Cumulative Attribute Space for Age and Crowd Density Estimation

Ke Chen, Shaogang Gong, Tao Xiang
Queen Mary, University of London
London E1 4NS, UK
cory,sgg,txiang@eecs.qmul.ac.uk

Chen Change Loy
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
ccloy@ie.cuhk.edu.hk

Abstract

A number of computer vision problems such as human age estimation, crowd density estimation and body/face pose (view angle) estimation can be formulated as a regression problem by learning a mapping function between a high dimensional vector-formed feature input and a scalar-valued output. Such a learning problem is made difficult due to sparse and imbalanced training data and large feature variations caused by both uncertain viewing conditions and intrinsic ambiguities between observable visual features and the scalar values to be estimated. Encouraged by the recent success in using attributes for solving classification problems with sparse training data, this paper introduces a novel cumulative attribute concept for learning a regression model when only sparse and imbalanced data are available. More precisely, low-level visual features extracted from sparse and imbalanced image samples are mapped onto a cumulative attribute space where each dimension has clearly defined semantic interpretation (a label) that captures how the scalar output value (e.g. age, people count) changes continuously and cumulatively. Extensive experiments show that our cumulative attribute framework gains notable advantage on accuracy for both age estimation and crowd counting when compared against conventional regression models, especially when the labelled training data is sparse with imbalanced sampling.

1. Introduction

A number of computer vision problems concern with the estimation of a scalar value given a high dimensional feature input vector. Examples of such problems include age estimation from facial images [10, 12, 15, 16, 33, 35], crowd counting [4, 5, 8, 25], and human body/face pose (view angle) estimation [14, 27, 34]. Such a scalar value can vary continuously within a certain range but is often assumed to be discrete (e.g. human age and people count), and its estimation can be obtained by solving a multi-class classification problem [13, 21]. Such a multi-class labelling

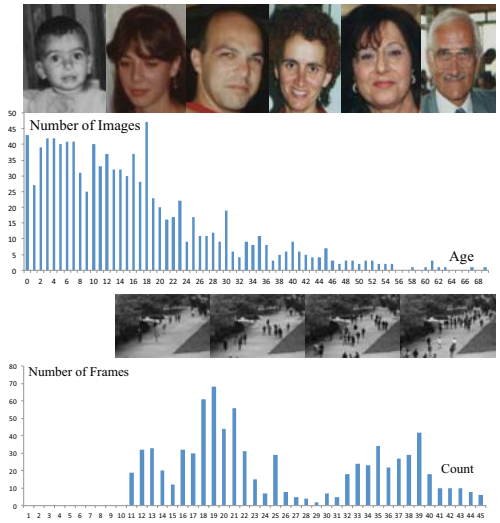


Figure 1. Age estimation and crowd counting both suffer from sparse and imbalanced training data distribution. Top: FG-NET facial age dataset. Bottom: UCSD crowd dataset.

treatment of scalar value estimation assumes implicitly that each scalar output value (a label) is independent from other possible values (labels). On the contrary, human age and people-count are strongly correlated and neighbouring values have closer similarities than those further apart, e.g. a human face of 50 years old is more similar to that of 49 than that of 10. To exploit this observation, most existing approaches to the problem consider a regression solution in which a mapping function is learned explicitly between high dimensional feature input vectors and scalar output values [4, 5, 8, 10, 12, 15, 16, 33, 35]. However, there are two major challenges for learning a good regression function for solving such a problem: (1) inconsistent and incomplete features, (2) sparse and imbalanced training data.

In general, regression based interpretation suffers from large feature variations caused by both viewing conditions and visual inconsistency in interpretation. For instance, people of the same age can appear visually very different, e.g. the images were taken under very different lighting conditions (extrinsic condition change) or images of

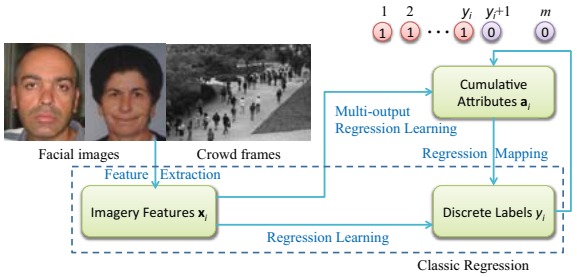


Figure 2. The pipeline of our framework compared with conventional regression framework.

very different people of the same age (intrinsic condition change). In addition to lighting and viewing angles, occlusion can also cause crowd frames of the same people-count to appear significantly different. Existing regression techniques have mostly focused on addressing the challenge of feature inconsistency by constructing a low-level feature representation robust against both the intrinsic and extrinsic condition changes [16, 34]. There are less efforts on addressing the second challenge on sparse and imbalanced data.

Accurately labelled facial images for human age estimation and public space video data for crowd counting are generally sparse and imbalanced due to inherent ambiguities in annotation and a lack of sufficient samples for covering the data distribution. For example, despite large quantities of facial images available publically, e.g. from Flickr, annotating the true age of a facial image can be very unreliable [10, 28]. As a result, benchmarking datasets such as FG-NET [7, 13, 15, 35] and MORPH [7, 13] contain very limited samples of each age group and consist of faces of true ages rather than annotated age. Figure 1 shows that in the FG-NET dataset, at most 46 images are available for each age group and the distribution is highly imbalanced across the age groups. This is rather sparse given that the faces belong to different genders and ethnical groups (therefore compounded by inconsistent visual features). Even though annotating crowd images can be made more reliable, annotating people count exhaustively for all possible values is laborious and often practically infeasible, e.g. a public place as shown in Figure 1 may never exhibit fewer than 10 people or greater than 50 people in any realistic time gap. Consequently existing crowd benchmarking datasets such as UCSD [4, 5, 8] are also sparse. Moreover, the sparseness in training data also implies that there are often gaps in training samples where no imagery sample is available for mapping onto certain output values causing difficulties in learning the regression mapping function.

In this work, we consider that the two challenges above are related in the sense that the feature inconsistency problem is compounded by sparse and imbalanced training data and vice versa, and they need be tackled jointly in modelling and explicitly in representation. To that end, we propose a novel cumulative attribute based representation for learning a regression model. Attributes have been successfully applied for solving various computer vision problems by classification [11, 20, 22, 24], but have never been used for regression to the best of our knowledge. Attribute models are designed to solve the data sparsity problem by exploiting shared characteristics between different classes. These common characteristics are either defined manually by human a priori knowledge [20, 22] or discovered automatically from data [11, 24]. Existing attribute learning methods cannot be directly applied to our regression problem because: (1) Attributes need be discriminative to be useful. For classification, it is natural to identify discriminative attributes for differentiating classes. Discriminative attributes can also be discovered by learning a discriminative model [24]. However, for learning a regression model it is much less clear what is discriminative and more importantly what can be shared across different scalar output values when those values change continuously. (2) Existing attribute definitions do not reflect nor exploit the unique characteristic of neighbouring scalar output values sharing more similarities than those further apart.

Our notion of cumulative attributes aims to explore the spirit of the conventional discriminative attribute for addressing sparse training data, whilst is specifically designed for addressing the regression problem. More specifically, each attribute is not only discriminative but also cumulative in constraining all other attribute values depending on its relative positioning in value: each attribute separates all training images into two groups (binary) by a label (e.g. an age). For instance, for learning a regression model for age estimation, if there are 70 age groups, there will be 69 binary attributes, each separating facial images above certain age from all those below. By cumulative attributes, we consider each attribute cumulatively conditioning all other attributes. That is, for a person of 50, not only the corresponding attribute 50 is positive, but also from 1 all the way to 49 are conditionally positive. This is designed specifically to capture the unique correlation of data samples so that those with neighbouring scalar output values share more than those further away in our cumulative attribute space. Critically, this cumulative nature is also able to cope with sparse and imbalanced data distribution more effectively. In particular, by utilising all data samples for discriminating each attribute regardless the availability of labelled data for that attribute (value) alone, sparsity problem is mitigated. The cumulative nature of the attribute also greatly reduce the ill-effect of imbalanced data, e.g. even if there was no sample for a certain age value (attribute), that attribute is positively assigned by any samples of lower age than the considered value, thus can be learned indirectly using plenty of neighbouring samples.

The pipeline of our framework is illustrated in Figure 2.

Once cumulative attributes are constructed from the scalar values of training samples, a two-layers regression framework is employed. Firstly, given any low-level feature presentation of the image, we learn a multi-output regression model to map the feature inputs to an intermediate attribute space. To that end, a single structured output model is learned to correlate explicitly different attributes. Secondly, another regression model is learned to estimate the scalar output using the attribute representation as input. Extensive experiments are carried out using benchmarking age estimation and crowd counting datasets and show that (1) our cumulative attribute representation improves generally the age estimation and crowd counting accuracy over the state-of-the-art with standard image feature representations, (2) the improvement is particularly significant when the training data is sparse and imbalanced.

2. Related Work

Age estimation – Most existing techniques for age estimation from facial images fall into three categories: multi-class classification [13], regression [16], and hybrid [15] of the two, with regression models being the most widely used. Guo *et al*. [15] proposed a locally adjusted regression method to search local regions for adjusting. They further introduced BIF features for regression [16]. Recently, Zhang *et al*. [35] proposed a multi-task wrapped Gaussian Process Regression for personalized age estimation that jointly learns personalized characteristics and common changes shared between people. Our approach is designed to utilise any low-level features and regression models, with the key difference being that the input to the regression model is represented by cumulative attributes instead of the low-level features directly. More recently, a ranking based age estimation method is proposed [7]. For each age group, a ranker (a binary classifier) is learned to separate people into two groups, older or younger than the said age group. Given a testing image the output of the rankers are aggregated directly for estimating the age. This method shares similar spirit to our model in that learning each ranker uses all the data in the dataset in order to mitigate any sparsity problem. However, different from our method, the rankers are not cumulative therefore do not share mutual information, and they do not benefit from an intermediate representation. Moreover, such a ranking based model is extremely expensive to both learn and apply (see Section 4.6 on computational cost).

Crowd counting – Similar to age estimation, crowd counting can be solved by either classification and regression with most recent work adopting the regression approach. Despite the low-level features being very different, the same regression models such as support vector machine regression and Gaussian Processes have been employed for both problems [4, 5, 8, 25]. Crowd counting in images may be considered somewhat less ambiguous than age estimation because the latter has to cope with different people of any gender and race but with the same age, whilst most existing crowd counting models are scene specific, equivalent to learning a person specific age estimator. Our cumulative approach is shown to improve on existing methods on both problems.

Attribute learning – Visual attributes have received increasing interests in the past three years for classification problems ranging from image categorisation [20, 29], person re-identification [22], to action and video event recognition [11]. Attributes are either user defined based on prior knowledge [20, 22] or data driven or latent and discovered from data [11, 24]. The former has clear semantic meaning and the latter not necessarily so. On the other hand, manually defined attributes may not be computable consistently nor discriminative sufficiently despite additional human annotation, from which data driven attributes do not suffer. Our cumulative attributes are unique such that each attribute has clear semantic meaning and by definition being discriminative, yet no additional annotation is required. They are specifically designed for learning a regression model whilst none of the existing attribute representations is suitable. Moreover, it is more desirable to learn attributes jointly as they are typically correlated [26]. However, computationally learning a large number of attributes and modelling their correlation explicitly is a challenge. In this paper, a multi-output regression model is formulated to learn all attributes in a single model that is also extremely efficient to compute. Note that recently proposed notion of relative attribute [19, 29] defines attribute as the real-valued strength of the presence of visual properties. However, relative attributes are learned as a ranking problem rather than a regression problem because only pairwise-comparison data are available [19, 29].

Contributions – Our contributes are three-fold: (1) For the first time, an attribute representation is constructed for learning a regression model. (2) A novel concept of cumulative attributes is proposed with both clear semantic meaning and also discriminative, with added advantages of efficiently computable and requiring no additional annotation. (3) Extensive experiments on both age estimation and crowd counting benchmark datasets demonstrate the superiority of our method over the state-of-the-arts, especially when the data is sparse and imbalanced.

3. Methodology

As shown in Figure 2, our cumulative attributes can be considered as an intermediate-level semantic representation that bridges the gap between any low-level features and a regression model given sparse annotation. During training

our cumulative attribute based regression framework consists of the following steps:

1. Given a set of training images, we extract low-level imagery features and the scalar output value (e.g. age or people count) is converted into a binary cumulative attribute vector (Section 3.1).

2. A cumulative attribute representation is computed so that given an image, its cumulative attributes can be assigned and used as an intermediate representation of the image. Specifically, a single multi-output regression model is learned to evaluate and assign all attributes simultaneously (Section 3.2).

3. A second layer single output regression model is learned to map the attribute representation to the scalar output value (Section 3.3).

During testing, given an unseen image, the cumulative attribute vector is first computed using the multi-output regression model with the low-level imagery features as input. The cumulative attribute vector is then fed into the single output regression model to estimate the scalar output value.

3.1. Cumulative Attribute

Given a training image/frame i, where $i = 1,2\ldots\mathrm{N}$ and N denotes the total number of training images/frames, we firstly extract low-level imagery features $\mathbf{x}_i$ from the whole image/frame. This can be Active Appearance Model features [9] for age estimation and foreground & edges & GLCM features [4, 8] for crowd counting. Any other features in the literature can be equally applied. Secondly, normalization on the feature data including scale normalization and extra perspective normalization [4] for crowd counting are carried out.

Now for the ith training data point, the known scalar value y_i (e.g. age and people count) is converted into a cumulative attribute vector $\mathbf{a}_i$. The dimensionality of the vector $\mathbf{a}_i$, denoted as m, depends on the value range of y. Typically, for age or crowd count, there is an upper limit, e.g. 70 for a certain age dataset and 100 for a certain crowd scene. This upper limit will be used as the value of m. Formally, given N training sample $\{(\mathbf{x},y)\}_i, i = 1,2\ldots\mathrm{N}$, the jth element of the cumulative attribute vector for the ith sample assumes a binary value:

$$a_i^j = \begin{cases} 1, & \text{when } \ j \leqslant y_i, \\ 0, & \text{when } \ j > y_i, \end{cases}$$

where $j = 1,2,\cdots,m$. Evidently, for the ith attribute vector $\mathbf{a}_i$, the first y_i attribute elements are all "ones" and the rest $m - y_i$ elements are all "zeros".

In comparison , a non-cumulative attribute (NCA) is constructed as follows:

$$a_i^j = \begin{cases} 1, & \text{when } \ j = y_i, \\ 0, & \text{when } \ j \neq y_i. \end{cases}$$

Note, only one element of a non-cumulative attribute vector $\mathbf{a}_i$ is one and all the rest elements are zeros. There is thus a critical difference between our CA representation and the conventional NCA representation: with the CA representation, data points with neighbouring scalar values are represented by a very similar attribute set, whilst with conventional NCA representations, the difference between the attributes of two data points of any scalar value is the same. For example, a face of age 40 and another face of age 41 represented using a 69D CA vector will have only one element that is different, whilst the number of different attribute elements increases to 30 for a face of age 10. On the other hand, using a NCA representation, there is always a single element difference no matter how different the ages are and how the two faces look like. Our cumulative attributes thus capture a better representation of a continuously changing value for object appearance, corresponding directly to a scalar output value change continuously for learning a regression function. Our experiments in Section 4.3 show the distinct advantages of using CA over NCA for both age estimation and crowd counting.

3.2. Joint Attribute Learning

Now the training set is represented as $\{(\mathbf{x},\mathbf{a},y)\}_i, i = 1,2\ldots\mathrm{N}$. We need to learn the mapping relationships between both $\mathbf{x}$ and $\mathbf{a}$, and $\mathbf{a}$ and y. In this section we focus on the former. Most existing attribute learning methods aim to establish a mapping between $\mathbf{x}$ and each element of $\mathbf{a}$ independently using a binary classifier such as a support vector machine. However, this is not only making the false assumption that different attributes are independent from each other, but also computationally expensive. In our work, we estimate the mappings of all m attributes simultaneously by learning a multi-output regression function, in particular, a multivariate ridge regression function [1, 17]. In its conventionally form, a ridge regression function learns a single output mapping. Recently, multivariate ridge regression [1, 8] has been exploited for simultaneous output estimation. Following established design principle of multi-task learning [2, 3, 18, 30], we formulate the following multi-output attribute learning problem. Given $\mathbf{x}_i$ and a_i^j being low-level features of the ith image and the jth element of its corresponding attribute vector, the objective function for the jth attribute is written as:

$$\min \quad \frac{1}{2}\|\mathbf{w}^j\|_2^2 + C\sum_{i=1}^{\mathrm{N}} \mathrm{loss}(a_i^j, f^j(\mathbf{x}_i)),$$

where $f^j(\mathbf{u}) = \mathbf{w}^j\mathbf{u} + b^j$ and $\mathrm{loss}(\cdot)$ denotes the loss function. Hence, a joint attribute learning by multi-output regression is formulated as

$$\min \quad \sum_{j=1}^{m}(\frac{1}{2}\|\mathbf{w}^j\|_2^2 + C\sum_{i=1}^{\mathrm{N}}\mathrm{loss}(a_i^j, f^j(\mathbf{x}_i))).$$

For simplifying the above without losing generality, quadratic loss function is considered. The objective function of the joint attribute learning is then given as:

$$\min \quad \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^{\mathrm{N}}\|\mathbf{a}_i^T - (\mathbf{x}_i^T\mathbf{W} + \mathbf{b})\|_F^2, \qquad (1)$$

where $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \cdots, \mathbf{w}^j, \cdots, \mathbf{w}^m]$ is the weight matrix, $\mathbf{a}_i = [a_i^1, a_i^2, \cdots, a_i^m]^T$ is the training attribute vector, and $\mathbf{b} = [b^1, b^2, \cdots, b^m]$ is the bias term. The model parameters $\mathbf{W}$ are estimated by solving an equality-constrained Quadratic Programming Problem, which has a closed-form global optimal solution as follows:

$$\begin{bmatrix}\mathbf{W}\\ \mathbf{b}\end{bmatrix} = -(Q^TQ)^{-1}Q^TP,$$

where positive semi-definite matrix Q and matrix P are given as

$$Q = \begin{bmatrix} 2C\sum_{i=1}^{\mathrm{N}}\mathbf{x}_i\mathbf{x}_i^T + I & 2C\sum_{i=1}^{\mathrm{N}}\mathbf{x}_i \\ 2C\sum_{i=1}^{\mathrm{N}}\mathbf{x}_i^T & 2CN \end{bmatrix},$$
$$P = \begin{bmatrix} -2C\sum_{i=1}^{\mathrm{N}}\mathbf{x}_i\mathbf{a}_i^T \\ -2C\sum_{i=1}^{\mathrm{N}}\mathbf{a}_i^T \end{bmatrix}.$$

The trade-off parameter C is determined by cross validation.

The weight matrix $\mathbf{W}$ plays an important role in transferring information between tasks thus modelling the correlation between different attributes. In particular, with the same feature representation, for each attribute $a_i^j, j = 1, 2, \cdots, m$, we formulate our model to jointly weigh each attribute. In Equation (1), the jth column of matrix $\mathbf{W}$ is employed to weigh the imagery feature vector $\mathbf{x}_i$ for the jth binary attribute in corresponding attribute learning, i.e. the jth element of $\mathbf{a}_i$. Since the residual error of all attribute learning tasks are penalized jointly by the Frobenius-norm, this multi-output model can capture the correlation between different attributes explicitly.

3.3. Mapping Attributes to Scalar Output

To estimate the mapping between $\mathbf{a}$ and y, first the low-level feature $\mathbf{x}$ is mapped onto our cumulative attribute space using the learned multi-output regression model above. With each image now represented as $\hat{\mathbf{a}}_i \in \mathbb{R}^m$ and the corresponding label (ground truth) $y_i \in \mathbb{R}$, where $i = 1, 2 \ldots \mathrm{N}$, a second-layer output regression model is learned. Note, this regression model has a single scalar output and any existing regression models used in the literature for either age estimation or crowd counting can be readily applied.

4. Experiments

4.1. Datasets & Settings

Datasets – For age estimation, two widely used benchmarking datasets FG-NET [7, 13, 15, 35] and MORPH [7, 13] were used. Both datasets are designed primarily for learning person-independent age estimator and contain people of different ethnical origins. For crowd counting, experiments were conducted on the benchmarking UCSD [4, 5, 8] and the Mall [8] datasets which feature an outdoor and an indoor scene respectively. Details in Table 1 show that among the four datasets, FG-NET is the most sparse in terms of the average number of samples per scalar output value (MORPH is 5 times more densely sampled).

Data	$\mathbf{N_{i/f}}$	R
FG-NET [13]	1002	0–69
MORPH [7]	5475	16–77
UCSD [4]	2000	11–46
Mall [8]	2000	13–53

Table 1. Dataset details: $\mathbf{N_{i/f}}$ = number of images/frames, **R** = range of scalar output value.

Features – For age estimation, the low level image features are Active Appearance Model features [9]. This feature representation is widely used in recent approaches [7, 13, 15, 32, 33, 35]. For crowd counting, three types of image features, i.e. foreground segments, edge features, and local texture features, are adopted as in [4, 8]. Note that, to use these features, all frames of crowd databases were transformed to gray-scale prior to feature extraction.

Settings – For FG-NET, we followed the same leave-one-person-out setting as in [7, 15, 32, 33, 35]. For MORPH we randomly split the dataset into 80% training data and the rest 20% testing data and repeated the experiments 30 times as in [7]. For crowd counting, we followed the same training and testing partitions as in [8], *i.e.* we employed Frames $601-1400$ in UCSD dataset and Frames $1-800$ in Mall dataset respectively for training, while the rest frames were used for testing. For the single output regression model (Section 3.3), Support Vector Regression (SVR) with RBF kernel and Ridge Regression (RR) were employed for age estimation and crowd counting respectively, owing to their strong performance reported in the literature for age [15, 16] and crowd [8] respectively. However, any regression models can be used.

Evaluation Metrics – For age estimation, we employed two evaluation metrics, namely *mean absolute error* (mae) and *cumulative score* (cs), which was first defined in [13] and we set the same error level 5 as in [7]. Three metrics employed in [8], namely *mean absolute error* (mae), *mean squared error* (mse), and *mean deviation error* (mde) were employed for evaluating the performance of crowd counting. Among all five metrics, only for cs higher value means better performance.

4.2. Comparison with State-of-the-Arts

Method	FG-NET [13]		MORPH [7]	
	mae	cs	mae	cs
AGES [13]	6.77	–	8.83	–
RUN [33]	5.78	–	–	–
Ranking [32]	5.33	–	–	–
RED-SVM [6]	5.24	–	6.49	–
LARR [15]	5.07	–	–	–
MTWGP [35]	4.83	–	6.28	–
OHRank [7]	4.85	74.4%	**5.69**	56.3%
SVR [15]	5.66	68.0%	5.77	57.1%
CA-SVR	**4.67**	**74.5%**	5.88	**57.9%**

Table 2. Age estimation performance comparison.

Age estimation – Our model (CA-SVR) is compared with a number of recently published results in Table 2. Most of the methods compared are regression based except AGES [13], RED-SVM [6] and OHRank [7], and use the same AAM features except AGES [13]. For FG-NET dataset, our model obtained the best results so far on both mae and cs metrics. Note that compared with SVR [15], identical low level feature and single output regression models were used. The only difference is in the input to the regression model: low level feature directly for SVR and our cumulative attributes for CA-SVR. This change of representation brings an significantly improvement (17.5% decrease in mde and 9.6% relative increase in cs). The best performance reported so far on FG-NET is the Ordinal Hyperplane Rank model (OHRank) [7]. As discussed in Section 2, OHRank can also cope with the sparse data problem. However, as shown in Section 4.6, it is in the order of four magnitudes slower than our model in model training[1]. On the MORPH dataset, our CA-SVR gives comparable result to the best reported so far (OHRank) on mae, but best performance measured by cs. As the key difference between the FG-NET and MORPH dataset is data sparsity and the number of age groups without samples, it is evident from these results that the advantage of our cumulative attribute based regression model is more significant given sparse and imbalanced data. This is further supported by our missing data experiments reported in Section 4.4.

Method	UCSD [4]			Mall [8]		
	mae	mse	mde	mae	mse	mde
LSSVR [31]	2.20	7.29	0.107	3.51	18.2	0.108
KRR [1]	2.16	7.45	0.107	3.51	18.1	0.108
RFR [23]	2.42	8.47	0.116	3.91	21.5	0.121
GPR [4]	2.24	7.97	0.112	3.72	20.1	0.115
RR [8]	2.25	7.82	0.110	3.59	19.0	0.110
CA-RR	**2.07**	**6.86**	**0.102**	**3.43**	**17.7**	**0.105**

Table 3. Crowd counting performance comparison.

Crowd counting – Table 3 compares crowd estimation performances of six different methods, all based on regression, using the two benchmarking datasets. The result shows that the cumulative attribute based model (CA-RR) performs the best for both datasets and using all three metrics. The most direct effect of using our cumulative attribute representation can be seen by comparing RR [8] with CA-RR. CA-RR clearly outperforms RR using all three measures. Since both have the same low level feature input and use the same single output regression model, the performance gain can only be explained by the superior representation by our cumulative attribute space. Improved performance can also been seen by comparing CA-RR with a number of recently proposed models [1, 4, 23, 31], all of which use the same features as input and differ only in the regression model used.

[1]The results of OHRank were based on our implementation and are slightly lower than those reported in [7].

4.3. Cumulative vs. Non-Cumulative Attributes

Methods	FG-NET [13]		MORPH [7]	
	mae	cs	mae	cs
NCA-SVR	8.95	41.8%	7.28	44.2%
CA-SVR	**4.67**	**74.5%**	**5.88**	**57.9%**

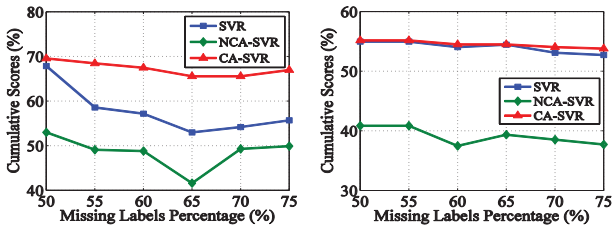
Table 4. Cumulative vs. non-cumulative attributes on age estimation.

Methods	UCSD [4]			Mall [8]		
	mae	mse	mde	mae	mse	mde
NCA-RR	2.85	11.9	0.137	4.31	25.8	0.131
CA-RR	**2.07**	**6.86**	**0.102**	**3.43**	**17.7**	**0.105**

Table 5. Cumulative vs. non-cumulative attributes on crowd counting.

A key novelty of our model is the cumulative attribute representation. As explained in Section 3.1, compared with the conventional non-cumulative (NCA) attributes, the unique characteristics of our cumulative attributes (CA) is that data points of neighbouring scalar value are designed to be close to each other in the attribute space. It is evident from Tables 4 and 5 that constructing such cumulative attributes is a significant advantage for a regression model that performs age estimation and crowd counting.

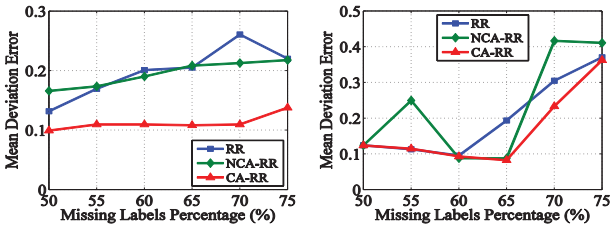
4.4. Against Sparse and Imbalanced Data



(a) FG-NET (b) MORPH

Figure 3. Age estimation performance with sparse and imbalanced data measured using cumulative scores (the higher the better).

Figures 3 and 4 evaluate our model when the training data become more and more sparse and imbalanced. Data of certain age groups and certain crowd counts were removed



(a) UCSD (b) Mall

Figure 4. Crowd counting performance measured by mean deviation error (the lower the better).

to make the data more sparse and imbalanced. For age estimation, since the two dataset have few missing age groups, we randomly selected a fixed number of age groups, each time to remove and then train the model. For the crowd counting dataset, this way of removing data would be less effective because the mapping between the low level features and the scalar count numbers is more linear. Therefore, a different strategy for removing samples is adopted. That is, we start from the middle of count number ($26-30$ for missing 10% count groups in our case) and then remove an entire chunk of count groups. It is evident from Figures 3 and 4 when more training data were removed, the performance of all the models degrades. However, our model's performance degraded more gracefully, resulting in the bigger performance gain over both the non-attribute based models (SVR and RR for age and crowd respectively) and non-cumulative attribute methods. These results further validate our early observation that the construction of a cumulative attribute space is uniquely effective for coping with sparse and imbalanced training data, a common problem in learning regression functions.

4.5. Learning Attributes Jointly vs. Independently

Methods	FG-NET [13]		UCSD [4]		
Original Dataset	mae	cs	mae	mse	mde
i-CA	4.73	73.7%	**2.07**	7.09	**0.102**
j-CA	**4.67**	**74.5%**	**2.07**	**6.86**	**0.102**
Missing 75% labels	mae	cs	mae	mse	mde
i-CA	6.45	55.6%	2.87	13.3	0.139
j-CA	**5.51**	**66.9%**	**2.79**	**12.6**	**0.137**

Table 6. Jointly learning cumulative attributes (j-CA) vs. independently learning cumulative attributes (i-CA).

Instead of learning all attributes jointly using our multi-out regression model, experiments were conducted to learn each attribute independently using a single out ridge regression model. Table 6 shows that comparing with the jointly learned attributes, the independently learned attributes led to poorer performance. In particular, for more imbalanced data with the removal of 75% labels from the original training dataset, our joint learning model yields more significant advantage on both the FG-NET age dataset and the UCSD crowd dataset. This is because that for sparse data, information sharing between attributes can contribute to improve robustness because of jointly penalizing the errors in different attributes.

4.6. Computational Cost

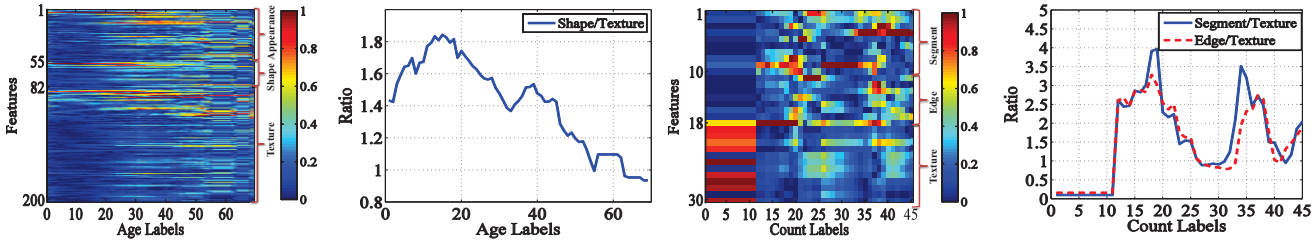
Methods	Age (mins)		Crowd (secs)	
	FG-NET [13]	MORPH [7]	UCSD [4]	Mall [8]
OHRank	1.30×10^4	3.02×10^4	–	–
SVR [15]	2.69×10^0	2.08×10^1	–	–
RR [8]	–	–	**0.70**	**0.67**
CA	$\mathbf{8.91 \times 10^{-1}}$	$\mathbf{6.10 \times 10^0}$	1.57	1.52

Table 7. Model training time required by different models.

Table 7 shows the training time for four different models. It is evident that the proposed cumulative attribute based model is extremely fast to learn owing to its closed form solution based on a multi-output regression model (see Section 3.2). For age estimation, it is even faster to train than the non-attribute based model with the same single output regression. The closest competitor for age estimation accuracy, OHRank [7] is four orders of magnitude (10^4) slower than our model (under 7 mins). This is because after mapping the low level image features to the cumulative attribute space, dimensionality reduction is achieved as a by-product resulting faster single output regression model training. For crowd counting, RR [8] is faster than CA. This is because the cumulative attribute space has a similar dimension as the original low-level feature and CA has the additional step of estimating the attribute values. Nevertheless, both are very fast to train (under 2 sec).

4.7. What is Learned by Cumulative Attributes?

To answer this question, Figures 5(a) and (c) visualise the weight matrix $\mathbf{W}$ in Formulation (1) which shows how different low level features are weighted for different scalar value groups. For age estimation, the AAM features capture the shape and texture characteristics of a human face. It is known [10] that at earlier ages, the human aging process is mainly reflected by the facial bone change (getting mature) resulting in shape changes. Entering adulthood, texture change gradually starts to play a more important role because aging is now more concerned with skin changes (e.g. having more wrinkles). Figures 5(a) and (b) show that our learned cumulative attribute indeed capture this phenomenon rather well. In particular, the shape features are the most important ones that separate attributes correspond to young ages (< 20), while texture features become more and more important for elder ages. For crowd counting, the 30 low level features contain foreground segment area, edge features and texture features. Segment and edge features would in general be more sensitive to the different crowdedness levels compared to the texture feature. That is, more



(a) Weight Heat Map of FG-NET (b) Feature-Ratio of FG-NET (c) Weight Heat Map of UCSD (d) Feature-Ratio of UCSD

Figure 5. Visualization of the importance of different features for cumulative attributes. Weights of each type of features were averaged for computing the weight ratio between different types of features.

people in the scene normally means larger foreground regions and more edges. This is also reflected by the learned weights shown in Figures 5(c) and (d).

5. Conclusion

We have introduced a novel cumulative attribute based framework for solving a number of computer vision problems invoking the need for regression estimation. Noisy and sparse low level visual features are mapped onto a cumulative attribute space where each dimension is designed specifically to give a clear semantic meaning that captures how the scalar output (e.g. age, people count) changes continuously. It requires no additional human annotation to assign attributes and can be estimated efficiently and robustly given sparse and imbalanced training data. Extensive experiments show the effectiveness and efficiency of the proposed model for both age estimation and crowd counting. This advantage of our approach is particularly significant when the training data is sparse and imbalanced.

References

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *CVPR*, 2007. 4, 6

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2006. 4

[3] A. Argyriou, T. Evgeniou, M. Pontil, A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *ML*, 2008. 4

[4] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: counting people without people models or tracking. In *CVPR*, 2008. 1, 2, 3, 4, 5, 6, 7

[5] A. B. Chan and N. Vasconcelos. Counting people with low-level features and bayesian regression. *TIP*, 2012. 1, 2, 3, 5

[6] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. A ranking approach for human ages estimation based on face images. In *ICPR*, 2010. 6

[7] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, 2011. 2, 3, 5, 6, 7

[8] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012. 1, 2, 3, 4, 5, 6, 7

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001. 4, 5

[10] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: a survey. *TPAMI*, 2010. 1, 2, 7

[11] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012. 2, 3

[12] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *TMM*, 2008. 1

[13] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *TPAMI*, 2007. 1, 2, 3, 5, 6, 7

[14] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Head pose estimation: classification or regression? In *ICPR*, 2008. 1

[15] G. Guo, Y. Fu, T. S. Huang, and C. R. Dyer. Image-based human age estimation by manifold learning and locally adjusted robust regression. *TIP*, 2008. 1, 2, 3, 5, 6, 7

[16] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, 2009. 1, 2, 3, 5

[17] Y. Haitovsky. On multivariate ridge regression. *Biometrika*, 1987. 4

[18] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 4

[19] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *TPAMI*, 2011. 3

[20] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2, 3

[21] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *TSMC*, 2004. 1

[22] R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In *BMVC*, 2012. 2, 3

[23] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2002. 6

[24] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*. IEEE, 2011. 2, 3

[25] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *ICCIT*, 2010. 1, 3

[26] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 3

[27] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: a survey. *TPAMI*, 2009. 1

[28] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *ACM MM*, 2009. 2

[29] D. Parik and K. Grauman. Relative attributes. In *ICCV*, 2011. 3

[30] M. Solnon, S. Arlot, and F. Bach. Multi-task regression using minimal penalties. *JMLR*, 2012. 4

[31] J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Automatic relevance determination for least squares support vector machine regression. In *IJCNN*, 2001. 6

[32] S. Yan, H. Wang, T. S. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *ICME*, 2007. 5, 6

[33] S. Yan, H. Wang, X. Tang, and T. S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV*, 2007. 1, 5, 6

[34] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang. Regression from patch-kernel. In *CVPR*, 2008. 1, 2

[35] Y. Zhang and D. Yeung. Multi-tasks warped gaussian process for personalized age estimation. In *CVPR*, 2010. 1, 2, 3, 5, 6