Spectral Attribute Learning for Visual Regression

Ke Chen[a,*], Kui Jia[b], Zhaoxiang Zhang[c], Joni-Kristian Kämäräinen[a]

[a]*Department of Signal Processing, Tampere University of Technology, Tampere, Finland*
[b]*Faculty of Science and Technology, University of Macau, Macau, China*
[c]*Institute of Automation, Chinese Academy of Sciences, Beijing, China*

Abstract

A number of computer vision problems such as facial age estimation, crowd counting and pose estimation can be solved by learning regression mapping on low-level imagery features. We show that visual regression can be substantially improved by two-stage regression where imagery features are first mapped to an attribute space which explicitly models latent correlations across continuously-changing output. We propose an approach to automatically discover "spectral attributes" which avoids manual work required for defining hand-crafted attribute representations. Visual attribute regression outperforms direct visual regression and our spectral attribute visual regression achieves state-of-the-art accuracy in multiple applications.

Keywords: Facial age estimation, Crowd counting, Head Pose Estimation, Spectral learning, Attributes, Regression

1. Introduction

Visual regression maps imagery features to a continuous output space and is therefore a suitable tool for a number of computer vision applications such as facial age estimation, pedestrian counting and object pose estimation. These vision problems can also be formulated as multi-class classification problems (*e.g.* [1]), but recently visual regression based methods that naturally exploit continuous scalar-valued output label spaces have achieved superior results in important applications (*e.g.* age estimation [2, 3, 4], pedestrian density estimation [5, 4] and human body/face pose estimation [6, 7, 8]).

[*] Tel.: +358-466459305; Fax.: +358-33641352.
Email address: ke.chen@tut.fi (Ke Chen)

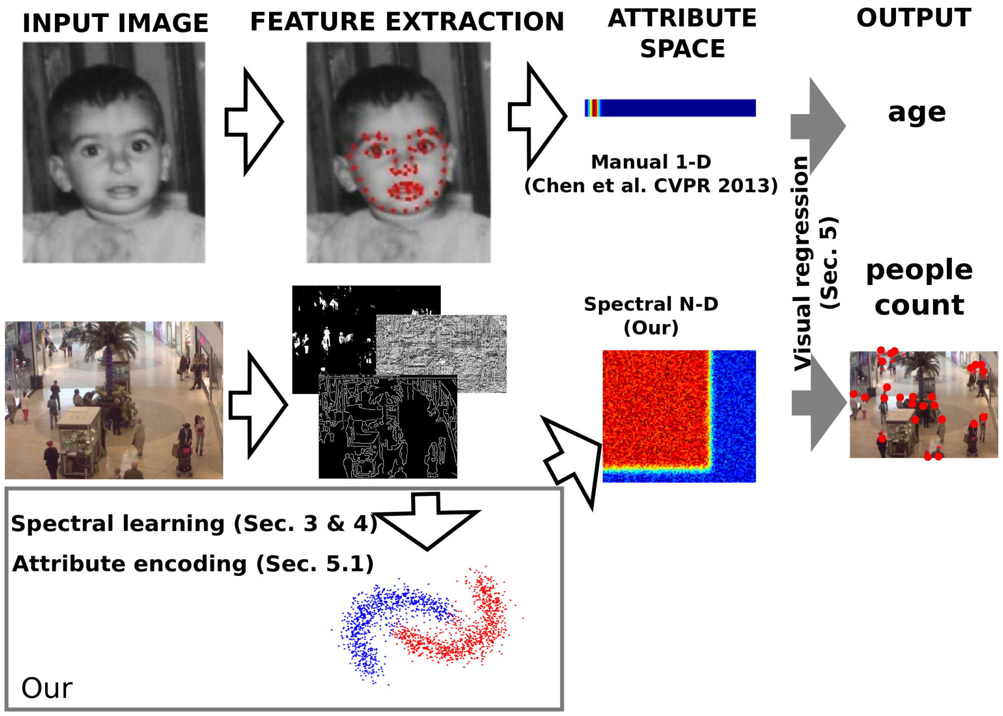


Figure 1: The proposed spectral attribute visual regression. First imagery features are extracted from input images. Imagery features are then mapped to an attribute space - in our method to automatically discovered spectral attributes - and final target value regression is made from the attribute space to the target output space.

The regression approaches exploit the inherent dependencies in the label space, but there are other major challenges in visual regression that degrade the performance: (1) large visual appearance variation in captured images and extracted features (background clutter, occlusions, illumination and view point changes); (2) ambiguous correspondence between observed imagery data and annotated labels (for example, the apparent difficulty to estimate the age from a face image); (3) data imbalance (for example, young people are over-presented in age estimation datasets produced by higher education institutes). Visual regression becomes extremely challenging as the above challenges co-exist. In the light of this, the best methods are often specifically tailored for the target application, *e.g.* by using advanced feature extraction [7, 9] or robustified estimation [2, 3]. However, the recent works [7, 10] on visual regression provided generic tools that boost and robustify regression with minor computational or manual tuning burden. Similar to the introduction of visual attributes in classification [11, 12, 13], a regression problem can be cast as a two-step process where inputs are mapped to the output space via hand-crafted attribute space that explicitly captures inherent label dependencies to improve the robustness against feature inconsistency and data imbalances and sparsity [4].

Alternatively, we adopt and extend the recent tools for visual regression to establish a unified visual regression approach that is free of manual work (attribute hand-crafting) and can be applicable to multivariate regression problems (Figure 1). We construct an effective attribute space, which we call *spectral attributes*, by using the available regression labels in spectral learning [14] which is a supervised extension of spectral clustering. Our spectral attribute learning for visual regression achieves state-of-the-art accuracy on multiple facial age estimation, crowd counting, and head pose estimation benchmarks.

2. Related Work

Visual regression – Visual regression is suitable for computer vision applications where the output space is continuous such as facial age estimation [2, 3, 4], pedestrian counting [5, 4] and human body or face pose estimation [6, 7, 8]. The existing approaches can be divided to two categories: *estimation-by-classification* (*e.g.* [1]) and *estimation-by-regression* (*e.g.* [9]). There are several works based on other approaches (*e.g.* relative ranking in [15]), but our focus is on generic visual regression approaches. Due to the well-known challenges with visual data the application specific solutions often perform the best, *e.g.* tailored feature extraction stage [9] or the whole regression pipeline [3]. Our approach is inspired by processing stages and methods that have been proposed for specific tasks [4, 10, 16], but which seem to provide generic tools as indicated by the experimental results in our work. We adopt the two-stage hierarchy introduced by Chen *et al.* [4]. However, our work differs from theirs rather substantially in the sense that our attribute space construction requires no manual hand-crafting of suitable mappings, but learns from training data. We also extend our mid-level spectral attribute representation to multiple output variables.

Attribute learning – Visual attributes were introduced by Ferrari and Zisserman [11] who trained classifiers for manually selected visual attributes. Their seminal work inspired various attribute learning methods [12, 17, 18], but attribute learning has received much less attention. Several works propose to use auxiliary information for attribute construction. For example, Berg *et al.* [19] mine attributes from text descriptions on the Web page. Computational attributes have been used in classification tasks based on support vector machines [20, 21] and random fields [22, 23]. To the authors' best knowledge, our work is the first attempt to automatic attribute discovery and exploitation for visual regression.

Contributions – The novel contributions of our work are illustrated in Figure 1 and listed as:

- We propose a novel spectral attribute, which for the first time investigates and utilizes global statistics in the continuous label space.
- To capture latent label ordinal dependency, we develop an ordinal label constrained spectral learning for single-variate regression and also its multivariate variant in this paper.
- We report experimental results on several popular benchmarks for which our method achieves state-of-the-art performance.

3. Feature Similarity

Unsupervised attribute construction in our work is implemented using spectral learning (Sec. 4) which requires computation of an affinity matrix. We compute a $N \times N$ affinity matrix from pairwise visual similarities. There are recent works on computational similarity functions by discriminative learning [24] and generative learning [21], but the computational (learning) similarity metrics require dedicated training process. More explicit similarity measures have been successfully used in object matching [25], searching sketches from images [26] and in non-rigid matching of image sequences [27]. Feature-based similarity of two images I_a and I_b is measured in two ways: how similar feature points are to their corresponding feature points ("matching"), and how much the spatial arrangement of the feature points is changed ("distance"). The total similarity is formulated as a cost function to be minimized:

$$C(I_a, I_b) = \lambda_1 C_{match}(F^{(a)}, F^{(b)}) + \lambda_2 C_{dist.}(\boldsymbol{x}^{(a)}, \boldsymbol{x}^{(b)}) \tag{1}$$

where $F^{(\cdot)}$ denote features and $x^{(\cdot)}$ their spatial locations. In regression problems a simpler form of the above can succeed as images are pre-aligned (facial age estimation) or the camera is fixed (crowd counting) and therefore the geometric term can be omitted. The cost function reduces to the C_{match} for which the standard distance functions provide computationally attractive solutions

$$C(I_a, I_b) = D^F(F^{(a)}, F^{(b)}) = ||F^a - F^b|| \ . \tag{2}$$

We can map the distance (cost) (2) in $[0, \infty]$ to similarity values in $[0, 1]$ by the exponential function

$$\boldsymbol{S}(I_a, I_b) = e^{-\lambda_1 \boldsymbol{C}(I_a, I_b)} = e^{-\lambda_1 \boldsymbol{D}^F(F^{(a)}, F^{(b)})} \tag{3}$$

where λ_1 defines the similarity decay rate from the exact match. The decay term depends on the selected feature space and affects to the success of spectral clustering [28]. It is noteworthy, that despite the fact that we adopt the simplified similarity in (3) it can be easily extended to spatial dimension by the feature pyramid [29] or we can switch to the explicit definition in (1) by adapting [25, 26, 27].

4. Spectral Attributes

Spectral clustering [28] is one of the commonly used clustering algorithms, which is based on the spectral analysis of a pairwise affinity (similarity) matrix. With careful parameter and similarity function selection the spectral clustering is a powerful unsupervised tool [28, 30], but not readily applicable in our case since it does not exploit the continuous label values available for regression. However, a solution for the supervised case - “spectral learning” - was proposed by Kamvar *et al.* [14] and in more recent spectral learning frameworks [31, 32, 33, 34, 35, 36, 37]. We adopt the supervised spectral clustering principle for the spectral attribute construction (Algorithm 1).

Algorithm 1: Spectral learning for clustering

```
Given data $\boldsymbol{X}$, form the affinity matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$
Define $\boldsymbol{D}$ to be the diagonal matrix with $D_{ij} = \sum_j A_{ij}$
Normalize: $\boldsymbol{N} = (\boldsymbol{A} + d_{max}\boldsymbol{I} - \boldsymbol{D}) / d_{max}$
Find $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ the largest eigenvectors and form the matrix $\boldsymbol{X}' = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k]$
Normalize the rows of $\boldsymbol{X}'$ to be unit length
Treating each row of $\boldsymbol{X}'$ as a point in $\mathbb{R}^k$, cluster into $c$ clusters using k-means or any other sensible algorithm
Assign the original point $\boldsymbol{x}_i$ to the cluster $j$ if row $i$ of $\boldsymbol{X}'$ was assigned to the cluster $j$
```

The supervised spectral learning is implemented in the line 1 of Algorithm 1 by using the data labels in constructing a “constrained affinity matrix” $\boldsymbol{A}$, in our case of visual regression, ordinal label constrained affinity. In the following, we use $\boldsymbol{X}_{N\times|F|}$ to denote the data matrix that is constructed from the features F^i such that $\boldsymbol{X}_{:,i} = \boldsymbol{x}_i^T = F^i \in \mathbb{R}^d$ and the corresponding data labels are denoted by $y_i \in \mathbb{R}$. In our experiments, the number of clusters k is set via cross-validation.

As the first step, we sort the training examples to the ascending order using the data labels:

$$\begin{array}{c} \{\hat{y}_i\} = sort\_ascend(\{y_i\}) \\ \boldsymbol{X}(:,i) = \boldsymbol{X}(:,index(\hat{y}_i)) \end{array} . \tag{4}$$

In the sorted data matrix the adjacent samples in $\boldsymbol{X}$ represent the same or neighboring label values. Next, we assign the following label constrained ordinal similarity value to each affinity matrix element (note that integer y_i assumed for notation simplicity):

$$\boldsymbol{A}(i,j) = \begin{cases} 1, & \text{if } y_i = y_j \\ S(I_i,I_j), & \text{if } y_j = y_{i\pm 1,\dots,K} \\ 0, & \text{otherwise} \end{cases} . \tag{5}$$

After this step we proceed according to Algorithm 1. K defines the “active” label neighborhood and in all our experiments this is set to $K = 1$, *i.e.* the nearest neighbor only. Moreover, $\lambda_1 = 1$ in (3) is used and the results indicate that the label constrained similarity in (5) seems to provide robustness against sub-optimal similarity decay [28]. The output of our spectral learning step are a number of ordinal constrained label clusters/groups $\{G_1, G_2, \dots, G_m\}$, *i.e.* the labels of instances in each cluster are enforced to be ordinal adjacent. We denote the m group centers $p_j,\ j = 1,2,\dots,m$.

Inspired by the recent success of the Fisher scores [38] in visual class learning [39, 40] we encode each training sample with the first order (or additionally second-order) difference between the instance label y and the m group centuries $p_j,\ j = 1,2,\dots,m$ in the attribute space:

$$\boldsymbol{a}_i = \left[(y_i' - p_1)\ (y_i' - p_2)\ \dots\ (y_i' - p_m)\right]^T \ \in \mathbb{R}^m. \tag{6}$$

For multivariate regression problems, the aforementioned method to generate ordinal label constrained spectral attributes is not directly applicable due to multi-dimensional order of label variates. In the light of this, we take a simple solution by generating a set of cluster centers for each label output and concatenating their first order statistics to form spectral attribute vector $\boldsymbol{a}$. In our encoding each training sample $\boldsymbol{x}_i$ is assigned with a spectral attribute vector $\boldsymbol{a}_i$ and a scalar-valued label y_i or multivariate label $\boldsymbol{y}_i$ forming a tuple $\langle \boldsymbol{x}, \boldsymbol{a}, y/\boldsymbol{y} \rangle_i$ which is used in our regression.

5. Hierarchical Visual Regression

The core component of our hierarchical visual regression are the two mappings: from the feature space to the spectral attribute space $f_1 : \boldsymbol{x} \to \boldsymbol{a}$ and from the low-level and mid-level feature (attribute) space to the continuous label space $f_2 : [\boldsymbol{x}, \boldsymbol{a}] \to y/\boldsymbol{y}$. In the testing phase, given an unseen image, the low-level features are extracted and image attributes estimated using the f_1 regression mapping. Then the mapping f_2 is used to map the estimated attribute scores together with low-level features to the output label space. These stages are next explained in Sec. 5.1 and 5.2 respectively.

Let us consider the simpler single-variate regression problems. Specifically, given training pairs $\langle \boldsymbol{x}, \boldsymbol{a}, y \rangle_i^{i=1,2,\ldots,N}$, the goal of the proposed algorithm is to learn a robust regression mapping $f(\boldsymbol{x}, \boldsymbol{a}) = [\boldsymbol{x}^T, \boldsymbol{a}^T] * \boldsymbol{w}$ between concatenated low-level features and spectral attributes and scalar-valued labels, where $\boldsymbol{w} \in \mathbb{R}^{\mathrm{d}+m}$ is the weighting vector to be learned. The objective function can thus be formulated as:

$$\min \sum_{i=1}^{N} \mathrm{loss}([\boldsymbol{x}_i^T, \boldsymbol{a}_i^T]^T, y_i, \boldsymbol{w}) + C\|\boldsymbol{w}\|^2, \tag{7}$$

where C is a trade-off scalar parameter to be determined by cross-validation. The first term is an appropriate loss function and the second term is for regularization. If we simply assume $\boldsymbol{a}$ is independent of $\boldsymbol{x}$ in Eq. (7), it can be re-written as the following:

$$\begin{aligned}\min \sum_{i=1}^{N} \mathrm{loss}(\boldsymbol{x}_i, y_i, \boldsymbol{w}_x) + C\|\boldsymbol{w}_x\|^2 \\ + \sum_{i=1}^{N} \mathrm{loss}(\boldsymbol{a}_i, y_i, \boldsymbol{w}_a) + C\|\boldsymbol{w}_a\|^2,\end{aligned}$$

with $\boldsymbol{w} := [\boldsymbol{w}_x^T, \boldsymbol{w}_a^T]^T$. Without consideration of the regularisation on $\boldsymbol{w}_a$, $\sum_{i=1}^N \text{loss}(\boldsymbol{a}_i, y_i, \boldsymbol{w}_a)$ can be viewed as equality constraints for Euclidean loss function [41] and as inequality constraints for other loss functions (*e.g.* the slack variables in Support Vector Regression). As a result, $\boldsymbol{a}$ plays an important role in improving the robustness of regression function learning.

5.1. Multi-Output Attribute Encoding

We use training pairs $\{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1}^N$ to learn to regress $\boldsymbol{a}$. Given $\boldsymbol{x}_i$ and a_i^j being low-level features of the ith instance and the jth element of its corresponding attribute vector, the objective function for independently learning a regressor for the jth element is written as:

$$\min \quad \frac{1}{2}\|\boldsymbol{w}_x^j\|_2^2 + C\sum_{i=1}^N \text{loss}(a_i^j, g^j(\boldsymbol{x}_i)),$$

where $g^j(\boldsymbol{u}) = \phi(\boldsymbol{u})^T(\boldsymbol{w}_x^j) + b^j$ with $\phi(\cdot)$ being the kernel function to project $\boldsymbol{x}$ to a high dimensional Hilbert space. To realize a joint learning, we follow the established design principle of multi-task learning [42, 43, 44, 45], and write the objective function of the multi-output regression learning as

$$\min \quad \psi(\boldsymbol{W}_x) + C\sum_{i=1}^N \text{loss}(\boldsymbol{a}_i, f_1(\boldsymbol{x}_i)) \tag{8}$$

where $\boldsymbol{W}_x = [\boldsymbol{w}_x^1, \boldsymbol{w}_x^2, \cdots, \boldsymbol{w}_x^j, \cdots, \boldsymbol{w}_x^m]^T \in \mathbb{R}^{m\times d}$ is the weight matrix, $\boldsymbol{a}_i = [a_i^1, a_i^2, \cdots, a_i^m]^T$ is the training attribute vector, and $\boldsymbol{b} = [b^1, b^2, \cdots, b^m]^T \in \mathbb{R}^{\mathrm{m}}$ is the bias vector, and $f_1(\boldsymbol{x}_i) = \phi(\boldsymbol{x}_i)^T\boldsymbol{W} + \boldsymbol{b}$. $\psi(\cdot)$ could be chosen as a norm of $\|\cdot\|_{1,2}$ that encourages sparse use of low-level imagery features, for which optimization can be done by many sparse optimization algorithms such as ADMM [46]. Note that, when $\psi(\cdot)$ is chosen as squared Frobenius norm, Eq. (8) is equivalent to m independent single-variable regression problems, with optimal solutions given in a closed form. In experiments reported in this paper, we find the two choices of norms give similar performance and we use the later one for efficiency. However, Eq. (8) gives a more general regression formulation that may apply better in other datasets.

In this sense, Eq. (8) using Euclidean norm for jointly learning m spectral attributes can be written as

$$\min \ \frac{1}{2}\sum_{j=1}^{m}||\boldsymbol{w}^{j}||_2^2+C\sum_{i=1}^{N}\mathrm{loss}(\boldsymbol{a}_i, f_1(\boldsymbol{x}_i)). \tag{9}$$

Eq. (9) is the general formulation for multi-output regression. Adopting different loss functions leads to different regressors. In our paper, two popular multi-output regressors, *i.e.* multi-output support vector regression (MSVR) and multi-output ridge regression (MRR), are adopted for jointly attribute learning.

Multi-output ridge regression (MRR) – is aimed to minimize quadratic loss function:

$$\min \ \frac{1}{2}\sum_{j=1}^{m}||\boldsymbol{w}^{j}||_2^2+C\sum_{i=1}^{N}\|\boldsymbol{a}_i-(\phi(\boldsymbol{x}_i)^T\boldsymbol{W}+\boldsymbol{b})\|_2^2. \tag{10}$$

It has a closed-form solution based on matrix inversion [4, 47].

Multi-output support vector regression (MSVR) – employs a ε-sensitive loss function [48] as:

$$\mathrm{loss}(a_i^j, f(\boldsymbol{x}_i)) = \begin{cases} 0, & \text{if } |a_i^j - g(\boldsymbol{x}_i)| < \varepsilon \\ |a_i^j - g(\boldsymbol{x}_i)| - \varepsilon, & \text{if } |a_i^j - g(\boldsymbol{x}_i)| \geqslant \varepsilon \end{cases}$$

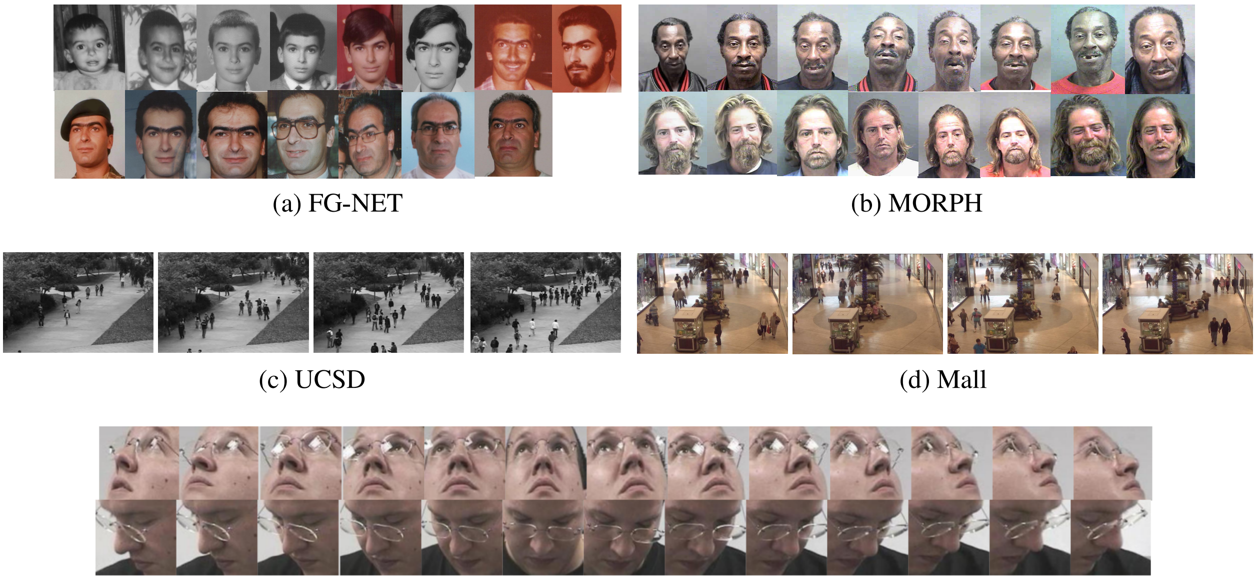
where ε in the formulation controls the insensitivity to output bias. Multi-output support vector regression is solved by using cutting-plane strategies [49].

5.2. Robust Regression Learning

With the optimized parameters $\boldsymbol{W}$ and $\boldsymbol{b}$, low-level features are applied to generating the additional features $\hat{\boldsymbol{a}} \in \mathbb{R}^{\mathrm{m}}$ for robust regression learning in Eq. (9). With each image, represented now by concatenating x_i and $\hat{\boldsymbol{a}}_i$, and the corresponding scalar-valued single variate label (*e.g.* age and person count) y_i or multivariate label $\boldsymbol{y}_i$ (*e.g.* head pose), $i = 1, 2 \ldots N$, a simple off-the-shelf regression model can be learned. It is also worth pointing out that any existing regression methods can readily incorporate spectral attributes to improve their robustness.

Table 1: Details of the datasets used in our experiments (two age estimation, two crowd counting and one 2D face pose).

Data	*Type*	*# of imgs*	*y range*
FG-NET [1]	age est.	1002	0–69
MORPH [15]	age est.	5475	16–77
UCSD [50]	crowd count.	2000	11–46
Mall [51]	crown count.	2000	13–53
Pointing'04 [52]	face pose ($\theta_{yaw/pitch}$)	2790	$-90°$–$90°$



(a) FG-NET

(b) MORPH

(c) UCSD

(d) Mall

(e) Pointing'04

Figure 2: Example images from the datasets in Table 1.

6. Experiments

6.1. Datasets and Settings

For age estimation, the two frequently used public benchmarks, FG-NET [3, 4, 15] and MORPH [1, 4, 15], were used. Both datasets are designed for evaluating accuracy of facial age estimators and contain many natural sources of variation such as different racial groups (Figure 2a and 2b). For crowd counting, the experiments were conducted on the widely-used UCSD [4, 5, 50, 53] and Mall [4, 53] benchmarks which feature outdoor and indoor scenes (Figure 2c and 2d). To test our multi-variate attribute extension, we also include the popular 2D face pose benchmark Pointing'04 [52] which contains face images of 15 persons captured with varying appearance and under controlled indoor environment.

For low level features we adopted the Active Appearance Model (AAM) features [54] used in many state-of-the-art works [1, 3, 4, 15, 55, 56] and for crowd counting we used the three different kind of gray-level features used in the recent works [4, 50, 51]: foreground segments, edge features, and local texture features. For the face pose estimation we extracted Histogram of Oriented Gradient (HoG) features [57] which have been widely employed in the recent works [58, 59].

With the FG-NET dataset the leave-one-person-out setting was used in evaluation as in [3, 4, 15, 55, 56] and with the MORPH dataset we randomly split it into 80% for training and the remaining 20% for testing repeated 30 times as in [4, 15]. For crowd counting, we followed the same training and testing partitions as in [4], *i.e.* , the frames $601 - 1400$ in the UCSD dataset and the frames $1 - 800$ in the Mall dataset were used for training and the remaining frames for testing. With the Pointing'04 benchmark [52], we adopt even data-split cross-validation adopted in [6, 59]. For multi-output ridge regression and support vector regression, linear kernel is adopted in our experiments. The free parameters were tuned using 4-fold cross-validation.

For age estimation we report the two evaluation metrics from [15, 4]: *mean absolute error* (mae) and *cumulative score* (cs) with the error level 5. For crowd counting we report the all three found metrics from the recent works: *mean absolute error* (mae), *mean squared error* (mse), and *mean deviation error* (mde). For evaluating the performance of head pose estimation, we employed the mean absolute error (mae) of yaw and pitch angles separately and combined. Out of the aforementioned metrics, only for cs a higher value means better performance and a lower value is better for the rest.

6.2. State-of-the-Art Comparison

6.2.1. Age Estimation

Table 2a shows the comparative results for our model and a number of recently-published algorithms. Only the AGES [1] method uses other than the AAM features. AGES [1], RED-SVM [61] and OHRank [15] are classification or ranking based methods while the rest are estimation-by-

Table 2: (a) Age estimation performance comparison; (b) Crowd counting performance comparison.

(a)

	FG-NET		*MORPH*	
Method	mae	cs	mae	cs
AGES [1]	6.77	–	8.83	–
RUN [55]	5.78	–	–	–
Ranking [60]	5.33	–	–	–
RED-SVM [61]	5.24	–	6.49	–
LARR [56]	5.07	–	–	–
MTWGP [3]	4.83	–	6.28	–
OHRank [15]	4.85	**74.4%**	**5.69**	56.3%
LDL [16]	5.23	69.4%	5.94	56.5%
SVR [56]	5.66	68.0%	5.77	57.1%
SAL-SVR	**4.61**	73.5%	5.90	**57.7%**

(b)

	UCSD			Mall		
Method	mae	mse	mde	mae	mse	mde
GPR [50]	2.24	7.97	0.112	3.72	20.1	0.115
WRR [62]	2.11	7.11	0.105	**3.58**	19.0	**0.110**
LDL [16]	2.25	7.89	0.111	3.70	21.2	0.117
RR [51]	2.25	7.82	0.110	**3.59**	19.0	**0.110**
SAL-RR	**2.02**	**6.65**	**0.100**	**3.59**	**18.9**	**0.110**

regression similar to us. To remove the effect of the used feature space, we implemented OHRank [15], LDL [16], and SVR [56] using the same AAM features as our method. Our method with spectral attribute space construction obtained the best results on the mae metric for FG-NET and on the cs metric for MORPH, while its performance on cs with FG-NET and mae with MORPH are also comparable. Our direct competitor, SVR [56], used the identical low-level features and the same single-output regression model, and thus the significant performance leap for our method (18.5% decrease on mae and 8.09% relative increase on cs for FG-NET) can only be explained by the effectiveness of the spectral attributes. On the other hand, we observe that our method achieved comparable results with SVR on mae for MORPH, which can be explained by inconsistent feature representations of MORPH as compared to FG-NET. Moreover, our method that does not require "attribute engineering" performed better than most of the state-of-the-art methods with both benchmarks. Also the Ordinal Hyperplane Rank model (OHRank) [15] performed well, but its computational cost is extremely high as indicated in [4] (five orders of magnitude slower than our method in Table 7 of Section 6.7).

6.2.2. Crowd Counting

In Table 2b, using the identical low-level features, we compare our method with the four recent methods considered as state-of-the-art counting-by-regression. For UCSD, our proposed algorithm significantly outperformed other methods on all of three performance metrics and results are comparable with WRR and RR with the Mall dataset. Similar to the age estimation experiment, the spectral attribute mapping explains the superiority to the RR method as all other parameters are the same. We found that there is larger feature variations for the Mall dataset and therefore the spectral clustering provides more variance to the attribute space as compared to that for UCSD. However, SAL still performs comparably to the state-of-the-arts.

6.3. Multivariate Regression

Table 3: Head pose estimation performance comparison on the Pointing dataset (even data-split cross-validation).

	Mean angular error		
	Yaw	Pitch	Yaw+Pitch
SVR [48]	7.84°	9.37°	–
LARR [6]	9.23°	7.69°	–
PLS [63]	9.83° ± 0.38°	9.95° ± 0.05°	16.68° ± 0.26°
KPLS [63]	7.82° ± 0.36°	7.33° ± 0.04°	13.06° ± 0.41°
MLD [58]	5.64° ± 0.16°	4.62° ± 0.10°	9.41° ± 0.26°
SAL-MLD	**5.55° ± 0.18°**	**4.41° ± 0.12°**	**9.06° ± 0.28°**

Our extension to multivariate regression was proposed in Sec. 4 and it was experimented with the Pointing'04 face pose estimation dataset where regression should output two dimensional variables, *i.e.* the yaw and pitch angles of the faces in input images. The results for our method and the state-of-the-arts are illustrated in Table 3. Our method produces the best accuracy for the pitch, yaw and combined pitch+yaw measures.

6.4. Hand-crafted vs. Spectral Attributes

Table 4: (a) Hand-crafted cumulative attributes (CA) vs. our spectral attributes; (b) Performance of unconstrained spectral learning (spectral clustering) vs. constrained spectral learning.

(a)

	FG-NET		UCSD		
Method	mae	cs	mae	mse	mde
CA [4]	4.67	**74.5%**	2.07	6.86	0.102
SAL (ours)	**4.61**	73.5%	**2.02**	**6.65**	**0.100**

(b)

	FG-NET		UCSD		
Method	mae	cs	mae	mse	mde
SAL (const.)	**4.61**	**73.5%**	**2.02**	**6.65**	**0.100**
SAL (unconst.)	5.66	67.3%	2.17	7.10	0.105

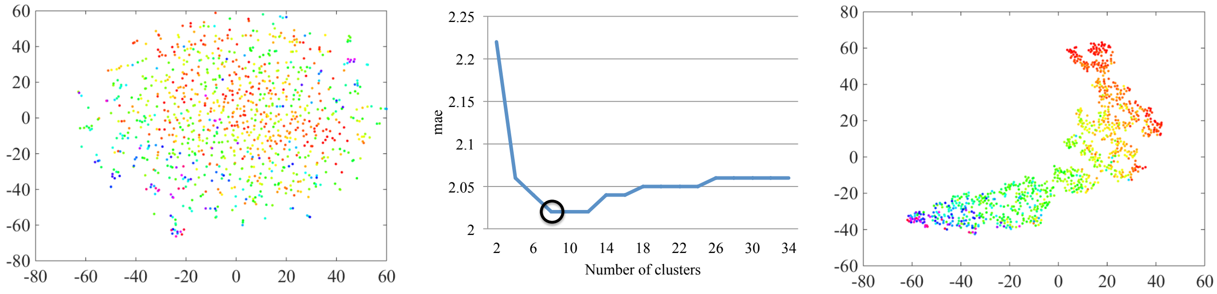


Figure 3: *Left:* 2D embedding of the feature space (colour temperature encodes the crowd density); *Middle:* Evaluation of the effect of the number of spectral clusters c for the UCSD dataset (smaller mae is better). The number surrounded by the black circle (8) denotes the number determined by self-tuning. *Right:* 2D embedding of our encoded spectral attribute space.

Table 4a shows the results for the hand-crafted cumulative attributes in [4] and our learned spectral attributes. Rather strikingly the spectral attribute space is able to construct attribute representation that is superior or at least comparable to the dedicated manual attribute space mapping. This clearly indicates that the spectral attribute learning is a strong extension for visual attribute learning regression.

6.5. *Unconstrained vs. Constrained Learning*

In Section 4 we proposed ordinal label constrained spectral learning that uses the continuous target label space in learning. The results in Table 4b verify the advantages of exploiting latent ordinal dependency of regression labels on both age estimation and crowd counting. In details, spectral learning that constraints the traditional spectral clustering can encode better the latent characteristics of the attribute space via continuous label relations. The results for the unsupervised approach are good as well, but this is partly due to the overly optimistic setting where the optimal number of attributes was manually set.

6.6. *Evaluation of the Attribute Encoding*

We introduced our spectral attribute encoding inspired by the Fisher vector encoding [39, 64]. In the next two experiments, we evaluated the robustness of the encoding procedure.

6.6.1. Number of Spectral Attributes

The number of spectral attributes can be optimized by the cross-validation as was done in our experiments. However, we wanted to investigate the robustness of that approach and experimented with varying number of spectral attributes. The result is illustrated Figure 3 that shows that the number is an important factor for the success of the method and the "self-tuning" procedure succeeds to find an optimal value. Due to the limited space, the results are shown only for the UCSD dataset, but hold for all datasets. The low level embeddings in Figure 3 visualize how spectral learning re-maps original features to the spectral attribute space where regression label relevant features are made stronger.

6.6.2. First-Order vs. Second-Order Coding

Table 5: 1st order vs. 2nd-order difference spectral attribute encoding.

	FG-NET		UCSD		
Methods	mae	cs	mae	mse	mde
1st-order	**4.61**	**73.5%**	**2.02**	**6.65**	**0.100**
2nd-order	5.12	70.0%	2.18	7.34	0.108

In Table 5, the results for first-order and second-order difference encoding of spectral attributes are listed using the same framework and settings. The results show that first-order difference encoding outperforms the second-order encoding. Both encoding schemes still beat the ordinary regression models (SVR for FG-NET in Table 2a and RR for UCSD in Table 2b).

Table 6: Multi-output ridge regression (MRR) vs. multi-output support vector regression (MSVR) for spectral attribute encoding.

	FG-NET		UCSD		
Methods	mae	cs	mae	mse	mde
MRR	**4.61**	**73.5%**	**2.02**	**6.65**	**0.100**
MSVR	4.63	**73.5%**	2.12	7.15	0.105

6.6.3. Multi-Output Ridge Regression vs. Multi-Output Support Vector Regression

We evaluate two popular regressors (*i.e.* multi-output ridge regression and multi-output support vector regression investigated in Sec. 5) for joint attribute encoding. Experiment results are shown in Table 6. Multi-output ridge regression (MRR) achieves better performance than multi-output support vector regression (MSVR) in both experiments. For age estimation on the FG-NET dataset, the results by MRR is almost identical with those of MSVR, while the difference for crowd counting on the USCD dataset becomes significant. The explanation for performance gap in crowd counting is that the feature-target relationship in crowd counting is more linear than that in age estimation. In this sense, ridge regression based attribute encoding is more suitable for crowd counting, which is also verified in recent crowd counting frameworks [4, 51].

6.7. Computational Cost

Table 7: Training times required by the different methods.

	Age (*mins*)		Crowd (*secs*)	
Method	FG-NET	MORPH	UCSD	Mall
SVR [56]	2.69×10^{0}	2.08×10^{1}	–	–
RR [51]	–	–	**0.011**	**0.009**
CA [4]	8.91×10^{-1}	6.10×10^{0}	0.065	0.065
LDL [16]	1.26×10^{2}	1.03×10^{3}	50.32	42.97
SAL (ours)	$\mathbf{2.18 \times 10^{-1}}$	$\mathbf{1.11 \times 10^{0}}$	0.018	0.022

The regression methods are efficient once trained and if the feature extraction time is not included. However, there are significant differences in training the different methods. Table 7 compares the computational costs of the four popular methods with our method which achieves the best overall efficiency in training, *i.e.* at least 4 times faster to its closest competitor. The RR method is efficient, but we need to take into account that the time spend for manual attribute engineering is not included while our method does that automatic manner. Moreover, the comparison was made using a linear kernel and the differences are even more favouring to our approach with the higher order polynomial kernels.

7. Conclusion

We proposed a novel approach to visual regression by combining two generic regression tools, attribute learning and kernel ridge regression. We extended the attribute learning stage by spectral learning that produces “spectral attributes” which are learned as opposed to the previously proposed manual construction and achieves superior accuracy. Moreover, we extended spectral attribute learning to multivariate regression. The proposed visual regression approach is not sensitive to parameter tuning, but the meta-parameters are optimized by spectral learning and cross-validation. On multiple important benchmarks our generic visual regression achieved superior accuracy to the existing techniques with favourable computational complexity.

Acknowledgements

This work is funded by the Academy of Finland under the Grant No. 267581 and 298700. The authors also acknowledge the donations of Nvidia Tesla K40 from Nvidia Corporation to the first author’s research group.

References

[1] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Transactions on pattern analysis and machine intelligence 29 (12) (2007) 2234–2240.

[2] Y. Fu, G. Guo, T. S. Huang, Age synthesis and estimation via faces: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (11) (2010) 1955–1976.

[3] Y. Zhang, D. Yeung, Multi-tasks warped Gaussian process for personalized age estimation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[4] K. Chen, S. Gong, T. Xiang, C. C. Loy, Cumulative attribute space for age and crowd density estimation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[5] A. B. Chan, N. Vasconcelos, Counting people with low-level features and Bayesian regression, IEEE Transactions on Image Processing 21 (4) (2012) 2160–2177.

[6] G. Guo, Y. Fu, C. Dyer, T. Huang, Head pose estimation: Classification or regression?, in: Proceedings of International Conference on Pattern Recognition, 2008.

[7] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, T. Huang, Regression from patch-kernel, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[8] E. Murphy-Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (4) (2009) 607–626.

[9] G. Mu, G. Guo, Y. Fu, T. S. Huang, Human age estimation using bio-inspired features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 112–119.

[10] S. An, W. Liu, S. Venkatesh, Face recognition using kernel ridge regression, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[11] V. Ferrari, A. Zisserman, Learning visual attributes, in: Proceedings of Advances in Neural Information Processing Systems, 2007.

[12] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[13] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Attribute learning for understanding unstructured social activity, in: Proceedings of European Conference on Computer Vision, 2012.

[14] K. Kamvar, S. Sepandar, K. Klein, D. Dan, M. Manning, C. Christopher, Spectral learning, in: Proceedings of International Joint Conference of Artificial Intelligence, 2003.

[15] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[16] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, IEEE transactions on pattern analysis and machine intelligence 35 (10) (2013) 2401–2412.

[17] D. Parikh, K. Grauman, Relative attributes, in: Proceedings of International Conference on Computer Vision, 2011.

[18] G. Patterson, J. Hays, SUN attribute database: Discovering, annotating, and recognizing scene attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[19] T. L. Berg, A. C. Berg, J. Shih, Automatic attribute discovery and characterization from noisy web data, in: Proceedings of European Conference on Computer Vision, 2010.

[20] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

[21] I. Endres, D. Hoiem, Category independent object proposals, in: Proceedings of European Conference on Computer Vision, 2010.

[22] M. Kim, V. Pavlovic, Structured output ordinal regression for dynamic facial emotion intensity prediction, in: Proceedings of European Conference on Computer Vision, 2010.

[23] W. Chen, C. Xiong, R. Xu, J. J. Corso, Actionness ranking with lattice conditional ordinal random fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[24] T. Malisiewicz, A. Gupta, A. Efros, Ensemble of exemplar-SVMs for object detection and beyond, in: Proceedings of International Conference on Computer Vision, 2011.

[25] A. C. Berg, T. L. Berg, J. Malik, Shape matching and object recognition using low distortion correspondences, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[26] S. Bagon, O. Brostovski, M. Galun, M. Irani, Detecting and sketching the common, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[27] M. Torki, A. Elgammal, One-shot multi-set non-rigid feature-spatial matching, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[28] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Proceedings of Advances in neural information processing systems, 2002.

[29] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[30] L. Zelnik-manor, P. Perona, Self-tuning spectral clustering, in: Proceedings of Advances in Neural Information Processing Systems, 2005.

[31] D. Tao, L. Jin, Y. Wang, Y. Yuan, X. Li, Person re-identification by regularized smoothing kiss metric learning, IEEE Transactions on Circuits and Systems for Video Technology 23 (10) (2013) 1675–1685.

[32] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, Y. Y. Tang, Person re-identification by dual-regularized kiss metric learning, IEEE Transactions on Image Processing 25 (6) (2016) 2726–2738.

[33] D. Tao, L. Jin, W. Liu, X. Li, Hessian regularized support vector machines for mobile image annotation on the cloud, IEEE Transactions on Multimedia 15 (4) (2013) 833–844.

[34] L. Yang, S. Yang, S. Li, R. Zhang, F. Liu, L. Jiao, Coupled compressed sensing inspired sparse spatial-spectral lssvm for hyperspectral image classification, Knowledge-Based Systems 79 (2015) 80–89.

[35] R. Shang, Z. Zhang, L. Jiao, C. Liu, Y. Li, Self-representation based dual-graph regularized feature selection clustering, Neurocomputing 171 (2016) 1242–1253.

[36] R. Shang, Z. Zhang, L. Jiao, W. Wang, S. Yang, Global discriminative-based nonnegative spectral clustering, Pattern Recognition 55 (2016) 172–182.

[37] L. Jiao, F. Shang, F. Wang, Y. Liu, Fast semi-supervised clustering with enhanced spectral embedding, Pattern Recognition 45 (12) (2012) 4358–4369.

[38] T. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: Proceedings of Advances in neural information processing systems, 1998.

[39] J. Sanchez, F. Perronnin, T. Mensink, J. Verbeek, Image classification with the Fisher vector: Theory and practice, International Journal of Computer Vision 105 (3) (2013) 222–245.

[40] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: Proceedings of British Machine Vision Conference, 2011.

[41] K. Chen, L. Zhang, Y. Zhang, Cyclic motion generation of multi-link planar robot performing square end-effector trajectory analyzed via gradient-descent and zhang et als neural-dynamic methods, in: ISSCAA, 2008.

[42] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: NIPS, 2006.

[43] A. Argyriou, T. Evgeniou, M. Pontil, A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Machine Learning 73 (2008) 243–272.

[44] Z. Kang, K. Grauman, F. Sha, Learning with whom to share in multi-task feature learning, in: Proceedings of International Conference on Machine Learning, 2011.

[45] M. Solnon, S. Arlot, F. Bach, Multi-task regression using minimal penalties, JMLR.

[46] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn.

[47] H. Borchani, G. Varando, C. Bielza, P. Larrañaga, A survey on multi-output regression, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (5) (2015) 216–233.

[48] A. J. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing 14 (3) (2004) 199–222.

[49] T. Joachims, T. Finley, C.-N. J. Yu, Cutting-plane training of structural svms, Machine Learning 77 (1) (2009) 27–59.

[50] A. B. Chan, Z.-S. J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: counting people without people models or tracking, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[51] K. Chen, C. C. Loy, S. Gong, T. Xiang, Feature mining for localised crowd counting, in: Proceedings of British Machine Vision Conference, 2012.

[52] N. Gourier, D. Hall, J. L. Crowley, Estimating face orientation from robust detection of salient facial structures, in: Proceedings of International Conference on Pattern Recognition, 2004.

[53] C. C. Loy, S. Gong, T. Xiang, From semi-supervised to transfer counting of crowds, in: Proceedings of International Conference on Computer Vision, 2013.

[54] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al., Active appearance models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[55] S. Yan, H. Wang, X. Tang, T. S. Huang, Learning auto-structured regressor from uncertain nonnegative labels, in: Proceedings of International Conference on Computer Vision, 2007.

[56] G. Guo, Y. Fu, C. R. Dyer, T. S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Transactions on Image Processing 17 (7) (2008) 1178–1188.

[57] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE Computer Vision and Pattern Recognition, 2005.

[58] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[59] K. Hara, R. Chellappa, Growing regression forests by classification: Applications to object pose estimation, in:

Proceedings of European Conference on Computer Vision, 2014.

[60] S. Yan, H. Wang, T. S. Huang, Q. Yang, X. Tang, Ranking with uncertain labels, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2007.

[61] K.-Y. Chang, C.-S. Chen, Y.-P. Hung, A ranking approach for human ages estimation based on face images, in: Proceedings of International Conference on Pattern Recognition, 2010.

[62] K. Chen, J.-K. Kämäräinen, Learning to count with back-propagated information, in: Proceedings of International Conference on Pattern Recognition, 2014.

[63] M. A. Haj, J. Gonzalez, L. S. Davis, On partial least squares in head pose estimation: How to simultaneously deal with misalignment, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[64] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Fisher networks for large-scale image classification, in: Proceedings of Advances in Neural Information Processing systems, 2013.