FACS3D-Net: 3D Convolution based Spatiotemporal Representation for Action Unit Detection

Le Yang*, Itir Onal Ertugrul†, Jeffrey F. Cohn‡, Zakia Hammal†, Dongmei Jiang*§, Hichem Sahli¶‖
*School of Computer Science, Northwestern Polytechnical University, Xian, China
†Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
‡Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
§Peng Cheng Laboratory, Shenzhen, Guangdong, China
¶Department of Electronics & Informatics, Vrije Universiteit Brussel, Brussels, Belgium
‖Interuniversity Microelectronics Centre, Heverlee, Belgium
Email: yangle.cst@gmail.com

***Abstract*—Most approaches to automatic facial action unit (AU) detection consider only spatial information and ignore AU dynamics. For humans, dynamics improves AU perception. Is same true for algorithms? To make use of AU dynamics, recent work in automated AU detection has proposed a sequential spatiotemporal approach: Model spatial information using a 2D CNN and then model temporal information using LSTM (Long-Short-Term Memory). Inspired by the experience of human FACS coders, we hypothesized that combining spatial and temporal information simultaneously would yield more powerful AU detection. To achieve this, we propose FACS3D-Net that simultaneously integrates 3D and 2D CNN. Evaluation was on the Expanded BP4D+ database of 200 participants. FACS3D-Net outperformed both 2D CNN and 2D CNN-LSTM approaches. Visualizations of learnt representations suggest that FACS3D-Net is consistent with the spatiotemporal dynamics attended to by human FACS coders. To the best of our knowledge, this is the first work to apply 3D CNN to the problem of AU detection.**

***Index Terms*—CNN, CNN-LSTM, FACS3D-Net, spatiotemporal information, multi-label AU detection**

I. Introduction

Facial expression is a powerful channel of emotion, intention, and non-verbal communication more broadly. To annotate facial expression, the Facial Action Coding System (FACS) [1] decomposes facial actions into anatomically-based action units (AU) that individually or in combinations can describe nearly all possible facial expressions. Automatic AU detection is increasingly deployed in a range of applications that include psychiatry, advertising, and health. While there has been much progress, there remains significant need for more accurate AU detection.

The recent development of computer vision technology has promoted the application of deep learning methods for AU detection [2], [3]. Convolutional neural networks (CNN), the most frequently used deep-learning approach, have shown excellent performance in image-related tasks because of their remarkable spatial representation ability. 2D CNN-based approaches often outperform traditional shallow approaches to AU detection [4], [5]. Li *et al.* [4], for instance, developed a deep, 2D CNN-based approach, referred to as EAC-Net, that enhances and crops regions of interest for AU detection and outperforms shallow and some deep approaches.

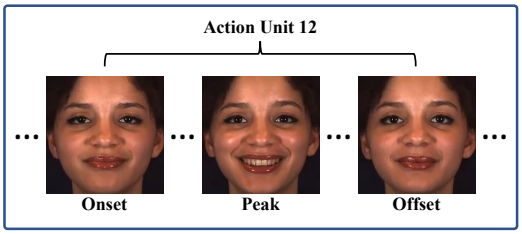

Fig. 1. Temporal variation in AU 12

Almost all recent work in AU detection, whether shallow or deep, ignores motion information or dynamics. Each video frame is considered independently and outside of its temporal context. While temporal context may matter little for strong AU, for subtle AU lack of attention to temporal features impairs detection. Human observers have difficulty perceiving subtle AU when motion information is unavailable [6]. The same may be true for automatic AU detection. AU have an onset, one or more peaks, and an offset. The correlation among proximal frames may be a critical feature for automatic AU detection.

To address the issue of temporal information, some investigators have proposed adding temporal information to spatial information [7]–[10]. Spatial features are learned first, and then temporal. The most common method of adding temporal information is Long Short-Term Memory (LSTM) [9]. LSTM when combined with 2D CNN is referred to as 2D CNN-LSTM. 2D CNN-LSTM learns spatial information prior to learning temporal information.

Simultaneous integration of spatiotemporal information, on the other hand, has been proposed in the literature on action recognition [11], [12]. Ji *et al.* [11] developed a novel 3D CNN model for action recognition. This model extracts features from both spatial and temporal dimensions by performing 3D convolutions on the video segment, thereby capturing the motion information encoded in multiple adjacent frames. 3D CNN approaches have outperformed both 2D CNN and 2D CNN-LSTM approaches for action recognition [13] [14].

We propose a 3D CNN for the problem of AU detection. Our apporach, referred to as FACS3D-Net integrates 3D and 2D

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

Convolutional Neural Networks, as shown in Figure 2. A 3D CNN learns spatiotemporal representations. Simultaneously, a 2D CNN learns spatial representations for each frame. A fully connected layer combines the spatiotemporal and spatial representations to achieve multi-label AU detection for each video frame. We hypothesize that 3D CNN will outperform both 2D CNN and 2D CNN-LSTM.

Inspired by recent work in AU visualization [2], [4], [9], we also explore the specificity of 3D CNN to hypothesized regions of interest. We find that the 3D CNN has good interpretability.

With a multi-label training strategy, the network can perform detection of multiple AUs at one time. We trained FACS3D-Net on the 12 most frequent AU in our database. In addition to analysis of the model's performance, we visualize what is learnt by our model and manipulate the learnt representations.

The contributions of this paper are two-fold:

1) A multi-label AU detection method based on both 3D and 2D convolutions is proposed. This method outperforms both 2D CNN and 2D CNN-LSTM.

2) Occlusion Sensitivity Maps visualize what is learnt by our model. From the maps we can infer that for most AUs the proposed architecture correctly learns the expected facial regions on the input frames.

II. Related Work

AU detection from spatial information. Convolutional neural networks (CNN) generally outperform shallow approaches, especially for large training data [15]. Using 2D CNN, Li and colleagues [4] combined enhancement and cropping layers in a pre-trained model. Cropping layers were used to obtain related facial areas corresponding to individual AU; then independent convolutional layers were applied to these facial areas to learn features. In the enhancing step, an attention map based on facial landmark features was designed and applied to a pre-trained neural network to conduct enhanced learning. Similar ideas have been used in [3] [16] [17]. Zhao *et al.* [3] proposed Deep Region Learning, in which a novel region layer was proposed. One crucial aspect of this network is that it uses feed-forward functions to induce important facial regions, forcing the learned weights to capture structural information of the face. These methods all ignore temporal information. Each video frame is analyzed independently.

AU detection from temporal information. To incorporate temporal information, Valstar *et al.* [7] combined Support Vector Machines and Hidden Markov Models. In subsequent work, Gonzalez *et al.* [8] exploited efficient duration modeling of the temporal behavior of AUs. They proposed a hidden semi-Markov model (HSMM) and variable duration semi-Markov model (VDHMM) to recognize AU dynamics.

For deep architectures, Long Short Term Memory (LSTM) has been proposed to describe the temporal cues of AU by virtue of its ability to model time series. In [9], Chu *et al.* proposed a structure composed of a 2D Convolutional Neural Network and a Long Short Term Memory Network, in which 2D CNN was used to learn spatial representations, and LSTM was used to model temporal dependencies among them. This work suggested that AU detection from spatiotemporal information was more accurate than traditional 2D CNN. A similar idea is presented in [17] [18]. It can be seen that in the algorithms involving AU detection from temporal cues, some networks will adopt the combination structure of 2D CNN and LSTM.

As noted above, both shallow and deep approaches (e.g., SVM and 2D CNN) alike combine spatial and temporal information sequentially. They fail to use temporal information. Temporal representation is added only after the fact. In contrast, manual FACS coders as well as people more generally perceive spatiotemporal information concurrently. Inspired by human perception, we propose an integrated spatiotemporal model for AU detection.

Integrated spatiotemporal approaches. While not previously proposed for AU detection, several investigators in the fields of video summarization and action recognition have proposed integrated spatiotemporal approaches. Ji *et al.* [11] proposed a 3D convolutional kernel to the cube that is formed by stacking multiple contiguous frames together. On the strength of its performance for video summary, 3D CNN has been explored for emotion recognition [19] [20].

When contiguous frames are highly correlated, there may be little loss of temporal information from sequentially down-sampling the video sequence. Down-sampling reduces computational cost. Jing [12] found that 3D CNN is relatively robust to random down-sampling of fixed length.

III. Multi-label AU Detection Using FACS3D-Net

A. Dataset

Data were an expanded version of BP4D+ [21], referred to as EB+ [22]. EB+ is a manually FACS-annotated database of spontaneous behavior. Video is 2D with resolution of 1040 by 1392 pixels. Average video duration is about 44 seconds. Average number of annotated frames is about 328 and standard deviation is 91. Well-designed tasks (e.g. interviews, physical activities) initiated by an experimenter are used to elicit varied emotions. Face orientation is nearly frontal with relatively little out-of-plane head rotation. EB+ contains videos from a total of 200 subjects (140 subjects from BP4D+ and 60 additional ones) associated with 5 to 8 tasks. Positive samples are defined as ones with intensities equal to or greater than B-level; the remaining ones are negative samples. [21] [22].

FACS3D-Net requires that each input video have the same number of frames. To satisfy this requirement, one could sample equal numbers of consecutive frames from each video, but the resulting video segment would fail to be representative of the longer video from which it is taken.

To address this problem, we randomly sampled each second or third frame to obtain equal number of segments from each video. Because adjacent frames are highly correlated, randomly sampling every second or third frame resulted in minimal loss of dynamic information. 126 frames from each video were sampled in this way, which maximized the number of video segments that could be included. The final data were 1215 of the 1261 possible videos with (153,090 frames).

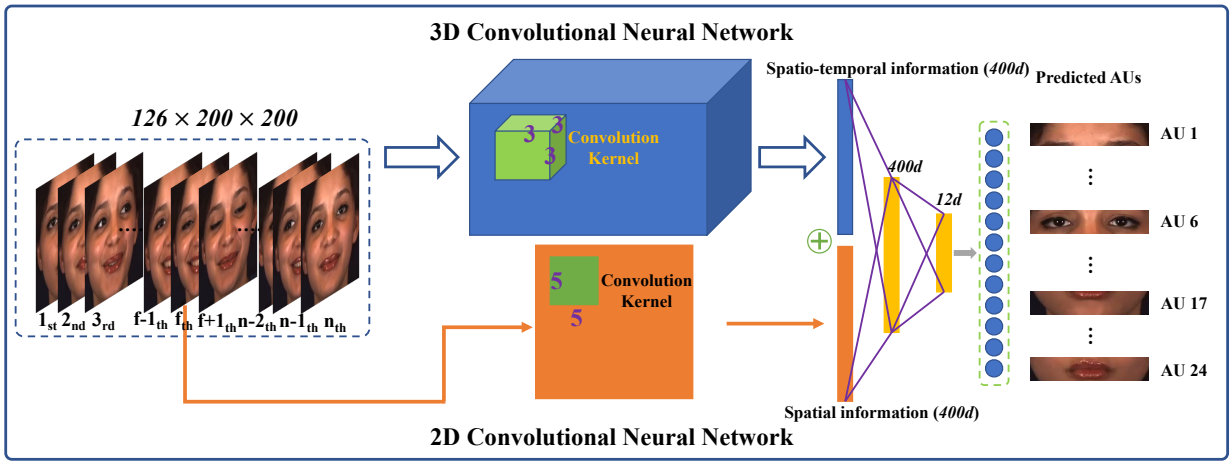

Fig. 2. Overview of the proposed FACS3D-Net for multi-label AU detection. The video clip is fed into 3D convolutional network to learn spatiotemporal information, then concatenated with the spatial information of f_{th} frame obtained from 2D convolutional network. After 2 dense layers, we will obtain the detection results of 12 AUs in the f_{th} frame.

B. Proposed AU Detection Architecture

We introduce a novel architecture for multi-label AU detection tasks that integrates 2D and 3D convolution. A 2D convolution network learns the spatial information of each facial image while a 3D convolution network captures the spatiotemporal information of AU from onset to offset. For convenience, FACS3D-Net will be used to represent the proposed model.

In 2D convolution, convolution kernels are applied to generate feature maps of the corresponding layer. Usually, the outputs of the last convolution layer are connected through several fully connected layers to the final output of the model. Formally, we define the value at position (x, y) of j_{th} feature map in i_{th} layer as V_{ij}^{xy}, which is given by:

$$V_{ij}^{xy} = \sigma\Big(\sum_{m} \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} V_{(i-1)m}^{(x+p)(y+w)} w_{ijm}^{hw} + b_{ij}\Big) \quad (1)$$

where m denotes the number of feature maps in the previous layer that are directly connected to the current feature map, and W_i and H_i are the height and width of the kernel of i_{th} layer, respectively. w_{ijm}^{hw} indicates the weight value of the convolutional kernel at position (h, w) which is connected the m_{th} feature map in the previous layer. $\sigma(\cdot)$ is the activation function and b_{ij} is the bias.

When applying convolution operation to video analysis problems, we seek to capture the variation in temporal dimension, that is, the information between contiguous image frames. Naturally, in [11], the 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. Compared with 2D convolutional kernel $K^2(c, k_w, k_h)$, it extends the time dimension in 3D convolutional kernel, that is, $K^3(c, k_t, k_w, k_h)$. Correspondingly, the value at position (x, y, z) of j_{th} feature map in i_{th} layer is given by:

$$V_{ij}^{xyz} = \sigma\Big(\sum_{m} \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} \sum_{t=0}^{T_i-1} V_{(i-1)m}^{(x+p)(y+w)(z+t)} w_{ijm}^{hwt} + b_{ij}\Big) \quad (2)$$

where T_i is the size of the 3D kernel along the temporal dimension.

Based on the Convolutional Neural Network described above, we propose a hybrid FACS3D-Net architecture for AU detection. As shown in Figure 2, the whole video segment is first fed into 3D convolutional neural network. After a series of 3D convolutions, the network outputs a global representation, which represents the spatiotemporal information of the corresponding segment. Meanwhile, the f_{th} frame is input into the 2D convolutional neural network for spatial information learning. Finally, spatiotemporal representations obtained by 3D CNN and spatial representations obtained by 2D CNN are concatenated in the fully connected layer which are followed by 2 dense layers. Through the activation layer, we will get the detection results of 12 AUs in the f_{th} frame.

The whole structure is similar to the AU annotation process by human coder. 3D CNN is first adopted for video summary, thereby generating a basic impression. In turn, this basic impression is concatenated with the spatial information of an individual frame to perform frame-level AU detection.

C. Network Optimization

Since AU detection can be considered a multi-label binary classification, the network is optimized by minimizing the following loss function:

$$\mathbb{L} = \sum_{i=1}^{N} [y_i \cdot \log \sigma(x)_i + (1 - y_i) \cdot \log(1 - \sigma(x)_i)] \quad (3)$$

Where N is the number of AUs, which is 12 in our case. y_i denotes the ground-truth of occurrence for the i^{th} AU. $\sigma(x)_i$

represents the output of FACS3D-Net for i^{th} AU when the input is x, that is, the corresponding predicted occurrence probability of the i^{th} AU. Since the distribution of AUs is skewed, we assign each AU with unique error cost and obtain the final weighted loss function:

$$\mathbb{L}' = W\mathbb{L} \tag{4}$$

Specifically, following [16], for AU_i, we set:

$$W_i = \frac{(1/r_i)N}{\sum_{i=1}^{N}(1/r_i)} \tag{5}$$

where r_i indicates the occurrence rate of i^{th} AU in the training set. An end-to-end training manner is adopted for FACS3D-Net. 2D and 3D convolution networks are updated simultaneously according to the generated loss.

IV. Experiments

A. Data Preprocessing and Experiments Setting

*1) **Face tracking and registration**:* Video was tracked and normalized using ZFace [23], a real-time face alignment software that accomplishes dense 3D registration from 2D videos and images without requiring person-specific training. Face images were normalized in terms of rotation and scale and then centred, scaled, and normalized to the average interocular distance (IOD) of the participants, which is about 80 pixels. After this step we obtain 200 by 200 pixel images of faces with 80 pixels IOD. Since the presence of AU is independent of facial color, and to increase training efficiency, the normalized RGB images are converted to grayscale images.

*2) **Data split**:* Experiments were performed with a subject-exclusive 5-fold protocol. Each fold contained 40 subjects. We iteratively trained a model using four partitions and evaluated on the remaining one until all subjects were tested. The average performance of the five models was taken as the final evaluation indicator.

*3) **Evaluation metrics**:* Evaluation metrics vary in what aspects of performance they quantify. They differ as well in how they behave when classes are imbalanced. Some are robust to class imbalance, others are not [24]. For these reasons, we report multiple metrics.

F1 score, which is the most commonly used metric in AU detection, considers both precision (P) and recall (R) of the model. It quantifies the performance on correct predictions on positive samples. F1 is calculated by the harmonic average of precision (P) and recall (R) as $F1 = \frac{2RP}{R+P}$. When categories are imbalanced, F1 is attenuated [24].

Negative agreement (NA) is the complement of F1. NA evaluates the solution by the harmonic agreement of samples not containing AUs. Contrary to F1 score, it reflects the performance on correct prediction of negative samples and is attenuated when categories are imbalanced.

Area under the Receiver Operating Characteristics Curve (AUC) illustrates the diagnostic ability of a binary classifier system. It quantifies the extent to which a model is capable of distinguishing between different classes. In our case, the higher the AUC, the better the model distinguishes between AU present and absent. AUC is robust to imbalanced data.

Accuracy quantifies how well a binary classification correctly identifies or excludes a condition. Accuracy is robust to imbalanced data, but for infrequent categories (i.e., AU that have low base rates) accuracy typically is inflated by high chance agreement for negative occurrences.

S score or "free-marginal kappa coefficient". To control for chance agreement, free-marginal kappa estimates chance agreement by assuming that each category is equally likely to be chosen at random [25]. Class imbalance has relatively mild influence on the measure. When applied to two annotators assigning facial actions to dichotomous categories, S score is calculated by Equation (6), where N is the number of samples, n is the number of annotators, k is the number of rating categories, and n_{ij} indicates the number of annotators who assigned the i_{th} sample to the j_{th} category:

$$\begin{gathered} \kappa_{free} = \frac{P_o - P_e}{1 - P_e}, \\ P_o = \frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn), \quad P_e = \frac{1}{k} \end{gathered} \tag{6}$$

*4) **Network and training setting**:* Figure 3 illustrates the 3D convolutional structure in FACS3D-Net. Following [22], we employ three convolutional layers with 64, 128 and 128 filters, respectively. For the 3D convolution kernel, we follow [12] and set all kernel sizes as $3 \times 3 \times 3$ (3×3 pixels in the spatial dimension and 3 frames in the temporal dimension), with a stride of 2, 1 and 1, respectively. Each 3D convolutional layer is followed by a 3D BatchNorm, Rectified Linear Unit (ReLU) and 3D MaxPooling layer. For the 3D MaxPooling, the pooling kernel size is set as $3 \times 3 \times 3$, with stride = 2. Finally, a fully connected layer with 400 units is connected to the last MaxPooling layer.

For the 2D convolutional part in FACS3D-Net, we adopt the same structure as that of 3D. The only difference is the convolution kernel size in 2D network is 5×5 pixels, and the Maxpooling kernel size is 2×2. Similarly, a fully connected layer with 400 units is also connected to the last MaxPooling layer. Finally, the outputs of 3D and 2D networks are concatenated to form an 800 (400+400)-dimensional feature vector, which is followed by 2 fully connected layers with 400 and 12 neurons respectively.

FACS3D-Net is trained using Pytorch with stochastic gradient descent (SGD) optimization algorithm, a momentum of 0.9, a mini-batch size of 20. The whole framework is trained from scratch and optimized with 30 epochs with a learning rate of 1e-3.

B. Multi-label AU Detection Results

Following [22], we include the 12 AUs that occurred in 3% or more of video frames. These are AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU6 (cheek raiser), AU7 (lid tightener), AU10 (upper lip raiser),

TABLE I
MULTI-LABEL AU DETECTION RESULTS (%) OF THREE APPROACHES.

AU	BR	k	F1			ACC			NA			AUC			S		
-	-	-	2D	LSTM	3D	2D	LSTM	3D	2D	LSTM	3D	2D	LSTM	3D	2D	LSTM	3D
1	9.11	88.00	34.64	39.09	**42.94**	84.33	87.90	**89.78**	88.29	92.25	**94.39**	76.81	79.53	**82.39**	68.65	75.79	**79.57**
2	6.77	90.00	32.59	32.93	**38.05**	88.67	89.96	**89.99**	92.19	**93.75**	93.22	77.12	78.40	**82.13**	77.34	79.93	**79.98**
4	7.94	89.00	44.06	44.05	**49.84**	91.92	91.04	**92.23**	**96.40**	95.16	96.09	83.08	81.66	**85.45**	83.85	82.07	**84.45**
6	42.59	69.00	82.13	81.48	**82.26**	**84.04**	83.47	83.99	**81.56**	81.20	80.61	92.10	91.51	**92.19**	**68.06**	66.95	67.97
7	62.99	70.00	85.31	**85.42**	85.08	80.26	**80.88**	79.92	62.28	**65.43**	59.89	**88.24**	88.17	87.36	60.52	**61.76**	59.83
10	58.64	78.00	**87.55**	87.01	87.15	**85.43**	84.57	84.58	**80.76**	77.97	76.99	**92.29**	91.60	91.66	**70.87**	69.14	69.16
12	52.82	76.00	87.18	86.34	**87.45**	86.34	85.17	**86.61**	83.49	80.13	**83.75**	93.58	92.93	**94.06**	72.67	70.34	**73.21**
14	39.34	65.00	65.86	65.68	**65.95**	75.28	74.91	**75.69**	67.85	67.45	**68.31**	78.52	77.83	**79.39**	50.56	49.82	**51.37**
15	10.46	81.00	44.01	44.05	**48.41**	86.88	86.97	**89.46**	91.24	91.37	**94.29**	82.36	82.47	**83.76**	73.75	73.94	**78.92**
17	14.89	83.00	44.28	44.09	**47.44**	78.09	75.42	**80.33**	81.52	77.25	**84.07**	79.15	78.84	**80.43**	56.17	50.84	**60.65**
23	14.10	80.00	44.78	44.27	**50.03**	81.56	81.85	**84.15**	86.08	86.65	**88.44**	78.62	78.46	**81.78**	63.12	63.70	**68.29**
24	2.99	91.00	29.59	25.30	**31.94**	**95.40**	94.82	95.31	**97.28**	96.69	97.05	84.97	82.17	**87.55**	**90.81**	89.64	90.62
Ave	26.89	80.00	56.83	56.64	**59.71**	84.85	84.75	**86.00**	84.08	83.78	**84.76**	83.90	83.63	**85.68**	69.70	69.49	**72.00**

BR = base rate; k = free-margin kappa for inter-observer agreement; 2D = 2D CNN; LSTM = 2D CNN-LSTM; 3D = FACS3D-Net.

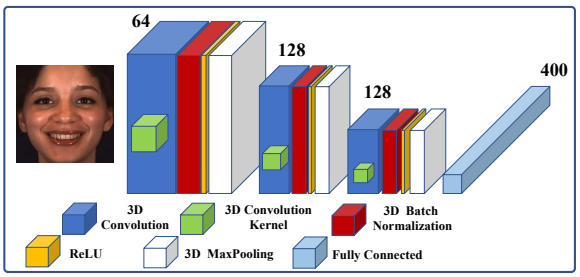

Fig. 3. Structure of 3D CNN in FACS3D-Net.

AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU23 (lip tightener) and AU24 (lip pressor).

Table I lists the detection results for each AU and the average performance across all AUs. We compare 2D CNN, 2D CNN-LSTM, and FACS3D-Net. For economy of presentation, we use 2D, LSTM and 3D to represent 2D CNN, 2D CNN-LSTM and FACS3D-Net, respectively. So that results may be comparable, all three share the same CNN structure (i.e., that of FACS3D-Net). The 2D CNN-LSTM consists of 2D CNN and one layer of LSTM with 256 neurons.

Because model performance may be influenced by reliability of the ground truth used in training (i.e., inter-observer agreement of manual FACS annotators) and also AU base rates, these data are reported in the table as well. Note that 7 of 12 AUs occur in fewer than 15 percent of frames. Consistent with [24], model performance for F1 and AUC tends to be lower for highly imbalanced frames. AU 24 is an exception. AUs with less class imbalance (i.e., higher base rates) are associated with higher F1 scores. Examples include AU 6, AU 7, AU 10, and AU 14. For 10 of 12 AUs, FACS3D-Net achieved the highest F1 score. For accuracy and S score, FACS3D-Net achieved the highest score for 8 of 12 AUs. For all metrics, FACS3D-Net achieved the highest average performance.

For F1 score, FACS3D-Net achieved best performance for all but two AUs (Table I). For ACC, NA, AUC and S score, FACS3D-Net has slight advantages over 2D CNN and 2D CNN-LSTM. This shows that the FACS3D-Net is better than the other two models in detecting positive samples, especially for the AUs with lower base rates.

TABLE II
SIGNIFICANCE OF DIFFERENCES BY T-TEST. * IS $p < 0.05$, ** IS $p < 0.01$, *** IS $p < 0.001$.

AU	FACS3D-Net v.s. 2D CNN					FACS3D-Net v.s. 2D CNN-LSTM				
-	F1	ACC	NA	AUC	S	F1	ACC	NA	AUC	S
1	***	***	***	***	***	***	*	*	**	*
2	***	n.s.	n.s.	**	*	***	n.s.	n.s.	**	n.s.
4	***	n.s.	n.s.	n.s.	n.s.	***	n.s.	n.s.	*	n.s.
6	**	n.s.	n.s.	**	n.s.	**	n.s.	n.s.	**	n.s.
7	n.s.	n.s.	n.s.	*	n.s.	n.s.	n.s.	n.s.	**	n.s.
10	n.s.	n.s.	n.s.	n.s.	n.s.	*	n.s.	n.s.	n.s.	n.s.
12	**	n.s.	*	n.s.	n.s.	***	**	***	n.s.	**
14	**	n.s.	*	n.s.	n.s.	**	n.s.	*	n.s.	n.s.
15	***	***	***	**	***	***	***	***	**	***
17	***	**	*	n.s.	**	***	***	***	n.s.	***
23	***	**	*	n.s.	**	***	**	*	n.s.	**
24	***	n.s.	n.s.	n.s.	n.s.	***	n.s.	n.s.	n.s.	n.s.

As noted above, LSTM is also used for modeling temporal information. Results for 2D CNN-LSTM are reported in Table I. 2D CNN-LSTM appears to have little or no significant advantages over CNN for most AUs.

To evaluate statistical significance of the findings reported in the previous table, we conducted paired t-tests. Table II reports the results of statistical tests between FACS3D-Net and 2D CNN and FACS3D-Net and 2D CNN-LSTM. F1 score reveals that FACS3D-Net performs significantly better than 2D CNN and 2D CNN-LSTM. Among the other four metrics, FACS3D-Net results were less consistent but better overall, especially in AUs with lower base rates.

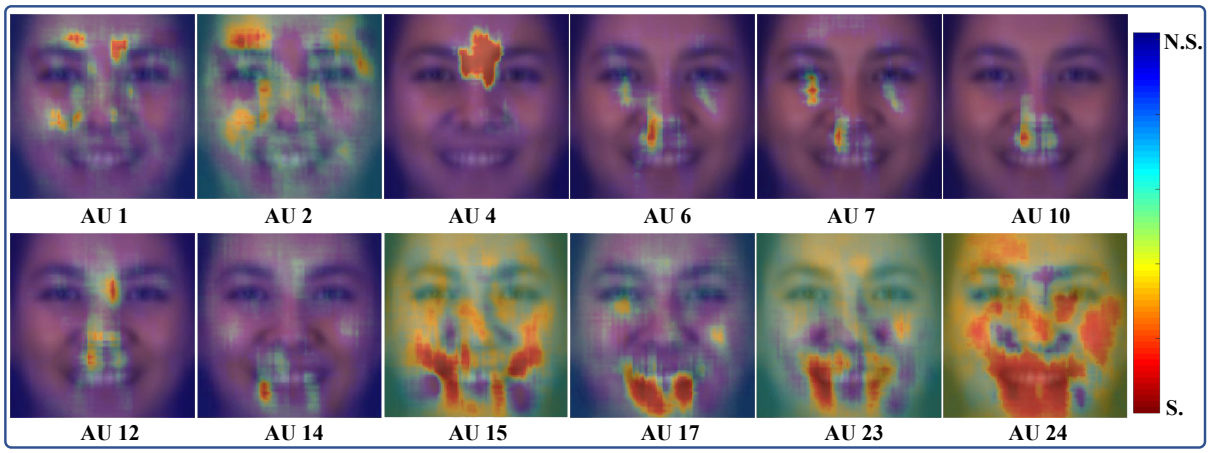

Fig. 4. Occlusion Sensitivity Maps overlaid on a mean face.

C. Occlusion Sensitivity Maps

To visualize our results and verify if the proposed FACS3D-Net model has learned specific facial areas for different AUs, we generate Occlusion Sensitivity Maps [26] for each AU. We select 40 subjects as the test set. A patch with 15×15 pixels is utilized to modify the original gray image value to 0.5 for all testing frames. The modified images are fed into the trained model to obtain an accuracy value for positive samples. Then, we slide the patch over the image of size 200×200 with a stride 3. Therefore, for different regions of patches, we obtain different accuracies. The lower accuracy of region, the more important this region is. Finally, after an interpolation step, we obtain an accuracy map which can be further transformed into occlusion sensitivity map shown in Fig. 4. In these maps, darker red colors indicate the lower accuracy of correctly estimating positive samples, while darker blue colors mean the parts have little effect on accuracy of positive samples. Therefore, the significant regions for each AU are the ones with dark red.

From the maps we can infer that for most AUs, our FACS3D-Net model correctly learns the corresponding facial areas, i.e. AU1 focuses on inner brow region, AU2 focuses on outer brow region, AU4 focuses on brow region, AU7 focuses on lid region, AU10 focuses on upper lip region, AU17 focuses on chin region and AU23 focuses on lip region.

Although the lip regions have been contained, one can see from the sensitivity maps that some cheek regions are also included for AU15 and AU24, the reason could be that while classifying AU15 and AU24, our model also considers the other AUs that co-occur with AU15 and AU24. Furthermore, we visualize the co-occurrence matrix of AUs computed using Jaccard index in Fig. 5. It can be observed that AU6, AU10 and AU12 have strong correlations with each other, meaning that they co-occur frequently. Therefore, for AU6 and AU12, we can see in Fig. 4 that our model also focuses on some regions that are important for AU10.

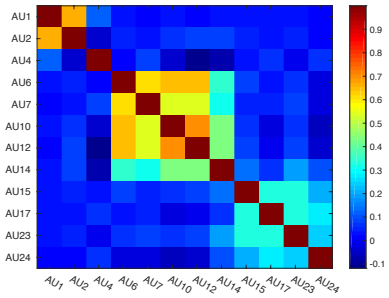

Fig. 5. Co-occurrence matrix of AUs computed with Jaccard index.

V. Conclusions

We propose a novel multi-label AU detection architecture named FACS3D-Net, which provides a new perspective to combine spatial and temporal information. FACS3D-Net learns spatiotemporal information first, then combines it with spatial information. Compared with 2D CNN and 2D CNN-LSTM, the AU detection process of FACS3D-Net considers spatial and temporal information simultaneously, and has a better modelling ability for larger time interval than 2D CNN-LSTM. FACS3D-Net has better performance for the AUs with lower base rates. In order to further analyse the model performance from different perspectives, we recommend ACC, NA, AUC and S score as the evaluation metrics.

Finally, we adopt Occlusion Sensitivity Maps to visualize the learnt representations by FACS3D-Net. One can see that the maps are generally consistent with the expected facial regions for most AUs. We can conclude that, employing FACS3D-Net for AU detection both provides promising results and brings more interpretability of model.

The future work will focus on combining attention mechanism to FACS3D-Net and considering AU intensity and AU occurrence together.

Acknowledgment

This research was supported in part by NIH awards NS100549 and MH096951, NSF award CNS-1629716, the

Shaanxi Provincial International Science and Technology Collaboration Project (grant 2017KW-ZD-14), and the VUB Interdisciplinary Research Program through the EMO-App project.

REFERENCES

[1] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[2] I. Onal Ertugrul, L. A. Jeni, and J. F. Cohn, "Facscaps: Pose-independent facial action coding with capsules," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2130–2139.

[3] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3391–3399.

[4] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 103–110.

[5] I. Onal Ertugrul, L. A. Jeni, W. Ding, , and J. F. Cohn, "Afar: A deep learning based tool for automated facial affect recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.

[6] Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions," *Psychological science*, vol. 16, no. 5, pp. 403–410, 2005.

[7] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden markov models for modeling facial action temporal dynamics," in *International workshop on human-computer interaction*. Springer, 2007, pp. 118–127.

[8] I. Gonzalez, F. Cartella, V. Enescu, and H. Sahli, "Recognition of facial actions and their temporal segments based on duration models," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 10 001–10 024, 2015.

[9] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning facial action units with spatiotemporal cues and multi-label sampling," *Image and vision computing*, vol. 81, pp. 1–14, 2019.

[10] J. He, D. Li, B. Yang, S. Cao, B. Sun, and L. Yu, "Multi view facial action unit detection based on cnn and blstm-rnn," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 848–853.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[12] L. Jing, X. Yang, and Y. Tian, "Video you only look once: Overall temporal convolutions for action recognition," *Journal of Visual Communication and Image Representation*, vol. 52, pp. 58–65, 2018.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[14] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[15] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3515–3522.

[16] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 705–720.

[17] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1841–1850.

[18] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–8.

[19] Y.-H. Byeon and K.-C. Kwak, "Facial expression recognition using 3d convolutional neural network," *International journal of advanced computer science and applications*, vol. 5, no. 12, 2014.

[20] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.

[21] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[22] I. Onal Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Cross-domain au detection: Domains, learning approaches, and measures," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8.

[23] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d video for real-time use," *Image and Vision Computing*, vol. 58, pp. 13–24, 2017.

[24] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *ACII*. IEEE, 2013, pp. 245–251.

[25] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (gft) spontaneous facial expression database," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 581–588.

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.