry for age 40 class in label distribution can also be trained when training samples with the positive value on the element for age 40 class in LDL (e.g. 39 or 41). In the light of this, label distribution learning can mitigate the suffering from sparse data.

Assuming a fixed age label distribution, the conventional label distribution learning (LDL) framework [14] is limited

by intrinsic complex relationship across age classes because of inconsistent ageing patterns as well as sensitiveness to personal identities and head poses. Adaptive label distribution learning [11] was proposed to model the inconsistent ageing patterns by iteratively updating dynamic label distribution for each age group, but it still adopted and thus suffered from hard boundaries between chronological age groups. We observe that facial images for age estimation have inter-group correlation during different ageing periods (age groups), which have not been exploited in label distribution learning. Specifically, inter-group correlation can be considered to reflect the *global* piecewise-like relationship of age labels in terms of localised chronological age groups corresponding to varying ageing periods. Moreover, conventional hard age group boundaries shown in Figure 1 were widely employed [11, 20, 21, 22], which relied heavily on the accuracy of age group classification. In other words, misclassification on coarse age groups can lead to significant performance reduction on the following fine age estimation. A naive solution [22] to improve the robustness led by age group misclassification is to employ the overlapping age ranges but such a strategy is data-specific sensitive to the size of the overlapped age ranges.

Motivated by the aforementioned observation, we introduce a novel concept of robust label distribution for single-variate scalar-valued problems, e.g. facial age estimation, which can simultaneously address both challenges in a unique framework. In the proposed framework, we attempt to jointly minimise the estimation errors along both age class and group distribution in order to improve the robustness of the conventional label distribution learning framework. Owing to the introduction of an additional group distribution dimension in our framework, both suffering of cross-group dependency and hard group boundaries are alleviated in addition to the capability to cope with data sparsity inherited from conventional LDL framework. Experiments on two popular age estimation benchmarks verify the effectiveness of our framework with superior performance over conventional algorithms.

## 2. RELATED WORK

The recent methods for estimating human age given facial images can be divided into three categories: classification based [15, 18], regression based [16, 17, 27], and ranking based [4, 23]. Facial age estimation was firstly formulated into a multi-class classification problem, but those classification based algorithms [15, 18] omitted the latent correlation across age labels, i.e. the closer age labels of facial images are, the more visually similar they are as well. Regression based frameworks [5, 16, 17, 27] are usually considered more suitable for age estimation owing to cumulative dependent nature of regression labels. Recently, a ranking based method, namely OHRank, was proposed [4] by utilising a series of ordinal hyperplane (i.e. binary classifiers) with all the data in the dataset in order to mitigate data sparsity problem, but it is extremely expensive to train such a model. Chen *et al.* [5] exploited the cumulative nature across age classes to achieve robust performance in a two-layer attribute learning framework. Geng *et al.* [14] designed a framework by learning from label distributions in the manner of multi-label learning. For each instance, a label distribution (e.g. Gaussian distribution) was first generated to give the degree that the neighbouring class labels describe

the instance. Such a learning paradigm designed for data sparsity problem can improve the performance owing to the reduction of ambiguity in the label space by using multi-labels instead of a single label. However, learning with label distribution [14] (or termed, learning with soft labels) only considered increasing the global discrimination by incorporating consistent local label relations, which is not valid because of dynamic ageing procedure. An advanced attempt of label distribution learning with adaptively updating label distribution for each age group was introduced in [11], but hard group boundaries can still have a negative effect on performance. As a result, to capture complex relationship across ages, we propose a framework to learn with over-complete robust label distribution (RLD) for age estimation in the fashion of soft group boundaries to incorporate inter-dependency across groups and improve the robustness of age grouping. Our performance is superior to the conventional approaches on two public benchmarking datasets justifying the usage of robust label distributions.

## 3. METHODOLOGY

Given imagery feature representation  $\mathbf{x}$  and its corresponding age label  $y$ , training samples consist of  $\{\mathbf{x}_i, y_i\}_{i=1,2,\dots,N}$ , where  $N$  denotes the number of training samples. The pipeline of the proposed algorithm is given as follows:

- We first construct a two-dimensional robust label distribution  $\mathbf{L}$  for each training sample according to its relative position  $y$  in a “chronological age map”, which adopts different Gaussian distributions along age labels and groups (See Sec. 3.1).
- Learning the mapping between imagery feature representation  $\mathbf{x}$  and a robust label distribution  $\mathbf{L}$  is achieved by employing multivariate distribution learning in Sec. 3.2.

During testing, imagery features of an unseen image are fed into the trained model to estimate the person’s age with the maximum predicted description degree along age classes in robust label distribution.

### 3.1 Robust Label Distribution

We first investigate the framework about one dimensional label distribution along age classes proposed by Geng *et al.* [14] to lay the basis for further discussion. Given a scalar-valued age label  $y \in \mathbb{R}$ , a label distribution vector  $\mathbf{y} \in \mathbb{R}^K$  with the sum of description degree  $d_{\mathbf{x}}$  (real values) of all the labels equal to one can be obtained, where  $K$  denotes the size of age range. The assumption of label distribution is two-fold: a) true labels have the highest description degree in  $\mathbf{y}$ ; and b) the farther labels are away from true labels, the lower description degree they have. It is noted that non-zero description degree in a label distribution vector reflects ambiguity of feature-label relationship, which also indicates the importance of neighbouring labels contributing to the exact label  $y$  associated to instance  $\mathbf{x}$ . In this sense, even without sufficient training samples, label distribution learning can also achieve robust performance. Typical label distributions are Gaussian and triangle distributions [14], and Gaussian distribution can keep consistently superior performance to triangle distribution, which encourages us to employ Gaussian distribution in the experiments of this paper.

Aiming to incorporate inter-correlation across different ageing patterns and improve the robustness against hard

age grouping, we propose an over-complete two-dimensional label distribution, namely robust label distribution, for a scalar-valued target. Conventional label distribution for human age estimation in [11, 14] is designed only along the chronological age dimension, but the proposed RLD extend an extra dimension along age groups. To construct a RLD  $\mathbf{L} \in \mathbb{R}^{K \times D}$  for each instance  $\mathbf{x}$  with  $D$  denoting the number of age groups, we first divide the whole age range into  $D$  subsets by their chronological ages via either manually-defined or clustering. Here we adopt manual definition as it is straightforward to define age groups  $\mathcal{G} = \{G_1, G_2, \dots, G_D\}$  (typically each  $[-5, +5]$  interval forms a group). Consequently, the scalar label  $y$  have extended to a two-dimensional  $\bar{\mathbf{y}}$  composed by its age label and group label. As mentioned before that Gaussian distribution is preferred in [14], we adopt a discretized bivariate Gaussian distribution [12] anchored at the extended label  $\bar{\mathbf{y}}$  in RLD:

$$d_{\mathbf{x}}(\mathbf{z}) = \frac{1}{2\pi\sqrt{|\Sigma|}F} \exp\left(-\frac{1}{2}(\mathbf{z} - \bar{\mathbf{y}})^T \Sigma^{-1}(\mathbf{z} - \bar{\mathbf{y}})\right),$$

where the variable  $\mathbf{z}$  is other labels in RLD besides  $\bar{\mathbf{y}}$  and  $F$  is a normalisation factor to make the sum of  $d_{\mathbf{x}}$  be one. Considering different characteristics of distribution along age classes and groups, we set the covariance matrix  $\Sigma \in \mathbb{R}^{2 \times 2}$  as a diagonal matrix with two finest granularity for age classes and groups respectively. Evidently, with the introduction of a group distribution in RLD, the following advantages are achieved: 1) utilise interdependency across age groups reflecting different ageing patterns with multi group labels instead of a single label; and 2) attempt to replace hard group boundaries with soft group distribution to improve robustness.

### 3.2 Learning with RLD

With the generated RLD for each image, the training set becomes  $\{\mathbf{x}, \mathbf{L}\}_i, i = 1, 2, \dots, N$ . Element  $\mathbf{L}_{jk}, j = 1, 2, \dots, K, k = 1, 2, \dots, D$  of  $\mathbf{L} \in \mathbb{R}^{K \times D}$  consists of  $j$ th age class label and  $k$ th group label. The aim is to learn a conditional density function  $p(\bar{\mathbf{y}}|\mathbf{x}; \boldsymbol{\theta})$  to minimise the distance between the predicted  $\hat{\mathbf{L}}$  generated by  $\boldsymbol{\theta}$  and the ground truth  $\mathbf{L}$ , where  $\boldsymbol{\theta}$  is the parameter vector to be optimised. It is evident that the learning with robust label distribution becomes a multi-variate distribution learning problem, which has been well presented in [12].

The object function for robust label distribution learning can be written as:

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

where weighted Jeffrey's divergence  $P(\mathbf{L}_a || \mathbf{L}_b)$  is to measure the similarity between two distributions  $\mathbf{L}_a$  and  $\mathbf{L}_b$  with considering inter-element relationship, which can be formulated as the following:

$$P(\mathbf{L}_a || \mathbf{L}_b) = \sum_{w,h} \lambda_{w,h} (\mathbf{L}_a^w - \mathbf{L}_b^h) \ln \frac{\mathbf{L}_a^w}{\mathbf{L}_b^h}, \quad (2)$$

where  $\mathbf{L}_a^w$  and  $\mathbf{L}_b^h$  denote the  $w$ th and  $h$ th element in  $\mathbf{L}_a$  and  $\mathbf{L}_b$  respectively, and  $\lambda$  is used to weight the relationship across elements. Evidently  $\lambda$  is an important factor to model underlying correlation of neighbourhood among ages. In this paper, we adopt the same  $\lambda$  setting in [12] to reflect localised correlation of the nine neighbouring points in robust label

distribution space as

$$\lambda_{jkmn} = \frac{1}{F} \exp\left(\frac{\|\bar{\mathbf{y}}_{jk} - \bar{\mathbf{y}}_{mn}\|^2}{-\delta}\right), \quad (3)$$

where  $F$  is normalised term,  $\bar{\mathbf{y}}_{jk}$  and  $\bar{\mathbf{y}}_{mn}$  are  $jk$ th corresponding label in the ground truth RLD  $\mathbf{L}$  and  $mn$ th label in the predicted RLD  $p(\bar{\mathbf{y}}_i | \mathbf{x}_i; \boldsymbol{\theta})$  respectively.  $\delta$  is a parameter to adjust influence of the distance between  $\bar{\mathbf{y}}_{jk}$  and  $\bar{\mathbf{y}}_{mn}$ .

Let us assume a maximum entropy model [1] as

$$p(\bar{\mathbf{y}}_{mn} | \mathbf{x}_i; \boldsymbol{\theta}) = \frac{\exp(\sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r)}{\sum_{m,n} \exp(\sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r)}, \quad (4)$$

where  $\mathbf{x}_i^r$  denotes the  $r$ th entry of feature  $\mathbf{x}_i$  and  $\boldsymbol{\theta}_{mn,r}$  is the element of  $\boldsymbol{\theta}$  associated to the label  $\bar{\mathbf{y}}_{mn}$  and  $r$ th feature element. Substituting Equations (2) and (4) into (1) reformulates the object function as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_i \ln \left( \sum_{m,n} \exp \left( \sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r \right) \right) + \\ & \sum_{i,j,k,m,n} \lambda_{jkmn} \left[ \frac{\exp(\sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r)}{\sum_{m,n} \exp(\sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r)} \times \right. \\ & \left. \left( \sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r - \ln \left( \sum_{m,n} \exp \left( \sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r \right) \right) - \right. \right. \\ & \left. \left. \ln d_{\mathbf{x}_i}(\bar{\mathbf{y}}_{jk}) \right) - d_{\mathbf{x}_i}(\bar{\mathbf{y}}_{jk}) \sum_r \boldsymbol{\theta}_{mn,r} \mathbf{x}_i^r \right], \end{aligned}$$

which can be solved by using limited-memory quasi-Newton method L-BFGS introduced in [19]. The advantages of L-BFGS is to avoid the expensive Hessian matrix inversion but iteratively approximate its inverse.

## 4. EXPERIMENTS

**Datasets and Settings** – For evaluating the proposed framework, two public benchmarks, FG-NET [4, 5, 15, 16, 27] and MORPH [4, 5, 15], were used. Specifically, the FG-NET dataset contains 82 persons varying from age 0 to age 69 with 1002 images in total, while the MORPH dataset covers the range of age 16 to age 77 with 5475 images<sup>1</sup>. Active Appearance Model (AAM) feature [8] is adopted as low-level imagery features because of its popularity in the recent works [4, 5, 15, 16, 25, 26, 27]. We followed the same leave-one-person-out setting as in [4, 5, 16, 25, 26, 27] for FG-NET, while the MORPH dataset was randomly split into 80% data for training and the remaining 20% for testing and we repeated the experiments 30 times as in [4, 5]. We employed two evaluation metrics, namely *Mean Absolute Error* (MAE) and *Cumulative Score* (CS) with the error level 5 as in [4, 5, 15].

**Comparison to State-of-the-Art** – Comparative results on two benchmarks are given in Table 1. Among those algorithms, RUN [26], Ranking [25], LARR [16], LSLR [24], SVR [16], CA-SVR [5] are regression based and RED-SVM [3] and OHRank [4] are ranking based, while the rest are classification based. All comparative evaluation are using the identical AAM features except AGES [15]. Evidently, our robust label distribution learning can consistently achieve the best performance over state-of-the-arts especially label distribution learning based frameworks including IIS-LLD, CPNN,

<sup>1</sup>The size of MORPH has increased to 55,608 images (MORPH-II), but we use the original MORPH dataset for fair comparison with the existing algorithms [3, 4, 5, 15, 16, 27].

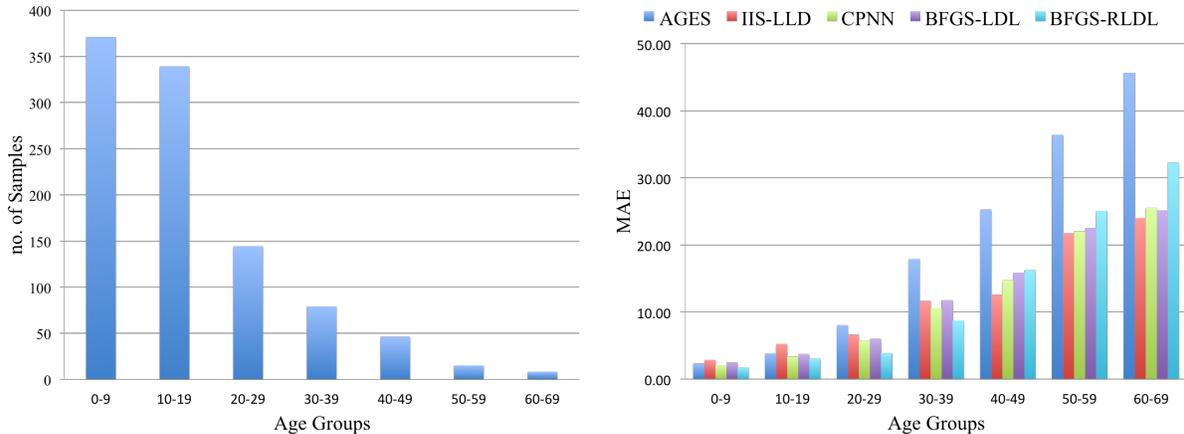


Figure 2: Data distribution (left plot) and comparative performance (right plot) on the FG-NET benchmark of the proposed BFGS-RLDL with AGES [15], IIS-LLD [13], CPNN [14], and BFGS-LDL [10] in different age range.

Table 1: Age estimation performance comparison.

Methods	FG-NET [15]		MORPH [4]	
	MAE	CS	MAE	CS
AGES [15]	6.77	–	8.83	–
RUN [26]	5.78	–	–	–
Ranking [25]	5.33	–	–	–
RED-SVM [3]	5.24	–	6.49	–
LARR [16]	5.07	–	–	–
MTWGP [27]	4.83	–	6.28	–
LSLR [24]	5.25	–	–	–
OHRank [4]	4.85	74.4%	<b>5.69</b>	56.3%
SVR [16]	5.66	68.0%	5.77	57.1%
CA-SVR [5]	4.67	74.5%	5.88	57.9%
IIS-LLD [13]	5.77	–	–	–
CPNN [14]	4.76	–	–	–
BFGS-LDL [10]	5.23	69.4%	5.94	56.5%
BFGS-RLDL	<b>4.27</b>	<b>76.2%</b>	<b>5.69</b>	<b>59.2%</b>

and BFGS-LDL on both benchmarks. The most direct effect of using robust label distribution can be observed by comparing between BFGS-LDL [10] and BFGS-RLDL, while BFGS-RLDL significantly beat BFGS-LDL. Since using the identical low-level AAM features and using the same optimisation algorithm BFGS, the performance improvement can only be explained by superior robust label distribution proposed in this paper.

**Evaluation on Different Age Ranges** – We visualise the data distribution and also comparative performance of a number of algorithms on the FG-NET benchmark in Figure 2. It can be observed that, compared to LDL based algorithms (IIS-LLD [13], CPNN [14], and BFGS-LDL [10]), the proposed RLDL can reduce errors when the size of training samples for a specific group is sufficient and yet achieve comparable performance for sparse data. The explanation for such a phenomenon is that soft group distribution for sparse data can be different from the dense data. With imbalanced data, our model is enforced to drift to the young ages during optimisation for the lower mean error over all samples. This observation inspired us to improve the design for age grouping to reflect the dynamic ageing patterns.

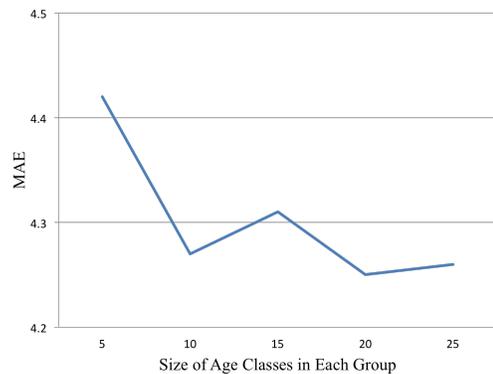


Figure 3: Effect with varying size of age classes in each group on the FG-NET dataset.

**Effect with Varying Age Group Size** – The size of age groups is an important parameter to decide the complexity of label distribution in view of the dimension of RLD linearly proportional to the size of age groups. Figure 3 illustrates the effect on the varying age group size. Figure 3 shows that the size of age groups plays an important role on the success of the algorithm. However, even the worst results of RLDL in Figure 3 are still superior to the existing algorithms, which further demonstrates its effectiveness.

## 5. CONCLUSIONS

This paper proposes a novel over-complete label distribution learning framework for visual regression. For facial age estimation our framework achieves superior accuracy as compared to conventional methods with hard age class and group boundaries. In our future work, unsupervised construction of the label distribution and more advanced design on incorporating group information into label distribution will be investigated.

## 6. ACKNOWLEDGMENTS

This work was funded by the Academy of Finland under Grant No. 267581.

## 7. REFERENCES

- [1] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 1996.
- [2] A. B. Chan and N. Vasconcelos. Counting people with low-level features and bayesian regression. *TIP*, 2012.
- [3] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. A ranking approach for human ages estimation based on face images. In *ICPR*, 2010.
- [4] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, 2011.
- [5] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013.
- [6] K. Chen and J.-K. Kämäräinen. Learning to count with back-propagated information. In *ICPR*, 2014.
- [7] K. Chen and J.-K. Kämäräinen. Pedestrian density analysis in public scenes with spatio-temporal tensor features. *TITS*, 2016.
- [8] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012.
- [9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001.
- [10] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: a survey. *TPAMI*, 2010.
- [11] X. Geng and R. Ji. Label distribution learning. In *ICDMW*, 2013.
- [12] X. Geng, Q. Wang, and Y. Xia. Facial age estimation by adaptive label distribution learning. In *ICPR*, 2014.
- [13] X. Geng and Y. Xia. Head pose estimation based on multivariate label distribution. In *CVPR*, 2014.
- [14] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. In *AAAI*, 2010.
- [15] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *TPAMI*, 2014.
- [16] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *TPAMI*, 2007.
- [17] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer. Image-based human age estimation by manifold learning and locally adjusted robust regression. *TIP*, 2008.
- [18] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009.
- [19] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *TSMC*, 2004.
- [20] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 1989.
- [21] K.-H. Liu, S. Yan, and C.-C. J. Kuo. Age estimation via grouping and decision fusion. *TIFS*, 2015.
- [22] K. Luu, K. Ricanek Jr, T. D. Bui, and C. Y. Suen. Age estimation using Active Appearance Models and support vector machine regression. In *BTAS*, 2009.
- [23] J. K. Pontes, A. S. Britto, C. Fookes, and A. L. Koerich. A flexible hierarchical approach for facial age estimation based on multiple features. *PR*, 2015.
- [24] S. Wang, D. Tao, and J. Yang. Relative attribute SVM+ learning for age estimation. *TC*, 2015.
- [25] C. Yan, C. Lang, and S. Feng. Facial age estimation based on structured low-rank representation. In *ACM MM*, 2015.
- [26] S. Yan, H. Wang, T. S. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *ICME*, 2007.
- [27] S. Yan, H. Wang, X. Tang, and T. S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV*, 2007.
- [28] Y. Zhang and D. Yeung. Multi-tasks warped Gaussian process for personalized age estimation. In *CVPR*, 2010.