Soft Facial Landmark Detection by Label Distribution Learning

Kai Su, Xin Geng*
MOE Key Laboratory of Computer Network and Information Integration,
School of Computer Science and Engineering,
Southeast University, Nanjing 210096, China
{sukai, xgeng}@seu.edu.cn

Abstract

Most existing facial landmark detection algorithms regard the manually annotated landmarks as precise hard labels, therefore, the accurate annotated landmarks are essential to the training of these algorithms. However, in many cases, there exist deviations in manual annotations, and the landmarks marked for facial parts with occlusion and large poses are not always accurate, which means that the "ground truth" landmarks are usually not annotated precisely. In such case, it is more reasonable to use soft labels rather than explicit hard labels. Therefore, this paper proposes to associate a bivariate label distribution (BLD) to each landmark of an image. A BLD covers the neighboring pixels around the original manually annotated point, alleviating the problem of inaccurate landmarks. After generating a BLD for each landmark, the proposed method firstly learns the mappings from an image patch to the BLD of each landmark, and then the predicted BLDs are used in a deformable model fitting process to obtain the final facial shape for the image. Experimental results show that the proposed method performs better than the compared state-of-the-art facial landmark detection algorithms. Furthermore, the proposed method appears to be much more robust against the landmark noise in the training set than other compared baselines.

Introduction

Facial landmark detection aims to localize feature points on a face image, such as the nose, chin, eyes and mouth. It is a prerequisite of many automatic facial analysis systems, e.g., face recognition (Zhao et al. 2003) and facial age estimation (Geng, Yin, and Zhou 2013). Thus, this task has attracted more and more attention in recent years. A large number of approaches have been proposed for facial landmark detection, which can be roughly classified into two families, i.e., model based methods and regression based methods.

Active Shape Models (ASM) (Cootes et al. 1995) and Active Appearance Models (AAM) (Matthews and Baker 2004) are two early typical model based methods. ASM applies Principal Component Analysis (PCA) to a set of aligned training shapes to build its shape model. AAM is an extension of ASM, generating both shape and appearance models for an image. Constrained Local Models (CLM) (Cristinacce and Cootes 2008; Zhu and Ramanan 2012) is another widely-used class of model based approaches for the facial landmark detection. The shape model of CLM is the same point distribution model as the one used by ASM and AAM. Unlike AAM building holistic appearance model, CLM uses a set of local appearance patches cropped around each current landmark to represent a face. Generally speaking, model based methods attempt to optimize the model parameters by maximizing the probability of a facial image being reconstructed by their deformable models. However, building powerful deformable models requires a massive amount of training images with carefully annotated landmarks (Sagonas et al. 2013b), while most existing model based methods have not taken into consideration the issue of inaccurate "ground truth" landmarks.

The most representative way of regression based methods is the Cascaded Shape Regression (CSR) framework (Cao et al. 2014; Ren et al. 2014; Trigeorgis et al. 2016; Xiong and De la Torre 2013; Zhang et al. 2014), which directly learns a set of regressors to update the estimated shape iteratively in a coarse-to-fine manner. For example, ESR (Cao et al. 2014) tries to directly learn a regression function with shape-indexed features to infer the whole facial shape for an image. SDM (Xiong and De la Torre 2013) learns a cascaded descent direction to minimize the shape residuals on the hand-crafted SIFT features. LBF (Ren et al. 2014) learns a set of local binary features for each landmark independently, and then uses these features to jointly learn a linear regressor to minimize errors between the predicted and ground truth shape. In recent years, deep learning techniques have also been applied to the CSR framework. For example, CFAN (Zhang et al. 2014) uses cascaded Auto-Encoder networks with different resolution image inputs to predict accurate landmarks. MDM (Trigeorgis et al. 2016) adopts powerful CNN-based features and RNN-based memory units to perform coarse-to-fine shape refinement. The optimization target of the shape regression methods is to directly minimize the residuals between the predicted and ground truth shapes in a cascaded manner. That is to say, accurate ground truth landmarks are essential to their training process.

However, obtaining accurate facial landmarks for a face image may be a serious issue. Currently, there are two kinds of landmark annotation methods (Sagonas et al. 2013b): manual and semi-automatic. Manual annotation of facial im-

*Corresponding author.

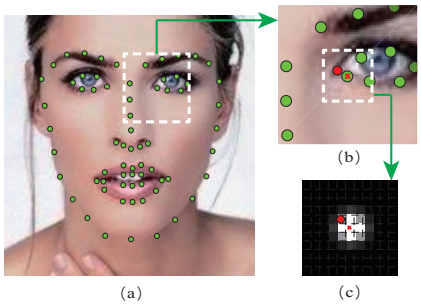


Figure 1: (a) The example of the inaccurate annotated landmarks in the 300-W database. (b) The annotated point with a red cross is inaccurate, which is significantly deviated from the ground truth (the red point). (c) The example of a BLD is assigned to the red cross point. Higher intensity (brighter) in the BLD means stronger relevance. Better view in color.

ages requires trained human experts. The heavy annotation workload often makes people tired, which will cause subjective deviations in the manual annotations. Moreover, it is usually difficult for the annotators to mark the landmarks for the facial parts with occlusions, large illumination or pose variations. Semi-automatic annotation methods usually contain three steps (Sagonas et al. 2013b): first of all, use the existing annotated facial image subsets to train an Active Orientation Models (AOM), and then fit the trained AOM to the non-annotated facial image subsets. The fitting results are manually classified into the "Bad" and "Good" subsets. A new AOM for the non-annotated subsets is trained from the "Good" fitting subsets, and the remaining images with the "Bad" results are re-fit using the re-trained AOM until convergence. However, semi-automatic methods still can not completely avoid the subjective errors of manual classification. Moreover, it may introduce errors caused by the AOM. In short, either manual or semi-automatic, the landmark annotation methods cannot avoid inaccurate points that are significantly deviated from the ground truth. For example, Fig. 1 shows a typical facial image with 68 annotated landmarks from the 300-W (Sagonas et al. 2013a) database. The green points are the manually annotated landmarks from the dataset. A close observation of the enlarged left eye patch (Fig. 1(b)) reveals that the original landmark for the eye corner (the point with a red cross) is actually significantly deviated from the ground truth (the red point). Such case is unfortunately very common in the current datasets. However, most existing facial landmark detection algorithms pay little attention to this issue, which might cause serious performance deterioration.

Based on the above observation, we propose a *soft* facial landmark detection method in this paper. As shown in Fig. 1, without further information, we can only assume that the annotated point with the red cross is the most relevant label to the ground truth red point. Meanwhile, the neighboring points around the red cross point can also be regarded as candidates for the ground truth. Of course, with the basic assumption that the ground truth point should not be far away from the annotated point, the possibility of the neighboring point being the ground truth will decrease with the increase of the distance to the annotated point. This will create a data structure matching a recently proposed machine learning paradigm called *Label Distribution Learning* (LDL) (Geng 2016). The label distribution covers a certain number of labels, each label has its own description degree, representing the degree to which each label describes the instance. In this paper, a label distribution is assigned to each annotated landmark. The label in the label distribution refers to the candidate point for the ground truth landmark, and the corresponding description degree is explained as the degree to which the point can describe the ground truth landmark. The description degree of a point fades away when the Euclidean distance between this point and the annotated landmark increases. In the two-dimensional image space, the description degrees of all possible points form a bivariate probability distribution, which is called *bivariate label distribution (BLD)*. As shown in Fig. 1, the example of a BLD can be seen in (c), which is generated from the red cross point in the image patch (b). If we crop a patch centered at the red cross point, all the pixels in the patch form the label space. Then, the BLD assigned to the red cross point is generated via a bivariate Gaussian distribution centered at the red cross point. Higher intensity in Fig. 1(c) means higher possibility of being the ground truth. In this way, we obtain a soft facial landmark covering a small neighborhood around the original annotation. As long as the ground truth landmarks are not far away from the annotated points, the BLDs assigned to the annotated points can cover the likelihoods of the ground truth landmarks, which alleviates the effects of the inaccurate annotated landmarks. After generating a BLD for each landmark, our proposed method firstly learns the mappings from an image patch to the BLD of each landmark, and then the predicted BLDs are used in a deformable model fitting process to obtain the final predicted facial shape.

The rest of this paper is organized as follows. First, prior works on the LDL and the CLM framework are reviewed. Second, Soft Facial Landmark Detection by LDL is proposed. After that, the experimental results are reported. Finally, a conclusion and future work are drawn.

Related Work

Label Distribution Learning

Label Distribution Learning (LDL) (Geng 2016) is a novel learning paradigm, which mainly focuses on the ambiguity at the label side. The label distribution covers a certain number of labels, each label has its own description degree, representing the degree to which each label describes the instance. The description degrees of all the labels sum up to 1. LDL is a more general learning framework which includes both single-label and multi-label learning (Tsoumakas and Katakis 2006) as its special cases. LDL has been successfully applied to many real applications, such as age estimation (Geng, Yin, and Zhou 2013), facial expression recognition (Zhou, Xue, and Geng 2015) and action detection in videos (Geng and Ling 2017). In this paper, a bivariate label distribution (BLD) is assigned to each landmark, modeling the likelihoods of the neighboring points.

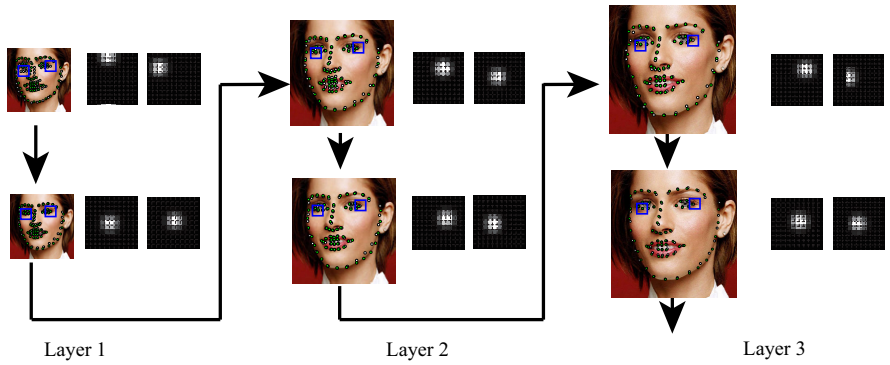


Figure 2: Overview of our Multi-scale Cascaded BLD Regression (MCBR). White points are the currently estimated shapes. Green points are the annotated shapes. The blue rectangles are the image patches centered at two selected example white landmarks. Their BLDs are on the right-side of each image. Higher intensity (brighter) means stronger relevance. Better view in color.

Constrained Local Model

Deformable model fitting is often performed in the Constrained Local Model (CLM) framework (Cristinacce and Cootes 2008). More specifically, the CLM framework is mainly composed of three parts: a point distribution model (PDM), local patch experts and the deformable model fitting approach.

Point Distribution Model Point Distribution Model (PDM) (Cootes et al. 1995) is a typical parameterized shape model, which applies PCA to obtain a linear approximation about the shape variations (e.g., facial expressions, head poses). In order to place a shape in the image frame, the PCA model is composed with a 2D global similarity transform (translation $\boldsymbol{t}$, in-plane rotation $\boldsymbol{R}$ and scale s):

$$\boldsymbol{x}_l = s\boldsymbol{R}(\overline{\boldsymbol{x}}_l + \boldsymbol{\Phi}_l\boldsymbol{q}) + \boldsymbol{t}. \tag{1}$$

The parameters describing the PDM is denoted as $\boldsymbol{p} = \{s, \boldsymbol{R}, \boldsymbol{t}, \boldsymbol{q}\}$, where $\boldsymbol{q}$ is the shape variation parameter vector. $\overline{\boldsymbol{x}}_l$ denotes the mean location of the l-th facial landmark, and $\boldsymbol{\Phi}_l$ denotes the related sub-matrix of the shape eigenvectors $\boldsymbol{\Phi}$. Generally, assume the shape variation parameters follow a Gaussian distribution, and the global similarity transform parameters have a uniform prior, then PDM parameters have the following prior (Saragih, Lucey, and Cohn 2011):

$$\begin{aligned} &f_{\mathcal{N}}(\boldsymbol{q}; \boldsymbol{0}, \boldsymbol{\Lambda}) = \frac{1}{\sqrt{2\pi\boldsymbol{\Lambda}}} exp(-\frac{\boldsymbol{q}^2}{2\boldsymbol{\Lambda}^2}), \\ &p(\boldsymbol{p}) \propto f_{\mathcal{N}}(\boldsymbol{q}; \boldsymbol{0}, \boldsymbol{\Lambda}), \end{aligned} \tag{2}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues associated to the shape eigenvectors.

Local Patch Experts Local patch experts are a very important part of the CLM framework. For the l-th landmark, a local patch expert evaluates the probability of the landmark being aligned at point $\boldsymbol{y}$, i.e., $p(g_l = 1|\boldsymbol{y})$, where $g_l \in \{-1, 1\}$ is a variable denoting whether l-th landmark is misaligned or aligned at point $\boldsymbol{y}$. There have been a number of different methods proposed as local patch experts, e.g., Logistic Regression (LR) (Saragih, Lucey, and Cohn 2011), Minimum Output Sum of Squared Errors (MOSSE) filters (Bolme et al. 2010).

Deformable Model Fitting The goal of deformable model fitting is to register a parameterized shape model (e.g., PDM) to a face image $\boldsymbol{I}$ such that landmarks reconstructed by the model is as close to the consistent locations in the image as possible (Saragih, Lucey, and Cohn 2011). Fitting process can be viewed as a search for the model parameters $\boldsymbol{p}$, that jointly maximizes the probability of all landmarks being well aligned, with the regularizations over $\boldsymbol{p}$. Assuming the alignments for each landmark (L landmarks totally) are conditionally independent, a Maximum A Posterior (MAP) estimation function of $\boldsymbol{p}$ is:

$$p(\boldsymbol{p}|\{g_l = 1\}_{l=1}^{L}, \boldsymbol{I}) \propto p(\boldsymbol{p}|\boldsymbol{\Lambda}) \prod_{l=1}^{L} p(g_l = 1|\boldsymbol{x}_l, \boldsymbol{I}, \boldsymbol{p}), \tag{3}$$

This can be solved using various methods, and the most popular one is the RLMS approach proposed in (Saragih, Lucey, and Cohn 2011):

$$\begin{aligned} &\Delta\boldsymbol{p} = -(\rho\widetilde{\boldsymbol{\Lambda}}^{-1} + \boldsymbol{J}^T\boldsymbol{J})^{-1}(\rho\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{p} - \boldsymbol{J}^T\boldsymbol{v}), \\ &\boldsymbol{p}^* = \boldsymbol{p} + \Delta\boldsymbol{p}, \end{aligned} \tag{4}$$

where $\boldsymbol{J} = [\boldsymbol{J}_1; ...; \boldsymbol{J}_L]$ is the Jacobian of PDM, $\boldsymbol{v} = [\boldsymbol{v}_1; ...; \boldsymbol{v}_L]$ is the concatenation of the mean shift vectors of each landmark:

$$\boldsymbol{v}_l = (\sum_{\boldsymbol{y} \in \boldsymbol{\Psi}_l} \frac{p(g_l = 1|\boldsymbol{y}, \boldsymbol{I}) f_{\mathcal{N}}(\boldsymbol{x}_l; \boldsymbol{y}, \rho\boldsymbol{E})}{\sum_{\boldsymbol{z} \in \boldsymbol{\Psi}_l} p(g_l = 1|\boldsymbol{z}, \boldsymbol{I}) f_{\mathcal{N}}(\boldsymbol{x}_l; \boldsymbol{z}, \rho\boldsymbol{E})} \boldsymbol{y}) - \boldsymbol{x}_l, \tag{5}$$

where ρ is a free parameter denoting the variance of the PCA reconstructed noise, $\boldsymbol{E}$ denotes the identity matrix, $\boldsymbol{x}_l$ is the currently estimated position of the l-th landmark, $\boldsymbol{\Psi}_l$ denotes all integer pixels in the l-th cropped image patch centered at $\boldsymbol{x}_l$. For a detailed derivation, the interested reader is referred to (Saragih, Lucey, and Cohn 2011). Then, the mean shifts are calculated and the update of the PDM parameter $\Delta\boldsymbol{p}$ is computed iteratively until convergence.

Soft Facial Landmark Detection by LDL

Given a face Image $\boldsymbol{I}$, the purpose of facial landmark detection is to estimate a shape $\boldsymbol{s}$, which is as close as possible to the ground truth shape $\boldsymbol{s}^*$. Formally, a 2D face shape $\boldsymbol{s} = [\boldsymbol{x}_1; ...; \boldsymbol{x}_L]^T$ consists of L facial landmarks, where $\boldsymbol{x}_l = [x_l, y_l]$ denotes the coordinate of the l-th landmark.

Multi-scale Cascaded BLD Regression

In this paper, we propose an architecture named Multi-scale Cascaded BLD Regression (MCBR), as illustrated in Fig. 2. Given a facial image, our method starts from a low resolution image with an initial estimated shape $\boldsymbol{s}^0 = [\boldsymbol{x}_1^0, ..., \boldsymbol{x}_L^0]^T$, to recover the ground truth shapes $\boldsymbol{s}^*$ progressively. In our implementation, we first conduct the BLD regression at the low resolution layer. At each later layer, we double the image resolution and conduct the BLD regression stage by stage (e.g., we perform the BLD regression twice in the later experiments), each stage with an updated shape. As shown in Fig. 2, the cropped image patch (the blue rectangle) of the same size for the annotated landmark at the lower resolution face covers more context information and constrains the BLD in a larger search region. On the other hand, the cropped image patch of the same size at the higher resolution face then constrains the BLD within a small region, which leads to finer adjustments. Thus, adopting the multi-scale strategy can accelerate the shape convergence, meanwhile avoid the trapping in local optimum.

Our training is conducted at m different resolution layers, each layer has n stages. Therefore, there are totally $T = m \times n$ stages in our training. The optimization target of each stage is to learn the mappings from an image patch to the BLD of each landmark independently. For example, during the t-th training stage, firstly, we crop an image patch (the blue rectangle in Fig. 2) centered at each currently estimated landmark $\boldsymbol{x}^t$ (the white point). Then, based on the annotated shapes $\hat{\boldsymbol{s}}$ (the green points), we can obtain the BLD for each image patch. The mapping parameters $\boldsymbol{\Theta}^t$ are optimized to generate a predicted BLD most similar to the true BLD.

During the test process, given an unseen image, we start from a coarse initial shape $\boldsymbol{s}^0$, and crop the image patch centered at the currently estimated shape. Then we use the trained mapping parameter matrix $\boldsymbol{\Theta}$ at this stage to obtain the predicted BLD for each image patch. These predicted BLDs are used in a deformable model fitting process to obtain the refined predicted facial shape, which will be used as an input shape for the next stage.

Training of Our Model

Bivariate Label Distribution Initialization At the t-th training stage of the Multi-scale Cascaded BLD Regression (MCBR), the first step is to initialize the BLD for each landmark independently. Assume the currently estimated shape is $\boldsymbol{s}^t = [\boldsymbol{x}_1^t, ..., \boldsymbol{x}_L^t]^T$. If we crop an image patch centered at the l-th estimated landmark $\boldsymbol{x}_l^t$, then, the label space $\mathcal{Y}_l = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_C\}$ is obtained for the l-th landmark, where $\boldsymbol{y}_c$ represents the c-th pixel in the cropped image patch, i.e., a label in the label space, C is the number of all pixels in the patch. The BLD at the t-th stage for the l-th landmark of the image $\boldsymbol{I}$ is defined as a vector $\boldsymbol{d}_{l,\boldsymbol{I}}^t$, which contains the description degrees of all labels $\boldsymbol{y}_c$ in $\mathcal{Y}_l$. Suppose the description degree of a point $\boldsymbol{y}_c$ in the cropped space to the image $\boldsymbol{I}$ is represented by $d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t$, and the l-th annotated point is $\hat{\boldsymbol{x}}_l$, then, $d_{l,\boldsymbol{I},\hat{\boldsymbol{x}}_l}^t$ should be the highest among all possible pixels in the l-th cropped label space. The description degree $d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t$ decreases with the increase of the distance between $\boldsymbol{y}_c$ and $\hat{\boldsymbol{x}}_l$, i.e., the farther a point $\boldsymbol{y}_c$ is away from $\hat{\boldsymbol{x}}_l$, the lower $d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t$ is. The desired bivariate facial landmark label distribution should satisfy two criteria. First is $d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t \in [0, 1]$, and the second is $\sum_{\boldsymbol{y}_c \in \mathcal{Y}_l} d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t = 1$.

In order to generate a reasonable BLD for the l-th facial landmark of the image $\boldsymbol{I}$, one way is to use a discretized bivariate Gaussian distribution $\mathcal{N}(\boldsymbol{y}_c; \hat{\boldsymbol{x}}_l, \boldsymbol{\Sigma})$ centered at the l-th annotated landmark $\hat{\boldsymbol{x}}_l$, i.e.,

$$d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}Z} exp(-\frac{1}{2}(\boldsymbol{y}_c - \hat{\boldsymbol{x}}_l)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_c - \hat{\boldsymbol{x}}_l)), \tag{6}$$

where $\boldsymbol{\Sigma}$ is a 2×2 covariance matrix, Z is a normalization factor that makes sure $\sum_{\boldsymbol{y}_c} d_{l,\boldsymbol{I},\boldsymbol{y}_c}^t = 1$.

Fig. 2 shows some examples of the BLDs. Blue rectangles are the cropped image patches centered at the currently estimated landmarks (the white points). Then in the cropped label space, the annotated landmark (the green point) has the highest description degree, neighboring pixels around the annotated landmark have a lower degree. The description degrees of the points far away from the annotated landmark are nearly zero, which displays black color in the BLD of Fig. 2.

After generating the BLD for each landmark, the training set at the t-th stage becomes $G^t = \{G_1^t, G_2^t, ..., G_L^t\}$, where $G_l^t = \{(\boldsymbol{I}_1, \boldsymbol{d}_{l,\boldsymbol{I}_1}^t), ..., (\boldsymbol{I}_N, \boldsymbol{d}_{l,\boldsymbol{I}_N}^t)\}$ is the training set for the l-th landmark, N is the number of the total training images.

Bivariate Label Distribution Learning The description degree $d_{l,\boldsymbol{I},\boldsymbol{y}}$ can be represented by the form of conditional probability, i.e., $d_{l,\boldsymbol{I},\boldsymbol{y}} = p_l(\boldsymbol{y}|\boldsymbol{I})$. It can be explained as that the probability of $\boldsymbol{y}$ equals to its description degree. Assume $p_l(\boldsymbol{y}|\boldsymbol{I})$ to be a parametric model $p_l(\boldsymbol{y}|\boldsymbol{I}; \boldsymbol{\Theta}_l)$, where $\boldsymbol{\Theta}_l \in \mathbb{R}^{D \times C}$ is the parametric matrix for the l-th landmark, D is the dimensions of image features extracted from the cropped patch, and C is the number of all pixels in the cropped label space $\mathcal{Y}_l$. Our target is the optimization of the parameter matrix $\boldsymbol{\Theta}_l$. This problem matches the Label Distribution Learning (Geng 2016). There are many criteria to measure the similarity between the ground truth and predicted distributions. If Kullback-Leibler (KL) divergence is used to measure the distance between the true and predicted BLD, then, the best parameter $\boldsymbol{\Theta}_l^t$ at the t-th stage is determined by

$$\begin{aligned} \boldsymbol{\Theta}_l^t &= \arg\min_{\boldsymbol{\Theta}} \sum_{i,c} (d_{l,\boldsymbol{I}_i,\boldsymbol{y}_c}^t ln(\frac{d_{l,\boldsymbol{I}_i,\boldsymbol{y}_c}^t}{p_l(\boldsymbol{y}_c|\boldsymbol{I}_i; \boldsymbol{\Theta}_l)})) \\ &= \arg\max_{\boldsymbol{\Theta}} \sum_{i,c} (d_{l,\boldsymbol{I}_i,\boldsymbol{y}_c}^t ln(p_l(\boldsymbol{y}_c|\boldsymbol{I}_i; \boldsymbol{\Theta}_l))). \end{aligned} \tag{7}$$

As to the form of $p_l(\boldsymbol{y}_c|\boldsymbol{I}_i;\boldsymbol{\Theta}_l)$, similar to (Geng 2016), we use the maximum entropy model to embody the mapping from the image patch to the corresponding BLD:

$$p_l(\boldsymbol{y}_c|\boldsymbol{I}_i;\boldsymbol{\Theta}_l) = \frac{1}{\Gamma_i} exp(\sum_r \boldsymbol{\Theta}_{c,r}\boldsymbol{\varphi}_i^r), \tag{8}$$

where $\Gamma_i = \sum_c exp(\sum_r \boldsymbol{\Theta}_{c,r}\boldsymbol{\varphi}_i^r)$ is the normalization factor, $\boldsymbol{\varphi}_i^r$ is the r-th element in the image features $\boldsymbol{\varphi}_i$, $\boldsymbol{\Theta}_{c,r}$ is an element in $\boldsymbol{\Theta}_l$ corresponding to the label $\boldsymbol{y}_c$ and the r-th image feature. For the image features $\boldsymbol{\varphi}$, we apply multi-scale HOG features to the cropped image patches centered at each currently estimated landmark, and then concatenate all the features into a long vector.

Substituting Equation (8) to (7) yields:

$$\begin{aligned}\boldsymbol{\Theta}_l^t = \arg\max_{\boldsymbol{\Theta}} &\sum_{i,c} d_{l,\boldsymbol{I}_i,\boldsymbol{y}_c}^t \sum_r \boldsymbol{\Theta}_{c,r}\boldsymbol{\varphi}_i^r \\ &-\sum_i ln \sum_c exp(\sum_r \boldsymbol{\Theta}_{c,r}\boldsymbol{\varphi}_i^r).\end{aligned} \tag{9}$$

The limited-memory quasi-Newton method L-BFGS (Liu and Nocedal 1989) is used to optimize Equation (9). After obtaining the optimal parameter $\boldsymbol{\Theta}_l$, given an unseen image $\boldsymbol{I}$, we can have the predicted BLDs in the cropped patch for each landmark by $p_l(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\Theta}_l)$, which are used in a deformable fitting process to obtain the predicted facial shapes.

Test of Our Model

At the t-th test stage, our target is to use the predicted BLD for each estimated landmark to refine the currently estimated shapes, with the help of the deformable model fitting approach. For an unseen image $\boldsymbol{I}$, first, we cropped the image patch $\boldsymbol{\Psi}_l$ centered at each currently estimated landmark $\boldsymbol{x}_l^t$, then, we extracted image features from the patch, and use $p_l(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\Theta}_l^t)$ to predict the BLD for the l-th landmark. The predicted BLDs are used to calculate the mean shifts of each landmark, i.e., substitute $p_l(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\Theta}_l^t)$ to Equation (5) yields

$$\boldsymbol{v}_l = (\sum_{\boldsymbol{y}\in\boldsymbol{\Psi}_l} \frac{p_l(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\Theta}_l^t) f_{\mathcal{N}}(\boldsymbol{x}_l^t;\boldsymbol{y},\rho\boldsymbol{E})}{\sum_{\boldsymbol{z}\in\boldsymbol{\Psi}_l} p_l(\boldsymbol{z}|\boldsymbol{I};\boldsymbol{\Theta}_l^t) f_{\mathcal{N}}(\boldsymbol{x}_l^t;\boldsymbol{z},\rho\boldsymbol{E})}\boldsymbol{y}) - \boldsymbol{x}_l^t. \tag{10}$$

Using Equation (8), (10) and (4) iteratively to update $\boldsymbol{p}$ until convergence. Then, we have the updated PDM parameter $\boldsymbol{p}^t = \boldsymbol{p}^{t-1} + \Delta\boldsymbol{p}$ at the t-th stage, simultaneously obtaining refined estimated shape $\boldsymbol{s}^t = [\boldsymbol{x}_1^t, ..., \boldsymbol{x}_L^t]$ using Equation (1). $\boldsymbol{s}^t$ is then sent to the next cascaded stage, until t is not less than T.

Experiments

There are three parts in our experiments. The first experiment compares the accuracy of our proposed MCBR method with respect to other baselines based on the CLM framework. Second, we will test the performance of our proposed MCBR method, compared with the state-of-the-art methods. Finally, we will test the robustness against the increase of the annotated landmark noise.

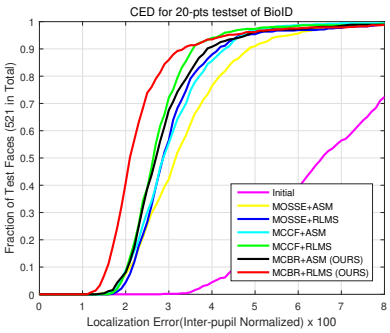


Figure 3: Cumulative Error Distributions over 20 landmarks on the BioID database (Jesorsky, Kirchberg, and Frischholz 2001).

Comparison with CLMs

We perform an experiment to see how our proposed MCBR method outperforms other CLM methods. The experiment is performed on the BioID database (Jesorsky, Kirchberg, and Frischholz 2001), consisting of 1521 frontal and close to frontal images with 20 landmarks. 1000 images are randomly selected for the training and the rest 521 images are used for the test. In this experiment, two effective and popular patch experts are evaluated: MOSSE (Bolme et al. 2010) and MCCF (Galoogahi, Sim, and Lucey 2013) filters. During the training of the MOSSE and MCCF filters, each aligned patch sample is represented using DSIFT, and then requires a normalization step, and finally it is multiplied by a cosine window. For our proposed MCBR method, the number of iterations in L-BFGS is set to 60. All methods are conducted at $m = 3$ resolution layers (e.g., $64, 128, 256$), and each layer contains $n = 1$ stage. For all methods, the size of local patches and desired output is set to 21×21, and the standard deviation in the desired output is set to 2. The deformable model fitting approach that best evaluates the local patch experts is the method that relies the most on the output of the patch experts, i.e., the Active Shape Model (ASM) (Cootes et al. 1995). The other fitting approach used is the RLMS (Saragih, Lucey, and Cohn 2011) method. The average run time for our proposed MCBR method on an Intel Core i7 2.20-GHz machine is 140 ms per image with 20 landmarks.

The Cumulative Error Distributions (CED) for this experiment is presented in Fig. 3, which shows the percentage of faces that achieved a given inter-pupil distance normalized landmark error (Ren et al. 2014) amount. The presented Initial curve represents the initial estimate, provided by the mean shape. This experiment shows that our proposed MCBR method always outperforms the others when using the same fitting approach, and maximum performance can be achieved by using our proposed MCBR method and the RLMS fitting approach.

Comparison with state of the art

Dataset Evaluations are conducted on the 300-W dataset (Sagonas et al. 2013a), which is a well known database with

Table 1: The inter-pupil distance normalized landmark error (%) of the compared methods on the 300-W (Sagonas et al. 2013a) dataset.

Method	Full Set	Common SubSet	Challenging SubSet
Zhu *et al.*	10.20	8.22	18.33
DRMF	9.22	6.65	19.79
MOSSE+RLMS *	8.48	6.89	15.00
GN-DPM	-	5.78	-
RCPR	8.35	6.18	17.26
CFAN	7.69	5.50	16.78
ESR	7.58	5.28	17.00
MCCF+RLMS *	7.39	5.99	13.11
SDM *	6.67	5.39	11.93
LBF Fast *	6.61	5.24	12.25
MDM *	6.35	5.23	**10.95**
MCBR (ours)	**6.33**	**5.07**	11.50

68 annotated facial points for robustness evaluation of facial landmark detection algorithms. Following the same dataset configuration as in (Ren et al. 2014) , our training set consists of the training set of LFPW (Belhumeur et al. 2013) and Helen (Le et al. 2012), the whole AFW set (Zhu and Ramanan 2012), totally 3148 training images. Our test set consists of the test set of LFPW and Helen, and the whole IBUG set, totally 689 test images.

Implementation Details In this experiment, we conduct our proposed MCBR method at m = 3 different resolution layers, each layer contains n = 2 stages. The standard deviation in Equation (6) to compute the BLDs is set to 2. And the size of cropped patch is set to 31×31. The number of iterations in L-BFGS is set to 66.

To provide a better initial shape for an image, we divide the training set into three view-specific subsets, i.e., left $(-30°, -0°)$, frontal $(-15°, 15°)$ and right $(0°, 30°)$. The overlaps between adjacent views are considered for fault tolerance. We decide which view range a face image belongs to by the distance between the pose of this image and the central poses of the three subsets, i.e., $-15°$, $0°$, $15°$ [1]. Then, we assign the mean shape of the corresponding view subsets as an initial shape for an image during the test phase. The average run time for our proposed MCBR method using unoptimized Python implementations on an Intel Core i7 2.20-GHz machine is 800 ms per image with 68 landmarks.

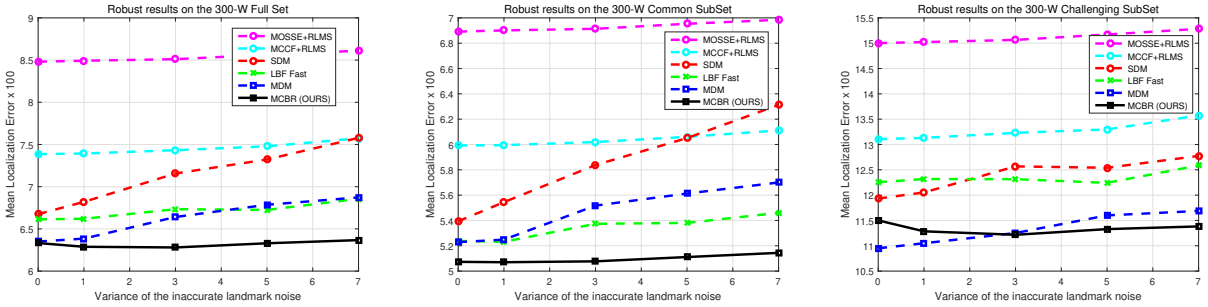
During training, we use data augmentation to enlarge the training data. For each training image, we randomly select the shape of other training images in the same view-specific subset as the initial shape four times. In this way, we generate 4 perturbed initial shapes for each training image. During test, we only use the view-specific mean shape as the initial shape without multiple initializations. Note that we only use view-specific subsets to provide better initial shapes, rather than training our method separately in each view.

[1]We apply the open source pose estimator to the 300-W dataset: https://github.com/mpatacchiola/deepgaze

Baselines We choose several existing state-of-the-art facial landmark detection algorithms for comparison, including Zhu *et al.* (Zhu and Ramanan 2012), DMRF (Asthana et al. 2013), GN-DPM (Tzimiropoulos and Pantic 2014), RCPR (Burgos-Artizzu, Perona, and Dollár 2013), CFAN (Zhang et al. 2014), ESR (Cao et al. 2014), SDM (Xiong and De la Torre 2013), LBF Fast (Ren et al. 2014) and MDM (Trigeorgis et al. 2016). Also, we add two best CLM methods, i.e., MOSSE+RLMS and MCCF+RLMS. The mean landmark errors (Ren et al. 2014) of different methods are reported in Table 1. The results of methods marked with * are obtained by our implementation, and others are directly obtained from the corresponding papers. It is worth noting that we only sample initial shapes for each training image 4 times to augment the training set, which is far less than the amount of training data for other trained models mentioned in their paper. MOSSE+RLMS, MCCF+RLMS, SDM and LBF Fast are all conducted in the Multi-scale Cascaded manner. The size of local patches in MOSSE/MCCF+RLMS is set to 31×31. SDM uses the same HOG descriptors as ours. For LBF Fast, we set the number of trees at each stage to $68 \times 6 = 408$, and each tree depth is set to 5. The radiuses of two stages at each resolution layer are set to [0.3, 0.2], and the number of the randomly sampled candidate features in the local region is set to 500. For MDM, we set all parameters the same as described in their paper. Bounding boxes provided by 300-W set are used for all implemented methods. Initial shapes for the test images are set the same for all implemented methods.

Results We compare our method with the baseline methods in Table 1. The MOSSE/MCCF+RLMS method performs worst in all implemented methods. MDM adopts the powerful CNN to learn the data-driven features and RNN to impose the memory constraint on the descent directions, so it performs better than LBF Fast and SDM on the Full Set. LBF Fast performs better than SDM, which benefits from its highly discriminative local binary features. While all baseline methods do not consider the issue of the inaccurate annotated landmarks, our method uses soft landmarks, i.e., BLD, to deal with the landmark noise, and thus achieve the best performance on the Full Set.

To further analyze, as the previous work (Ren et al. 2014) did, we split the full test set into two parts. First is the 554 face images from the test set of LFPW and Helen, which is called Common Subset. Second is the 135 face images from the whole IBUG, which is called Challenging Subset. It is worth noticing that IBUG dataset is extremely challenging as its face images have large variations in head poses, facial expression, illumination, *etc*. For the Common Subset, our method shows a greater superiority to the baseline methods. For the Challenging Subset, since MDM uses the end-to-end CNN features in the shape regression, which is much effective than the hand-crafted features for the quite challenging cases, it performs better than our method. However, our method still performs better than other compared methods.



(a) Full Set (b) Common Subset (c) Challenging Subset

Figure 4: Results on the 300-W dataset (Sagonas et al. 2013a) when increasing the inaccurate annotated landmark noise.

Robustness against annotation noise

Our model appears robust against the increase of the inaccurate facial landmarks annotations since it associates a BLD for each landmark, which considers the neighboring pixels around the originally annotated landmark. To further demonstrate this, we design an experiment that gradually increases the annotated landmarks noise in the training set. In order to observe the performance of the algorithm, we keep the test set unchanged while the noise of the training landmarks gradually increases.

For the training images in the 300-W database, we impose annotation noise on the original landmarks. Annotation noise can be modeled as Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\rho}^2)$, which can be explained as the deviation pixel distance from the original landmark. $\boldsymbol{\rho}^2$ denotes the variance of the noise, which reflects the inaccuracy of the annotated landmarks. In our experiments, we conduct 5 different variances of the noise, i.e., $\boldsymbol{\rho}^2$ = [0, 1, 3, 5, 7]. We choose five compared methods from Table 1, the top three baseline algorithms, i.e., SDM (Xiong and De la Torre 2013), LBF Fast (Ren et al. 2014) and MDM (Trigeorgis et al. 2016), and two CLM method, i.e., MOSSE/MCCF+RLMS. The implementation details of all algorithms are set to the same as mentioned above.

The performances of different algorithms against the increase of the landmark noise are shown in Fig. 4. Our model deals with facial landmark detection well not only in the value of the averaged landmark error, but also the robustness against more and more training annotated landmark noise. Generally speaking, for the Full Set in Fig. 4(a), although MOSSE/MCCF+RLMS does not seem to be sensitive to the landmark noise, their averaged landmark error is too high, which makes no sense. SDM deteriorates quickly with the increase of the inaccurate landmark noise. Compared with SDM, LBF Fast considers the shape constraints between the landmarks, the curve of it shows a relatively gentler tendency. Since MDM adopts effective convolutional features, it starts from a lower mean landmark error than LBF Fast. However, the data-driven MDM deteriorates faster than LBF Fast, revealing that the performance of MDM is sensitive to the landmark noise. Furthermore, our method deteriorates most slowly with the increase of the landmark noise.

For the Common Subset shown in Fig. 4(b), our method shows the most slowest rising trend among all compared baselines against the increased landmark noise. For the Challenging SubSet shown in Fig. 4(c), since hand-crafted features are sub-optimal compared with convolutional features, our method starts from a higher position than MDM. However, with the increased landmark noise, our method gradually performs better than MDM. Note that there exist a phenomenon that some methods achieve a slightly better performance when increasing the landmark noise. For example, in Fig. 4(c), our method shows a slight improvement in the experiments of noise 1 than in noise 0. The reason might be that the performances in the Challenging Subset are sensitive to the initial shapes, and when increasing landmark noise, the initial shapes calculated from the training set for the test images also changed, which may cause experimental random improvements.

In summary, performance curves of our method on the different test subsets are rather stable than all compared baselines, which validates the robustness of our algorithm against the increased training landmark noise.

Conclusion and Future Work

This paper is motivated by the inaccurate manually annotated facial landmarks. Towards this, we propose a soft facial landmark detection algorithm by Label Distribution Learning. By associating a bivariate label distribution (BLD) to each landmark of an image, we consider the neighboring pixels around the original manually annotated landmark, which can alleviate the effects of the inaccurate landmarks. By minimizing the Kullback-Leibler (KL) divergence between the true and predicted BLD, we can obtain mapping functions from an image patch to the BLD of each landmark, which are used to generate the predicted BLDs for the landmarks of unseen images. Then, these BLDs are used in a deformable model fitting process to achieve the final facial shape. Experimental results show that our proposed MCBR method performs better than compared state-of-the-art algorithms and appears more robust against the increase of inaccurate landmark noise than compared baselines. As the results of the Challenging Subset in Table 1 and Fig. 4(c) shown, in the future work, we will adopt powerful end-to-end features in our method to achieve greater performance.

Acknowledgments

This research was supported by the National Key Research & Development Plan of China (No. 2017YFB1002801), the National Science Foundation of China (61622203), the Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Collaborative Innovation Center of Wireless Communications Technology.

References

Asthana, A.; Zafeiriou, S.; Cheng, S.; and Pantic, M. 2013. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3444–3451.

Belhumeur, P. N.; Jacobs, D. W.; Kriegman, D. J.; and Kumar, N. 2013. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence* 35(12):2930–2940.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2544–2550. IEEE.

Burgos-Artizzu, X. P.; Perona, P.; and Dollár, P. 2013. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 1513–1520.

Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *International Journal of Computer Vision* 107(2):177–190.

Cootes, T. F.; Taylor, C. J.; Cooper, D. H.; and Graham, J. 1995. Active shape models-their training and application. *Computer vision and image understanding* 61(1):38–59.

Cristinacce, D., and Cootes, T. 2008. Automatic feature localisation with constrained local models. *Pattern Recognition* 41(10):3054–3067.

Galoogahi, H. K.; Sim, T.; and Lucey, S. 2013. Multi-channel correlation filters. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 3072–3079. IEEE.

Geng, X., and Ling, M. 2017. Soft video parsing by label distribution learning. In *AAAI*, 1331–1337.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(10):2401–2412.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748.

Jesorsky, O.; Kirchberg, K. J.; and Frischholz, R. W. 2001. Robust face detection using the hausdorff distance. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, 90–95. Springer.

Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; and Huang, T. S. 2012. Interactive facial feature localization. In *European Conference on Computer Vision*, 679–692. Springer.

Liu, D. C., and Nocedal, J. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming* 45(1):503–528.

Matthews, I., and Baker, S. 2004. Active appearance models revisited. *International journal of computer vision* 60(2):135–164.

Ren, S.; Cao, X.; Wei, Y.; and Sun, J. 2014. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1685–1692.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013a. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 397–403.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013b. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 896–903.

Saragih, J. M.; Lucey, S.; and Cohn, J. F. 2011. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision* 91(2):200–215.

Trigeorgis, G.; Snape, P.; Nicolaou, M. A.; Antonakos, E.; and Zafeiriou, S. 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4177–4187.

Tsoumakas, G., and Katakis, I. 2006. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3(3).

Tzimiropoulos, G., and Pantic, M. 2014. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1851–1858.

Xiong, X., and De la Torre, F. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 532–539.

Zhang, J.; Shan, S.; Kan, M.; and Chen, X. 2014. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, 1–16. Springer.

Zhao, W.; Chellappa, R.; Phillips, P. J.; and Rosenfeld, A. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35(4):399–458.

Zhou, Y.; Xue, H.; and Geng, X. 2015. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1247–1250. ACM.

Zhu, X., and Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2879–2886. IEEE.