Soft-Margin Mixture of Regressions

Dong Huang*
Carnegie Mellon University
donghuang@cmu.edu

Longfei Han*
Beijing Institute of Technology
hanlongfei@hotmail.com

Fernando De la Torre
Carnegie Mellon University
ftorre@cs.cmu.edu

Abstract

Nonlinear regression is a common statistical tool to solve many computer vision problems (e.g., age estimation, pose estimation). Existing approaches to nonlinear regression fall into two main categories: (1) The universal approach provides an implicit or explicit homogeneous feature mapping (e.g., kernel ridge regression, Gaussian process regression, neural networks). These approaches may fail when data is heterogeneous or discontinuous. (2) Divide-and-conquer approaches partition a heterogeneous input feature space and learn multiple local regressors. However, existing divide-and-conquer approaches fail to deal with discontinuities between partitions (e.g., Gaussian mixture of regressions) and they cannot guarantee that the partitioned input space will be homogeneously modeled by local regressors (e.g., ordinal regression). To address these issues, this paper proposes Soft-Margin Mixture of Regressions (SMMR), a method that directly learns homogeneous partitions of the input space and is able to deal with discontinuities. SMMR outperforms the state-of-the-art methods on three popular computer vision tasks: age estimation, crowd counting and viewpoint estimation from images.

1. Introduction

Nonlinear regression is a common statistical tool to solve many computer vision applications such as age estimation [21], crowd counting [2] or pose estimation [27, 18]. These methods typically learn a mapping from hand-crafted features (e.g., Histogram of Oriented Gradients(HoG), Scale-Invariant Feature Transform(SIFT)) to the desired output (e.g., facial attributes, pose angles, landmarks). Recently, deep learning architectures (e.g., [33, 35]) directly learn a convolutional nonlinear mapping from images to outputs and achieved state-of-the-art results. Despite the exciting advances in deep learning, it is unclear how optimal these techniques are in the case of naturalistic input data that is heterogeneous, non-uniformly sampled, or discontinuous. Recall that in many real problems it is labor intensive to collect well-sampled training data and heuristic to select training batches [13]; Moreover, the features learned are difficult to be shared among different databases. To address these issues in the standard regression and deep learning methods, novel nonlinear regression approaches are needed to deal with heterogeneous and discontinuous data.

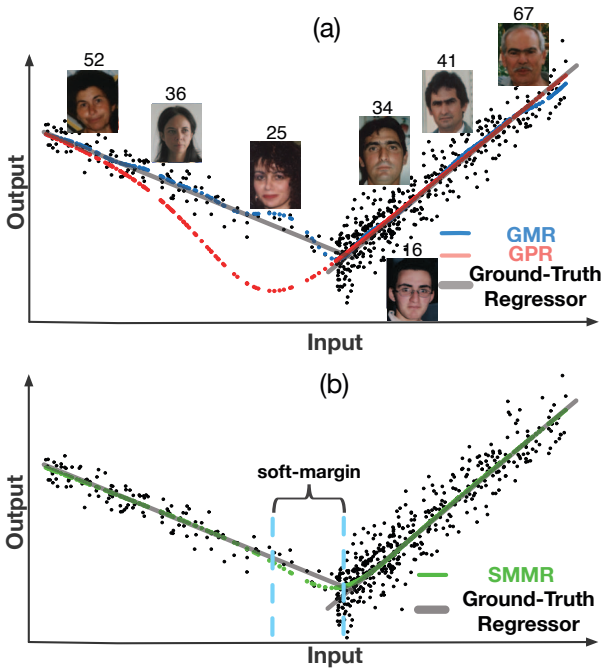

Figure 1. Nonlinear regression for the problem of age estimation (number on top of the face image). Gray dots denote data points. Gray lines denote the ground truth nonlinear regressor between input features and outputs. The colored dotted curves denote the predictions for different methods. (a) Gaussian Process Regression (GPR), the red dotted curve and Gaussian Mixture Regression (GMR), the blue dotted curve. (b) Soft-Margin Mixture of Regressions (SMMR), the green dotted curve (**Best viewed in color**).

Existing approaches to nonlinear regression fall into two main categories: (1) Universal approaches and (2) Divide-and-conquer approaches. Universal approaches find a implicit or explicit global non-linear mapping. Examples of

*These authors contributed equally to this work.

1063-6919/17 $31.00 © 2017 IEEE
DOI 10.1109/CVPR.2017.432

this category are Kernel Ridge Regression (KRR) [1], Kernel Support Vector Regression (KSVR) [21], Kernel Partial Least Square Regression (PLSR) [11], the covariance function in Gaussian Process Regression (GPR) [2], the Logistic function used in Bayesian approaches, or the Sigmoid and Rectifier functions used in neural networks. Given discontinuous and heterogeneous training data, an universal regressor is inevitably biased by the data distribution. That is, the model will incur in low regression error in densely sampled space while high error in everywhere else. To model heterogeneous data, divide-and-conquer approaches learn multiple local regressors. For instance, the hierarchical-based [14] and tree-based regression [15] make hard partitions according to outputs, and the subsets of samples may not be homogeneous for learning local regressors. Finite Mixture of Regressions (FMR), such as Gaussian Mixture of Regressions (GMR) [26], distributes regression error among local regressors by maximizing likelihood in the joint input-output space. The output of an input feature is then computed as a weighted combination of outputs by all local regressors. In existing FMR approaches[34, 19], regressors learned with more training data tend to dominate the final output estimation.

Figure 1 (a) illustrates above disadvantages in the problem of age estimation from images. Each gray "·" denotes a data sample, that corresponds to an input image feature (e.g., a BIF vector [21] and an output scalar (i.e., the age of the subject). The gray lines denote the true regressor from inputs to outputs. We trained regressors using an example of universal approaches (GPR) and an example of divide-and-conquer approach (GMR) on all data samples. Then we plotted their output predictions as colored dotted curves in Figure 1. In this example the data is (1) *heterogeneous*: the data cloud on the left is drawn from a Gaussian distribution and the data cloud on the right drawn from exponential distribution; (2) *non-uniformly sampled*. Using GPR (Universal approach), the output predictions (the red dotted curve in Figure 1 (a)), fit the true regression (the gray line in Figure 1 (a)) poorly when the data samples are scattered and discontinuous. Using GMR, two local regressors were learned, one for the point cloud on the left and the other for the point cloud on the right. The output prediction of GMR (the blue dotted curve in Figure 1 (a)) was computed as a weighted sum of predictions by the two local regressors. Because the densely sampled data on the right produces much higher weights, the output prediction of GMR on the left was dominated by the prediction on the right. As a result, the output prediction of GMR (the blue dotted curve in Figure 1 (a)) is far from the true regression.

To address the above mentioned problems, we propose Soft-Margin Mixture of Regressions (SMMR) for solving heterogeneous regression problems. SMMR simultaneously finds homogeneous partitions in the joint input-output space using max-margin classification, and learns a local regressor for each partition. Using the hinge-loss as mixing proportions, SMMR minimizes the regression error in the soft margin between partitions. SMMR uses the hinge loss to model the transition between regressors, it produces "0" weights within partitions and smooth weights between partitions. This property prevents one regressor from dominating other regressors in estimating final outputs. By contrast, SVR just uses the hinge loss on the regression error to exclude outliers in training samples. Our approach effectively reduces the overall regression error. Observe that output predictions made by SMMR (the green dotted curve in Figure 1 (b)) accurately fit the true regression (the gray line in Figure 1 (b)). We applied SMMR to three computer vision tasks: facial age estimation, crowd counting and viewpoint estimation, and in all of the problems SMMR outperformed state-of-the-art results.

2. Related Work

This section reviews the regression approaches developed for facial age estimation, crowd counting, and object viewpoint estimation.

Facial age estimation from images has extensive applications in visual surveillance, access control, and demographics analysis. Using hand-craft features (e.g., AAM [4] or BIF [21]) as inputs, a variety of universal regression approaches were used: Gaussian Process Regression (GPR) [37], Kernel Support Vector Regression (KSVR) [21], Kernel Partial Least Square Regression (KPLS) [11]. However, Human face matures in *non-stationary* patterns [22] at different age. Facial aging effects appear as changes in the shape of the face during childhood and changes in skin texture during adulthood. Dividing data by ages, hierarchical models [14] and group-specific regression have produced good results. While this hierarchical model tries to overcome the error mitigation by empirically splitting the label space with overlapping ranges, it may not find homogeneous subsets for learning local regressors. Ordinal regression [5, 22] performs a series of binary classification to partition the samples according to ages, and estimates ages by summing over classifier outputs. Moreover, ordinal regression is limited to scalar outputs.

Crowd counting [2] is an essential video analysis tool for public surveillance, security assessment and traffic statistics. The goal of crowd counting is to estimate number of persons from images. Using hand-crafted features [3], universal approaches, e.g. Ridge Regression [7], Gaussian Process Regression [2] and kernel-based regression [1] have been applied to crowd counting. Many divide-and-conquer approaches outperform above universal approaches. Chan *et al*. [2] firstly segmented the image into components of homogeneous patches, then used Gaussian Process Regression to estimate the number of people per segment. Zhang

et al. [35] adapted a CNN model for the sampling differences between training and testing data. To compute the similarity between the training and testing data, perspective maps were required between training and test scene.

Object viewpoint estimation was solved as either a classification problem and a regression problem. Here, we focus on the regression problem. Two strategies have been explored: (1) Using 3D information to model the configuration of object parts[38, 32, 28]. For instance, Pepik *et al*. [24] extended the 2D Deformable Part Models (DPM) to 3D DPM, and performed interpolation inference to produce the continuous estimation. (2) Using 2D image features to estimate Object viewpoints. Torki *et al*. [29] built a manifold representation of an object class and a regression from manifold to object viewpoint. Recent work performed divide-and-conquer: He *et al*. [16] partitioned data under different viewpoints by classification, and performed refined estimation by regression. Fenzi *et al*. [9] learned an object class representation by aggregating local features using spectral clustering. Local regressors were then learned for each cluster and used to estimate viewpoint by weighted combination. Hara and Chellappa[15] used a K-Clusters Regression Forest: k-means clustering + regression forest. These partition approaches relies solely on either inputs or output. In a partitioned subset, the input-output correlation may not be homogeneous making it difficult to learn accurate local regressors.

3. Finite mixture of regression

Finite Mixture of Regression (FMR) is the most basic divide-and-conquer approach. Most other approaches can be understood as extensions of FMR, including our SMMR approach.

Denote[1] $\mathbf{x}_i \in \Re^{d_x}$ the vector representation of the i^{th} input feature and $\mathbf{y}_i \in \Re^{d_y}$ the output vector of $\mathbf{x}_i$ ($i = 1, \cdots, n$). Finite Mixture of Regression (FMR) splits the n pairs of samples $\{\mathbf{x}_i, \mathbf{y}_i\}$s into k subsets, and learn a local regressor for each subset. Without loss of generality, the regressor of the j^{th} subset ($j = 1, \cdots, k$) is a linear mapping

$$f(\mathbf{y}|\mathbf{x}, z = j) = \phi(\mathbf{y}; \boldsymbol{\beta}_j^T \hat{\mathbf{x}}, \boldsymbol{\sigma}_j^2), \tag{1}$$

where $\boldsymbol{\beta} \in \Re^{d_y \times (d_x+1)}$ is the regression coefficients, $\hat{\mathbf{x}} = [1, \mathbf{x}]$. z is a latent variable that denotes the affiliation of $\{\mathbf{x}, \mathbf{y}\}$ to a subset. $\phi(\cdot)$ is a density function of regression error, e.g., Gaussian error $\mathcal{N}(0, \boldsymbol{\sigma}^2)$. The conditional density of FMR is computed by summing over local regressors:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \sum_z p(z|\mathbf{x}) p(\mathbf{y}|\mathbf{x}, z) \\ &= \sum_{j=1}^{k} \pi_j \phi(\mathbf{y}|\boldsymbol{\beta}_j^T \hat{\mathbf{x}}, \boldsymbol{\sigma}_j^2), \end{aligned} \tag{2}$$

where π_j is the mixing proportions of the j^{th} regressor to the FMR output. π_js are positive and $\sum_{j=1}^{k} \pi_j = 1$. The output $\mathbf{y}_i$ is estimated as a weighted combination over all local regressors

$$\mathbf{y} = \sum_{j=1}^{k} \pi_j \boldsymbol{\beta}_j^T \hat{\mathbf{x}}. \tag{3}$$

Note the standard FMR makes no assumption on the marginal distribution of $\mathbf{x}$. The mixing proportions π_js are constants to samples, meaning an universal regression.

Extensions of FMR compute input-dependent π_js to handle heterogeneous data. Given a test input $\mathbf{x}_t$, Young *et al*. [34] compute π_js as the sum of similarities between $\mathbf{x}_t$ and all training $\mathbf{x}_i$s belong to the j^{th} local regression; Huang *et al*. [19] grid to fit kernel-based function; Gaussian Mixture Regression (GMR) assumes $\mathbf{x}_i$s of a local regression follows the Gaussian distribution, and π_j is the likelihood of $\mathbf{x}_t$ to the j^{th} Gaussian distribution. All these approaches suffer from the same limitation: Local regressors trained with more data and/or larger variance always produces larger π_js. As a result, the output prediction is dominated by local regressors with large π_js (See Figure 1 (a)).

Other extensions avoid above problem by hard-partition of input space according to outputs, and compute output estimation only in the partitioned space, e.g., Hierarchical Mixture of Experts (HME) [20] and Ordinal regression [5]. These approaches rely on perfect hard-partition (classification), and do not necessarily learn accurate local regressors. This is because only the outputs were used to supervise the partition, the local input-output correlation may not be homogeneous.

4. Soft-Margin Mixture of Regressions

SMMR overcomes above limitations by jointly learning soft-margin partition and local regressors.

The training algorithm of SMMR consists of four main ***operations***: ***(1)*** partitions the input space with multi-class Max-margin classification; ***(2)*** computes soft-margin mixing proportions using the hinge-loss in classification; ***(3)*** learns local regressors using training samples fall in each class where the regression errors weighted by soft-margin mixing proportions; ***(4)*** assigns samples to the class whose regressor produces the smallest regression error. These operations are alternated until convergence.

[1] Bold capital letters denote matrices $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$. $\mathbf{x}_j$ represents the j^{th} column of the matrix $\mathbf{X}$. All non-bold letters represent scalar variables. x_{ij} denotes the scalar in the row i and column j of the matrix $\mathbf{X}$ and the scalar i-th element of a column vector $\mathbf{x}_j$. $\mathbf{I}_k \in \Re^{k \times k}$ denotes the identity matrix. $\|\mathbf{x}\|_2$ denotes the L2-norm of the vector $\mathbf{x}$. $tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$. $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}^T\mathbf{A}) = tr(\mathbf{A}\mathbf{A}^T)$ designates the Frobenious norm of matrix $\mathbf{A}$.

Two critical observations can be made from above operations: (i) Because the soft-margin mixing proportions computed by the hinge-loss is zero outside of margin, by ***operation (2)***, mixture of local regressors only take place in the margin between classes. (ii) By ***operation (4)***, local regressors are updated using homogeneous subsets of data. The heterogeneous portion of the input space are contained in the margin.

Formally, given a training dataset $\mathbf{x}_i, \mathbf{y}_i$ $(i = 1, \cdots, n)$ with latent variable z, where $z \in \{1, \ldots, k\}$. The mapping from $\mathbf{x}$ to z is a classifier function

$$p(z|\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b, \tag{4}$$

Where $\varphi(\cdot)$ is a mapping the input space into high/infinite dimensional feature space, e.g., Reproducing Kernel Hilbert Space(RPHS). The mapping from x to y is a regression function

$$\mathbf{y} = \boldsymbol{\beta}_j^T \hat{\mathbf{x}}, \tag{5}$$

In ***operation (1)***, following, the max-margin partition is solved with one-against-one multi-class classification [17]:

$$\begin{aligned} \min_{\mathbf{w}^{jj'}, b^{jj'}, \xi^{jj'}} \quad & \frac{1}{2}||\mathbf{w}^{jj'}||^2 + C\sum_i \xi_i^{jj'} \\ & (\mathbf{w}^{jj'})^T \varphi(\mathbf{x}_i) + b^{jj'} \geq 1 - \xi_i^{jj'}, \; if \; z_i = j, \\ & (\mathbf{w}^{jj'})^T \varphi(\mathbf{x}_i) + b^{jj'} \leq -1 + \xi_i^{jj'}, \; if \; z_i = j', \\ & \xi_i^{jj'} \geq 0, \end{aligned} \tag{6}$$

where j and j' are index of classes $(j, j' = 1, \ldots, k; j \neq j')$.

In ***operation (2)***, the mixing proportion function $\pi_j(\mathbf{x}, \mathbf{w}, b)$ for sample $\mathbf{x}_i$ (denoted as π_{ij}) is computed as follows: for any $j' \neq j$, if $\mathbf{x}_i$ belongs to the j^{th} class, $(\mathbf{w}^{jj'})^T \varphi(\mathbf{x}_i) + b^{jj'} > 0$, then $r_{ij} = \sum_{j' \neq j}^k [(\mathbf{w}^{jj'})^T \varphi(\mathbf{x}_i) + b^{jj'}]_+$ quantifies the certainty of assigning $\mathbf{x}_i$ to the j^{th} class. The operator $[f]_+ = f$ when $f > 0$, and $[f]_+ = 0$ when $f \leq 0$. The mixing proportion π_{ij} is computed by normalizing r_{ij} with the soft-max technique:

$$\pi_{ij} = \frac{exp(r_{ij}^{(t)}/2h^2)}{\sum_{j=1}^k exp(r_{ij}^{(t)}/2h^2)}, \tag{7}$$

Where each class $\pi_{ij} = p(z = j|\mathbf{x}) \sim [0, 1]$, and $\sum_j^k \pi_{ij} = 1$. h is the bandwidth of the soft-max operator. The larger h, the "softer" the assignment for each $\mathbf{x}_i$.

In ***operation (3)***, to jointly learn classifiers and local regressors, the objective of SMMR is written as the log-likelihood function

$$\ell(\mathbf{w}, b, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_{i=1}^n log \sum_{j=1}^k \pi(\mathbf{x}_i, \mathbf{w}, b) \phi(\mathbf{y}|\boldsymbol{\beta}_j^T \hat{\mathbf{x}}_i, \boldsymbol{\sigma}_j^2), \tag{8}$$

where $\mathbf{w}$ and b are global notation of classifier parameters $\mathbf{w}^{jj'}$ and $b^{jj'}$. $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ global notation of regressor parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\sigma}_j$. We proposed a modified EM algorithm to solve Eq. 8. In the t^{th} iteration of the EM algorithm, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\sigma}^{2(t)}$ and $\mathbf{w}^{(t)}, b^{(t)}|\mathbf{x}_i$ is updated as follows:

In E-step, fixing $\boldsymbol{\beta}_j$, $\boldsymbol{\sigma}_j^2$ and π_{ij}, and compute the expectation of component identities for Eq. 8

$$p_{ij}^{(t+1)} = [1 + \sum_{j' \neq j} \frac{\pi_{ij} \phi(\mathbf{y}_i|\boldsymbol{\beta}_{j'}^{T(t)} \hat{\mathbf{x}}_i, \boldsymbol{\sigma}_{j'}^{2(t)})}{\pi_{ij} \phi(\mathbf{y}_i|\boldsymbol{\beta}_j^{T(t)} \hat{\mathbf{x}}_i, \boldsymbol{\sigma}_j^{2(t)})}]^{-1}. \tag{9}$$

In M-step, update $\boldsymbol{\beta}_j$, $\boldsymbol{\sigma}_j^2$ and π_{ij},

$$r_{ij}^{(t+1)} = \sum_{j' \neq j}^k [(\mathbf{w}^{jj'})^T \varphi(\mathbf{x}_i) + b^{jj'}]_+, \tag{10}$$

$$\pi_{ij}^{(t+1)} = \frac{exp(r_{ij}^{(t)}/2h^2)}{\sum_{j=1}^k exp(r_{ij}^{(t)}/2h^2)} + \epsilon, \tag{11}$$

$$\boldsymbol{\beta}_j^{t+1} = (\mathbf{X}\mathbf{G}_j^{(t+1)}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{G}_j^{(t+1)}\mathbf{Y}^T, \tag{12}$$

$$\boldsymbol{\sigma}_j^{2(t+1)} = \frac{\| \mathbf{Y} - \boldsymbol{\beta}_j^{T(t+1)}\mathbf{X} \|^2 \boldsymbol{G}_j^{(t+1)}}{tr(\boldsymbol{G}_j^{(t+1)})} + \epsilon \mathbf{I}_{d_y} \tag{13}$$

where

$$\boldsymbol{G}_j^{(t+1)} = diag(p_{1j}^{(t+1)}, p_{2j}^{(t+1)}, \ldots, p_{nj}^{(t+1)}). \tag{14}$$

Here, ϵ is a small positive scalar that avoids numeric instability in E-step (Eq. 9). The pseudo code of the SMMR training algorithm is summarized in Algorithm 1.

Algorithm 1 SMMR Training Algorithm

Require: Given training samples $\mathbf{x}_i, \mathbf{y}_i$ $(i = 1, \cdots, n)$:

1. Initialize the number of the regression k.
2. Do k-means clustering on $\{\mathbf{x}_i, \mathbf{y}_i\}$s to get k subsets, and initialize π_{ij} according to cluster assignment, i.e. $\pi_{ij} = 1$ if $\mathbf{x}_i$ is assigned to the j^{th} cluster, otherwise $\pi_{ij} = 0$.
3. Use samples in each subsets to learn the local regressors, and initialize the coefficient $\boldsymbol{\beta}_j$, and $\boldsymbol{\sigma}_j^2$.
4. implement the modified EM algorithm:

for each iteration **do**

a. Calculate the p_{ij} in E-Step (Eq. 9);

b. Train the multi-class SVM model based on p_{ij}, and compute the mixing proportion by formula r_{ij};

c. Update the $\pi_{ij}, \boldsymbol{\beta}_j$ and $\boldsymbol{\sigma}_j^2$ in M-Step (Eq. 10-13).

end for

5. Repeat until convergence.

The M-step optimizes max-margin classifier and weighted linear regression, which simply increases the completed likelihood function. For maximizing log-likelihood function (Eq. 8), the SMMR uses a generalized

EM algorithm and is guaranteed to reach local minima. In training SMMR above, there are three parameters to be selected: the number of components k, the weight on classification error C and the kernel bandwidth h. We selected these parameters that achieved the best results on training data by cross-validation.

Given a test input $\mathbf{x}_t \in \Re^{d_x}$, the output of SMMR, $\mathbf{y}_t \in \Re^{d_y}$, is computed as

$$\begin{aligned}\mathbf{y}_t &= \mathbb{E}[\mathbf{y}_t|\mathbf{x}_t] \\ &= \mathbb{E}[\sum_{j=1}^{k} \pi_{tj}\phi(\mathbf{y}|\boldsymbol{\beta}_j^T\hat{\mathbf{x}}_t, \boldsymbol{\sigma}_j^2)] \\ &= \sum_{j=1}^{k} \pi_{tj} \cdot \boldsymbol{\beta}_j^T\hat{\mathbf{x}}_t,\end{aligned} \tag{15}$$

where

$$\pi_{tj} = \frac{exp(r_{tj}/2h^2)}{\sum_{j=1}^{k} exp(r_{tj}/2h^2)}, \tag{16}$$

and

$$r_{tj} = \sum_{j' \neq j}^{k} [(\mathbf{w}^{jj'})^T\varphi(\mathbf{x}_t) + b^{jj'}]_+. \tag{17}$$

5. Experiments

5.1. Benchmark Datasets

For facial age estimation, the most frequently used benchmark is the Longitudinal Morphological Face Database (MORPH) [25] database. The MORPH database contains $55,132$ face images from more than $13,000$ subjects. The ages of the subjects range from 16 to 77 with a median age of 33. The faces are from different races, among which the African faces account for about 77 percent, the European faces account for about 19 percent, and the remaining 4 percent includes Hispanic, Asian, Indian, and other races.

For crowd counting, we used the widely-used University of California, San Diego pedestrian dataset (UCSD-ped)[2]. This dataset contains 2000 frames selected from two hours of video. The video was collected from a surveillance camera in the UCSD campus. The selected frames contain on average 25 pedestrians moving in two directions along a walkway. The resolution of the frame is 158×238.

For viewpoint estimation, the EPFL Multi-view Car (EPFL-car) dataset [23] was used. The dataset contains 20 sequences of cars under various viewpoints. There are 2299 images in the dataset. Each image comes with a bounding box specifying the location of the car. The ground truth viewpoint angles of the cars were estimated based on the shooting time of images. The viewpoint angle ranges from 0° to 360°.

Figure 2 show some example images in the UCSD and EPFL-car datasets. For the MORPH database, we only used the Bio-inspired Features(BIF) [21] from Dr. Guodong Guo and do not have the original images.

5.2. State-of-the-Art Comparison

Faical age Estimation Most recent results on the MORPH dataset [25] were obtained using three main approaches: AAM features [4] + nonlinear regression, BIF features[21] + nonlinear regression [13], and CNN-based approaches [33, 30, 22]. SMMR uses the BIF features for it usually performs better than AAM features, see Table 1. The original BIF feature for one face image is 4376 dimensions. We also followed the approaches in [10] to reduced the original BIF to 200 dimensional vectors using marginal Fisher analysis [31].

Many papers reported results under their own experimental protocols making it very difficult to perform a fair comparison with other approaches. For instance, [11, 12] divided the data into three subsets, used one subset for training, and the rest two subsets for testing. They reported 4.43 and 3.98 in MAE respectively. [6] only selected 5475 samples in their experiment. In our experiment, we compared approaches that follows the same experimental protocols: randomly divide the whole dataset into two parts: 80% of the data is used for training, and the other 20% of the data is used for testing. There is no overlap between the training and testing data. For statistical analysis, this procedure is done with 5-fold cross validation. All results were evaluated by the variance of Mean Absolute Error (MAE) [2].

Table 1. Facial Age Estimation on the MORPH dataset [25].

Method	Feature	MAE
RED-SVM [4]	AAM	6.49
MTWGP [37]	AAM	6.28
CA-SVR [6]	AAM	5.88
CPNN [10]	BIF	4.87
DLA+KSVR [30]	CNN	4.77
CCA [12]	BIF	4.73
KPLS [11]	BIF	4.43
LSVR [21]	BIF	4.31
OHRank [5]	BIF	3.82
CPLF [33]	CNN	3.63
HSVR [14]	BIF	3.6
OR-CNN [22]	CNN	3.27
SMMR(ours)	BIF	**3.24**

[2]CCA and LSVR results in Table 1 under above experimental protocol were provided by [22].

Figure 2. Examples images in the UCSD database for crowd counting: pedestrian scenes recorded by a surveillance camera on UCSD campus, and the EPFL-car dataset for viewpoint estimation: image sequences of rotating cars.

Four observations can be made from Table 1: (1) For standard nonlinear regressions, divide-and-conquer outperforms universal non-linearity: see Hierarchical SVR (HSVR) [14] $\rightarrow$ LSVR [21] and KPLS[11]; (2) For CNN-based approach, divide-and-conquer high-layer objectives outperforms universal objectives: Ordinal objective (OR-CNN [22])$\rightarrow$ Multi-scale objective (CPLF [33] $\rightarrow$ Universal nonlinear objective (DLA+KSVR [30]); (3) using divide-and-conquer techniques, standard nonlinear regression outperforms CNN without divide-and-conquer (HSVR $\rightarrow$ DLA+KSVR), and only after using the Ordinal objective (OR-CNN [22]) the CNN-based approach regained better results; (4) our SMMR approach produced the best result. Note here with only **RBF kernel in partition** and **linear local regressors**, our soft-margin technique is powerful enough to outperform other approaches. The the best result was obtained with 6 partitions.

Divide-and-conquer is the key factor in learning *non-stationary* age changes in human face. Our SMMR approach has two advantages over the existing divide-and-conquer approaches: (1) SMMR jointly learns overlapping partitions and minimizes regression error in each partition. HSVR manually selects overlapping ranges (*e.g.* $\Delta = 5$ in [14]) which may lead to heterogeneous local partitions and high local regression errors. (2) SMMR confined mixture of regressions in the soft-margin, which prevented the output estimation from being dominated by locally dense sampling. Both HSVR and Ordinal regression (OHRank and OR-CNN) classify input data by ages, and inevitably bias to the age range with dense samples. Moreover, ordinal regression trains $m-1$ binary classifiers, m being the integer number of ages, which is time-consuming to tune the parameters.

To illustrate the effectiveness of above advantages, we visualized the training and testing results of SMMR in Figure 3. Figure 3 (a) displays histogram of data samples (the vertical axis) with respect to age (the horizontal axis). The data was sampled mostly below age 60, and densely concentrated around 20's and 40's. Figure 3 (b) plots the samples fall in the 6 partitions learned in training (partitions

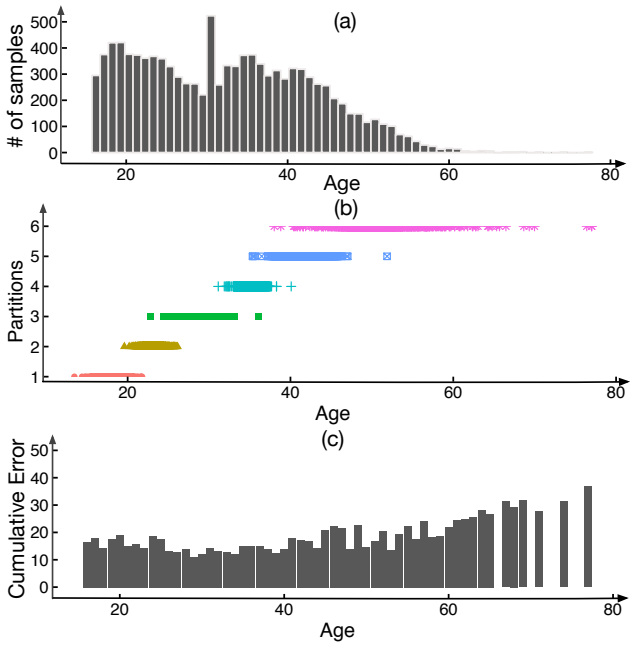

Figure 3. Visualization of SMMR on Morph dataset for Age Estimation. (a) Histogram of data samples with respect to age; (b) Samples fall in the 7 partitions learned in training (**best viewed in color**); (c) Cumulative age prediction error in testing.

are colored). Note that the learned partitions are heavily overlapped in ages. Although SMMR learns homogeneous local partitions, the number of samples are not necessarily uniformly distributed among partitions. Figure 3 (c) shows the cumulative prediction error in testing. The height of each bar denotes the sum of predicted age errors in one-year range. Despite the non-uniformly distributed samples (Figure 3 (a)), and imbalanced overlapping partitions (Figure 3 (b)), SMMR was able to produce uniform error through out the age range (from 0 to 80).

Crowd Counting Most recent works on the UCSD pedestrian (UCSD-ped) dataset were obtained by universal regression approaches. These approaches used the 29-dimensional low-level features (Segment features, inter-

Table 2. Crowd Counting on the UCSD-ped dataset[2].

Method	Feature	MAE	MSE
KRR [1]	SET	2.16	7.45
RR [7]	SET	2.25	7.82
GPR [2]	SET	2.24	7.97
CA-RR [6]	SET	2.07	6.86
Crowd-CNN [35]	CNN	1.60	3.31
SMMR(ours)	SET	**1.38**	**2.94**

Table 3. Viewpoint Estimation on the EPFL-car dataset [23].

Method	Feature	MAE	MedAE
FER [29]	Blur	33.98°	11.3°
CGFM [8]	SIFT	31.27°	-
KRF [15]	HOG	24.24°	-
HMA [36]	HOG	24.00°	-
EGGM [9]	SIFT	23.28°	-
SMMR(ours)	HOG	$\mathbf{12.61}^\circ$	$\mathbf{3.52}^\circ$

nal Edge features and Texture features (SET)) provided by [3]. We compared five recent regression results reported under the same the experimental setting in [35]. In Table 2, these regression methods are Kernel Ridge Regression (KRR)[1], Ridge Regression(RR) [7], Gaussian Process Regression(GPR)[2] , Cumulative Attribute based Ridge Regression (CA-RR)[6], and Crowd-Convolution Neural Networks (Crowd CNN)[35]. For all approaches, frames from index 601 to 1400 were used as training data, and the remaining 1200 frames were used as test data. The Mean Average Error (MAE) and Mean Square Error (MSE) metrics are used for evaluating the performance.

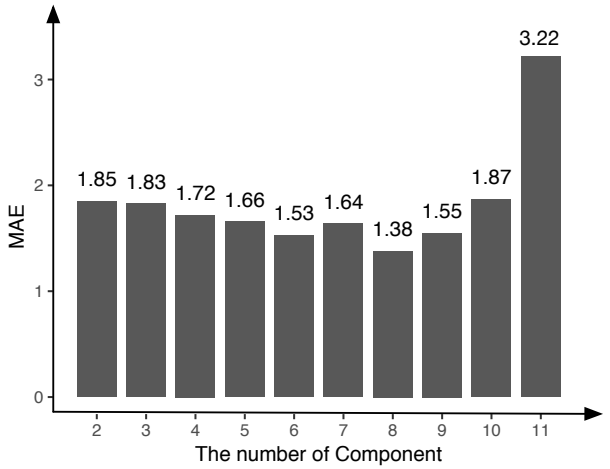

Figure 4. Sensitivity of SMMR to number of components k in the crowd counting experiment. The lowest MAE, 1.38, was achieved when $k = 8$.

As shown in Table 2, SMMR significantly outperformed the others on both MAE and MSE. Note here SMMR used **RBF kernel in partition** and **linear local regressors** and obtained the best result with 8 partitions. The sensitivity of parameter k is shown in Fig. 4. It is very interesting to note that CA-RR and Crowd-CNN performed better than other universal regression approaches by partially addressing the nonuniform sampling problem in data. CA-RR projected original features into a (0-1 valued) cumulative feature space partitioned by crowd counts. This procedure normalizes the variance among crowd counts ranges, and results in a weighted Ridge Regression (RR). Similarly, Crowd-CNN normalized the data density among crowd counts ranges.

Viewpoint Estimation Many recent work on EPFL-car dataset [23] combines bounding box detection and viewpoint estimation. To focus on the regression results, we only compared five approaches using the ground-truth bounding boxes. As shown in Table 3, these approaches are Feature Embedding based Regression (FER)[29], Class Generative Feature Model (CGFM)[8], K-Clusters Regression Forest (KRF)[15], Homeomorphic Manifold Analysis (HMA) [36], and Embedding Geometry based Generative Model (EGGM)[9]. All the compared approaches used the first 10 image sequences for training and the remaining 10 sequences for testing. Two evaluation metrics were reported in the compared approaches: Mean Angular Error (MAE) and Median Angular Error (MedAE).

Our approach used the same input features proposed in [15]: given bounding box of a car, a image patch was cropped and re-sized to 64×64. Multi-scale Histogram of Oriented Gradients (HOG) feature was computed with cell size $\{8, 16, 32\}$ and 2×2 cell blocks. The orientation histogram of each cell is computed with signed gradients in 9 orientation bins. The resulted HOG feature is 2124-dimensional; finally, the HOG feature is reduced to 50 dimensions by marginal Fisher analysis [31]. Considering the view angle ranges (from 0° to 360°), the Euclidean distance is inappropriate to measure the regression error. For instance, the distance between 0° to 350° should be smaller than that between 0° to 50°. Hence, in our experiments, the 1D viewpoint angle output space is represented by the 2-dimensional coordinates on a unit circle. Given viewpoint angle a, the 2D coordinates is $(y_1, y_2) = (\sin a, \cos a)$. The 2D outputs were used for both training and testing. After a 2D output is estimated in testing, its viewpoint angle is computed using $\arctan 2$ function.

As shown in Table 3, SMMR significantly outperformed the others on both MAE and MedAE. Note here SMMR used **RBF kernel in partition** and **linear local regressors** and obtained the best result with 7 partitions. Three ob-

servations can be make: (1) divide-and-conquer approaches outperform universal approaches: KRF and HMA $\rightarrow$ FER. HMA builds a homeomorphic manifold mapping for each color channel (*e.g*. RGB), and combine the estimation from three manifolds to compute the viewpoint of a car. KFR is a hierarchical approach which recursively split the output space into a set of disjoint partitions, and uses random forest as the final regression model; (2) Approaches with soft partition are better than those with hard partition: EGGM and SMMR $\rightarrow$ KRF. EGGM builds a graph of object parts, and computes soft matching scores of test object parts to the graph. The soft matching scores are then used as weights to combine the viewpoint of parts to viewpoints of a test object. SMMR uses soft-margin error to weight local regressors. Whereas KRF fixes the clustered partitions and performs hard partition of the input features by classification; (3) Partition in the joint input-output space is better than output space only partition: SMMR $\rightarrow$ KRF. KRF uses k-means clustering only in the output space making it very difficult to correctly classify a test input feature. On the other hand, SMMR learns partitions in the joint input-output space, such that homogeneous and accurate local regressors can be jointly learned.

6. Conclusion

We proposed SMMR to learn nonlinear regression with heterogeneous, non-uniformly sampled and discontinuous feature space that typically emerges from computer vision tasks. SMMR is a mixture of regressions approach that finds homogeneous partitions of data while minimizing the error in partition boundaries using soft-margin mixtures. Experiments on three computer vision tasks showed that SMMR outperformed standard nonlinear regression approaches using the same hand-craft features as well as many CNN-based regression approaches. Note that in this paper, we achieved superior results using only the standard RBF kernel max-margin classifier and linear local regressors. To explore the full potential of SMMR, we plan to try more sophisticate local regressors and extend SMMR as a higher-layer objective of CNN.

7. Acknowledgement

Research reported in this publication was supported in part by the National Science Foundation under the grants RI-1617953. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

References

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[2] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

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

[4] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. A ranking approach for human ages estimation based on face images. In *International Conference Pattern Recognition*, pages 3396–3399, 2010.

[5] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–592, 2011.

[6] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *IEEE conference on computer vision and pattern recognition*, pages 2467–2474.

[7] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *British Machine Vision Conference*, volume 1, page 3, 2012.

[8] M. Fenzi, L. Leal-Taixé, B. Rosenhahn, and J. Ostermann. Class generative models based on feature regression for pose estimation of object categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 755–762, 2013.

[9] M. Fenzi and J. Ostermann. Embedding geometry in generative models for pose estimation of object categories. In *British Machine Vision Conference*, volume 1, page 3, 2014.

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

[11] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 657–664, 2011.

[12] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *Automatic Face and Gesture Recognition Workshops*, pages 1–6, 2013.

[13] G. Guo and C. Zhang. A study on cross-population age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4257–4263, 2014.

[14] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161, 2015.

[15] K. Hara and R. Chellappa. Growing regression forests by classification: Applications to object pose estimation. In *European Conference on Computer Vision*, pages 552–567, 2014.

[16] K. He, L. Sigal, and S. Sclaroff. Parameterizing object detectors in the continuous pose space. In *European Conference on Computer Vision*, pages 450–465, 2014.

[17] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[18] D. Huang, M. Storer, F. De la Torre, and H. Bischof. Supervised local subspace learning for continuous head pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[19] M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012.

[20] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[21] G. Mu, G. Guo, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009.

[22] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016.

[23] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785, 2009.

[24] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In *European Conference on Computer Vision*, pages 356–370, 2012.

[25] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[26] H. G. Sung. *Gaussian mixture regression and classification*. PhD thesis, Rice University, 2004.

[27] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793, 2014.

[28] D. Teney and J. Piater. Multiview feature distributions for object detection and continuous pose estimation. *Computer Vision and Image Understanding*, 125:265–282, 2014.

[29] M. Torki and A. Elgammal. Regression from local features for viewpoint and pose estimation. In *International Conference on Computer Vision*, pages 2603–2610, 2011.

[30] X. Wang, R. Guo, and C. Kambhamettu. Deeply-learned feature for age estimation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 534–541, 2015.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.

[32] L. Yang, J. Liu, and X. Tang. Object detection and viewpoint estimation with auto-masking neural network. In *European Conference on Computer Vision*, pages 441–455, 2014.

[33] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision*, pages 144–158, 2014.

[34] D. S. Young and D. R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010.

[35] C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015.

[36] H. Zhang, T. El-Gaaly, A. M. Elgammal, and Z. Jiang. Joint object and pose recognition using homeomorphic manifold analysis. In *Association for the Advancement of Artificial Intelligence*, volume 2, page 5, 2013.

[37] Y. Zhang and D.-Y. Yeung. Multi-task warped gaussian process for personalized age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2622–2629, 2010.

[38] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2608–2623, 2013.