TRANSFORMING 3D FACIAL LANDMARKS FOR ACTION UNIT PREDICTION

Saurabh Hinduja and Shaun Canavan

University of South Florida

ABSTRACT

We propose a method for facial action unit (AU) detection using 3D facial landmarks. Specifically, we train a 2D convolutional neural network (CNN) on transformed 3D facial landmarks for binary and multi-class AU detection. We show that the proposed approach is able to accurately model AU occurrences, as the movement of the facial landmarks corresponds directly to the movement of the AUs. By training a CNN on this transformed 3D representation of the landmarks, we can achieve accurate AU detection on two state-of-the-art emotion datasets, namely BP4D and BP4D+. Using the proposed method, we detect multiple AUs on over 330,000 frames, reporting improved results over state-of-the-art methods.

Index Terms— Action units, 3D, landmarks, affective computing, deep learning

1. INTRODUCTION

Facial geometry has a lot of information about an individual and has been used for various applications [1], [2]. It can also convey expressions, such as happy, sad, pain, and embarrassment [3], which can vary between subject to subject. Considering this, the Facial Action Coding System [4] was developed, which represents fundamental facial activity in terms of Action Units (AUs). During an expression (i.e. facial activity), a single AU can occur or multiple at the same time. This allows for FACS to represent the large variety of facial expressions that exist between subjects.

Recently, there has been encouraging progress in automatically detecting AUs. Zeng et al. [5] developed a confidence preserving machine (CPM) for the task. In their proposed method, the CPM learns two classifiers. First the positive classifier separates all positive classes, and the negative classifier does the same for the negative classes. The CPM then learns a person-specific classifier to detect the AUs. Chu et al. [6] used a combination of convolutional neural networks (CNN) with long short-term networks to learn the spatial and temporal cues from images. Their proposed approach achieved promising results on the GFT [7] and BP4D [8] datasets. Li et al. [9] used temporal fusing for AU detection. They developed a deep learning framework where regions of interest are learned independently so each sub-region has a local CNN; multi-label learning is then used.

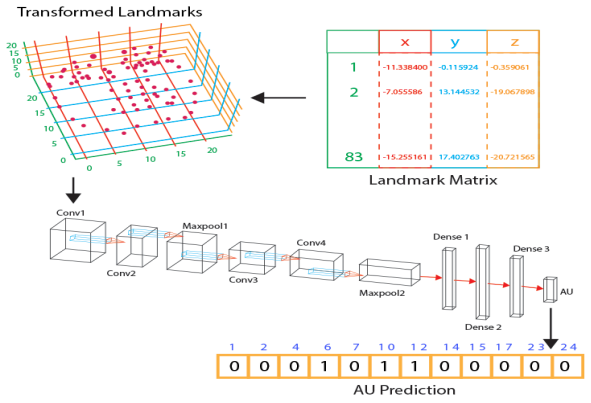


Fig. 1. Overview of proposed method. The raw 3D landmarks are stored as a 2D matrix, which is then transformed to a 3D face representation, which is then used to train a CNN for AU prediction.

Current work largely focuses on 2D images, however, it has been shown that 3D facial landmarks are important when understanding facial geometry for recognizing emotion [10]. Motivated by this, we propose a method for transforming 3D facial landmarks for AU detection. The proposed method transforms 3D facial landmarks to a 3D representation that is uniform across all faces (Figure 1). The contributions of this work are 3-fold and can be summarized as follows.

1. We propose a method for transforming 3D facial landmarks for AU detection.
2. We show the proposed method outperforms current state of the art on BP4D [8].
3. To the best of our knowledge, this is the first work presenting AU detection results on all frames, that are labeled with AUs, in the BP4D+ [3], resulting in a baseline for the community.

2. TRANSFORMING 3D LANDMARKS

We propose a method for AU detection from 3D facial landmarks, by transforming them into a new 3D representation, as we detail in the following subsections.

2.1. Landmark normalization and scaling

Given a set of 3D facial landmarks, *L*, of size *N*, where

$$L = (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ..., (X_N, Y_N, Z_N), \quad (1)$$

we first represent *L* as a 2D matrix of size $N \times 3$ (Figure 1). We then normalize the 2D matrix of landmarks to be within the range between [0,1] using *min-max* normalization. This is done, independently, for each of the *(X,Y,Z)* axes of the frame as $X_{norm} = \frac{(X_i - X_{min})}{(X_{max} - X_{min}}$; $Y_{norm} = \frac{(Y_i - Y_{min})}{(Y_{max} - Y_{min}}$; $Z_{norm} = \frac{(Z_i - Z_{min})}{(Z_{max} - Z_{min}}$. These normalized values of X_{norm}, Y_{norm}, Z_{norm} are then multiplied by a constant value *C-1* to scale the landmarks into the range of *[0,C-1]*, giving us $X_{scaled}, Y_{scaled}, Z_{scaled}$. This is done to scale all faces into the uniform range of *[0,C-1]*.

For the CNN to learn features for AU detection, we need $X_{scaled}, Y_{scaled}, Z_{scaled}$ to be scaled properly. Considering this we used a value of $C = 24$ as we have empirically found that this gives a good representation of the face (Figure 2 (d)). If we try to map the normalized (i.e. non-scaled) coordinates directly to a 3D array then the array will be of size $2 \times 2 \times 2$ each axis going from $[0, 1]$. For example, if two or more normalized landmarks $(X_{norm}, Y_{norm}, Z_{norm})$ are in the range $[0, 0.5)$ then they all set the same cell of $(0, 0, 0)$ to 1, which will lead to loss in information of landmarks (Figure 1 (b)). Also, if *C* is not large enough, it can lead to loss of information (Figure 2 (c)).

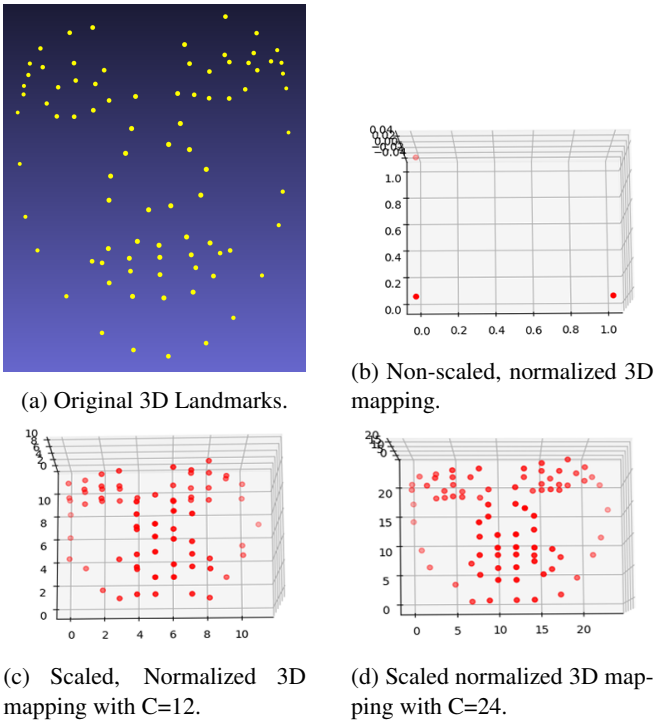
2.2. Scaled landmark mapping to 3D

Given a set of scaled 3D facial landmarks, we then create a 3D representation of them, which can be used to train CNNs (Section 3.2). To achieve this we create a 3D array *A* of size $C \times C \times C$ initialized with all zeroes. We then set the locations where the landmarks are present to 1; $A[X_{scaled_i}][Y_{scaled_i}][Z_{scaled_i}] = 1$. Scaling and mapping the landmarks creates a 3D array which is representative of the 3D locations of the landmarks (Figure 2 (d)), which allows us to explicitly model the 3D shape of the landmarks, and subsequent AUs. As can be seen in Figure 2, the normalized 3D mapping does have some visual differences compared to the original 3D landmarks, however, the general shape, and more importantly AU activation, of the original 3D landmarks is the same. As we will show, this 3D representation of each set of 3D landmarks (i.e. face) can train CNNs to detect changes to the AUs across subjects, achieving high detection accuracy.

3. EXPERIMENTS

3.1. Facial expression databases

BP4D was used in the Facial Expression Recognition and Analysis (FERA) challenge in 2015 [11] and 2017 [12]; containing 2D and 3D data. It contains 41 subjects with eight



(a) Original 3D Landmarks.

(b) Non-scaled, normalized 3D mapping.

(c) Scaled, Normalized 3D mapping with C=12.

(d) Scaled normalized 3D mapping with C=24.

Fig. 2. Comparison of original 3D landmarks with (a)non-scaled normalized 3D mapping (b) and with scaled normalized 3D Mapping for both C=12 (c) and C=24 (d).

dynamic expressions plus neutral. The dataset contains 18 male and 23 female subjects ages 18-29 years of age, with a range of ethnicities. For each sequence in this database, approximately 500 frames contain AU occurrences; we used all labeled frames, which is over 140,000.

BP4D+ consists of 140 subjects (58 males and 82 females); ages 18-66, each with highly varied emotional responses. It includes thermal, physiological, 2D, and 3D data. It was also used in the FERA challenge 2017 [12]. Like BP4D, approximately 500 frames per sequence contain AU occurrences; we used all labeled frames which is over 190,000.

3.2. Experimental Design

We detected 83 3D facial landmarks on all AU labeled frames from BP4D and BP4D+, using a shape-index-based statistical shape model [13]. We then normalized the 83 landmarks to [0,23] ($C = 24$). While any number of landmarks that represent a face with AU occurrences can be used, we chose 83 to be consistent with other related works [3], [8], [12], [10]. We then transform the normalized landmarks to a binary representation in a 3 dimensional array of $24 \times 24 \times 24$ (Figure 1 (d)).

We consider both binary [9] and 3-class [6] classification. For binary classification, the presence/absence of an action unit was coded as 1/0. For the 3-class classification, the pres-

ence/absence of an action unit was coded as 1/-1 and 0 when no information was available [6].

For the binary-class, our CNN consists of 8 layers; two convolutional layers followed by a max pooling layer, two more convolutional layers, another max pooling layer and finally three fully connected layers. The output layer has 12 binary outputs for the 12 AUs being predicted. The activation function for each of the hidden layers is relu [14] and the activation function for the output layer is sigmoid [15], the loss function for this network was binary cross-entropy and the optimizer used was adam [16] (Figure 1). All training is done with 250 epochs.

For the 3-class, we use a similar network with common convolutional layers, but different fully connected layers for each AU (a common output layer can not predict the 3-classes for all the AUs). The loss function was categorical cross-entropy and the activation function for the initial layers was still relu, but for the output layers it was softmax [17], as it performs well for multi-class classification [18]. Again, all training is done with 250 epochs.

For binary-class, we used 3-fold and 10-fold cross-validation, for 3-class problem we used 3-fold cross-validation. We also balanced the distribution of positive and negative samples (i.e. occurrence of AUs), as the distribution of AUs is not consistent (Table 1). We chose this experimental design to be consistent with related works [6], [9]. As can be seen from Table 1, some of the AUs have a small number of frames where the AU occurred, especially in BP4D+.

AU	BP4D	BP4D+
1	21.07	9.54
2	17.04	8.01
4	20.22	0.67
6	46.10	65.88
7	54.90	3.46
10	59.39	57.37
12	56.18	59.99
14	46.60	32.46
15	16.96	12.96
17	34.37	0.67
23	16.56	2.35
24	15.16	-

Table 1. Percentage of AU labeled frames with occurrences in BP4D and BP4D+.

For binary classification, we conducted 4 experiments for each 3-fold and 10-fold: (1) training and testing on BP4D; (2) training and testing on BP4D+; (3) training on BP4D and testing on BP4D+; and (4) training on BP4D+ and testing on BP4D. For our 3-class problem, we performed 2 experiments: (1) training and testing on BP4D; and (2) training and testing on BP4D+. For BP4D we detected 12 AUs, and 11 for BP4D+, as AU 24 did not occur in the labeled frames of this dataset. We refer the reader to Tables 2, 3, or 4 for the list of AUs detected for each dataset.

With this type of classification, especially with imbalanced data (Table 1), F1 score can be a better indicator of performance compared to classification accuracy [11]. Considering this, we calculated the frame-based F1 score as $(F1 - frame = \frac{2RP}{R+P})$ where R is recall and P is precision. This approach is also consistent with related works, allowing us to compare our results.

4. RESULTS AND ANALYSIS

4.1. Binary classification

For binary classification on BP4D, we achieved an average F1 binary score of 92.90 and 94.08 for 3-fold and 10-fold, respectively. On BP4D+, we achieved an average F1 binary score of 86.01 and 87.63 for 3-fold and 10-fold, respectively. See Table 2 for more details.

AU	BP4D		BP4D+	
	3 Fold	10 Fold	3 Fold	10 Fold
1	91.16	92.68	82.78	85.24
2	90.31	92.11	82.82	85.21
4	93.12	94.14	75.57	78.07
6	96.24	97.01	97.18	97.50
7	96.40	96.98	80.62	83.67
10	97.59	97.95	96.97	97.35
12	97.89	98.31	95.85	96.40
14	95.47	96.29	91.22	92.31
15	87.63	89.66	83.26	85.19
17	91.14	92.35	72.39	73.62
23	85.87	88.23	87.40	89.39
24	91.93	93.19	-	-
Avg	**92.90**	**94.08**	**86.01**	**87.63**

Table 2. Binary F1 scores for 3-fold and 10-fold on BP4D and BP4D+.

We also investigated cross-database validation between BP4D and BP4D+. When training on BP4D and testing on BP4D+, we achieved an average F1 score of 42.9 and 42.84 for 3-fold and 10-fold, respectively. When training on BP4D+ and testing on BP4D, we achieved an average F1 score of 39.1 and 40.02 for 3-fold and 10-fold, respectively (Table 3). We found that the best performing AUs were 6, 10, and 12; while the worst performing were 4, 17, and 23. This difference in F1 scores can be explained by the disparity in the occurrence of AUs. The best performing AUs are present in a similar percent of the frames whereas the poor performing AUs have a large difference between BP4D and BP4D+ (Table 1). The biggest difference in occurrence is for AU 17; BP4D has 34.37% and BP4D+ has just 0.67% frames with the AU.

AU	Trained BP4D+ / Tested BP4D		Trained BP4D/ Tested BP4D+	
	3 Fold	10 Fold	3 Fold	10 Fold
1	47.30	54.27	54.28	53.18
2	43.50	47.11	44.05	42.53
4	12.99	11.44	11.83	13.06
6	70.05	69.79	79.24	79.05
7	23.85	20.72	26.73	26.74
10	72.54	73.33	80.46	80.41
12	75.61	75.21	76.60	76.77
14	36.49	36.30	41.54	41.81
15	23.55	24.30	22.53	22.72
17	10.03	12.32	14.61	14.26
23	14.26	15.38	20.06	20.68
Avg	**39.10**	**40.02**	**42.90**	**42.84**

Table 3. Cross-database F1 scores for 3-fold and 10-fold.

4.2. 3-class classification

We performed 3-fold validation on BP4D and BP4D+, reporting the F1-macro and micro scores (Table 4). F1-macro is the average of the F1 scores of the 3 classes; where as F1-micro is the weighted average of the 3 F1 scores. On BP4D, we achieved an average F1-micro score of 94.55, and F1-macro score of 96.09. For BP4D+, an F1-micro score of 87.90 and F1-macro score of 97.03 was achieved. Some of the lowest performing AUs with BP4D+ were again 4, 17, and 23. This is consistent with the cross-database validation and can be explained by the low number of AU occurrences (Table 1).

AU	BP4D		BP4D+	
	F1 Macro	F1 Mirco	F1 Macro	F1 Mirco
1	93.81	96.01	86.90	96.57
2	93.67	96.52	87.06	97.07
4	95.12	96.93	85.83	99.61
6	96.18	96.21	91.69	96.02
7	95.66	95.71	86.62	98.60
10	96.79	96.90	90.51	96.22
12	97.33	97.37	90.80	94.66
14	95.60	95.62	89.17	94.15
15	91.88	95.65	87.39	95.54
17	92.80	93.58	82.19	99.54
23	90.79	95.17	88.71	99.30
24	94.94	97.46	-	-
Avg	**94.55**	**96.09**	**87.90**	**97.03**

Table 4. F1 scores for 3-class, 3-fold cross validation.

4.3. Comparisons to state of the art

On BP4D, many works use 2D images for 3-fold binary [9], 10-fold binary [5], and 3-fold 3-class classification [6]. For each of these experimental designs, the proposed method outperforms the state of the art, detailing the power of explicitly using 3D facial landmarks compared to 2D images.

For 3-fold binary classification, the proposed method achieves a significant increase in the average F1 score compared to current state of the art (Table 5). For 10-fold binary classification, the proposed method achieved an average F1 score of 94.08, compared to Zeng et al. [5], that achieved an average F1 score of 56.5. We also compare our 3-fold, 3-class results to state of the art. The proposed method achieved an average F1-micro and macro score of 96.06 and 94.55, respectively, compared to the work from Chu et al. [6] that reported an average F1 score of 82.5. These increases can be attributed to the proposed 3D representation of the landmarks.

Method	Avg F1 Score
Proposed	**92.9**
R-T1[9]	66.1
FERA[19]	61.4
CNN+LSTM[20]	53.2
CPM[5]	50.0
DRML[21]	48.3
JPML[22]	45.9

Table 5. Comparison of proposed method with state of the art on BP4D, for 3-fold binary classification.

For BP4D+ a subset of data was used in FERA 2017 [12]. BP4D was used as training data, and BP4D+ was used as development and testing sets. They report results, using maximum likelihood, on both of these sets. When training on BP4D and testing on BP4D+, the proposed method achieved an average F1 score of 42.9 and 42.84, for 3-fold and 10-fold cros-validation, respectively. This compares to an average F1 score of 41.6 and 45.2 on the FERA development and test sets, respectively. As the two experimental design are *not* the same, we *do not* claim this as a direct comparison. It is included for clarification of results reported on BP4D+ for AU detection.

5. CONCLUSION

We have proposed a method for transforming 3D landmarks and using them to detect AUs. We have shown that the proposed method is powerful for AU detection on multiple challenging, state-of-the-art databases, including cross-database validation. We report state of the art results on BP4D using 3-fold and 10-fold cross-validation, as well as binary and 3-class classification. To the best of our knowledge, this is the first work to report AU detection results for all AU labeled frames in BP4D+, resulting in a baseline for the community.

6. REFERENCES

[1] S. Jeng et al., "Facial feature detection using geometrical face model: an efficient approach," *Pattern recognition*, vol. 31, no. 3, pp. 273–282, 1998.

[2] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing*, vol. 16, no. 1, pp. 172–187, 2007.

[3] Z. Zheng et al., "Multimodal spontaneous emotion corpus for human behavior analysis," in *CVPR*, 2016, pp. 3438–3446.

[4] P. Ekman and E. Rosenberg, "What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs)," *Oxford University Press*, 1997.

[5] J. Zeng et al., "Confidence preserving machine for facial action unit detection," in *ICCV*, 2015.

[6] W. Chu et al., "Learning spatial and temporal cues for multi-label facial action unit detection," in *Face & Gesture Recognition*, 2017, pp. 25–32.

[7] J. Cohn and M. Sayette, "Spontaneous facial expression in a small group can be automatically measured: An initial demonstration," *Behavior Research Methods*, vol. 42, no. 4, pp. 1079–1086, 2010.

[8] X. Zhang et al., "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[9] W. Li et al., "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *CVPR*, 2017, pp. 6766–6775.

[10] D. Fabiano and S. Canavan, "Spontaneous and non-spontaneous 3d facial expression recognition using a statistical model with global and local constraints," in *ICIP*, 2018.

[11] M. Valstar et al., "Fera-2015 - second facial expression recognition and analysis challenge," in *FG*, 2015.

[12] M Valstar et al., "Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge," in *Face and Gesture*, 2017.

[13] S. Canavan, P. Liu, X. Zhang, and L. Yin, "Landmark localization on 3d/4d range data using a shape index-based statistical shape model with global and local constraints," *CVIU*, vol. 139, pp. 136–148, 2015.

[14] K. Jarrett et al., "What is the best multi-stage architecture for object recognition?," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2146–2153.

[15] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*. Springer, 1995, pp. 195–201.

[16] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] J. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, pp. 227–236. Springer, 1990.

[18] K. Duan et al., "Multi-category classification by softmax combination of binary classifiers," in *International Workshop on Multiple Classifier Systems*. Springer, 2003, pp. 125–134.

[19] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Winter Conference on Applications of Computer Vision*, 2016, pp. 1–8.

[20] W. Chu et al., "Modeling spatial and temporal cues for multi-label facial action unit detection," *arXiv preprint arXiv:1608.00911*, 2016.

[21] K. Zhao et al., "Deep region and multi-label learning for facial action unit detection," in *CVPR*, 2016.

[22] K. Zhao et al., "Joint patch and multi-label learning for facial action unit detection," in *CVPR*, 2015.