Leveraging Datasets with Varying Annotations for Face Alignment via Deep Regression Network

Jie Zhang[1,2] Meina Kan[1] Shiguang Shan[1,3] Xilin Chen[1]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{jie.zhang,meina.kan,shiguang.shan,xilin.chen}@vipl.ict.ac.cn

Abstract

Facial landmark detection, as a vital topic in computer vision, has been studied for many decades and lots of datasets have been collected for evaluation. These datasets usually have different annotations, e.g., 68-landmark markup for LFPW dataset, while 74-landmark markup for GTAV dataset. Intuitively, it is meaningful to fuse all the datasets to predict a union of all types of landmarks from multiple datasets (i.e., transfer the annotations of each dataset to all other datasets), but this problem is nontrivial due to the distribution discrepancy between datasets and incomplete annotations of all types for each dataset. In this work, we propose a deep regression network coupled with sparse shape regression (DRN-SSR) to predict the union of all types of landmarks by leveraging datasets with varying annotations, each dataset with one type of annotation. Specifically, the deep regression network intends to predict the union of all landmarks, and the sparse shape regression attempts to approximate those undefined landmarks on each dataset so as to guide the learning of the deep regression network for face alignment. Extensive experiments on two challenging datasets, IBUG and GLF, demonstrate that our method can effectively leverage multiple datasets with different annotations to predict the union of all landmarks.

1. Introduction

Facial landmark detection is a key component of many computer vision tasks, such as face recognition, face animation, video editing, *etc*. In the past few decades, many efforts are devoted to learn robust models for accurate face alignment under the controlled and uncontrolled setting [10, 34, 26, 4, 29, 28, 40, 2, 12, 35, 17]. At the same time, a lot of datasets under laboratory condition or wild condition are published for extensive evaluations [3, 21, 40, 27, 19, 18, 23]. These datasets have abundant variations in head pose, expressions, partial occlusions, *etc*. However, one dataset usually only focuses on one or several types of variations, and besides these datasets usually have different annotations, *e.g*., 68-landmark markup for LFPW [3], while 74-landmark markup for GTAV [18]. So the face alignment model learnt from one dataset can only predict those landmarks defined on that dataset, and different datasets can only be used independently. Naturally, it is quite anticipating to predict the union of all types of landmarks by leveraging multiple datasets.

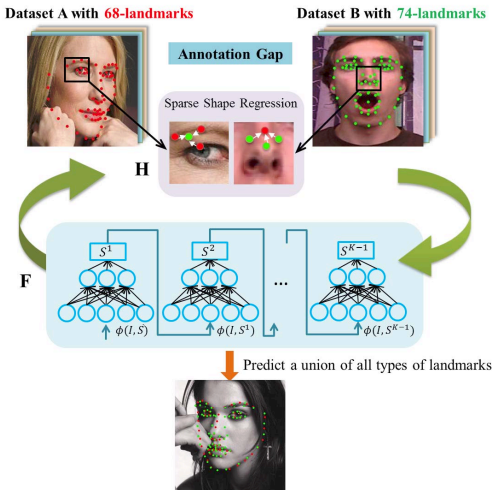

Figure 1. Overview of our DRN-SSR for face alignment by leveraging datasets with varying facial landmark annotations. **F** denotes the unified cascade deep regression networks which can predict a union of all types of landmarks, by taking shape-indexed feature $\phi(I, S)$ as input. **H** denotes the sparse shape regression model which approximates those undefined landmarks for each dataset, so as to guide the learning of deep networks **F**.

One straightforward solution is to train multiple detection models on each dataset, and therefore the union of all types of landmarks can be achieved by uniting the pre-

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.433

dictions of all models. However, the model trained on one dataset cannot get promising results on other datasets due to the distribution discrepancy between varied data. So ideally, all datasets are re-labeled with a union of all landmarks, based on which a unified alignment model of predicting the union of landmarks can be learnt. Nevertheless it is quite cost-consuming to re-label all datasets, and even impossible in some scenarios. The goal of this study is to predict a union of all types of landmarks on multiple datasets by taking together the variations and annotations from all datasets.

In order to predict the union of landmarks defined on multiple datasets, we propose a deep regression network coupled with sparse shape regression (DRN-SSR), which can take advantages of the variations (*e.g*., pose, expression, *etc*.) and annotations from multiple datasets without re-labelling them, as shown in Fig. 1. Specifically, for each dataset, a sparse shape regression model that characterizes the shape correlations across different datasets is introduced to approximate those undefined landmarks on this dataset, *i.e*., landmarks defined on other datasets and thus guides the learning of deep regression networks.

The main contributions can be summarized as: 1) By leveraging multiple datasets with varying annotations, a unified deep regression network is achieved, which can predict the union of all types of landmarks. 2) Benefited from the sparse shape regression, the proposed DRN-SSR can take advantages of both variations and annotations from all datasets, leading to a robust face alignment model that can characterize all variations from multiple datasets. Extensive experiments shows that our method achieves impressive performance for predicting all types of landmarks.

2. Related Work

2.1. Typical Face Alignment Methods

The early popular face alignment methods, such as ASMs [10, 11, 24] and AAMs [9, 22], employ Principal Component Analysis (PCA) to build statistical models of face shape and appearance. Generally, these parametric models achieve promising results on favorable images. However, their performances severely degenerate when tested on face images from unseen domain. For example, even with a large dataset for training, AAM generalizes poorly across datasets. Recently, in [31] a new framework is proposed to do fast and exact AAM fitting and achieves promising results on real-world face alignment.

Lately, regression based methods achieve great success for both controlled and uncontrolled face alignment [14, 34, 2, 7, 4, 26, 15, 13]. Dollar et al. [14] pioneer cascade shape regression algorithm for fast and accurate shape estimation of general objects by using shape-indexed features. Then Cao et al. [7] improve this method by simultaneously regressing all landmarks and it achieves better performance for face alignment in the wild. In [4], Burgos-Artizzu et al. employ interpolated shape-indexed feature and smart restart strategy to improve the robustness to large shape variations. In [34], SDM, as a highly effective and efficient face alignment method, cascades several linear regression models to predict shapes with shape-indexed SIFT feature and achieves impressive results for face alignment. In [13], Dantone et al. present a real-time face alignment method based on conditional regression forests, which achieves close-to-human accuracy on LFW [19]. Furthermore, Chen et al. [26] learn local binary features for robust shape regressions, which achieves both better accuracy and efficiency for face alignment in the wild.

Besides regression based methods, deep models also make great progress on face alignment [29, 33, 36, 37] and human pose estimation [30, 25]. Sun et al. [29] employ a three-level deep convolutional neural networks (DCNN) for facial landmark detection. Zhang et al. [37] also use cascade DCNN for face alignment and further improve the detection accuracy by simultaneously optimizing multiple correlated tasks. Besides, Zhang et al. [36] design a coarse-to-fine auto-encoder network for robust face alignment. Benefited from the favorable ability of characterizing nonlinearity, all these deep methods have achieved impressive results for real-world face alignment. In [30], Toshev et al. cascade deep neural networks to jointly estimate human pose and achieve state-of-art performance on real-world images. Pfister et al. [25] present a deep convolutional neural network to estimate human pose in videos, which exploits temporal information from videos and significantly outperforms the state-of-the-art methods on the BBC TV Signing dataset [8].

2.2. Face Alignment Across Datasets

Considering the existence of dataset bias for face alignment, Zhu et al. [39] extend the original SDM [34] to transductive SDM for transferring landmark annotations across datasets. By exploiting common facial landmarks as guidance, densely labeled landmarks are transferred from source dataset to target images and a more robust model is achieved with a combined training set of source and target. This method achieves good performance for face alignment of cross-dataset or unseen dataset. In another interesting work, Smith et al. [28] integrate nonparametric appearance model, affine-invariant shape constraint [38] and graph matching to get a prediction of the union of all types of landmarks. Firstly, landmarks from each source dataset are independently transferred to each target image. Then the individual landmark predictions are integrated into a single result by doing joint face alignment on the target dataset. It is the first effort to combine multiple datasets for effectively predicting a union of all types of landmarks. However, it suffers from high computation problem and may be inapplicable for single target image which is popular in online applications.

3. Our Approach

In this section, we will firstly give an overview of our DRN-SSR for predicting the union of landmarks defined on multiple datasets, and then illustrate the details about the formulation and optimization, followed by a discussion about the differences with the existing works.

3.1. Methodology

3.1.1 Overview

Suppose we have n datasets $\{D_1, D_2, ..., D_n\}$, and each dataset defines an individual type of landmarks, denoted as $S_i \in \mathbf{R}^{p_i \times 2}$ with p_i landmarks. In other words, we have n types of landmarks $S_1, S_2, ..., S_n$ that are defined on n datasets respectively. These n types of landmarks may or may not have common ones. For clear description of the formulation, we assume there are no common points between the n types of landmarks. But it should be noted that our method is also applicable with common landmarks.

As shown in Fig. 1, our goal is to build a deep regression network $\mathbf{F}$ that can predict the union of all types of landmarks $S = S_1 \cup S_2 ... \cup S_n, S \in \mathbf{R}^{p \times 2}$, by leveraging multiple datasets $\{D_1 \cup D_2 \cup ... \cup D_n\} \triangleq D$ as follows:

$$S = \mathbf{F}(\phi(I, \bar{S})) + \bar{S}, \quad S \in \mathbf{R}^{p \times 2}, \tag{1}$$

where $p = \sum_{i=1}^{n} p_i$ denotes the numbers of all landmarks, ϕ is a feature extraction function, and $\bar{S}$ is an initial shape.

If all images from D are labeled with n types of landmarks S, the face alignment model $\mathbf{F}$ can be achieved by minimizing the residual between the prediction from $\mathbf{F}$ and the ground truth S with an initial shape $\bar{S}$ as below:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \sum\nolimits_{I \in D} ||S - \big(\mathbf{F}(\phi(I, \bar{S})) + \bar{S}\big)||_2^2, \tag{2}$$

however, for image I only one type of landmarks are defined. That is, for images from the i^{th} dataset D_i, only the i^{th} type of facial landmarks are available while the other $n-1$ types of landmarks $\{S_1, \cdots, S_{i-1}, S_{i+1}, \cdots, S_n\}$ are missing. Fortunately, landmarks from different datasets have strong correlation as they usually have related semantics. Inspired by this, a sparse shape regression model $\mathbf{H}$ is introduced to approximate those undefined landmarks, and the approximation of S obtained from $\mathbf{H}$ is denoted as $S_{\mathbf{H}}$.

Briefly speaking, based on all images from D with all types of landmarks that are defined or approximated from the sparse regression model $\mathbf{H}$, the overall objective function of the unified deep regression model can be reformulated as follows:

$$\{\mathbf{F}^*, \mathbf{H}^*\} = \arg\min_{\{\mathbf{F},\mathbf{H}\}} \sum_{I \in D} ||S_{\mathbf{H}}(I) - \big(\mathbf{F}(\phi(I, \bar{S})) + \bar{S}\big)||_2^2. \tag{3}$$

Here, if $I \in D_i$, the i^{th} type landmarks in $S_{\mathbf{H}}$ are available and the other $n-1$ types of landmarks in $S_{\mathbf{H}}$ are approximated from $\mathbf{H}$.

In the following, we will give the details about the deep regression model $\mathbf{F}$ and the sparse shape regression model $\mathbf{H}$ for approximating the ground truth shape.

3.1.2 Deep Regression Network Coupled with Sparse Shape Regression

Considering that mapping from image to shape is a complex nonlinear process and inspired by the impressive nonlinear deep networks, here $\mathbf{F}$ is modeled as a deep regression network with $l-1$ hidden layers:

$$\Delta S \triangleq \mathbf{F}(I) = (f_l(f_{l-1}(...f_1(\phi(I, \bar{S})))), \tag{4}$$

$$a_q \triangleq f_q(a_{q-1}) = \sigma(W_q a_{q-1} + b_q), q \in [1, l-1], \tag{5}$$

$$f_l(a_{l-1}) = W_l a_{l-1} + b_l, \tag{6}$$

where $\phi(I, \bar{S})$ denotes the shape-indexed feature extracted around the initial shape $\bar{S}$ and f_q denotes the nonlinear mapping in the q^{th} layer parameterized with W_q and b_q, $q = 1, ..., l-1$. A sigmoid function σ is employed to characterize the nonlinearity mapping at the first $l-1$ layers, and $\{a_1, a_2, ..., a_{l-1}\}$ denotes the feature representations from each hidden layer. For the last layer, linear regression f_l is exploited to predict the shape deviation ΔS between the ground truth S and a initial shape $\bar{S}$.

As mentioned above, the S in Eq. (2) represents the ground truth shape for all types of landmarks. However, for any image $I \in D$ only one type of landmarks are defined, so we introduce a sparse shape regression model $\mathbf{H}$ to approximate the other $n-1$ types of landmarks as below.

Specifically, for the i^{th} dataset, only the i^{th} type of landmarks S_i are defined while the other $n-1$ types of landmarks S_j are undefined. Although the landmarks defined on different datasets are varied, they usually have strong correlation between each other, *e.g.*, a landmark on the upper eyelid from the j^{th} type can be approximated by those landmarks on eyes and eyebrows from the i^{th} type. So naturally, those undefined landmarks in S can be represented by several locally relevant ones defined in S_i, leading to the approximated $S_{\mathbf{H}} = [\widehat{S}_1; \widehat{S}_2; \cdots; \widehat{S}_{i-1}; S_i; \widehat{S}_{i+1}; \cdots; \widehat{S}_n] \in \mathbf{R}^{p \times 2}$. Formally, for images from D_i, sparse shape regression is employed to approximate the j^{th} type of undefined landmarks as follows:

$$\widehat{S}_j = H_{ij} * S_i, \;\; s.t., |h_{ij}^r|_1 < \tau, r = 1, \cdots, p_j, \tag{7}$$

where $H_{ij} \in \mathbf{R}^{p_j \times p_i}$ is a sparse matrix to approximate the undefined landmarks $\widehat{S}_j \in \mathbf{R}^{p_j \times 2}$ for those images in D_i, and h_{ij}^r is the r^{th} row of H_{ij} corresponding to one landmark. The principle of H_{ij} being sparse is that all landmarks of S_i would span a large subspace containing lots of landmarks, while sparse regression model tends to select a very small subset of S_i that are relevant to the undefined landmarks, which can span a very compact subspace around

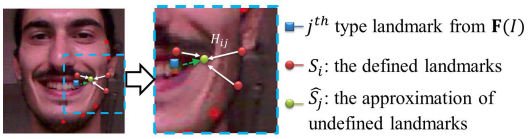

Figure 2. The j^{th} type of undefined landmarks are approximated by defined landmarks S_i with sparse shape regression H_{ij}.

the ground truth, as shown in Fig. 2. Besides, the sparse regression model is shared by all samples from one dataset, so it can tolerate the outliers benefitted from the statistics of subspace. With Eq. (7), all types of landmarks in $S_{\mathbf{H}}$ for those images from $D_i|_{i=1}^n$ can be reformulated as:

$$S_{\mathbf{H}} = \boldsymbol{H}_i * S_i = \begin{bmatrix} H_{i1} \\ \cdots \\ H_{i(i-1)} \\ E \\ H_{i(i+1)} \\ \cdots \\ H_{in} \end{bmatrix} * S_i = \begin{bmatrix} H_{i1} * S_i \\ \cdots \\ H_{i(i-1)} * S_i \\ S_i \\ H_{i(i+1)} * S_i \\ \cdots \\ H_{in} * S_i \end{bmatrix} \quad (8)$$

where $\boldsymbol{H}_i = [H_{i1}; \cdots; H_{i(i-1)}; E; H_{i(i+1)}; \cdots; H_{in}] \in \mathbf{R}^{p \times p_i}$ is the regression matrix for images from D_i, $E \in \mathbf{R}^{p_i \times p_i}$ is an identity matrix.

Overall, the objective function of the deep regression network for predicting the union of landmarks on multiples datasets can be reformulated as below:

$$\begin{aligned} &\arg\min_{\mathbf{F},\mathbf{H}} \sum_{i=1}^{n} \sum_{I \in D_i} \| S_{\mathbf{H}}(I) - \left(\mathbf{F}(I, \bar{S}) + \bar{S}\right) \|_2^2 \\ \Leftrightarrow &\arg\min_{\mathbf{F},\mathbf{H}} \sum_{i=1}^{n} \sum_{I \in D_i} \| \boldsymbol{H}_i * S_i(I) - \left(\mathbf{F}(I, \bar{S}) + \bar{S}\right) \|_2^2 \\ &s.t., |h_{ij}^r|_1 < \tau, r = 1, \cdots, p_j. \end{aligned} \quad (9)$$

$S_i(I)$ denotes its defined landmarks for image $I \in D_i$. $\mathbf{F}$ is a deep regression network which can predict a union of all types of landmarks. $\mathbf{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_n\}$ denotes a combination of sparse regression matrices for all datasets. Details for optimizing Eq. (9) are presented in Sec. 3.2.

3.1.3 Cascade Deep Regression Model

The above deep regression model will give a prediction $S^1 \in \mathbf{R}^{p \times 2}$ for the union of all types of landmarks. However, it is usually not close enough to the ground truth. So we cascade K successive deep regression models $\mathbf{F}^k, k = 1, 2, ..., K$ to further refine the shapes in higher and higher resolution. Specifically, after obtaining the shape prediction S^{k-1} from stage $k-1$, the stage k further improves the alignment results by optimizing the following objective:

$$\arg\min_{\mathbf{F}^k,\mathbf{H}^k} \sum_{i=1}^{n} \sum_{I \in D_i} \| \boldsymbol{H}_i^k * S_i(I) - (\mathbf{F}^k(I, S^{k-1}) + S^{k-1}) \|_2^2 \quad (10)$$

where $\mathbf{F}^k(I, S^{k-1}) = f_l^k(f_{l-1}^k(...f_1^k(\phi(I, S^{k-1}))))$ denotes a deep regression model of l layers with the shape-indexed SIFT feature $\phi(I, S^{k-1})$ as input and $\mathbf{H}^k = \{\boldsymbol{H}_1^k, \boldsymbol{H}_2^k, ..., \boldsymbol{H}_n^k\}$ is constrained to be sparse as in Eq. (9).

3.2. Optimization

As seen, Eq. (9) is a non-convex optimization problem with both $\mathbf{F}$ and $\mathbf{H}$ are unknown, thus we solve it by using the alternating optimization method, *i.e.*, the deep face alignment model $\mathbf{F}$ and the sparse matrices $\mathbf{H}$ are iteratively updated until both converge.

3.2.1 Initialization for Deep Regression Model F

The face alignment model F is simply initialized by only using those defined landmarks for each image $I \in D$, and can be formulated as the following objective:

$$\arg\min_{\mathbf{F}} \sum\nolimits_{i=1}^{n} \sum\nolimits_{I \in D_i} ||S_i(I) - \delta_i(\mathbf{F}(I, \bar{S}) + \bar{S})||_2^2, \quad (11)$$

where δ_i is a selection function to pick the i^{th} type of landmarks. In other words, the $l-1$ hidden layers are optimized by using samples from n datasets while the parameters corresponding to the i^{th} type of landmarks in the last layer are optimized by merely using those samples from the i^{th} dataset. Eq. (11) can be easily optimized by employing L-BFGS [20].

3.2.2 Alternating Optimization for H and F

After the deep face alignment model is initialized, we optimize the objective function in Eq. (9) by using the alternating method, *i.e.*, iteratively optimizing $\mathbf{H}$ and $\mathbf{F}$.

Given F, optimize H. When $\mathbf{F}$ is fixed, the objective function in Eq. (9) degenerates as below:

$$\begin{aligned} &\mathbf{H}^* = \arg\min_{\mathbf{H}} \sum\nolimits_{i=1}^{n} \sum\nolimits_{I \in D_i} ||\boldsymbol{H}_i * S_i(I) - Y_I||_2^2, \\ &s.t., |h_{ij}^r|_1 < \tau, , r = 1, \cdots, p_j, \end{aligned} \quad (12)$$

where $Y_I = \mathbf{F}(I, \bar{S}) + \bar{S}$, $\mathbf{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, ..., \boldsymbol{H}_n\}$. As seen from Eq. (12), $\boldsymbol{H}_i$ is independent from each other, so they can be optimized separately as follows.

$$\begin{aligned} &\boldsymbol{H}_i^* = \arg\min_{\boldsymbol{H}_i} \sum\nolimits_{I \in D_i} ||\boldsymbol{H}_i * S_i(I) - Y_I||_2^2, \\ &s.t., |h_{ij}^r|_1 < \tau, , r = 1, \cdots, p_j, \end{aligned} \quad (13)$$

with $\boldsymbol{H}_i = [H_{i1}; \cdots; H_{i(i-1)}; E; H_{i(i+1)}; \cdots; H_{in}]$. Each row h_{ij}^r of $\boldsymbol{H}_i$ is also irrelevant to each other, and thus can be further optimized independently. Eq. (13) can be formulated as follows with the r^{th} row of Y_I denoted as y_{ij}^r:

$$\begin{aligned} &h_{ij}^{r*} = \arg\min_{h_{ij}^r} \sum\nolimits_{I \in D_i} ||h_{ij}^r S_i(I) - y_{ij}^r||_2^2, \\ &s.t., |h_{ij}^r|_1 < \tau, \ i \neq j, \ r = 1, \cdots, p_j, \end{aligned} \quad (14)$$

which can be efficiently optimized by using the least angle regression algorithm [16]. **Given H, optimize F**. When **H** is fixed, the objective in Eq. (9) can be re-formulated as:

$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \sum_{i=1}^{n} \sum_{I \in D_i} ||\Delta S(I) - \mathbf{F}(I, \bar{S})||_2^2 + \alpha \sum_{q=1}^{l} ||W_q||_F^2 \quad (15)$$

where $\Delta S(I) = \boldsymbol{H}_i * S_i(I) - \bar{S}$. Additionally, the weight decay term $\sum_{q=1}^{l} ||W_q||_F^2$ is included to prevent over-fitting and this problem can be optimized by using L-BFGS [20].

Finally, we can achieve a stable solution by iteratively optimizing **H** and **F** according to the above two steps until both converge. As our DRN-SSR has several stages, for the k^{th} stage, we initial $\mathbf{H}^k$ with the optimized $\mathbf{H}^{k-1}$ from $k-1$ stage, and then alternatively optimize $\mathbf{H}^k$ and $\mathbf{F}^k$ as above.

3.3. Discussions

Differences from [28]. Smith et al. [28] also propose a pipeline to predict a union of all landmarks across datasets, and our method differs from theirs in two aspects: 1) the application scenario is different: [28] can only jointly align all testing images together as the correlation between all testing images are employed to guide the collaborative transfer of annotations from the training set to testing set. On the contrary, ours only use the training set to learn the model which makes it applicable for even single image, and thus our method is more practical since in many scenarios it is unable to access all testing images at the same time. 2) [28] trains individual models on each dataset for detecting one type of landmarks and then get the predictions of the union of all landmarks by integrating the results from individual models. In other words, in [28] the predictions from multiple datasets are integrated to achieve the result of all landmarks, while in our DRN-SSR the data from multiple datasets are integrated to learn single model that can predicts the union of landmarks. Integrating data rather than the predictions can better characterize the variations from multiple datasets, leading to a more robust alignment model. Moreover, [28] takes more than 30 seconds per image on a powerful workstation while our DRN-SSR performs much more efficiently with 0.33 second on a I7 desktop.

Differences from [39]. Zhu et al. [39] propose a transductive alignment method (TCR) to transfer landmark annotations from one dataset to another. The differences between our method and [39] are as follows: 1) TCR [39] employs only those common facial points to guide the transfer between datasets, while our method takes all defined landmarks on each dataset as the guidance for transferring landmarks across datasets. 2) [39] employs transductive SDM to get the approximation of those undefined landmarks which stays unchanged as long as they are approximated, while ours refines the approximation and updates the face alignment model iteratively, resulting in a more accurate approximation of undefined landmarks and then achieving more robust alignment model.

4. Experiments

4.1. Experimental settings

To evaluate our method, seven public datasets are employed, *i.e.*, **LFPW** [3], **HELEN** [21], **AFW** [40], **IBUG** [27], **GTAV** [18], **LFW** [19] and **FaceWarehouse** [6]. Both FaceWarehouse and GTAV are collected under laboratory conditions while the others are collected in the wild. FaceWarehouse contains 5904 images of 150 individuals with various expressions and GTAV has large variations in pose, expressions, illuminations and partial occlusions. Both HELEN and AFW are collected from *Flickr*. HELEN contains 2330 high resolution images and AFW includes 205 images with 468 faces. Recently, LFPW, HELEN and AFW are relabeled with 68 landmarks and released in [1]. Besides, they release another dataset IBUG including 135 images with extreme pose and expressions. For GTAV, LFW and FaceWarehouse, the annotations of 74 landmarks are released by [5].

As illustrated, there are two types of annotations among these datasets, *i.e.*, 68 landmarks and 74 landmarks, with 29 common ones. So the problem is how to predict the union of all 113 landmarks (68+74-29=113) by leveraging two types of datasets respectively. We re-organize these datasets as follows: a training set consisting of two subsets, one with 68 landmarks and the other with 74 landmarks; two testing sets with 113 landmarks for evaluation.

The training subset with 68 landmarks consists of 3478 images from LFPW training set, HELEN and AFW datasets, recorded as **68-type training set**. The training subset with 74 landmarks consists of 14360 images from GTAV, LFW and FaceWarehouse datasets, recorded as **74-type training set**. One testing set is **IBUG** dataset, which is comprised 135 extremely challenging images in the wild. The other testing set named as **GLF** dataset, contains 100 images selected from GTAV, LFW and FaceWarehouse, which have no overlap with 74-type training subset. In order to evaluate all types of landmarks, we manually re-label IBUG and GLF to make them have 113 landmarks. The cumulative function (CDF) is employed to measure the performance based on the normalized root-mean-squared error (NRMSE) which is normalized by face size.

In our approach, we cascade 4 deep regression networks and each of them has four layers including three non-linear hidden layers and the last layer of linear regression. For all stages, the numbers of hidden units in each layer are respectively 1296, 784, 400. For the first two stages, the face images are normalized to 80×80 pixels and for the third and fourth stages, the face images are normalized to 140×140 pixels. The weight decay parameter α of each layer is set to

0.001. The sparsity parameter τ which controls the sparsity of matrices $\mathbf{H}$ and determines the number of landmarks selected as bases is set as 1.4 by exploring its performance w.r.t. different values (see Sec. 4.2.3 for details).

4.2. Analysis of DRN-SSR

4.2.1 Benefits of Coupling DRN with SSR

As seen in Eq. (9), our approach couples the deep alignment model with sparse shape regression together, so the approximations of those undefined landmarks are refined iteratively. To investigate the effectiveness of the coupling strategy, we compare it with another two strategies: one is to learn deep alignment model without sparse shape regression denoted as "Deep", another is a two-step approach which firstly does sparse shape regression and then learns deep alignment model with the approximated landmarks, denoted as "Deep+Sparse", and our method is denoted as "Deep Coupled with Sparse" which can iteratively refine the deep alignment model with the updated approximations for undefined landmarks from previous stage.

To simulate the scenario of multiple datasets with different annotations, 36 landmarks are assumed to be available on HELEN training set, and another 32 landmarks are assumed to be available on LFPW training set. For comparison, images from LFPW and HELEN testsets are evaluated in terms of 68 landmarks. The "Deep" strategy learns two separate alignment models on HELEN and LFPW to predict 36 and 32 landmarks respectively. In "Deep+Sparse" strategy, the 32 undefined landmarks of HELEN and 36 undefined ones of LFPW are firstly predicted by using the two models in "Deep" strategy, then the predictions are refined by sparse shape regression only once, and finally based on the defined and approximated landmarks, "Deep+Sparse" model is trained with both images from HELEN and LFPW to predict 68 facial points. Our "Deep Coupled with Sparse" models the deep alignment networks and sparse regression together, and the approximations for undefined landmarks can be refined iteratively, which can further refine the alignment model.

The evaluation results are shown in Fig. 3. Given a threshold t, "Data Proportion" denotes the percentage of faces whose NRMSE is below t. As seen, "Deep" performs the worst as each dataset is modeled independently which means only those variations from one dataset are captured to predict each facial landmark. The "Deep+Sparse" strategy performs better benefited from capturing all variations from both datasets to predict each point. However, the approximations of undefined landmarks on training data are not good enough leading to a biased model, as the predictions of those undefined landmarks on each training set may be far from the ground truth. On the contrary, in our coupled strategy, the alignment model and the approximations for undefined landmarks are iteratively refined, leading to a more robust alignment model which is learned with better and better approximations for undefined landmarks. As seen from Fig. 3, the accuracy is further improved up to 8% when NRMSE is 0.03, implying the necessity of coupling deep alignment model and sparse shape regression.

4.2.2 Jointly Predicting Landmarks vs. Uniting Data

In our approach, the data from multiple datasets are united together to predict the union of multiple types of landmarks, which means the performance gain might stem from the more variations of the united data, the jointly predicting multiple types of landmarks, or both. As stated in Sec. 4.2.1, the same datasets are used for this investigation.

The baseline method learns two separate alignment models on HELEN and LFPW with 36 and 32 landmarks respectively, and final detection of 68 landmarks are achieved by merging the predictions from these two models. This method is recorded as "LFPW_{32}+HELEN_{36}", in which neither the multiple types of landmarks are jointly predicted nor the data from multiple datesets are leveraged together. As illustrated in Fig. 4, it performs the worst as expected.

Furthermore, a method which only considers the joint prediction of multiple types of facial landmarks is evaluated, recorded as "LFPW_{68}+HELEN_{68}". This method also learns two separate models on LFPW training set and HELEN training set respectively, but with 68 landmarks for both models. As seen in Fig. 4, "LFPW_{68}+HELEN_{68}" performs slightly better than "LFPW_{32}+HELEN_{36}", as detection of different but relative annotations can benefit each other, which means jointly predicting multiple types of landmarks can improve the performance, but only slightly.

Moreover, the most favorable method should jointly predict multiple types of landmarks and leverage the multiple datasets, denoted as "$(\text{LFPW+HELEN})_{68}$", which learns a unified model on both LFPW and HELEN training sets with 68 landmarks. As seen, "$(\text{LFPW+HELEN})_{68}$" performs much better, which means more variations from multiple datasets can significantly improve the performance.

Yet, in the real world scenario, each dataset usually has only one type of annotations, and it is quite cost-consuming or even impossible to attain all types of annotations for multiple datasets to make "$(\text{LFPW+HELEN})_{68}$" applicable. Instead, our DRN-SSR can leverage all these datasets without relabelling them, to jointly predict the union of all types of landmarks. As seen, although our method is not as good as "$(\text{LFPW+HELEN})_{68}$" which needs all types of annotations, it still significantly outperforms "LFPW_{32}+HELEN_{36}" and "LFPW_{68}+HELEN_{68}" with an improvement up to about 13% and 10% respectively when NRMSE is 0.03.

These comparisons demonstrate that more variations from multiple datasets can significantly improve the performance of alignment, and the proposed DRN-SSR can effectively leverage multiple datasets, even if each dataset only

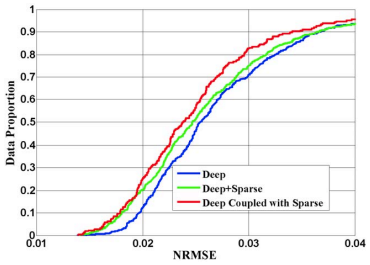

Figure 3. Benefits of coupling deep alignment model with sparse shape regression.

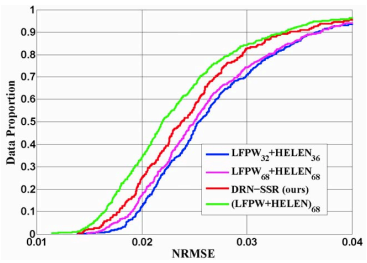

Figure 4. Jointly predicting landmarks vs. uniting data.

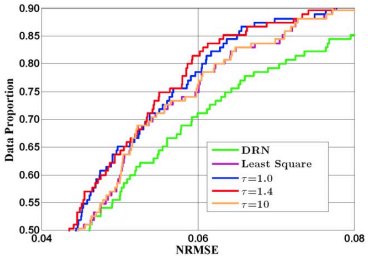

Figure 5. Results of the first stage on IBUG.

has one type of annotation, which makes our method a very practical solution for leveraging data as much as possible.

4.2.3 Sparsity Parameter of Sparse Shape Regression

As mentioned above, the sparse shape regression can characterize the shape correlations between datasets by constructing compact approximations for those undefined landmarks. To explore how the sparsity τ influences the accuracy of approximations of those undefined landmarks, we evaluate the performance of our method under different sparsity. Specifically, we evaluate our method with τ set as 1.0, 1.4, 10, least square regression (*i.e.*, $\tau \to +\infty$) and DRN, *i.e.*, a method without considering the approximation for undefined landmarks.

Fig. 5 shows the performance of the first stage from these methods by evaluating 113 landmarks on IBUG. As seen, DRN performs the worst due to no consideration of approximation for undefined landmarks. Furthermore, the deep regression network with least square regression to approximate those undefined landmarks performs better by leveraging both datasets with 68 landmarks and 74 landmarks together. When the sparsity is set to be large, *e.g.*, $\tau = 10$, deep regression network coupled with sparse shape regression performs similarly as that with the least square regression. This is because almost all landmarks are selected for sparse reconstruction, which approaches the least square regression. Although least square regression or sparse regression with a large sparsity can achieve better approximations for those undefined landmarks, they are not accurate enough as the approximations are from a relax subspace spanned by almost all defined landmarks. On the contrary, sparse regression with a small sparsity, e,g., $\tau = 1.4$ and 1.0, can form a compact subspace spanned by only several relevant landmarks, resulting in more accurate approximations of those undefined landmarks as shown in Fig. 6. Fig. 7 shows the selected landmarks (red dots) that are used to approximate the undefined landmarks (green dots) with $\tau = 1.4$. As seen, only those landmarks that have strong correlations with the undefined landmarks are selected as expected.

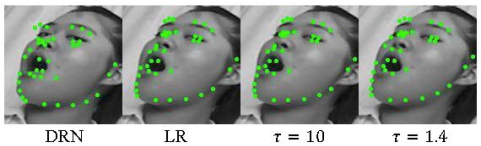

Figure 6. Approximations of undefined landmarks at the first stage on training dataset. LR denotes least square regression.

Figure 7. The selected landmarks (red dots) for approximating the undefined facial points (green dots), when $\tau = 1.4$.

4.3. Comparison with the Existing Methods

To compare with the existing methods, two types of training sets are used: 68-type training set and 74-type training set. The goal is to predict the union of both types of landmarks, *i.e.*, 113 landmarks with 29 common ones. We compare our DRN-SSR with a few state-of-the-art methods, *e.g.*, RCPR [4], SDM [34], FAST-SIC [32] and Smith et al. [28]. For RCPR and FAST-SIC, we use their off-the-shelf codes. We implement SDM which achieves comparable accuracy to the origin. For Smith et al. [28], we directly quote results from [28]. Besides, a Deep Regression Network (DRN) for face alignment is implemented as a baseline, which is also a deep method but without leveraging multiple datasets with varying annotations.

To our best knowledge, none of these methods except [28] and ours can utilize multiple datasets with varying annotations to predict a union of landmarks. To make these methods, *i.e.*, RCPR, SDM, FAST-SIC and DRN, predict a union of 113 landmarks, two models respectively predicting 68 and 74 landmarks are trained on 68-type and 74-types training sets separately. The predictions of these two models are merged together as the final output of the union of 113 landmarks, with that of those common points averaged. [28] and our proposed method can learn a unified model that directly predicts 113 landmarks by leveraging both 68-type and 74-type training sets.

4.3.1 Evaluations on IBUG Dataset

Firstly, we evaluate all methods on IBUG dataset which is extremely challenging due to extreme poses, exaggerat-

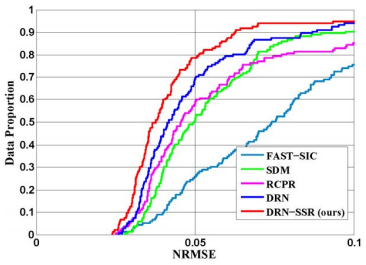

Figure 8. IBUG, 113 Landmarks.

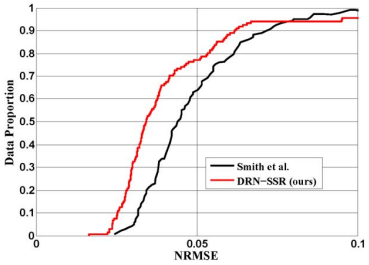

Figure 9. IBUG, 68 Landmarks.

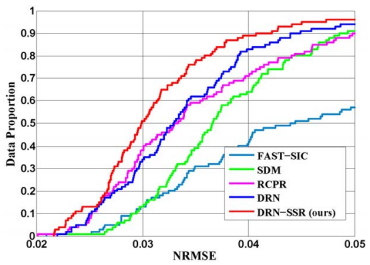

Figure 10. GLF, 113 Landmarks.

ed expressions, occlusions, *etc*. The performance on this dataset is reported in terms of 113 landmarks, as shown in Fig. 8. As a fast and exact AAM fitting method, FAST-SIC significantly outperforms project-out inverse compositional algorithm (POIC) [22] as evaluated in [32], however it degenerates on this challenging dataset, as the linear principal component analysis model cannot well characterize the complex variations in shape and appearance. Furthermore, SDM and RCPR performs better benefited from the effective shape regression pipeline with shape-indexed feature. Attributed to the favorable ability of modeling nonlinearity, DRN makes further improvement than SDM and RCPR. Our DRN-SSR outperforms DRN with an improvement up to 10% when NRMSE is 0.05, as more variations from multiple datasets with varying annotations are modeled together, promising a more robust alignment model. Tabel 1 reports the mean errors of all these methods on IBUG. Moreover, we compare our method to [28], which can also combines multiple datasets with different annotations to predict a union of all landmarks. Since the CDF curve of only 66 landmarks is reported in [28], the common 66 points are evaluated for fair comparison. As shown in Fig. 9, our method outperforms [28] and the improvement is even up to 13% w.r.t. NRMSE=0.05 even though [28] use manually labeled eye centers to remove the rotation and scale variations. This is possibly because [28] cannot simultaneously cover all variations from multiple training sets with varying annotations as it only integrates the alignment results from individual models rather than combining data, while ours can well capture all variations from multiple training sets simultaneously, leading to better predictions. Besides, [28] performs much slower than ours, and it is only applicable for image sets, but not single image scenario.

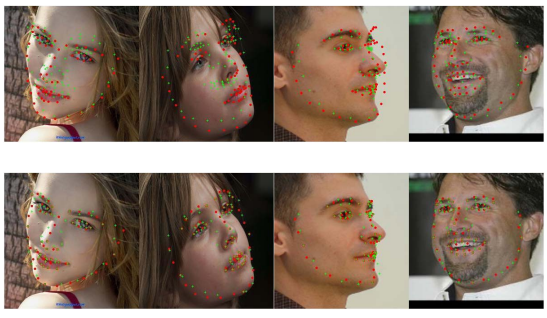

Figure 11. Fitting Results on IBUG and GLF. Top: DRN fitting results, Bottom: our fitting results.

Table 1. The mean errors (%) of 113 landmarks on IBUG and GLF datasets, which is normalized by face size.

	FAST-SIC	SDM	RCPR	DRN	DRN-SSR
IBUG	8.22	6.17	6.73	5.26	4.58
GLF	5.47	3.84	3.72	3.42	3.16

4.3.2 Evaluations on GLF Dataset

Secondly, we further evaluate our method on GLF dataset. GLF consists of 100 challenging images with large variations in pose, expression, partial occlusion, blur, *etc*. The comparison results are shown in Fig. 10, from which the similar conclusion can be obtained. As seen, RCPR and DRN performs the best among the existing methods and our DRN-SSR achieves a further improvement with 10% when NRMSE is 0.03 and 6% when NRMSE is 0.04 respectively. This demonstrates that it is beneficial to fuse the multiple datasets with varying annotations, and our method is an effective solution to leverage more variations from multiple datasets to predict the union of all types of landmarks. Results shown in Tabel 1 also support the conclusion. Fig. 11 shows the fitting results of some challenging samples from IBUG and GLF. As seen, the proposed DRN-SSR can accurately predict the union of all landmarks.

5. Conclusions

By leveraging the datasets with varying annotations, we present a unified deep regression network coupled with the sparse shape regression (DRN-SSR) to predict the union of all types of landmarks. With the shape correlations between different datasets bridging the annotation gap, DRN-SSR can utilize multiple datasets with different annotations, which can integrate more data variations. As evaluated on two challenging datasets, our method achieves impressive performance for predicting the union of all landmarks.

Acknowledgements

This work was partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61173065, 61222211, 61390511 and 61402443.

References

[1] 300 faces in-the-wild challenge. http://ibug.doc.ic.ac.uk/resources/300-W/.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[5] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *TOG*, 2014.

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 2013.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[8] J. Charles, T. Pfister, M. Everingham, and A. Zisserman. Automatic and efficient human pose estimation for sign language videos. *IJCV*, 2014.

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

[10] T. F. Cootes and C. J. Taylor. Active shape models. In *BMVC*. 1992.

[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995.

[12] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[13] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.

[14] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[15] C. Dong, R. Shaoqing, W. Yichen, C. Xudong, and S. Jian. Joint cascade face detection and alignment. In *ECCV*, 2014.

[16] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of Statistics*, 2004.

[17] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *IVC*, 2009.

[18] A. R. Francesc Tarrs. GTAV Face Database. http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm.

[19] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, 2007.

[20] Q. V. Le, A. Coates, B. Prochnow, and A. Y. Ng. On optimization methods for deep learning. In *ICML*, 2011.

[21] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*. 2012.

[22] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004.

[23] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended m2vts database. In *AVBPA*, 1999.

[24] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*. 2008.

[25] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *ACCV*. 2014.

[26] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, 2013.

[28] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *ECCV*. 2014.

[29] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[30] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[31] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, 2013.

[32] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, 2013.

[33] Y. Wu, Z. Wang, and Q. Ji. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *CVPR*, 2013.

[34] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[35] J. Zhang, M. Kan, S. Shan, X. Zhao, and X. Chen. Topic-aware deep auto-encoders (tda) for face alignment. In *ACCV*. 2014.

[36] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*. 2014.

[37] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*. 2014.

[38] F. Zhou, J. Brandt, and Z. Lin. Exemplar-based graph matching for robust facial landmark localization. In *ICCV*, 2013.

[39] S. Zhu, C. Li, C. C. Loy, and X. Tang. Transferring landmark annotations for cross-dataset face alignment. *arXiv preprint arXiv:1409.0602*, 2014.

[40] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.