Pattern Recognition 87 (2019) 29–37

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Cumulative attribute space regression for head pose estimation and color constancy

Ke Chen[a], Kui Jia[b], Heikki Huttunen[a,*], Jiri Matas[a,c], Joni-Kristian Kämäräinen[a]

[a] Laboratory of Signal Processing, Tampere University of Technology, Finland
[b] School of Electronic and Information Engineering, South China University of Technology, China
[c] Department of Cybernetics, Czech Technical University, Prague

ARTICLE INFO

Article history:
Received 11 April 2018
Revised 21 September 2018
Accepted 9 October 2018
Available online 10 October 2018

Keywords:
Multivariate regression
Cumulative attribute space
Head pose
Color constancy

ABSTRACT

Two-stage Cumulative Attribute (CA) regression has been found effective in regression problems of computer vision such as facial age and crowd density estimation. The first stage regression maps input features to cumulative attributes that encode correlations between target values. The previous works have dealt with single output regression. In this work, we propose cumulative attribute spaces for 2- and 3-output (multivariate) regression. We show how the original CA space can be generalized to multiple output by the Cartesian product (CartCA). However, for target spaces with more than two outputs the CartCA becomes computationally infeasible and therefore we propose an approximate solution - multi-view CA (MvCA) - where CartCA is applied to output pairs. We experimentally verify improved performance of the CartCA and MvCA spaces in 2D and 3D face pose estimation and three-output (RGB) illuminant estimation for color constancy.

1. Introduction

Multiple output regression predicts several continous variables simultaneously. One of the emerging topics within regression problems is *visual regression*. Regression has many applications in vision, such as 2D and 3D head pose estimation and landmark detection [1–3] (see Fig. 1), illumination estimation for color constancy [4], as well as apparent age estimation [5].

A straightforward solution is to learn individual regressors for each target variable separately using the traditional techniques (e.g. ridge regression, random forest regression [6] and support vector regression [7]). However, independent regressors discard the interdependence between the target variables, which can be substantial in vision problems. There are more advanced approaches for multivariate regression, such as joint learning of regressors in a multi-task fashion [8] and structured learning [9], but even these generic approaches cannot effectively model cross-target correlations of visual data and are often inferior to problem specific methods.

Most of the above methods apply the traditional single layer regression architecture, where the multivariate output is estimated either directly from image features, or by optimizing a tailored score function . During the recent years there have been multiple successful attempts to replace the single layer model with two layer (two stage) architectures [10–12]. The first layer output represents an “attribute space” where attribute features have an important semantic meaning for the regression or classification task solved by the second layer output.

In this work, we focus on the concept of cumulative attribute (CA) space mapping that was proposed in our previous work [12]. The main idea behind the cumulative attributes is the intuitive fact that low level features for certain vision tasks, such as age estimation or crowd counting, are *cumulative* by nature. In this work, we show that this hypothesis holds for a wider class of vision problems.

Inspired by the success of CA for scalar-valued regression [13], we extend CA to the multivariate output setting. A straightforward extension is to apply CA regression to each output variable independently. This approach is the baseline in our work – *Independent Cumulative Attribute space (IndepCA)*. The drawback of IndepCA is its limited ability to exploit the multi-dimensional nature of the target space thus omitting the correlations of the output variables (such as visual similarity of faces between adjacent pitch and yaw bins in Fig. 1).

To overcome this limitation we generalize the CA to 2-output case by adopting a mapping based on the Cartesian product (Fig. 1) – *Cartesian Cumulative Attribute space (CartCA)*. The CartCA divides the multi-dimensional space into disjoint regions. For a landmark

* Corresponding author.
E-mail address: heikki.huttunen@tut.fi (H. Huttunen).

https://doi.org/10.1016/j.patcog.2018.10.015

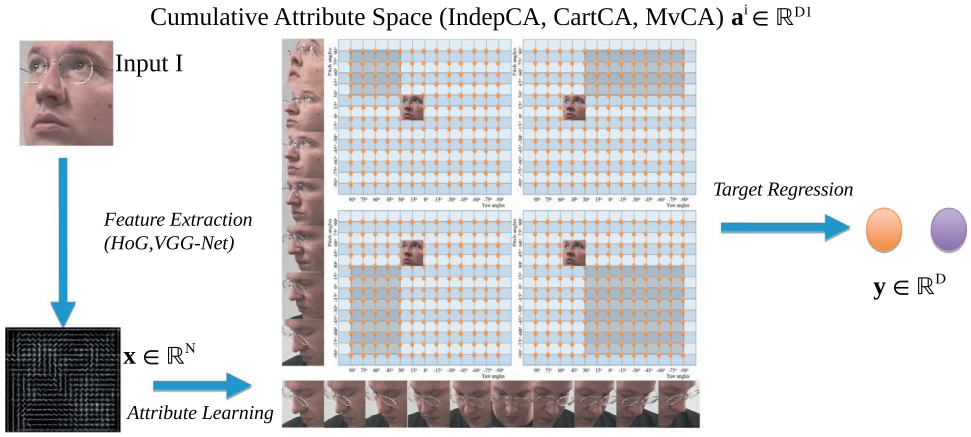

Fig. 1. Cartesian Cumulative Attribute space (CartCA) for 2-output regression. CA-based regression has three processing stages: i) feature extraction, ii) mapping from feature space to Cumulative Attribute space (*Attribute Learning*) and iii) mapping from CA space to a two-dimensional output space (*Target Regression*: head yaw and pitch angles).

point anchored in a multi-dimensional target space, *i.e.* a single regression label, CartCA forms uniquely different binary partitions of training samples. CartCA is a generalization of the original CA for two-dimensional target space. The number of binary partitions grows exponentially w.r.t. the label space dimensionality making CartCA impractical beyond more than two outputs.

To avoid the combinatorial explosion, we propose an approximation by projecting training samples into various 2D sub-spaces for which CartCA is applied. We call this approach *Multi-View Cumulative Attribute (MvCA)* regression. In the experimental part, we study these methods in three different multivariate visual regression problems: 2D head pose estimation, 3D head pose estimation and 3D illumination (RGB) estimation for color constancy. In all experiments, our method provides competitive performance and consistently outperforms methods that do not construct a cumulative attribute space layer for regression.

Our main contributions are summarized as follows:

- We extend the scalar value cumulative attribute (CA) regression to 2-output cumulative regression by adopting the Cartesian product to partition output spaces (CartCA).
- We propose an approximation approach for CA with $\geq$ 3 outputs by partitioning output spaces to multiple 2D views - Multi-view Cumulative Attribute (MvCA). This approximation avoids exponential growth of CartCA.
- We demonstrate effectiveness of multi-output CA regression in several computer vision applications (2D and 3D head pose estimation and RGB illumination estimation for color constancy) where CartCA and MvCA achieve competitive accuracies as compared to state-of-the-art.

2. Related work

In this section, we provide a short survey on the recent and related works in visual regression and attribute learning. Since our experiments are performed on 2D and 3D targets, we also survey related works on these applications (namely, head pose estimation and color constancy estimation).

Multivariate Regression – For the standard univariate regression problems in computer vision, we seek for a mapping $f: \mathbb{R}^N \mapsto \mathbb{R}$, where the input $\boldsymbol{x} \in \mathbb{R}^N$ corresponds to N extracted image features and the output $y \in \mathbb{R}$ is a real-valued regression target. Traditional methods include L_2 regularized (ridge) regression , L_1 regularized (LASSO) regression [14], random forest regression [6] and support vector regression [7], to name a few. These regression methods can be applied to multivariate regression problems $f: \mathbb{R}^N \mapsto \mathbb{R}^D$ by independently learning univariate regressors $f: \mathbb{R}^N \mapsto \mathbb{R}$ for each target variable $y_1, y_2, \ldots, y_D$ separately. This approach, however, omits interdependencies between output variables and for that purpose there are other generic approaches such as enforcing jointly learning regressors in a multi-task fashion [8] or structured learning methods [9]. For example, structured multivariate regression is applied in a number of computer vision applications [15].

Mid-layer attributes have been adopted in certain recent works [10–12,16–18]. These methods learn D_1-dimensional feature representation, which is used in a two-layer learning architecture $f: \mathbb{R}^N \mapsto \mathbb{R}^{D_1} \mapsto \mathbb{R}^D$ or (concatenation of features and attributes) $f: \mathbb{R}^N \mapsto \mathbb{R}^{D_1}, \mathbb{R}^N \mapsto \mathbb{R}^D$. Indeed, it has been shown in many cases that the two-layer structure improves the accuracy. Inspired by the success of cumulative attributes (CAs) for scalar-valued regression [13], we generalize CA to the 2-output ($D = 2$) and 3-output ($D = 3$) settings in this work. For this work, we adopt the Partial Least Squares (PLS) regression [19] and NIPALS [20] for estimating the regression score (and loading) matrices due to their simplicity (for more details see Section 3.3).

Attribute Learning – Visual attributes, which can be either manually defined according to prior knowledge [17,18] or discovered from data [10,16], have been widely applied to a number of classification problems in computer vision, e.g., image categorisation [11,17], person re-identification [18], and action and video event recognition [16]. The proposed classification problems, however, are different from the regression problems since they rarely establish natural and cumulative correlation, such as the person age or a number of people, and often require manual annotation. Yang et al. [21] proposed correlation analysis for two-view image reconstruction.

Recently, the concept of cumulative attributes [12] was proposed for regression problems, as those classification-oriented attributes cannot be utilized directly to explore the cumulative dependency across regression labels. However, CA developed for scalar-valued regression problems can only be applied to multivariate regression problems with the price of missing multi-dimensional nature of the target space (IndepCA in this work).

Head Pose Estimation – In this case, the regression target is either two-dimensional (yaw and pitch angles) or 3D (+roll). The challenges reside in feature inconsistency and label ambiguity. In particular, for the same head pose, feature variations between different persons are large due to varying facial appearance. Moreover, the pose labels are noisy as the exact ground truth is difficult to acquire. As head pose estimation is challenging due to uncertain labels, it is considered a good testbed for evaluating robustness of the proposed attributes. The recent algorithms for head

pose estimation can be categorized into two groups: classification-based [22] and regression-based [1,15,23,24]. Moreover, deep architectures have been proposed for human pose recovery [25].

If the head pose estimation problem is cast to a *classification* problem, the implicit assumption is that pose labels are independent, which discards the ordered dependency across the label space [22]. In the view of this, the regression-based algorithms have recently become more popular for both 2D [15,26,27] and 3D head pose estimation [23,24].

In [27], a partial least square regression model was adopted to cope with the misalignment problem when estimating the head pose. Foytik and Asari [26] introduced a two-layer regression framework in a coarse-to-fine manner, which first determines the range of prediction (*i.e.* coarse estimation to robustify against ambiguous labels) and then learns a regression function to estimate the final pose value. Recently, Geng et al. [1] introduced the concept of soft labelling by using adjacent labels around the true pose label in a multi-label learning fashion. This reduces the negative effect of ambiguous targets and helps to capture correlations between the neighbouring targets. However, the soft labelling suffers from the invalid assumption that label correlations exist only locally.

On the contrary, the goal of our CartCA and MvCA is to represent the target correlations globally across the whole pose space. Beyond multivariate label distribution, regression forests [23] and its variants [15,24] were proven their effectiveness and real-time efficiency in 2D and 3D head pose estimation.

Illumination Estimation **–** Another experimental case in our paper considers the estimation of illumination of color images. This is a 3-output regression problem, where the goal is to estimate the R, G and B values of scene illumination.

Existing algorithms for illumination estimation can be categorised into two main groups: statistics based [28,29] and learning based [30–32]. In [32], a five-layer ad-hoc CNN was designed combining feature generation and multi-channel regression to estimate illumination in an end-to-end manner. Qian et al. [4] employed an implicit structured output regression on the output of fully-connected layer of VGG-Net to discover inter-output correlation.

3. Methodology

This section first introduces cumulative attribute (CA) regression in [12] (Section 3.1). Next, a two-variate generalization of CA is proposed (CartCA) and then multi-view CA (MvCA) which is more practical for $D > 2$ target outputs (Section 3.2). In Section 3.3 the two-stage regression is discussed in more detail.

3.1. Cumulative attribute space

Consider a standard scalar value visual regression problem, with I training examples $\{\boldsymbol{x}_i, y_i\}$, where $\boldsymbol{x}_i \in \mathbb{R}^N$ are N extracted image features for the image indexed by i and $y_i \in \mathbb{R}$ is the corresponding scalar target. Chen et al. [12] introduces mid-level mapping to $\boldsymbol{a}_i \in \mathbb{R}^{D_1}$ which is termed as a "cumulative attribute" vector of $\boldsymbol{x}_i$.

The main workflow is based on two stage regression, where the first regressor provides attribute mapping $f_1 : \mathbb{R}^N \mapsto \mathbb{R}^{D_1}$ and the second regressor provides the target output mapping $f_2 : \mathbb{R}^{D_1} \mapsto \mathbb{R}$. It is noteworthy, that the best performance is achieved by concatenating the original features and the estimated attribute vector in the second stage, i.e. $f_2 : \boldsymbol{x}, \boldsymbol{a} \mapsto \mathbb{R}$.

During the training stage, the mid-level attribute values $\boldsymbol{a}_i \in \mathbb{R}^{D_1}$ are generated by thresholding the regression target $y_i \in \mathbb{R}$ using the following CA rule:

$$a_{i,j} = \begin{cases} 1, & \text{when } \; y_i \leqslant \tau_j, \\ 0, & \text{when } \; y_i > \tau_j, \end{cases} \tag{1}$$

for $j = 1, 2, \ldots, D_1$. In other words, the regression problem is decomposed into D_1 binary classification problems by thresholding the target at τ_j. The dimension of the attribute space D_1 and the corresponding thresholds are problem specific; for example, in age estimation an obvious choice is to set $\tau_1 = 1, \tau_2 = 2, \ldots, \tau_{99} = 99$ when $D_1 = 99$.

The attribute mapping f_1 is learned using ridge regression; meaning that we learn D_1 attribute functions corresponding to D_1 mid-level binary targets. Ideally the mapping should look like a step function with the change located at the true target value, but estimated attributes $\hat{\boldsymbol{a}}_i$ are actually real valued vectors that are not binarized but directly used in the next stage regressor f_2. This means that binary values are used only during the training stage and in the testing stage real value multiview cumulative attributes are used for the final regressor.

Alternative to the regression based attribute functions in our work, also any two-class (binary) classifier can be trained for the attribute assignments defined in (2). However, during our experiments we have found the real valued outputs of regressors, *soft attributes*, more effective. This can be explained by the fact that no information is lost in the binary decisions and the whole pipeline is regression based.

3.2. 2- and 3-output cumulative attribute spaces

We will now propose three variants of generalizing the univariate case to multivariate.

IndepCA— A straightforward multivariate ($D \geq 2$) extension of CA is to treat all output dimensions as independent and use the standard CA for each output variable. We denote this straightforward extension as *IndepCA*. If, for simplicity, we assume that all D output dimensions are similar, then their corresponding cumulative attribute spaces can be represented by D_1-dimensional attribute vectors. IndepCA learns D_1-dimensional attribute mapping for each D dimensions of the target space $\boldsymbol{y}_i \in \mathbb{R}^D$. For the final stage regression we concatenate D D_1-dimensional attribute vectors to a single vector of length $D_1 \times D$. The second stage regressor is a multi-variate regressor or D univariate regressors that provide the target output $\boldsymbol{y}_i = (y_1, y_2, \ldots, y_D)$. More details about the practical computation are in Section 3.3.

For scalar-valued regression, an important advantage of CA comes from its more effective use of the 1D target space than traditional regression learning settings. In particular, with all the available training samples, each attribute function in CA is trained to output either positive (*i.e.* one) or negative (*i.e.* zero) values, and a collection of such trained attribute functions, corresponding to a range of landmark points anchored in the 1D target space (e.g. integer ages), provides strong evidence for estimation of the target output. In contrast, regressors in traditional settings are trained to give a complete range of values in the target space, while regression fidelity for any specific target value is taken care of only by a (usually small) subset of training samples. This advantage of CA is particularly important for many regression problems in computer vision, such as human age estimation and crowd density estimation which often suffer from sparse and imbalanced training data.

The aforementioned collective evidence provided by trained attribute mapping functions and the attribute vector representation where each entry corresponds to a "landmark" (e.g. age) in a target space is intuitive and easy to manually select for 1D cases. However, the multivariate setting is more complex as there is no similarly unique way to divide the output space to "zeros" and "ones". We have already defined a multivariate model based on multiple

CA regressors (IndepCA), but its main weakness is that it does not exploit the multi-dimensional nature of the target space in multivariate regression, i.e. cross-correlations and interdependencies of output variables.

CartCA— The main problem in generalizing CA to multivariate cases is how to partition D-dimensional space such that it naturally represents the cumulative nature of attributes with their mutual dependency. As a novel solution, we propose a model termed *Cartesian Cumulative Attributes (CartCA)*.

Assume again that we have I training samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$. Considering a D-dimensional target $\boldsymbol{y}_i \in \mathbb{R}^D$, each component $y_{j=1,2,\ldots,D}$ will partition the training samples into two subsets as defined in (1). Now, if this is done for all j variables and their superpositions added by Cartesian product, the vector entries $\boldsymbol{y}_i$ collectively partition the training samples into 2^D subsets, which we denote as $\{\mathcal{S}_1, \ldots, \mathcal{S}_{2^D}\}$. These subsets of training samples suggest that we can learn 2^D different attribute functions anchored at the position $\boldsymbol{y}$ in the target space. For $k = 1, \ldots, 2^D$, CartCA assigns attribute labels $\{a_i^k\}$ to the training samples $\{\boldsymbol{x}_i\}$ based on the following rule

$$a_i^k = \begin{cases} 1, & \text{when } \boldsymbol{y}_i \in \mathcal{S}_k, \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Consider, for example, the particular case of two-dimensional targets, *i.e.,* $D = 2$. Then, the above rule for constructing the 2^D (in this case 4D) attribute tensors is given as follows

$$\begin{aligned}
a_{i,j}^{(1)} &= \begin{cases} 1, & \text{when } \tau_j^{(1)} \leqslant y_i^{(1)} \text{ and } \tau_j^{(2)} \leqslant y_i^{(2)}, \\ 0, & \text{otherwise}, \end{cases} \\
a_{i,j}^{(2)} &= \begin{cases} 1, & \text{when } \tau_j^{(1)} \geqslant y_i^{(1)} \text{ and } \tau_j^{(2)} \leqslant y_i^{(2)}, \\ 0, & \text{otherwise}, \end{cases} \\
a_{i,j}^{(3)} &= \begin{cases} 1, & \text{when } \tau_j^{(1)} \leqslant y_i^{(1)} \text{ and } \tau_j^{(2)} \geqslant y_i^{(2)}, \\ 0, & \text{otherwise}, \end{cases} \\
a_{i,j}^{(4)} &= \begin{cases} 1, & \text{when } \tau_j^{(1)} \geqslant y_i^{(1)} \text{ and } \tau_j^{(2)} \geqslant y_i^{(2)}, \\ 0, & \text{otherwise}, \end{cases}
\end{aligned} \tag{3}$$

where $\tau_j^{(1)}$ and $\tau_j^{(2)}$ are set similarly to the original CA and they have clear semantic meaning. For a training example, the two-dimensional output sets an anchor point to partition the $4D$ attribute tensor. An illustration of the above attribute label assignment rule is shown in Fig. 1, where the goal is to estimate the head pose yaw and pitch angles.

MvCA— One may notice that the number of attributes in CartCA increases exponentially with the dimensionality of target space, which makes learning of CartCA impractical in cases of high-dimensional target space and a small amount of data. In our experiments we found CartCA impractical for $D > 2$. As a remedy, we propose an approximate CartCA termed Multi-view Cumulative Attributes (MvCA). The MvCA attribute construction rule is based on CartCA which is still practical for $D = 2$ using (2).

More specifically, for training samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ in the D-dimensional target space, we first select an output dimension pair (j_1, j_2) with $j_1, j_2 \in \{1, \ldots, D\}$, $j_1 \neq j_2$, and project all the training samples into this CartCA subspace. For a fixed anchor point $\boldsymbol{y}_{i,\{j_1,j_2\}} \in \mathbb{R}^2$ in the CartCA sub-space, its entries partition the output space into 4 subsets (like those of Fig. 1), based on which MvCA uses 4 different "attribute planes" by following the rules in (3).

For studying complexity of CartCA and MvCA we may assume that the D_1 attribute spaces are similar. In this case, we have the total of D_1^2 possible anchor points in the attribute space. MvCA learns 4 attribute planes associated with each of the landmark points, and there are in total $D(D-1)/2$ such dimension pairs (j_1, j_2). MvCA learns attribute functions in the same way for each of the pairs, producing a total of $2D_1^2 D(D-1)$ attribute planes. For $D > 3$, this is significantly less than the corresponding number $(2D_1)^D$ for the CartCA.

In the case that the target space of multivariate regression is two-dimensional (a plane), *i.e.* $D = 2$, CartCA and MvCA are equivalent and give the same number of attribute features. In the case $D = 1$ all the original CA, IndepCA, CartCA and MvCA are equivalent. There are also recent works that could be used for dimensional reduction [33], but these are beyond the scope of this work.

Geometric Interpretation of CartCA and MvCA. We take CartCA as an example, but MvCA can be similarly analyzed. Attribute label assignment rule (2) suggests that each attribute function in CartCA is learned based on a unique binary partition of training samples. Each attribute function trained this way serves as a hyperplane,[1] gives an indicative measure of the position (*i.e.* multi-variate regression label) of test samples in the target space. In the following, we consider a particular test sample $\boldsymbol{x}$ with a ground-truth label $\boldsymbol{y} = \hat{\boldsymbol{y}}$.

- A group of 2^k attribute functions learned by the rule (2), (referring to rule (3) for samples on the boundary), which anchored at the position $\hat{\boldsymbol{y}}$ in the label space, *ideally* provide an *exact* indication on the target of $\boldsymbol{x}$: attributes given by these functions form a vector $\mathbf{1} \in \mathbb{R}^{2^k}$ with all entry values of 1 (any zero-valued entry in this vector indicates $\boldsymbol{y} \neq \hat{\boldsymbol{y}}$). When such a group of attribute functions are not available, attribute functions anchored at neighboring positions of $\hat{\boldsymbol{y}}$ form polytopes in the target space, which provide different levels of refined position information for the estimation of $\boldsymbol{y}$.
- Based on different (and unique) binary partitions of the target space, other attribute functions provide different half-space constraints for the estimation of $\boldsymbol{y}$. When these attributes are concatenated to the vector $\boldsymbol{a}_{CartCA}$, they collectively provide rich (and redundant) information for the estimation of $\boldsymbol{y}$.

An illustration of the above geometric interpretation is presented in Fig. 2. In summary, CartCA (or MvCA) encodes in the attribute vector $\boldsymbol{a}_{CartCA}$ (or $\boldsymbol{a}_{MvCA}$) strong information about the underlying position of any test sample in the target space, which can be exploited for final label estimation.

3.3. Two-stage regression

Given training samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ with input features $\boldsymbol{x}_i \in \mathbb{R}^N$ and output target vector $\boldsymbol{y}_i \in \mathbb{R}^D$, we construct the training attribute targets $\boldsymbol{a}_i \in \mathbb{R}^{D_1}$ based on the attribute construction rules in the previous sections.

To this end, we employ the Partial Least Squares (PLS) regression [19] for its capability to cope with multicollinearity problem, and which has recently been applied to a number of visual regression problems [27]. Typical solution for estimating the score (and loading) matrices is the NIPALS [20], which we adopt for its low computational complexity ($O(N^2)$). Alternatively, other multivariate regression models can also be employed such as multivariate ridge regression [12] and regression forests [6]. Partial least square regression is adopted owing to its simplicity in implementation and computational efficiency. PLS learns a mapping function $f : \mathbb{R}^N \mapsto \mathbb{R}^{D_1}$ from training data, which is used to estimate an attribute feature vector $\tilde{\boldsymbol{a}} \in \mathbb{R}^{D_1}$ for an unseen test sample $\boldsymbol{x}$ and is

[1] Alternative to the regression based attribute functions in our work, also any two-class (binary) classifier can be trained for the attribute assignments defined in (2). However, during our experiments we have found the real valued outputs of regressors, *soft attributes*, are more effective. This can be explained by the fact that no information is lost in the binary decisions and the whole pipeline is a regression pipeline.

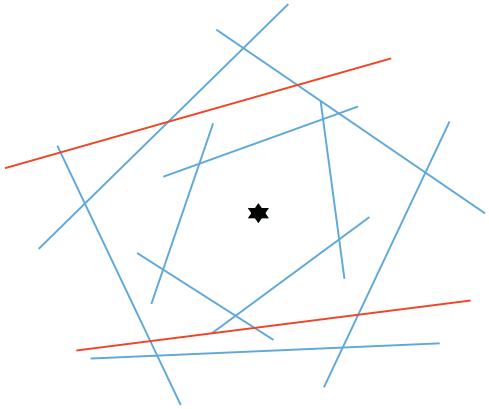

Fig. 2. Geometric intuition of the proposed Cartesian Cumulative Attributes. Attribute functions/hyperplanes (blue lines) form polytopes in the target space, which provide different levels of indicative position information on the target (dark star point) of a test sample. In the weaker form certain attributes provide half-space constraints (red lines) on the target of the test sample. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the first stage regressor in the proposed CartCA and MvCA regression methods.

To perform the second stage target estimation, we first estimate $\tilde{\boldsymbol{a}}_i = f(\boldsymbol{x}_i)$ and then concatenate $\boldsymbol{x}_i$ with $\tilde{\boldsymbol{a}}_i$. The concatenated vectors are used as the training data for the second stage multivariate regression. To learn a mapping function from the concatenated feature space to the multivariate target space, we adopt a few recent state-of-the-art methods, *e.g.* KPLS [27], KRF [15], and MLD [1] and compare them in our experiments. Our use of the existing methods is mainly to verify the effectiveness of our proposed CartCA and MvCA attribute features, by removing contributions from other factors.

4. Experiments

In the following, the proposed multi-output cumulative attribute space regression methods, IndepCA, CartCA and MvCA, are evaluated in multiple vision problems: 2D head pose estimation (2 pose angles), 3D head pose estimation (3 pose angles) and illumination estimation for color constancy (3 color correction terms for the red, green and blue channel).

4.1. Datasets and settings

Datasets— For 2D head pose estimation, we used the popular Pointing'04 benchmark dataset [34] which contains face images of 15 persons captured in varying appearance and under controlled indoor environment. For 3D head pose estimation, we used the Biwi Kinect Head Pose Estimation dataset [35], which contains depth images of 20 persons. As a distinct visual regression problem from head pose estimation we also evaluated our model with two illumination estimation datasets [30,36] where illuminant tri-stimulus value (Red, Green, Blue) is estimated to correct a color biased input image. The SFU Indoor dataset [36] contains 321 images captured in 11 different controlled lighting conditions. The SFU Color Checker dataset [30] contains 568 12-bits dynamic range images which all include the Macbeth Color Checker chart as groundtruth. Details of the datasets are given in Table 1.

Features— For 2D head pose estimation, after cropping the foreground of faces with manually-annotated bounding boxes, the facial images are normalized into 32×32 pixels from which we extract a 2511-dimensional histogram of oriented gradients (HoG) feature vector [37], which is widely employed in the recent works [1,15,26,27]. Encouraged by the significant advances with Convolutional Neural Networks (CNNs) in facial recognition [38], we also extract CNN features from the "*fc6*" layer of the pre-trained VGG-net 16 layer model [39].

For 3D head pose estimation, we first remove the background using the provided foreground masks by cropping 96×96 facial region anchored in the center of foreground masks. The cropped facial patches are then resized into 32×32 pixels. Inspired by the features used in [23,24], the depth value of each pixel in 32×32 patches were used as low-level features after which applying normalization of the non-zero pixel intensities (*i.e.* depth distance) into [0,1].

Finally, for the illumination estimation problem, we used the pre-trained 19-layer VGG-net without fine-tuning as described in [4]. For both SFU Indoor and Color Checker datasets, we follow the settings in [4] to extract 4096-dimensional CNN "*fc6*" features from images resized to 224×224.

Settings— For the Pointing'04 dataset, two experiments were conducted according to the settings of data split. In the first experiment, we followed the same training and testing partition as [1,15,26,27], *i.e.* five-fold cross-validation. An alternative setting, *i.e.* two image sequences of the same person evenly split into training and testing data, was adopted for the second experiment as in [15]. For the Biwi Kinect dataset, two experiments were conducted by 1) dividing the data into training part containing the images of the first 18 persons and testing part with the remaining images [23,24] and 2) by adopting five-fold cross-validation [23], respectively. For the SFU Indoor and Color Checker datasets, we followed the standard 3-fold cross-validation protocol in [4,29,31,32,40,41].

Comparative Methods— We collected most of the results of competitive approaches from corresponding papers. For ablation study with the 2D dataset we implemented several state-of-the-art methods including linear/kernel partial least square regression (PLS/KPLS) [27], k-cluster regression forests (KRF) [15], and multivariate label distribution learning (MLD) [1].

For 3D head pose estimation, we adopted standard regression forests (RF) [6] for the second layer multi-variate regression model owing to its strong performance in recent works [23,24].

For illumination estimation, we implement comparative multi-output support vector regression [4] in the light of its competitive performance. The number of factors for PLS and KPLS with RBF kernel is 25 and 40 respectively.

For KRF, we followed the setting in [15], the minimal size in each leaf node is 5 and we grew 20 regression trees. Following [1], MLD adopts weighted Jeffrey's divergence and two-dimensional Gaussian distribution with the finest granularity of head pose $\mu = 15$. Regression forests for 3D head pose estimation have at least the sample size of 5 in each leaf node and grow 20 regression trees. For illumination estimation, we used multi-output support vector regression (MSVR) [4] with the RBF kernel. Trade-off parameter C and γ of the RBF kernel were tuned by three-fold cross-validation.

We adopted the class labels to generate CartCA for 2D head pose estimation, while rounded values to nearest integers of 3D head pose angles are employed to generate CartCA and MvCA. For illumination estimation, we first normalised groundtruth illuminations into [0,255] levels, which are quantised into 64 bins in a cumulatively and continuously changing manner. The class label of each bin on each colour channel was adopted to generate CartCA and MvCA.

Performance Metrics— For evaluating the performance of head pose estimation, we employed two types of performance metrics, *i.e. regression metric* in Mean Absolute Error (MAE) and *classification metric*. Considering the different data characteristics in labels (*i.e.* integer angles in the Pointing'04 dataset and scalar values in the Biwi Kinect dataset), we report the classification accuracy of

Table 1
Details of the datasets used in the experiments. $D(i)$ = range of the ith output dimension (2D face pose: yaw, pitch; 3D face pose: +roll; color constancy: color corrections c_R, c_G, c_B).

Data	# of imgs	Resolution	$D(1)$	$D(2)$	$D(3)$	Note
Face pose						
Pointing'04 [34]	2790	384×288	$[-90°, 90°]$	$[-90°, 90°]$	–	13 yaw and 9 pitch angles
Biwi Kinect [35]	15,677	640×480	$[-67°, 77°]$	$[-84°, 54°]$	$[-70°, 63°]$	float values
Color constancy						
SFU Indoors	321	224×224	[0, 255]	[0, 255]	[0, 255]	RGB values
SFU Color Checker	568	224×224	[0, 255]	[0, 255]	[0, 255]	RGB values

Table 2
Comparison with state-of-the-art on 2D head pose estimation with the Pointing'04 dataset (5-fold cross-validation). For MAE smaller number is better and for classification accuracy larger number is better. Note that for 2-output regression CartCA and MvCA are equivalent.

Method	Regression Metric (MAE)			Classification Metric(Accuracy)		
	Yaw	Pitch	Yaw+Pitch	Yaw	Pitch	Yaw+Pitch
Various feature combinations						
Fenzi [43]	5.9°	6.7°	–	–	–	–
AKRF-V [44]	5.5°	2.8°	–	–	–	–
SDL [45]	$4.12° \pm 0.17°$	$2.09° \pm 0.12°$	–	–	–	–
PLS [27]*	$8.97° \pm 0.87°$	$9.27° \pm 0.41°$	$15.51° \pm 0.53°$	$49.25\% \pm 3.37\%$	$46.38\% \pm 3.19\%$	$23.15\% \pm 1.04\%$
HoG Features						
KPLS [27]*	$5.89° \pm 0.83°$	$5.76° \pm 0.25°$	$10.28° \pm 0.70°$	$64.87\% \pm 4.30\%$	$65.34\% \pm 2.08\%$	$44.34\% \pm 2.58\%$
KRF [15]*	$5.49° \pm 0.27°$	$3.90° \pm 0.65°$	$8.79° \pm 0.61°$	$64.52\% \pm 1.97\%$	$76.67\% \pm 3.73\%$	$47.53\% \pm 2.90\%$
MLD [1]*	$\mathbf{4.41° \pm 0.57°}$	$2.83° \pm 0.62°$	$6.74° \pm 0.70°$	$71.61\% \pm 3.12\%$	$84.98\% \pm 2.19\%$	$61.76\% \pm 3.84\%$
IndepCA	$4.31° \pm 0.83°$	$2.76° \pm 0.66°$	$6.53° \pm 0.76°$	$72.87\% \pm 4.30\%$	$85.34\% \pm 2.08\%$	$63.84\% \pm 4.34\%$
CartCA/MvCA	$\mathbf{4.09° \pm 0.70°}$	$\mathbf{2.60° \pm 0.69°}$	$\mathbf{6.22° \pm 0.80°}$	$\mathbf{74.01\% \pm 3.94\%}$	$\mathbf{86.95\% \pm 2.47\%}$	$\mathbf{65.59\% \pm 4.12\%}$
VGG-Net Features						
CNN	$4.81° \pm 0.23°$	$\mathbf{1.85° \pm 0.17°}$	$6.67° \pm 0.16°$	$68.96\% \pm 1.08\%$	$\mathbf{89.93\% \pm 1.24\%}$	$61.58\% \pm 1.22\%$
KPLS	$4.72° \pm 0.29°$	$4.45° \pm 0.39°$	$8.38° \pm 0.44°$	$71.25\% \pm 1.51\%$	$72.11\% \pm 2.26\%$	$51.79\% \pm 2.51\%$
KRF	$5.37° \pm 0.67°$	$3.76° \pm 0.51°$	$8.71° \pm 0.23°$	$65.60\% \pm 4.12\%$	$76.95\% \pm 2.76\%$	$48.52\% \pm 1.15\%$
MLD	$3.53° \pm 0.34°$	$2.13° \pm 0.22°$	$5.37° \pm 0.37°$	$77.49\% \pm 2.22\%$	$88.71\% \pm 1.25\%$	$69.10\% \pm 1.72\%$
IndepCA	$3.44° \pm 0.26°$	$2.18° \pm 0.31°$	$5.33° \pm 0.48°$	$77.81\% \pm 2.53\%$	$88.71\% \pm 2.23\%$	$69.32\% \pm 2.64\%$
CartCA/MvCA (ours)	$\mathbf{3.25° \pm 0.34°}$	$2.04° \pm 0.45°$	$\mathbf{5.01° \pm 0.69°}$	$\mathbf{78.96\% \pm 2.04\%}$	$89.21\% \pm 2.29\%$	$\mathbf{70.93\% \pm 2.90\%}$

*Results are slightly different from those reported in the paper because of using our own implementation

predicted poses with respect to the ground truth [1] for 2D head pose estimation and used *Cumulative Score* (CS) defined in [42] for 3D head pose estimation as the classification metrics, respectively. Following [30,36], for illumination estimation we measured the angular error (cosine distance) ε between estimated illumination $I \in \mathbb{R}^3$ and groundtruth $I_{gt} \in \mathbb{R}^3$:

$$\varepsilon_{I,I_{gt}} = \arccos\left(\frac{I^T I_{gt}}{\| I \| \| I_{gt} \|}\right),$$

where $\| \cdot \|$ is the Euclidean norm. We report median and mean value of $\varepsilon_{I,I_{gt}}$ of all test samples.

4.2. Comparative evaluation

2D Head Pose Estimation— We compared our IndepCA, CartCA and MvCA with a number of recent methods on the Pointing'04 datasets. The results of these experiments are shown in Table 2. Among the methods, PLS [27], KPLS [27], KRF [15], and MLD [1] use identical HoG and VGG-Net features as our approach. Since our models can use any general purpose regressor we selected MLD since it performed well both in the original paper and in our experiments. Interestingly, our multivariate baseline IndepCA is on par with the existing methods using traditional features (HoG) and clearly superior with the deep CNN features. However, in the both cases the proposed CartCA/MvCA is more accurate.

In order to further assess the significance of the feature set, we also fine-tuned the VGG-Net end-to-end in the same evaluation setting. More specifically, we used the VGG convolutional pipeline, with two output layers in place of the original 1000-class output-layer. The parallel output layers predict the yaw and the pitch angle encoded as two independent classification problems. The network was trained using the negative log-likelihood loss and softmax activations individually for both yaw and pitch targets. Moreover, we tested alternative network structures: the ResNet50 base network as well as alternative target encodings. It turned out that clearly the best results are obtained using the VGG-Net structure and classification encoding (each yaw and pitch angle is one class) instead of the regression target (the two output layers have linear activation and are directly predicting the yaw and pitch angles).

It can be seen that in most cases, the end-to-end network is inferior to the proposed approach. The network is able to predict the pitch (vertical) angle better than alternative methods, but performs poorly on yaw angle prediction rendering the yaw+pitch metric inferior, as well. The inferior performance in horizontal angle prediction may be due to the larger number of classes in this direction (13 yaw angles, 7+2 pitch angles), which decreases the number of training samples per class and causes the network to overfit to the relatively small training set.

Finally, in order to assess the general suitability of a CNN for multivariate regression problems, we also considered using the original VGG-Net features with a neural network classifier. More specifically, we trained the described network architecture with frozen convolutional layers, forcing the network to use exactly same features as the other methods. The results are discouraging as the errors are up to three times higher than the best ones in Table 2. This is an indication that a plain dense neural network may not be ideal for multivariate regression tasks (note, however, successful results in related tasks with *e.g.,* autoencoder structure [25]), and even better results could be obtained by coupling the fine-tuned convolutional pipeline with the proposed CartCA/MvCA.

Table 3
Comparison with state-of-the-art on 3D head pose estimation with the Biwi Kinect Database (data split 1: 18 persons for training and the remaining for testing; data Split 2: five-fold cross-validation).

Method	Data Split 1 (MAE)				Data Split 2 (MAE)			
	Yaw	Pitch	Roll	Y+P+R	Yaw	Pitch	Roll	Y+P+R
HF [46]*	3.79°	9.27°	6.62°	13.48°	8.9°	8.5°	7.9°	–
ADF [24]*	3.54°	7.87°	5.39°	11.48°	–	–	–	–
ARF [47]*	3.52°	8.18°	4.77°	11.17°	–	–	–	–
RF [35]	3.80°	3.50°	5.40°	–	–	–	–	–
KPLS [19]	1.90°	1.48°	1.80°	3.47°	2.01° ± 0.06°	1.63° ± 0.03°	1.80° ± 0.06°	3.65° ± 0.06°
RF-i**	1.95°	1.50°	1.94°	3.72°	2.00° ± 0.07°	1.49° ± 0.04°	1.96° ± 0.05°	3.77° ± 0.10°
RF-s**	1.59°	1.20°	1.39°	2.84°	1.79° ± 0.11°	1.31° ± 0.07°	1.47° ± 0.05°	3.11° ± 0.15°
IndepCA	1.51°	1.23°	1.37°	2.80°	1.77° ± 0.13°	1.34° ± 0.14°	1.45° ± 0.04°	3.10° ± 0.18°
CartCA	1.42°	1.29°	1.40°	2.74°	1.71° ± 0.15°	1.30° ± 0.11°	1.46° ± 0.06°	3.05° ± 0.18°
MvCA	**1.39°**	**1.15°**	**1.35°**	**2.64°**	**1.63° ± 0.10°**	**1.24° ± 0.06°**	**1.43° ± 0.06°**	**2.92° ± 0.14°**

* uses foreground detection; ** is based on our implementation of [6].

Table 4
Comparison with state-of-the-art on color constancy with the SFU Indoor and Color Checker datasets. Median and mean angular errors between estimated and ground truth illuminant (RGB) are reported as the errors (smaller is better). We use identical deep features to MSVR [4].

	SFU Indoor		SFU Color Checker	
	Median	Mean	Median	Mean
second-order Gray Edge (2stGE) [48]	2.7	5.2	4.4	5.1
Weighted Gray Edge (WGE) [49]	2.4	5.6	–	–
Gamut Mapping (GM-pixel) [50]	2.3	3.7	**2.3**	4.2
Natural Image Statistics (NIS) [40]	–	–	3.1	4.2
Exemplar [31]	–	–	**2.3**	**2.9**
Grey Pixel (std) [29]	2.5	5.7	3.2	4.7
Grey Pixel (edge) [29]	2.3	5.3	3.1	4.6
MSVR [4]	1.9	3.1	2.8	4.3
IndepCA	1.8	3.0	2.6	4.2
CartCA	1.8	3.0	2.7	4.2
MvCA	**1.6**	**2.8**	2.6	4.1

3D Head Pose Estimation— Two experiments were conducted using different settings for data splitting and the results are in Table 3. Since the original random forest regression (RF-i and RF-s) in [6] performed well with the selected depth features we used RF as the regressor with our methods as well. Similar to previous 2D head pose estimation, IndepCA is on pair (better results for 6 out of the 8 possible measure) with state-of-the-art (RF-i/s). However, the two proposed extensions better exploiting output interdependencies, CartCA and MvCA, provide the best results. MvCA performed better than CartCA which can be explained by the limited number of training data - 2D projections of MvCA seem to robustify regression as compared to full CartCA.

Illumination Estimation— Table 4 compares our methods with the state-of-the-art illumination estimation algorithms on the SFU Indoor and Color Checker datasets. Our method achieves the best performance on both performance metrics on the SFU Indoor dataset, and our result is comparable to state-of-the-art on the SFU color checker. It is noteworthy that our results are always better than MSVR [4] who use identical deep features. Again, IndepCA performed well and MvCA was the best of the three proposed methods.

Computational Cost— The additional complexity of the proposed CA models yields from the mid-layer presentation, attribute vector, for which two regressors need to be trained. In traditional visual regression there is a single regressor which maps N input variables to D output variables. The computational complexities (sized of the attribute vectors) and the actual numbers for the three problems are shown in Table 5.

Table 5
The CA space sizes for the proposed models. Note that only CartCA and MvCA can represent cross-correlations between the output dimensions.

	2D Head	3D Head	Color constancy
IndepCA	22	418	192
CartCA	186	$2.7 \cdot 10^7$	$2.0 \cdot 10^6$
MvCA	186	$2.3 \cdot 10^5$	$4.9 \cdot 10^4$

Table 6
Comparison of the proposed CA spaces with various regressors for the second regression stage. Results correspond to the Yaw+Pitch MAE and classification accuracies with the Pointing'04 benchmark.

Method	Pointing'04	
	MAE	Accuracy
KPLS [27] + HoG		
IndepCA	10.80° ± 0.68°	41.72% ± 3.62%
MvCA	**7.52°** ± 0.74°	**56.77%** ± 5.23%
KRF [15] + HoG		
IndepCA	8.85° ± 0.76°	51.54% ± 4.41%
MvCA	**7.92°** ± 0.69°	**53.33%** ± 3.15%
MLD [1] + HoG		
IndepCA_{MLD}	6.53° ± 0.76°	63.84% ± 4.34%
MvCA_{MLD}	**6.22°** ± 0.80°	**65.59%** ± 4.12%

4.3. Ablation study

CA Mapping— In order to validate the claim that the proposed Cartesian cumulative attribute multivariate regression (CartCA) and its multi-view projection based approximation (MvCA) provide accuracy improvement over the straightforward IndepCA we conducted an ablation study where the different CA spaces were compared using different regressors but with the same visual features. The results are shown in Table 6. In all cases the higher dimensional CA spaces provided superior accuracy. However, it is obvious that this finding is most evident with more traditional regressors such as KPLS [27]. The more advanced regressors, such as KRF [15] and MLD [1], exploit output correlations more efficiently and therefore differences between IndepCA and CartCA/MvCA are less significant.

Concatenating with Imagery Features— During the experiments, we found that the best peformance was achieved by concatenating original imagery features and cumulative attributes for the second stage regression. In this experiment this finding was verified with the both face pose and color constancy datasets. The results are shown in Table 7 that clearly indicates that concatenation provides small but systematic improvement in all cases.

Table 7
The proposed CA spaces with ($+\boldsymbol{x}$) vs. without the original input features concatenated in the second stage regression.

Method	Pointing'04		SFU Indoor	
	MAE	Accuracy	Median	Mean
IndepCA	$5.40° \pm 0.37°$	$69.05\% \pm 2.87\%$	1.9	3.0
IndepCA + $\boldsymbol{x}$	$\mathbf{5.33° \pm 0.48°}$	$\mathbf{69.32\% \pm 2.64\%}$	**1.8**	**3.0**
MvCA	$5.15° \pm 0.38°$	$70.55\% \pm 2.31\%$	1.7	2.9
MvCA + $\boldsymbol{x}$	$\mathbf{5.01° \pm 0.69°}$	$\mathbf{70.93\% \pm 2.90\%}$	**1.6**	**2.8**

5. Conclusions

In this work, we investigated Cumulative Attribute space regression that has been found effective in many computer vision regression problems. In particular, we studied how correlations in the target label space can be exploited for improved accuracy. To this aim, we extended CA to 2-output and 3-output regression problems by introducing a general Cartesian CA (CartCA) and its multivariate approximation using multi-view CartCA projections–MvCA. The proposed CartCA and MvCA models are generally applicable and demonstrate systematic performance boost in 2-output and 3-output computer vision regression problems.

In the experimental section we compared the proposed methodology with state of the art deep neural networks. It is noteworthy that the CNN does not excel in this domain, unlike most areas of machine learning today. This is likely due to the small training sample size as well as the challenges in encoding regression problems for neural networks. This highlights the key benefit of cumulative attributes: they divide the regression problem into a number of binary classification problems. This increases the amount of data for each task by several orders of magnitude.

Our future work will address higher dimensional generalizations of CartCA and MvCA and their applications in general multivariate regression. Moreover, integrating the idea of (multivariate) cumulative attributes with state-of-the-art classifiers–deep neural networks–would bring together the best of both worlds: data-hungry but accurate deep learning and economical cumulative attribute models.

Acknowledgments

This work is funded by the Academy of Finland under Grants No. 267581, 309903 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] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1837–1842.
[2] K. Chen, K. Jia, Z. Zhang, J.-K. Kämäräinen, Spectral attribute learning for visual regression, Pattern Recognit. 66 (2017) 74–81.
[3] Q. Liu, J. Yang, J. Deng, K. Zhang, Robust facial landmark tracking via cascade regression, Pattern Recognit. 66 (2017) 53–62.
[4] Y. Qian, K. Chen, J.-K. Kämäräinen, J. Nikkanen, J. Matas, Deep structured-output regression learning for computational color constancy, in: Proceedings of International Conference on Pattern Recognition, 2016.
[5] W.-L. Chao, J.-Z. Liu, J.-J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression, Pattern Recognit. 46 (3) (2013) 628–641.
[6] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[7] A. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[8] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: Proceedings of Advances in Neural Information Processing Systems, 2007, pp. 41–48.
[9] S. Nowozin, C.H. Lampert, Structured learning and prediction in computer vision, Found. Trends. Comput. Graph. Vis. 6 (3–4) (2011) 185–365.
[10] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3337–3344.
[11] D. Parikh, K. Grauman, Relative attributes, in: Proceedings of International Conference on Computer Vision, 2011, pp. 503–510.
[12] 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, pp. 2467–2474.
[13] Q. Tian, S. Chen, Cumulative attribute relation regularization learning for human age estimation, Neurocomputing 165 (2015) 456–467.
[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodological) (1996) 267–288.
[15] K. Hara, R. Chellappa, Growing regression forests by classification: applications to object pose estimation, in: Proceedings of European Conference on Computer Vision, 2014, pp. 552–567.
[16] 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, pp. 530–543.
[17] 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, pp. 951–958.
[18] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by support vector ranking, in: Proceedings of British Machine Vision Conference, 2010, pp. 21.1–21.11.
[19] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[20] H. Wold, Soft modeling by latent variables: the nonlinear iterative partial least squares approach, Perspectives in probability and statistics, papers in honour of MS Bartlett (1975).
[21] X. Yang, W. Liu, D. Tao, J. Chen, Canonical correlation analysis networks for two-view image recognition, Inf. Fusion 385–386 (2018) 338–352.
[22] C. Huang, X. Ding, C. Fang, Head pose estimation based on random forests for multiclass classification, in: Proceedings of International Conference on Pattern Recognition, 2010, pp. 934–937.
[23] G. Fanelli, J. Gall, L. Van Gool, Real time head pose estimation with random regression forests, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 617–624.
[24] S. Schulter, P. Wohlhart, C. Leistner, A. Saffari, P.M. Roth, H. Bischof, Alternating decision forests, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 508–515.
[25] J. Yu, C. Hong, Y. Rui, D. Tao, Multitask autoencoder model for recovering human poses, IEEE Trans. Ind. Electron. 65 (6) (2018) 5060–5068.
[26] J. Foytik, V.K. Asari, A two-layer framework for piecewise linear manifold-based head pose estimation, Int. J. Comput. Vis. 101 (2) (2013) 270–287.
[27] 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, pp. 2602–2609.
[28] H.R.V. Joze, M.S. Drew, G.D. Finlayson, P.A.T. Rey, The role of bright pixels in illumination estimation, in: Proceedings of Color Imaging Conference, 2012.
[29] K.-F. Yang, S.-B. Gao, Y.-J. Li, Efficient illuminant estimation for color constancy using grey pixels, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[30] P.V. Gehler, C. Rother, A. Blake, T. Minka, T. Sharp, Bayesian color constancy revisited, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[31] H.R.V. Joze, M.S. Drew, Exemplar-based color constancy and multiple illumination, IEEE Trans. Pattern Anal. Mach. Intell. 36 (5) (2014) 860–873.
[32] S. Bianco, C. Cusano, R. Schettini, Color constancy using cnns, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015.
[33] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Trans. Image Process. 27 (5) (2018) 2420–2432.
[34] N. Gourier, D. Hall, J.L. Crowley, Estimating face orientation from robust detection of salient facial structures, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition Net Workshop on Visual Observation of Deictic Gestures, 2004, pp. 1–9.
[35] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3D face analysis, Int. J. Comput. Vis. 101 (3) (2013) 437–458.
[36] K. Barnard, L. Martin, B. Funt, A. Coath, A data set for color research, Color Res. Appl. 27 (3) (2002) 147–151.
[37] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[38] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition., in: Proceedings of the British Machine Vision Conference, 2015.
[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[40] A. Gijsenij, T. Gevers, Color constancy using natural image statistics and scene semantics, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2011) 687–698.
[41] A. Chakrabarti, K. Hirakawa, T. Zickler, Color constancy with spatio-spectral statistics, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1509–1519.
[42] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, TPAMI 29 (12) (2007) 2234–2240.
[43] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, J. Ostermann, Class generative models based on feature regression for pose estimation of object categories, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 755–762.

[44] K. Hara, R. Chellappa, Growing regression tree forests by classification for continuous object pose estimation, Int. J. Comput. Vis. (2016) 1–21.
[45] X. Zhen, M. Yu, A. Islam, M. Bhaduri, I. Chan, S. Li, Descriptor learning via supervised manifold regularization for multioutput regression, IEEE Trans. Neural Netw. Learn. Syst. (2016).
[46] G. Fanelli, T. Weise, J. Gall, L. Van Gool, Real time head pose estimation from consumer depth cameras, in: Annual Symposium of the German Association for Pattern Recognition, 2011, pp. 101–110.
[47] S. Schulter, C. Leistner, P. Wohlhart, P.M. Roth, H. Bischof, Alternating regression forests for object detection and pose estimation, in: Proceedings of International Conference on Computer Vision, 2013, pp. 417–424.
[48] J. Van De Weijer, T. Gevers, A. Gijsenij, Edge-based color constancy, IEEE Trans. Image Process. 16 (9) (2007) 2207–2214.
[49] A. Gijsenij, T. Gevers, J. Van De Weijer, Improving color constancy by photometric edge weighting, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 918–929.
[50] K. Barnard, Improvements to gamut mapping colour constancy algorithms, in: Proceedings of European Conference on Computer Vision, 2000.

Ke Chen was born in Wuxi, China, in 1985. He received the B.E. degree in automationand the M.E. degree in software engineering from Sun Yat-sen University, Guangzhou, China, in 2007 and 2009, respectively, under the supervision of Prof. Y. Zhang, and the Ph.D degree in computer vision from the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K., under the supervision of Prof. S. Gong and Prof. T. Xiang. He is currently the Academy of Finland Post-Doctoral Research Fellow with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He has published over 40 peer-reviewed conference and journal papers in computer vision, neural networks, and robotics. His current research interests include computer vision, pattern recognition, neural dynamic modeling, and robotic inverse kinematics.

Kui Jia is currently a Professor with School of Electronic and Information Engineering,South China University of Technology. He received the B.Eng. degree in marine engineering from Northwestern Polytechnical University, China, in 2001, the M.Eng. degree in electrical and computer engineering from the National University of Singapore in 2003, and the Ph.D. degree in computer science from Queen Mary University of London, London, U.K., in 2007. His research interests are in computer vision, machine learning, and image processing. On these areas, he has authored many publications at prestigious journals and conferences, such as the IEEE T-PAMI, IJCV, T-IP, T-SP, CVPR, ICCV, and ECCV. His recent research focuses on theoretical deep learning and its applications in various computer vision problems, including object recognition, analysis of human activities, and deep learning of 3D data.

Heikki Huttunen is an associate professor at Tampere University of Technology, Finland. He holds M.Sc. and Ph.D degrees from University of Tampere and Tampere University of Technology in 1995 and 1999, respectively. He leads the Machine Learning Group and his research interests are in machine learning deployment, to bring real time machine learning into embedded and mobile devices.

Jiri Matas is a professor at the Center for Machine Perception, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic. He is author or coauthor of more than 250 papers in the area of computer vision and machine learning. His research interests include object recognition, image retrieval, tracking, sequential pattern recognition, invariant feature detection, and Hough transform and RANSAC-type optimization.

Joni-Kristian Kämäräinen received the M.Sc. and Ph.D. degrees from the Lappeenranta University of Technology, Lappeenranta, Finland, in 1999 and 2003, respectively. He is currently an Associate Professor of Signal Processing with the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, where he leads the Computer Vision Group. His current research interests include 2-D and 3-D scene analysis, object detection and recognition, signal processing, and machine intelligence.