EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

SHELF: Combination of Shape Fitting and Heatmap Regression for Landmark Detection in Human Face

Ngo Thi Ngoc Quyen[1], Tran Duy Linh [1], Vu Hong Phuc [1], Nguyen Van Nam[2,3,*]

[1]Viettel Cyberspace Center (VTCC), Viettel Group, 7 Alley, TonThatThuyet Street, CauGiay district, Hanoi, Vietnam
[2]Viettel Information of Technology Department (VITD), Viettel Group, D26, CauGiay New City area, 7 Alley, TonThatThuyet Street, CauGiay district, Hanoi, Vietnam
[3]Thuyloi University, 175 TaySon street, DongDa district, Hanoi, Vietnam

Abstract

Today, facial emotion recognition is widely adopted in many intelligent applications including the driver monitoring system, the smart customer care as well as the e-learning system. In fact, the human emotions can be well represented by facial landmarks which are hard to be detected from images, due to the high number of discrete landmarks, the variation of shapes and poses of the human face in real world. Over decades, many methods have been proposed for facial landmark detection including the shape fitting, the coordinate regression such as ASMNet and AnchorFace. However, their performance is still limited for real-time applications in terms of both accuracy and efficiency. In this paper, we propose a novel method called SHELF which is the first to combine the shape fitting and heatmap regression approaches for landmark detection in human face. The heatmap model aims to generate the landmarks that fit to the common shapes. The method has been evaluated on three datasets 300W-Challenging, WFLW, 300VW-E with 31557 images and achieved a normalized mean error (NME) of 6.67% , 7.34%, 12.55% correspondingly, which overcomes most existing methods. For the first two datasets, the method is also comparable to the state of the art AnchorFace with a NME of 6.19%, 4.62%, respectively.

Received on 10 September 2023; accepted on 23 September 2023; published on 26 September 2023

Keywords: facial landmarks, heatmap regression, shape fitting, coordination regression

Copyright © 2023 N. T. N. Quyen *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi:10.4108/eetinis.v10i3.3863

1. Introduction

Recent years, given the higher and higher demands in intelligent applications for human monitoring, the recognition of facial expressions from images becomes active research field in literature. With the advancements of deep learning, this can be formulated as an image classification problem which is addressed by a state of the art model such as ResNet[2], EfficientNet[3], MobileNetV2[4], ShuffleNetV2[5], VisionTransformer[6] with a large dataset of training images. However, these models are not efficient because many pixels in the image do not contribute to the facial emotions.

In fact, a more promising approach for facial emotion recognition is based on the detection of facial landmarks which is a group of important pixels locating around the eyes, the nose, the mouth and the boundary of the face as shown in Fig. 1. The facial emotions can be clearly recognized by only a small ensemble of landmarks if they are correctly detected. The task of facial landmark detection is to locate these points in a given face image as depicted in Fig. 2 [1]. This problem is very challenging due to the variation of facial appearance, the high the number of discrete

Corresponding author. Email: nvnam@tlu.edu.vn
*

[1]The original image is referred from https://www.dreamstime.com/photos-images/bus-driver.html

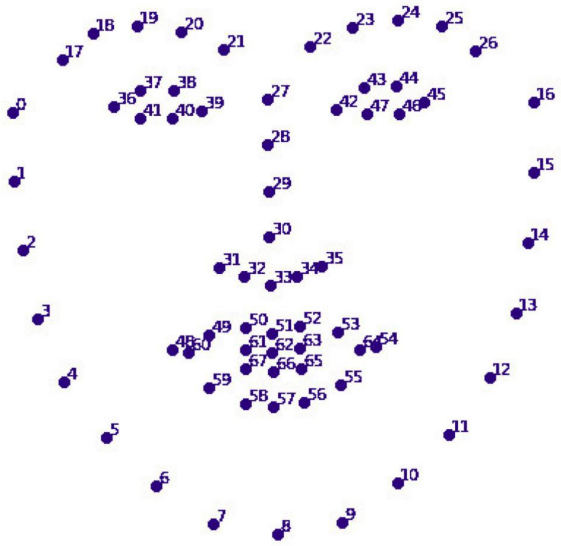


Figure 1. A facial image with 68 landmarks

Figure 2. Detection of facial landmark on an image: The left one is of a bus driver. The middle one denotes his angry face associated with 68 landmarks. The right one describes the resulted locations of the landmarks on the angry face.

landmarks as well as the complexity of facial shapes and poses.

Before, the facial landmarks are detected by shape fitting models which are composed of Active Shape Model (ASM)[7], Active Appearance Models (AAM)[8], Constrained Local Model (CLM)[9], Discriminative Response Map Fitting (DRMF)[10] as well as DeFA [11]. The models do not aware of facial appearance. In such cases, the landmarks are regressed from common facial shapes in a given dataset. These models are fast converged but usually under-fitted due to the high variation of facial shapes in the real world.

Recently, the regression networks are preferred thanks to advancements in convolutional neural networks (CNN). These include a CNN backbone for feature engineering and a regression head. For the direct regression methods like Style Aggregation Network (SAN) [12], the landmarks are directly regressed from the aggregated feature map of the CNN backbone. The feature map focus more on the facial appearance but less on the shape and pose. However, in case of faces that are occluded or of large pose, without appearance, certain landmarks may not be detected.

In heatmap regression models such as MobileFAN [13], the CNN backbone is replaced by a fully CNN (FCN) with additional de-convolutional layers. FCN is really an auto-encoder which encodes the face image and decodes to a corresponding heatmap highlighting the landmarks. Better than the direct regression methods, these models can reconstruct the facial landmarks in case of occlusion and of large pose. However, similar to generative models, these are less converged and suffered from the hallucination issue.

In this paper, we propose a novel method called **SHELF** which appropriately combines the **S**hape fitting and the **HE**atmap regression approaches for detection of **L**andmarks in human **F**ace. This is because their trade-off can compensate to each other. Our main contributions are therefore four-fold as follows:

- A heatmap generation neural network is built using a CNN with additional de-convolutional layers.
- A regression head is designed for determining the landmark with the highest probability using a softmax-argmax layer. Then, the shape fitting loss and the heatmap regression loss are combined in an efficient manner.
- A large dataset called 300VW-E of 31757 facial images, each labelled with 20 landmarks, has been prepared for recognition of emotions in human face. This is an extension of the 300VW public dataset.
- An evaluation of SHELF is effectuated on three datasets consisting of 300W-Challenging, WFLW and 300VW-E. The method achieved a low normalized mean error (NME) of 6.67%, 7.34% and 12.55%, respectively. These results outperform existing methods such as DeFA, SAN, MobileFAN, ASMNet and CFSS on all three datasets. SHELF is less performant than the state of the art AnchorFace[14] due to using less number of anchor shapes.

The rest of the paper is organized as follows. In Section 2, we introduce different approaches for facial landmark detection. Then, the proposed method SHELF is presented in detail in Section 3. We summarize the experimental results of SHELF on three datasets and study the ablation of SHELF in Section 4. Finally, Section 5 concludes our works.

2. Related Works

Over decades, many different approaches for facial landmark detection have been proposed. In this section,

we introduce state of the art methods relating to the shape fitting as well as the regression of landmarks.

2.1. Shape Fitting

Traditional template matching approaches such as ASM [7], AAM [8], CLM[9] and DeFA [11] detect the facial landmarks by learning their common distribution and from a mean shape, computed from certain active samples, regressing them. ASM is based on the dimension reduction method Principle Component Analysis (PCA) [15] for shape fitting. AAM improved the performance of ASM by combining both the shape and appearance models in iterative manner. CLM introduced another appearance sampling technique in which the pixel values in the texture patches are normalized with zero mean and unit variance. Using CNN, DeFA models the facial shape in 3D to not only aligns facial landmarks but also matches SIFT (Scale-Invariant Feature Transform) points as well as the facial contours. However, due to limited feature engineering, the performance of such approaches are limited especially in case of occluded face images.

2.2. Landmark Regression

As introduced, the neural networks for facial landmark detection usually include a CNN backbone and a regression head which is fed with a feature vector. The networks can be categorized as coordinate and heatmap regression according to the way such vector is built from the backbone.

Coordinate Regression. In case of coordinate regression networks, any CNN encoder can be used as their backbone. The regression head is directly fed with the flattened feature embedding of the backbone. Mnemonic Descent Method (MDM) [16] is a combined convolutional recurrent neural network which aims to cooperate the regressors of facial landmarks. DeepReg[17] is a deep regressor for gradual detection of facial landmarks with two-stage initialisation. In Wing [18], the wing regression loss was proposed for landmark localization rather than the L1 and L2 losses thanks to its ability to help the regression networks not only deal with large localization errors as L1 and L2, but treat also well the medium and small localization ones. Wing has been experimented with Resnet-50[2] backbone. However, such average loss for regression of a high number of positions on the whole face is unable to assure small prediction errors for individual landmarks.

Heatmap Regression. The heatmap regression networks such as AWing[19], MobileFAN [13], Gaussian Vector (GV)[20] and AdNet[21] are autoencoder backbone which is composed of a CNN encoder and a decoder to produce probability distributions in form of heatmaps corresponding to the facial landmarks. In each heatmap, the position with the highest probability is chosen for the respective landmark.

AWing proposed an adaptive Wing loss function for coordinate regression from facial boundary map for better conforming the heatmap pixels to the facial shape. Gaussian Vector (GV) converts heatmap in to a pair of vector for each landmark to preserve spacial information and simplify the post-processing. AdNet introduced anisotropic direction loss and anisotropic attention module for better learning the facial structure as well as the texture details and mitigating the error-bias of facial landmarks.

2.3. Joint Shape Fitting and Regression Networks

There are also few methods which combine the shape fitting approach and the regression network such as LAB[22], ASMNet[23] and AnchorFace[14]. LAB is a combination of the boundary fitting and the coordinate regression. Using a stacked Hourglass network [24] as an autoencoder backbone to produce facial boundary map, LAB then regresses the coordination of facial landmarks from the boundary in order to avoid the ambiguities of such key-points. ASMNet leveraged the light-weight MobileNetV2[4] as backbone and presented a multi-task loss which is the sum of the mean square error and the active shape model loss. This enables ASMNet to learn both the shape and the coordination of the facial landmarks with less parameters than LAB.

In AnchorFace, the authors introduced certain anchor templates and regress the offsets on each template. They then aggregates the predictions on every templates to produce the final results. AnchorFace utilized ShuffleNetV2 [5] as its backbone. AnchorFace can deal with face poses of large variations thanks to its anchor templates. Nevertheless, the anchor templates need to be carefully selected and the inference time must be improved. AnchorFace is also known as anchor-based method.

Such joint approaches are usually more performant than the separate ones. However, existing joint methods are only between coordinate regression and the shape fitting. In this paper, we propose SHELF, a facial landmark detection method based on shape fitting and heatmap regression to fill the gap as well as to leverage the robustness of such combination.

3. SHELF: the proposed model

Our proposed method SHELF consists of a heatmap regression network a training loss function including both the coordination and the shape matching errors. Two principal components of the heatmap regression network are the heatmap-generated backbone and the heatmap regression head.

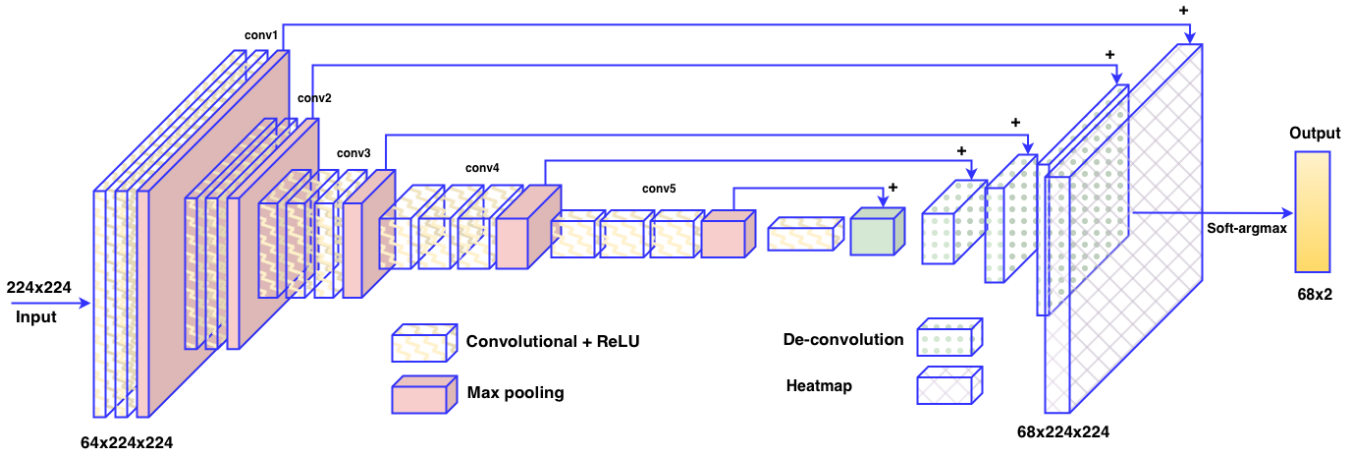


Figure 3. The network architecture of SHELF

3.1. The Heatmap-generated Backbone

As depicted in Fig.3, the backbone is an autoencoder which takes as input the face image of size 224x224 and produces a set of heatmaps. The encoder is composed of five multi-filter convolutional layers which are activated by ReLU function and dimensionally reduced by Max-Pooling. Meanwhile, the decoder is based on three deconvolutional layers to produce a set of heatmaps of the same size with the input image, each of which corresponds to a facial landmark.

3.2. The Heatmap Regression Head

Given a set of n heatmaps $H = \{H^i\}, i = \overline{1,n}$, each of size K x K (in this case K is qual to 224) and flattened to a vector of K^2 dimensions $h^i = (h^i_1, h^i_2, .., h^i_{K^2})$, the regression head of SHELF can predict the coordination for the respective facial landmarks using a soft arg-max function as follows:

$$\{\hat{x}_i, \hat{y}_i\} = softargmax_j(j \cdot f(j)) \tag{1}$$

where $f(j), j = \overline{1, K^2}$ is a probability distribution function defined as follows

$$f(j) = \frac{e^{\alpha \cdot h^i_j}}{\sum_{k=1}^{K^2} e^{\alpha \cdot h^i_k}} \tag{2}$$

in which $\alpha \geq 1$ is the temperature parameter. For the i^{th} heatmap H^i, the function $softargmax$ returns an index j^* where $f(j^*)$ is the maximal value of $\{f(j), \forall j = \overline{1, K^2}\}$. From j^*, we can calculate the coordination $(\hat{x}_i, \hat{y}_i)$ for the corresponding i^{th} facial landmark. This function can be differentiated that can be used in SHELF instead of the traditional $argmax$ and $softmax$ functions.

3.3. The Multitask Loss Function

As we aim to integrate the facial landmarks in to a given shape, we designed a multitask loss function for training our proposed network.

The Coordination Loss. The mean square error is used as the coordination loss as follows:

$$\mathcal{L}_{coord} = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \tag{3}$$

where n is the number of facial landmarks, $(x_i, y_i), (\hat{x}_i, \hat{y}_i), i = \overline{1,n}$ is the ground truth and predicted coordination of the i^{th} facial landmark, respectively.

The Shape Loss. Given a training set with m samples in which the $j^{th}, j = \overline{1,m}$ is represented as a vector of $2n$ dimensions $s^j = (x^j_1, y^j_1, x^j_2, y^j_2, ..x^j_n, y^j_n)$, using PCA (Principal Component Analysis)[7], this can be approximated by $\tilde{s}^j$ as follows:

$$\tilde{s}^j = \bar{s} + P \cdot b^j \tag{4}$$

where $\bar{s}$ is the mean shape

$$\bar{s} = \frac{1}{m}\sum_{j=1}^{m} s^j \tag{5}$$

and $P = (p_1|p_2|..|p_t)$ is a matrix constituted from t eigenvectors with the highest corresponding eigenvalues $\lambda_1, \lambda_1, .., \lambda_t$ of the following co-variance matrix:

$$S = \frac{1}{m-1}\sum_{j=1}^{m}(s^j - \bar{s})(s^j - \bar{s})^T \tag{6}$$

and b^j is a t-dimensional vector containing a set of parameters for a deformable model:

$$b^j = P^T(s^j - \bar{s}) \tag{7}$$

The shape loss is then calculated as follows

$$\mathcal{L}_{shape} = \frac{1}{2 \cdot n} \sum_{i=1}^{2n} (s_i^j - \tilde{s}_i^j)^2 \tag{8}$$

The Multitask Loss. For every training samples, the overall loss is the combination of the coordinate and the shape ones as the following

$$\mathcal{L} = \mathcal{L}_{coord} + \beta \cdot \mathcal{L}_{shape} \tag{9}$$

where β is the shape fitting rate which varies in reverse proportionally to the number of the training epochs for SHELF. This is because as many other convolutional neural networks, SHELF learns the shape before featuring the pixel-wise image. The ratio can then be defined as the following discrete function:

$$\beta = \begin{cases} 2 & \text{if } e \leq \frac{N_e}{5} \\ 1 & \text{if } \frac{N_e}{5} < e \leq 2 \cdot \frac{N_e}{5} \\ 0.5 & \text{if } 2 \cdot \frac{N_e}{5} < e \leq 3 \cdot \frac{N_e}{5} \\ 0 & \text{if } e > 3 \cdot \frac{N_e}{5} \end{cases} \tag{10}$$

where e, N_e is the current and total number of training epochs, respectively. At the initial steps of SHELF training where the shape features are important, the shape fitting rate β is also high enough. Reversely, at the final steps, β is set to zero since there exists mainly pixel featuring in the network.

4. Experiments

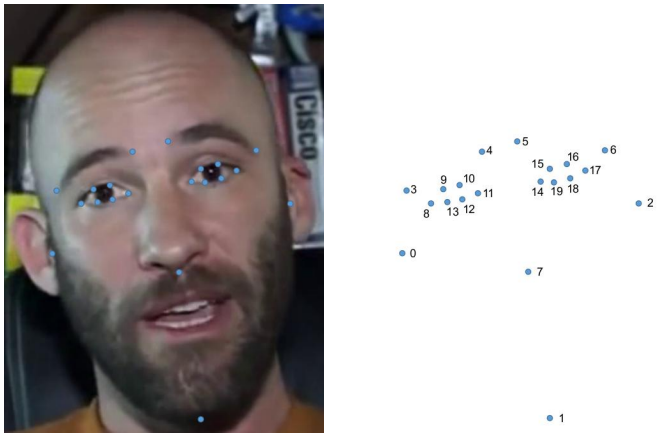
4.1. Datasets

Our proposed SHELF method is evaluated on two famous facial landmark datasets including 300W and WFLW. We also conducted experiments on our private dataset.

300W. The 300W dataset totally consists of 3837 facial images with 68 landmarks annotated. The training set includes 3148 images in which 2000 are from HELEN[25], 811 from LFPW [26] and 337 from AFW[27]. The full testing set is composed of 689 images which is divided in to a common set of 554 combining those from HELEN and LFPW and a challenging set with 135 images.

WFLW. The WFLW dataset [22] includes 10000 facial images which are annotated by 98 landmarks. Three fourths of the dataset are used for training and the rest for testing. This latter is composed of six subsets with different difficulties including 314 for expression, 326 for large pose, 206 for make-up, 736 for occlusion, 698 for illumination and 773 for blurring.

300VW-E. Our private dataset called 300VW-E include 31757 facial images which are extracted from videos in 300VW dataset[2] as well as from our driver-monitoring camera in the real world.



a) A facial image b)20 landmarks

Figure 4. A facial image and its corresponding 20 landmarks in the 300VW-E dataset.

These images are then annotated with only 20 landmarks locating mostly on the eyes of a human face as depicted in Fig. 4. This aims to clearly flash the facial emotions such as sleepy, tired, scared or distracted for DMS. Nearly 80% of these images are used for training, about 15% for validation and the rest for testing.

4.2. Evaluation Metrics

As commonly used for benchmarking of facial landmark detection methods, we also adopt the **normalized mean error (NME)** to evaluate our proposed method SHELF as follows:

$$NME = \frac{1}{N} \sum_{i=1}^{N} \frac{\frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_j^i - \hat{x}_j^i)^2 + (y_j^i - \hat{y}_j^i)^2}}{d} \tag{11}$$

where n is the number of landmarks, N is the number of images in the testing set, $(x_j^i, y_j^i), (\hat{x}_j^i, \hat{y}_j^i)$ correspond to the ground truth and predicted coordination of the j^{th} landmark on the i^{th} facial image of the testing set and d is the distance between the two outer eye corners (inter-ocular) specifically for each dataset. This is also the normalized factor used in the 300W and WFLW datasets.

The **failure rate (FR)** is also involved in this case to evaluate the robustness of the methods in term of NME. This indicates the rate of failed recognition in which NME is less than 10%. The smaller FR is, the more powerful the model is.

[2] https://ibug.doc.ic.ac.uk/resources/300-VW/

4.3. Model Training

The input images are all resized to 224x224 before training. SHELF used Resnet 50 as its backbone for better heatmap featuring and is implemented in Pytorch. The model is trained by 50 epochs using Adam optimizer with the learning rate of 10e-5, the decay of 10e-5 and batch size of 64 on a K80 GPU of Google Colaboratory.

4.4. Results

In this section, we present the experimental results of SHELF on 300VW-E, 300W and WFLW datasets.

Facial landmark detection with SHELF. After training, the model can be used to flash a given facial image to the landmarks thanks to their corresponding heatmaps, as visualized in Fig. 5. These visualizations prove the explainability of SHELF over other existing methods.

Evaluation results on 300VW-E dataset. SHELF is firstly evaluated on 300VW-E and achieved a NME of 12.55%. In fact, the dataset contains a high number of expressive facial images that makes the landmarks highly biased. However, as in Table 1, SHELF is much better than other coordinate regression and shape fitting methods such as SAN, CPM and ASMNet with NME of 13.05%, 15.58% and 18.47%, correspondingly. Clearly, the combination of heatmap regression and shape fitting makes SHELF more tolerant to such biases.

Table 1. NME(%) of SHELF and other comparative methods on 300VW-E dataset

Method	Category	NME
ASMNet [28]	Coordinate Regression, Shape Fitting	18.47
CPM [29]	Coordinate Regression	15.58
SAN [12]	Coordinate Regression	13.05
SHELF *(ours)*	Heatmap Regression, Shape Fitting	**12.55**

Evaluation results on 300W dataset. The results of SHELF on 300W dataset can be seen on the Table 2. Our model SHELF achieved a NME of 3.79%, 6.67% and 4.35% on the Common, Challenging and Full subset of 300W, respectively. These outperform most of the recent methods of coordinate regression, heatmap regression as well as shape fitting such as DeFA, MobileFAN, PCD-CNN, CPM, ASMNet especially on the Challenging subset. SHELF is a bit less accurate than the state-of-the-art AnchorFace but it runs faster at the rate of 43 frames per second (FPS) on NVIDIA Tesla K80 GPU than AnchorFace with 45 FPS on much more powerful NVIDIA GTX Titan Xp GPU. These results prove the efficiency of the combination between the heatmap regression and the shape fitting in our SHELF method.

Evaluation results on WFLW dataset. SHELF is also evaluated on the WFLW dataset using both NME and FR metrics as in Table 3. SHELF achieved the best performance and robustness with a NME of 7.34% and a FR of 17.08% in comparison with recent advanced methods such as ESR (with NME of 11.13%, FR of 35.24%), SDM (with NME of 10.29%, FR of 29.40%), CFSS (with NME of 9.07%, FR of 20.56%) and ASMNet (with NME of 10.77%, FR of 39.12%) on the full WFLW dataset. However, these results are far from those of AnchoFace with NME of 4.62% and FR of 4.2% on the full dataset. This is because SHELF is not efficient for the large pose, occlusion and blur subsets with a NME of 14.81%, 9.10%, 8.15% and a FR of 64.11%, 25.95% and 19.40%, respectively. In fact, AnchorFace is fine-tuned according to various shapes while our SHELF is relied on only one for a given dataset.

Ablation Study. Given the efficacity of the combination of heatmap regression and shape fitting through the variation of β coefficient in the loss function of SHELF, we go a further step to explore how relevant this coefficient is on a given dataset. We conducted an experiment of SHELF on the 300VW-E dataset, with different variation pattern of β including continuous, constant and stepped as demonstrated in Fig. 6.

The experimental results in Table 4 show that SHELF achieved the best NME of 12.55% with stepped variation of β, and exhibited a poor NME of 24.88% and 24.89% in case of constant and continuous ones. Notice that, in case of stepped pattern, the value of β is set to zero at a given training epoch. This confirms that the heatmap regression network learns the facial shapes only at the very beginning epoches of training.

4.5. Discussion

Facial landmark detection is an active research topic over many years because this can be more efficiently used to recognize the human facial emotion than relying on the whole human face. However, most recent methods focus more on the feature engineering of the individual facial landmarks but less on their distribution meaning the shape of the face. Although, the power of deep learning backbone networks has been thoroughly leveraged, the performance of such coordination and heatmap regression methods remains limited. ASMNet was the first to take in to account the shape fitting in to its coordination regression and initially gained positive results. However, the coordination regression approach aims to extract features at the cell level while the heatmap regression targets to the pixel level of the image which is closer to the facial landmarks in this case. Our proposed method SHELF is a combination of heatmap regression and shape fitting achieved a much better performance and

Figure 5. Generation of heatmaps corresponding to the 20 facial landmarks on a given image of 300VW-E dataset.

Table 2. NME(%) of SHELF and other comparative methods on 300W dataset

Model	Category	Common	Challenging	Full
CFSS[30]	Shape Fitting	4.73	9.98	5.76
DSRN [31]	Coordinate Regression	4.12	9.68	5.21
DeFA [11]	Shape Fitting	5.37	9.38	6.10
RDR [32]	Coordinate Regression and Shape Fitting	5.37	9.38	6.10
RCN [33]	Coordinate Regression	4.67	8.44	5.41
ASMNet [28]	Coordinate Regression and Shape Fitting	4.82	8.20	5.50
CPM [29]	Coordinate Regression	3.39	8.14	4.36
PCD-CNN [34]	Heatmap Regression	3.67	7.62	4.44
CPM+SBR [29]	Coordinate Regression	3.28	7.78	4.10
MobileFAN [13]	Heatmap Regression	4.22	6.87	4.74
ODN [35]	Coordinate Regression	3.56	6.67	4.17
SAN [12]	Coordinate Regression	3.34	6.60	3.98
AnchorFace [14]	Anchor-based Regression	**3.12**	**6.19**	**3.72**
SHELF (ours)	Heatmap Regression and Shape Fitting	*3.79*	*6.67*	*4.35*

robustness than ASMNet in both 300W and WFLW datasets which proved our judgements.

5. Conclusion

As discussed, the facial landmark detection is necessary for recognition of human emotion which can be applied in advanced driver assistance systems. This task is really hard due to the variation of facial appearance, shape, pose and the dispersion of high number of landmarks on the human face. Efficient methods such as ASMNet and AnchorFace all take in to account facial shapes and poses. However, these coordination regression methods extract the feature at the cell level which is less accurate than at the pixel level as in case of heatmap regression. In this paper, we proposed a novel facial landmark detection method called SHELF which is the first combination between heatmap regression and shape fitting. The evaluation on 300W, WFLW datasets and on the private one which is an extension of 300VW showed that SHELF outperforms many existing methods including SAN, ASMNet. SHELF can not be compared to AnchorFace due to using less number of anchor shapes. These results proved that such combination is reasonable and the SHELF can also be

Table 3. NME(%) and FR of SHELF and other comparative methods on WFLW dataset

Data	Metric	ESR[36]	SDM[37]	CFSS	ASMNet	AnchorFace	SHELF (ours)
Full	NME	11.13	10.29	9.07	10.77	**4.62**	*7.34*
	FR	35.24	29.40	20.56	39.12	**4.2**	*17.08*
Large Pose	NME	25.88	24.10	21.36	21.11	-	**14.81**
	FR	90.18	84.36	66.22	98.41	-	**64.11**
Expression	NME	11.47	11.45	10.09	12.02	-	**7.74**
	FR	42.04	33.44	23.25	59.87	-	**14.33**
Illumination	NME	10.49	9.32	8.30	9.93	-	**6.92**
	FR	30.80	26.22	17.34	33.38	-	**12.75**
Makeup	NME	11.05	9.38	8.74	10.55	-	**7.16**
	FR	38.84	27.67	21.84	38.34	-	**16.50**
Occlusion	NME	13.75	13.03	11.76	12.34	-	**9.10**
	FR	47.28	41.85	32.88	48.64	-	**25.95**
Blur	NME	12.20	11.28	9.96	11.62	-	**8.15**
	FR	41.40	35.32	23.67	46.31	-	**19.40**

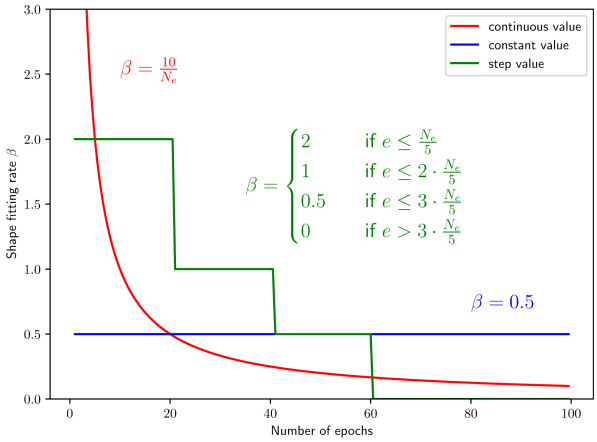


Figure 6. Different variation pattern of β

Table 4. NME(%) of SHELF with different variation pattern of β on 300VW-E dataset

The variation pattern of β	NME
Continuous	24.89
Constant	24.88
Stepped	12.55

better improved with more performant backbone and more facial priors.

References

[1] Nam, N.V. and Quyen, N.T.N. (2023) Flash: Facial landmark detection using active shape model and heatmap regression. In *The 9th EAI International Conference on Industrial Networks and Intelligent Systems*.

[2] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16 (IEEE): 770–778. doi:10.1109/CVPR.2016.90, URL http://ieeexplore.ieee.org/document/7780459.

[3] Tan, M. and Le, Q.V. (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. [eds.] *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA* (PMLR), *Proceedings of Machine Learning Research* **97**: 6105–6114. URL http://proceedings.mlr.press/v97/tan19a.html.

[4] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A. and Chen, L. (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018* (Computer Vision Foundation / IEEE Computer Society): 4510–4520. doi:10.1109/CVPR.2018.00474.

[5] Ma, N., Zhang, X., Zheng, H.T. and Sun, J. (2018) Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y. [eds.] *Computer Vision – ECCV 2018* (Cham: Springer International Publishing): 122–138.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. *et al.* (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (OpenReview.net). URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Cootes, T., Baldock, E. and Graham, J. (2000) An introduction to active shape models. *Image processing and analysis* **328**: 223–248.

[8] Cootes, T.F., Edwards, G.J. and Taylor, C.J. (1998) Active appearance models. In Burkhardt, H. and

Neumann, B. [eds.] *Computer Vision — ECCV'98* (Berlin, Heidelberg: Springer Berlin Heidelberg): 484–498.

[9] Cristinacce, D. and Cootes, T. (2006) Feature detection and tracking with constrained local models. **41**: 929–938. doi:10.5244/C.20.95.

[10] Asthana, A., Zafeiriou, S., Cheng, S. and Pantic, M. (2013) Robust discriminative response map fitting with constrained local models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13 (USA: IEEE Computer Society): 3444–3451. doi:10.1109/CVPR.2013.442, URL https://doi.org/10.1109/CVPR.2013.442.

[11] Liu, Y., Jourabloo, A., Ren, W. and Liu, X. (2017) Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*: 1619–1628.

[12] Dong, X., Yan, Y., Ouyang, W. and Yang, Y. (2018) Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 379–388.

[13] Zhao, Y., Liu, Y., Shen, C., Gao, Y. and Xiong, S. (2019) Mobilefan: Transferring deep hidden representation for face alignment. *Pattern Recognition* **100**: 107114. doi:10.1016/j.patcog.2019.107114.

[14] Xu, Z., Li, B., Yuan, Y. and Geng, M. (2021) Anchorface: An anchor-based facial landmark detector across large poses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**: 3092–3100.

[15] Jolliffe, I.T. and Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065): 20150202.

[16] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA: IEEE Computer Society): 4177–4187. doi:10.1109/CVPR.2016.453, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.453.

[17] Lv, J., Shao, X., Xing, J., Cheng, C. and Zhou, X. (2017) A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3691–3700. doi:10.1109/CVPR.2017.393.

[18] Feng, Z., Kittler, J., Awais, M., Huber, P. and Wu, X. (2018) Wing loss for robust facial landmark localisation with convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA: IEEE Computer Society): 2235–2245. doi:10.1109/CVPR.2018.00238, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00238.

[19] Wang, X., Bo, L. and Fuxin, L. (2019) Adaptive wing loss for robust face alignment via heatmap regression. In *The IEEE International Conference on Computer Vision (ICCV)*.

[20] Xiong, Y., Zhou, Z., Dou, Y. and Su, Z. (2021) *Gaussian Vector: An Efficient Solution for Facial Landmark Detection*, 70–87. doi:10.1007/978-3-030-69541-5_5.

[21] Huang, Y., Yang, H., Li, C., Kim, J. and Wei, F. (2021) Adnet: Leveraging error-bias towards normal direction in face alignment. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* : 3060–3070.

[22] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y. and Zhou, Q. (2018) Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*.

[23] Fard, A.P., Abdollahi, H. and Mahoor, M.H. (2021) Asmnet: A lightweight deep neural network for face alignment and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021* (Computer Vision Foundation / IEEE): 1521–1530. doi:10.1109/CVPRW53098.2021.00168.

[24] Newell, A., Yang, K. and Deng, J. (2016) Stacked hourglass networks for human pose estimation. In Leibe, B., Matas, J., Sebe, N. and Welling, M. [eds.] *Computer Vision – ECCV 2016* (Cham: Springer International Publishing): 483–499.

[25] Le, V., Brandt, J., Lin, Z., Bourdev, L. and Huang, T.S. (2012) Interactive facial feature localization. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. and Schmid, C. [eds.] *Computer Vision – ECCV 2012* (Berlin, Heidelberg: Springer Berlin Heidelberg): 679–692.

[26] 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. doi:10.1109/TPAMI.2013.23.

[27] Köstinger, M., Wohlhart, P., Roth, P.M. and Bischof, H. (2011) Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*: 2144–2151. doi:10.1109/ICCVW.2011.6130513.

[28] Fard, A.P., Abdollahi, H. and Mahoor, M. (2021) Asmnet: A lightweight deep neural network for face alignment and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*: 1521–1530.

[29] Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y. and Sheikh, Y. (2018) Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 360–368.

[30] Zhu, S., Li, C., Loy, C.C. and Tang, X. (2015) Face alignment by coarse-to-fine shape searching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 4998–5006. doi:10.1109/CVPR.2015.7299134.

[31] Miao, X., Zhen, X., Liu, X., Deng, C., Athitsos, V. and Huang, H. (2018) Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S. and Kassim, A. (2017) Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*: 1642–1651. doi:10.1109/ICCV.2017.181.

[33] Honari, S., Yosinski, J., Vincent, P. and Pal, C. (2016) Recombinator networks: Learning coarse-to-fine feature aggregation. In *Computer Vision and Pattern Recognition*

(CVPR), 2016 IEEE Conference on (IEEE).

[34] Kumar, A. and Chellappa, R. (2018) Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA: IEEE Computer Society): 430–439. doi:10.1109/CVPR.2018.00052, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00052.

[35] Ding, H., Zhou, P. and Chellappa, R. (2020) Occlusion-adaptive deep network for robust facial expression recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)* (IEEE Press): 1–9. doi:10.1109/IJCB48548.2020.9304923, URL https://doi.org/10.1109/IJCB48548.2020.9304923.

[36] Cao, X., Wei, Y., Wen, F. and Sun, J. (2012) Face alignment by explicit shape regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*: 2887–2894. doi:10.1109/CVPR.2012.6248015.

[37] Xiong, X. and De la Torre, F. (2013) Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*: 532–539. doi:10.1109/CVPR.2013.75.