Learning Spatial and Temporal Cues for Multi-label Facial Action Unit Detection

Wen-Sheng Chu[†] **Fernando De la Torre**[†] **Jeffrey F. Cohn**[†‡]
[†]Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213
[‡]Department of Psychology, University of Pittsburgh, Pittsburgh PA 15260

***Abstract*— Facial action units (AU) are the fundamental units to decode human facial expressions. At least three aspects affect performance of automated AU detection: *spatial representation*, *temporal modeling*, and *AU correlation*. Unlike most studies that tackle these aspects separately, we propose a hybrid network architecture to jointly model them. Specifically, spatial representations are extracted by a Convolutional Neural Network (CNN), which, as analyzed in this paper, is able to reduce person-specific biases caused by hand-crafted descriptors (*e.g.*, HOG and Gabor). To model temporal dependencies, Long Short-Term Memory (LSTMs) are stacked on top of these representations, regardless of the lengths of input videos. The outputs of CNNs and LSTMs are further aggregated into a fusion network to produce per-frame prediction of 12 AUs. Our network naturally addresses the three issues together, and yields superior performance compared to existing methods that consider these issues independently. Extensive experiments were conducted on two large spontaneous datasets, GFT and BP4D, with more than 400,000 frames coded with 12 AUs. On both datasets, we report improvements over a standard multi-label CNN and feature-based state-of-the-art. Finally, we provide visualization of the learned AU models, which, to our best knowledge, reveal how machines see AUs for the first time.**

I. Introduction

Facial actions convey information about a person's emotion, intention, and physical state, and are vital for use in studying human cognition and related processes. To encode such facial actions, the Facial Action Coding System (FACS) [10] is the most comprehensive. FACS segments visual effects of facial activities into action units (AUs), providing an essential tool in affective computing, social signal processing and behavioral science. Such AUs have shown powerful descriptions in universal expressions and led to discoveries in many areas such as marketing, mental health, and entertainment.

A conventional pipeline of automated facial AU detection compiles four major stages: face detection $\mapsto$ alignment $\mapsto$ representation $\mapsto$ classification. With the progress made in face detection and alignment, most research nowadays focuses on features, classifiers, or their combinations. However, due to slow-growing rate in the amount of FACS-coded data, it remains unclear how to pick the best combination that generalizes across subjects and datasets. At least three aspects affect the performance of automated AU detection: (1) *Spatial representation:* Engineered features, *e.g.*, HOG, induce person-specific biases in AU estimation, and hence often require sophisticated models to reduce such effects (*e.g.*, [3], [26], [37]). A good representation must generalize to unseen subjects, regardless of individual differences caused by behavior, facial morphology or recording environment.

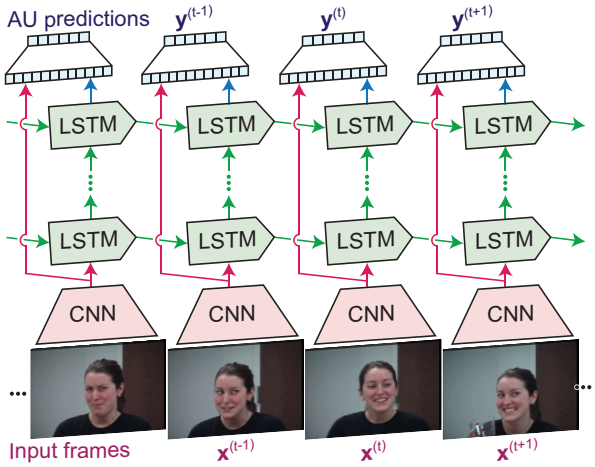

Fig. 1. An overview of the proposed hybrid deep learning framework: The proposed network possesses both strengths of CNNs and LSTMs to model and utilize both spatial and temporal cues. Then, a fusion network is employed to combine both cues to produce frame-based prediction.

(2) *Temporal modeling:* Temporal cues are crucial for identifying AUs, due to the ambiguity and dynamic nature of facial actions. However, it remains unclear how temporary context can be effectively encoded and recalled. (3) *AU correlation:* The presence of AUs influences each other due to the underlying use the same group of facial muscles. For instance, AU12 suggests a co-occurrence of AU6, and reduces the likelihood of AU15. Such correlation helps a detector determine one AU given others. Despite the seemingly unrelated nature of these aspects, this paper shows that it is possible and better to consider them jointly. One intuition is that a good representation helps learn temporal models and AU correlations, and knowing AU correlations could benefit learning of representation and temporal cues. Most existing studies, however, address these aspects separately, and thus are unable to fully capture their entangled nature.

To address the above issues, this paper proposes a hybrid network architecture that models both spatial and temporal relationships from multiple AUs. The proposed network is appealing for naturally modeling the three complementary aspects. Fig. 1 gives an overview of the proposed framework. To learn a generalizable representation, a CNN is trained to extract spatial features. As analyzed in this study, such features reduce the ubiquitous person-specific biases in hand-crafted features [3], [26], [37], and thus offer possibilities to reduce the burden of designing sophisticated classifiers. To capture temporal dependencies, LSTMs are stacked on top of the spatial features. Lastly, we aggregate the output

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.13

IEEE computer society

scores from both CNNs and LSTMs into a fusion network to predict 12 AUs for each frame. Extensive experiments were performed on two spontaneous AU datasets, GFT and BP4D, containing totally >400,000 frames. We report that the learned spatial features, further combined with temporal information, outperform a standard CNN and feature-based state-of-the-art methods. In addition, we visualize notions of each AU learned by the model, which, to our best knowledge, reveal how machines see facial AUs for the first time.

II. Related Work

Below we review contemporary issues in automated facial AU detection and success in deep networks.

Facial AU detection: Despite advances in features, classifiers, and their combinations [6], [23], [27], [32], three important aspects reside in automated AU detection. The first aspect is *spatial representation*, which is typically biased to individual differences such as appearance, behavior or recording environments. These differences produce shifted distributions in feature space (*i.e.*, *covariate shift*), hindering the generalizability of pre-trained classifiers. To reduce distribution mismatch, several studies merged into *personalization* techniques. Chu *et al.* [3] personalized a generic classifier by iteratively reweighting training samples based on relevance to a test subject. Along this line, Sangineto *et al.* [26] directly transferred classifier parameters from source subjects to a test one. Zeng *et al.* [39] adopted an easy-to-hard strategy by propagating confident predictions to uncertain ones. Yang *et al.* [37] further extended personalization for estimating AU intensities by removing a person's identify with a latent factor model. Rudovic *et al.* [25] interpreted the person-specific variability as a context-modeling problem, and propose a conditional ordinal random field to address context effects. Others sought to learn AU-specific facial patches to specialize the representation [41], [42]. However, while progress has been made, these studies still resort to hand-crafted features. We argue that person-specific biases from such features can be instead reduced by learning them.

Another aspect remains in *temporal modeling*, as modeling dynamics is crucial in human-like action recognition. To explore temporal context, graphical models have been popularly used for AU detection. A hidden CRF [1] classified over a sequence and established connections between the hidden states and AUs. These models made Markov assumption and thus lacked consideration of long-term dependencies. As an alternative, switching Gaussian process models [2] was built upon dynamic systems and Gaussian process to simultaneously track motions and recognize events. However, the Gaussian assumption unnecessarily holds in real-world scenarios. In this paper, we attempt to learn long-term dependencies to improve predicting AUs without the requirement to a priori of state dependencies and distributions.

Last but not least, it has attracted an increasing attention on how to effectively incorporate *AU correlations*. Due to the fact that AUs could co-occur simultaneously within a frame, AU detection by nature is a *multi-label* instead of a *multi-class* classification problem as in holistic expression recognition, *e.g.*, [9], [21]. To capture AU correlations, a generative dynamic Bayesian networks (DBN) [31] was proposed with consideration of their temporal evolutions. Rather than learning, pairwise AU relations can be statistically inferred using annotations, and then injected into a multi-task framework to select important patches per AU [41]. In addition, a restricted Boltzmann machine (RBM) [33] was developed to directly capture the dependencies between image features and AU relationships. Following this direction, image features and AU outputs were fused in a continuous latent space using a conditional latent variable model [11]. For the scenario with missing labels, a multi-label framework can be applied by enforcing the consistency between the prediction and the annotation and thetsmooth of label assignment [35]. Although improvements can be observed from predicting multiple AUs jointly, these approaches rely on engineered features such as HOG, LBP, or Gabor.

Deep networks: Recent success of deep networks suggests strategically composing nonlinear functions results in powerful models for perceptual problems. Closest to our work are the ones in AU detection and video classification.

Most deep networks for AU detection directly adapt CNNs. Gadi *et al.* [14] used a 7-layer CNN for estimating AU occurrence and intensity. Ghosh *et al.* [12] showed that a shared representation can be directly learned from input images using a multi-label CNN. To incorporate temporal modeling, Jaiswal *et al.* [15] trained CNNs and BLSTM on shape and landmark features to predict for individual AUs. Because input features were predefined masks and image regions, unlike this study, gradient cannot backprop to full face region to analyze per-pixel contributions to each AU. In addition, it ignored AU dependencies and temporal info that could improve performance in video prediction, *e.g.*, [29], [36]. On the contrary, our network simultaneously models spatial-temporal context and AU dependencies, and thus serves as a more natural framework for AU detection.

The construction of our network is inspired by recent studies in video classification. Simonyan *et al.* [29] proposed a two-stream CNN that considers both static frames and motion optical flow between frames. A video class was predicted by fusing scores from both networks using either average pooling or an additional SVM. To incorporate "temporally deep" models, Donahue *et al.* [8] proposed a general recurrent convolutional network that combines both CNNs and LSTMs, which can be then specialized into tasks such as activity recognition, image description and video description. Similarly, Wu *et al.* [36] used both static frames and motion optical flow, combined with two CNNs and LSTMs, to perform video classification. Video-level features and LSTM outputs were fused to produce a per-video prediction.

Our approach fundamentally differs from the above networks in several aspects: (1) Video classification is a *multi-class* classification problem, yet AU detection is *multi-label*. (2) Motion optical flow is usually useful in video classification, but *not* in AU detection due to large head movements. (3) AU detection requires per-frame detection; video classification produces *video-based* prediction.

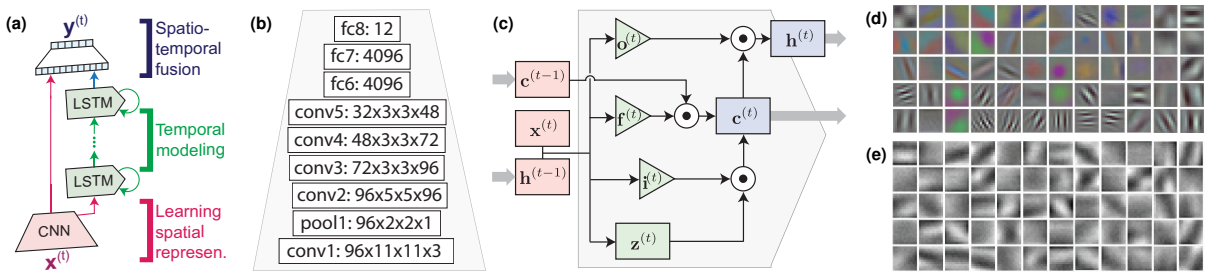

Fig. 2. The structure of the proposed hybrid network: (a) Folded illustration of Fig. 1, showing 3 components of *learning spatially representation*, *temporal modeling*, and *spatiotemporal fusion*, (b) our 8-layer CNN architecture, and (c) the schematic of an LSTM block. (d)-(e) Visualization of conv1 layers of models trained on ImageNet [19] and GFT datasets, respectively. As can be seen, filters learned on our face dataset contain less color blob detectors, suggesting color is less informative in AU detection. (best view in color)

III. The Hybrid Network for Multi-label Facial AU Detection

Fig. 2(a) shows a folded illustration of the hybrid network. Below we describe each component in turn.

A. Learning spatial representation

The literature has shown evidence that hand-crafted features impair generalization of AU detectors [3], [26], [37]. We argue that specialized representation could be learned to reduce the burden of designing sophisticated models, and further improve performance. On the other hand, some AUs co-occur frequently (*e.g.*, AUs 6+12 in a Duchenne smile), and some infrequently. Classifiers trained with such relation are likely to lead to more reliable results [11], [35], [41]. To these two ends, we train a multi-label CNN by modifying the AlexNet [19] as shown in Fig. 2(b). Given a ground truth label $\mathbf{y} \in \{-1, 0, +1\}^L$ (-1/$+1$ indicates absence/presence of an AU, and 0 missing label) and a prediction vector $\widehat{\mathbf{y}} \in \mathbb{R}^L$ for L AU labels, this multi-label CNN aims to minimize the multi-label cross entropy loss:

$$L_E(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{-1}{L} \sum_{\ell=1}^{L} [y_\ell > 0] \log \widehat{y}_\ell + [y_\ell < 0] \log(1 - \widehat{y}_\ell),$$

where $[x]$ is an indicator function returning 1 if x is true, and 0 otherwise. The outcome of the fc7 layer is L_2 normalized as the final representation, resulting in a 4096-D vector. We denote this representation "fc7" hereafter. Due to dropout and ReLu, fc7 feature contains ∼35% zeros out of 4096 values, resulting in a significantly sparse vector. The proposed multi-label CNN is similar to Ghosh *et al.* [12] and AlexNet [19], with slightly different architecture and purpose. Ghosh *et al.* [12] took 40×40 images as input, which, in our experience, can be insufficient for recognizing subtle AUs on the face. AlexNet [19] was designed for object classification, yet, for structured face images, the original design can be an overkill and cause overfitting. Instead, we train our modified network from scratch. Fig. 2(d) visualizes the learned kernels from the conv1 layer on the ImageNet and the GFT datasets. As can be seen, the kernels learned on GFT contain less color blob detectors than the ones learned on ImageNet [19]. This suggests color info is less useful in facial objects than in natural images. In Sec. IV, we will empirically evaluate fc7 against hand-crafted features such as HOG or Gabor.

B. Temporal modeling with stacked LSTMs

It is usually hard to tell an "action" by looking at only a single frame. Having fc7 extracted, we use stacked LSTMs [13] for learning such temporal context. Fig. 2(c) shows the schematic of a standard LSTM block. Due to an absence of theory in choosing the number of LSTM layers and size of each memory cell, we took an empirical approach by considering the tradeoff between accuracy and computational cost. It ended up with 3 stacks of LSTMs with 256 memory cells each. One benefit of LSTM is its ability of encoding crucial information during the transition between two frames. Unlike learning spatial representation on fixed and cropped images, videos can be difficult to be modeled with a fixed-size architecture, *e.g.*, [1], [18]. LSTM serves as an ideal model for avoiding the well-known "vanishing gradient" effect in recurrent models, and makes it possible to model long-term dependencies.

Recurrent LSTMs: Denote a sequence of input frames as $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$, and their labels as $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(T)})$, where superscripts indicate time steps. A recurrent model is expressed by iterating the equations from $t = 1$ to T:

$$\mathbf{h}^{(t)} = \mathcal{H}(\mathbf{W}_{xh}\mathbf{x}^{(t)} + \mathbf{W}_{hh}\mathbf{h}^{(t-1)} + \mathbf{b}_h), \quad (1)$$

$$\mathbf{y}^{(t)} = \text{softmax}(\mathbf{W}_{hy}\mathbf{h}^{(t)} + \mathbf{b}_y), \quad (2)$$

where $\mathbf{W}$ denotes weight matrices, $\mathbf{b}$ denotes bias vectors, $\mathcal{H}$ is the hidden layer activation function (typically the logistic sigmoid function), and the subscripts $\{x, h, y\}$ denote the (input,hidden,output) layers respectively. LSTM replaces the hidden nodes in the recurrent model with a memory cell, which allows the recurrent network to remember long term context dependencies. Given an input vector $\mathbf{x}^{(t)}$ at each time t and the hidden state from previous time $\mathbf{h}^{(t-1)}$, we denote a linear mapping as:

$$\phi_\star^{(t)} = \mathbf{W}_\star \mathbf{x}^{(t)} + \mathbf{R}_\star \mathbf{h}^{(t-1)} + \mathbf{b}_\star, \quad (3)$$

where $\mathbf{W}$ is the rectangular input weight matrices, $\mathbf{R}$ is the square recurrent weight matrices, and $\star$ denotes one of LSTM components $\{c, f, i, o\}$, *i.e.*, cell unit, forget gate, input gate, and output gate. Element-wise activation functions are applied to introduce nonlinearity. Gate units often use a *logistic sigmoid* activation $\sigma(a) = \frac{1}{1+e^{-a}}$; cell units are transformed with *hyperbolic tangent* $\tanh(\cdot)$. Denote the point-wise multiplication of two vectors as $\odot$,

LSTM applies the following update operations: (block input) $\mathbf{z}^{(t)} = \tanh(\phi_c^{(t)})$, (forget gate) $\mathbf{f}^{(t)} = \sigma(\phi_f^{(t)})$, (input gate) $\mathbf{i}^{(t)} = \sigma(\phi_i^{(t)})$, (output gate) $\mathbf{o}^{(t)} = \sigma(\phi_o^{(t)})$, (cell state) $\mathbf{c}^{(t)} = \mathbf{i}^{(t)} \odot \mathbf{z}^{(t)} + \mathbf{f}^{(t)} \odot \mathbf{c}^{(t-1)}$, and (block output) $\mathbf{h}^{(t)} = \mathbf{o}^{(t)} \odot \tanh(\mathbf{c}^{(t)})$. As seen in the update of cell states, an LSTM cell involves *summation* over previous cell states. The gradients are distributed over sums, and propagated over a longer time before vanishing. Because AU detection is by nature a *multi-label classification* problem, we optimize LSTMs to jointly predict multiple AUs according to the maximal-margin loss:

$$L_M(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{n_0} \sum_i \max(0, \lambda - y_i \widehat{y}_i), \tag{4}$$

where λ is a pre-defined margin, and n_0 indicates the number of non-zero elements in ground truth $\mathbf{y}$. Although typically $\lambda = 1$ (such as in regular SVMs), here we empirically choose $\lambda = 0.5$ because the activation function has squeezed the outputs into $[-1, 1]$, making the prediction value never go beyond $\lambda = 1$. During back propagation, we pass the gradient $\frac{\partial L}{\partial \widehat{y}_i} = -\frac{y_i}{n_0}$ if $y_i \widehat{y}_i < 1$, and $\frac{\partial L}{\partial \widehat{y}_i} = 0$ otherwise. At each time step, LSTMs output a vector indicating potential AUs.

Practical issues: There has been evidence that a deep LSTM structure preserves better descriptive power than a single-layer LSTM [13]. However, because fc7 features are of high-dimension (4096-D), our design of LSTMs can lead to a large model with >1.3 million parameters. To ensure that the number of parameters and the size of our datasets maintain the same order of magnitude, we applied PCA to reduce the fc7 features to 1024-D (preserving 98% energy). We set dropout rate as 0.5 to the input and hidden layers, resulting in a final model of ∼0.2 million parameters. More implementation details are in Sec. IV.

C. Frame-based spatiotemporal fusion

The spatial CNN performs AU detection from still video frames, while the temporal LSTM is trained to detect AUs from temporal transitions. Unlike video classification that produces video-based prediction (*e.g.*, [8], [29], [36]), we model the correlations between spatial and temporal cues by adding an additional fusion network. We modify the late fusion model [18] to achieve this goal. Fig. 1 shows an illustration. For each frame, two fully connected layers with shared parameters are placed on top of both CNNs and LSTMs. The fusion network merges the stacked L_2-normalized scores in the first fully connected layer. In experiments, we see this fusion approach consistently improves the performance compared to CNN-only results.

IV. Experiments

A. Datasets

We evaluated the proposed hybrid network on two large spontaneous datasets: BP4D [40] and GFT [4]. Each dataset was FACS-coded by certified coders. AUs occurring more than 5% base rate were included for analysis. In total, we selected 12 AUs to perform the experiments, resulting in >400,000 valid frames. Unlike previous studies that suffer from scalability issues and require downsampling of training data, the network is in favor of large dataset so we made use of all available data. Note that the CK+ benchmark [22] is not applicable because the AU annotations were given at video-level, while we aim at per-frame prediction.

BP4D [40] is a spontaneous facial expression dataset in both 2D and 3D videos. The dataset includes 41 participants associating with 8 interviews. Frame-level ground-truth for facial actions are obtained using the FACS. In our experiments, we used 328 2D videos from 41 participants, resulting in 146,847 available frames with AU coded. We selected positive samples as those with intensities equal or higher than A-level, and negative samples as the remaining.

GFT [4] contains 240 groups of three previously unacquainted young adults. Moderate out-of-plane head motion and occlusion are presented in the videos, making AU detection challenging. We used 50 participants with each containing one video of about 2 minutes (∼5000 frames), resulting in 254,451 available frames with AU coded. Frames with intensities equal or greater than B-level are used as positive, otherwise, intensities less than B-level are negative.

B. Settings

Pre-processing: We pre-processed all videos by extracting facial landmarks using IntraFace [5]. Tracked faces were registered to a reference face using similarity transform, resulting in 200×200 face images, which were then randomly cropped into 176×176 and/or flipped for data augmentation. Each frame was labeled $+1/-1$ if an AU is present/absent, and 0 otherwise (*e.g.*, lost face tracks or occluded face).

Dataset splits: For both datasets, we adopted a *3-fold* and a *10-fold* protocol. For 3-fold protocol, each dataset was evenly divided into 3 subject-exclusive partitions. We iteratively trained a model using two partitions and evaluated on the remaining one, until all subjects were tested. For 10-fold protocol, we followed standard train/validation/test splits as in the deep learning community (*e.g.*, [19], [29], [36]). In specific, we divided entire dataset into 10 subject-exclusive partitions, where 9 for training/validation and 1 for test. For both protocols, we used ∼20% training subjects for validation. By comparing 3-fold and 10-fold, we hope to examine the performance v.s. the number of training samples because the 10-fold protocol uses ∼30% more samples than the 3-fold. To measure the transferability of fc7 features, we also performed a *cross-dataset* protocol by training CNNs on one dataset and using it to extract spatial representations for training a classifier on another.

Evaluation metrics: We reported performance using frame-based F1-score (F1-frame=$\frac{2RP}{R+P}$) for comparisons with the literature, where R and P denote recall and precision, respectively. Other options of metrics include F1-norm, which computes a skew-normalized F1-frame by multiplying false negatives and true negatives by the factor of skewness (the ratio of positive samples over negative ones). In addition, because AUs occur as temporal signals, an event-based F1 (F1-event=$\frac{2ER \cdot EP}{(ER+EP)}$) can be used to measure detection performance at segment-level, where ER and EP are event-

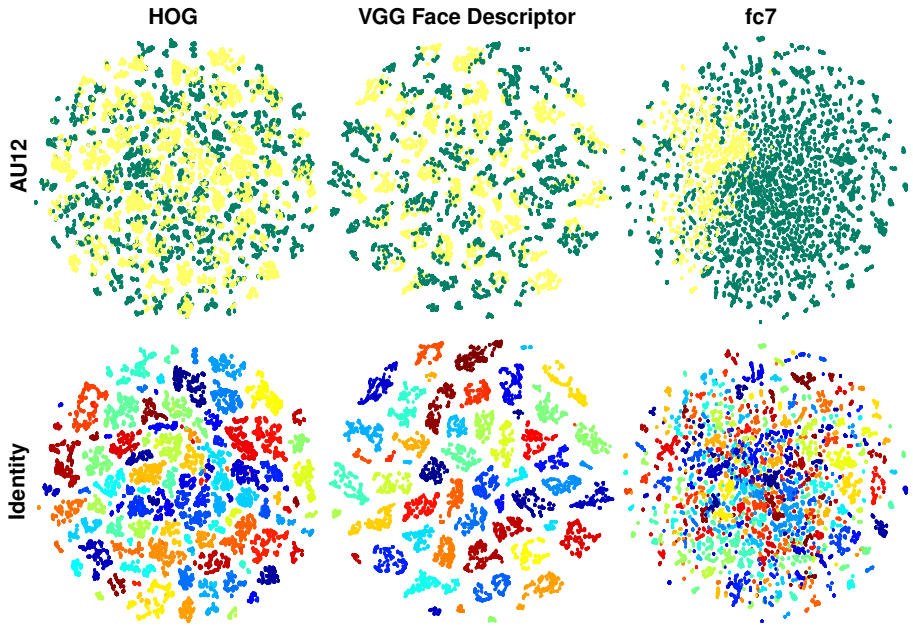

Fig. 3. A visualization of t-SNE embedding using HOG, VGG face descriptor [24] and fc7 features on the BP4D dataset [40] by coloring each frame sample in terms of AU12 (**top row**) and subject identities (**bottom row**). The clustering effect in HOG and VGG face descriptors reveal their encoded information about not only facial AUs but more subject identities. As can be seen, the separation between subjects of VGG descriptors is more clear than the separation of HOG, because VGG face descriptors were originally trained for face recognition. On the other hand, the learned fc7 features are optimized for multi-label AU classification, and thus reduce the influence caused by individual differences. More importantly, fc7 features maintain the grouping effect on samples of the same AU, implying its ability of capturing necessary information for AU classification. (best viewed in color)

based recall and precision as defined in [7]. Different metrics capture different properties about the detection performance. Choices of one or another metric depend on a variety of factors, such as purposes of the task, preferences of individual investigators, the nature of the data, etc. Due to space limitation, we only reported F1 in this paper.

Network settings and training: We trained the CNNs with mini-batches of 196 samples, a momentum of 0.9 and weight decay of 0.0005. All models were initialized with learning rate of 1e-3, which was further reduced manually whenever the validation loss stopped decreasing. The implementation was based on the Caffe toolbox [16] with modifications to support multi-label cross-entropy loss. For training LSTMs, we set an initial learning rate of 1e-3, momentum of 0.9, weight decay 0.97, and RMSProp for stochastic gradient descent. All gradients were computed using back-propagation through time (BPTT) on 10 subsequences randomly sampled from training video. All sequences were 1300 frames long, and the first 10 frames were disregarded during the backward pass, as they carried insufficient temporal context. In the end, our network went through about 10 passes over the full training set. The matrix $\mathbf{W}$ were randomly initialized within $[-0.08, 0.08]$. As AU data is heavily skewed, randomly sampling the sequences could cause LSTMs biased to negative predictions. As a result, we omitted training sequences with less than 1.5 active AUs per frame. We refer interested readers to the author's Ph.D. thesis for details about more sophisticated strategies on multi-label sampling. All experiments were performed using one NVidia Tesla K40c GPU.

C. Evaluation of learned representation

To answer the question whether individual differences can be reduced by feature learning, we first evaluated the fc7 features with standard features in AU detection, including shape (landmark locations), Gabor, and HOG features. Because such features for AU detection are unsupervised, for fairness, fc7 features for BP4D were extracted using CNNs trained on GFT, and vise versa. Fig. 3 shows the t-SNE embeddings of frames represented by HOG, VGG face descriptor [24] and fc7 features colored in terms of AU12 and subject identities. As can be seen in first and second columns, HOG and VGG face descriptors have strong distributional biases toward subject identity. On the other hand, as shown in the third column, although the network is learned on the other dataset, fc7 features show relative invariance to individual differences. More importantly, as shown in the plot of AU12, fc7 features maintain the grouping effect on samples of the same AU, implying its ability of capturing necessary information for classification.

As a quantitative evaluation, we treated the frames of the same subject as a distribution, and computed the distance between two subjects using Jensen-Shannon (JS) divergence [20]. Explicitly, we first computed a mean vector $\boldsymbol{\mu}_s$ for each subject s in the feature space, and then squeezed $\boldsymbol{\mu}_s$ using a logistic function $\sigma(a) = \frac{1}{1+e^{-a/m}}$ (m is median of $\boldsymbol{\mu}_s$ as the median heuristic) and unity normalization, so that each mean vector can be interpreted as a discrete probability distribution, *i.e.*, $\boldsymbol{\mu}_s \geq 0$, $\|\boldsymbol{\mu}_s\|_1 = 1, \forall s$. Given two subjects p and q, we computed their JS divergence as:

$$D(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q) = \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{\mu}_p || \mathbf{m}) + \frac{1}{2} D_{\mathrm{KL}}(\boldsymbol{\mu}_q || \mathbf{m}), \quad (5)$$

where $\mathbf{m} = \frac{1}{2}(\boldsymbol{\mu}_p + \boldsymbol{\mu}_q)$ and $D_{\mathrm{KL}}(\boldsymbol{\mu}_p, \mathbf{m})$ is the discrete KL divergence of $\boldsymbol{\mu}_p$ from $\mathbf{m}$. JS divergence is symmetric and smooth, and has been shown effective in measuring the dissimilarity between two distributions (*e.g.*, [34]). Higher value of $D(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q)$ tells larger mismatch given distributions for two subjects. Fig. 4 shows the statistics of distributional divergence over all subjects in one dataset, which was

Dataset	Shape	Gabor	HOG	fc7
BP4D	5.38±.40	4.63±.16	3.87±.12	3.58±.09
GFT	5.43±.39	4.74±.23	3.41±.25	0.89±.13

Fig. 4. Subject-invariance on the BP4D and GFT datasets in terms of a computed JS-divergence d normalized by $\log(d) \times 1e6$. (details in text)

computed by summing over $D(\boldsymbol{\mu}_p, \boldsymbol{\mu}_q), \forall q \neq p$. As can be seen, HOG consistently reached a lower divergence than Gabor, providing an evidence that local descriptor (HOG) is more robust to appearance changes compared to holistic ones (Gabor). This also serves as a possible explanation why HOG consistently outperformed Gabor (*e.g.*, [43]). Overall, fc7 yields much lower divergence compared to alternative engineered features, implying reduced individual differences.

D. Evaluation of detection performance

This section evaluates the performance of the proposed network on BP4D and GFT datasets. Below we summarize alternative methods, and then provide discussion.

Alternative methods: For evaluation, we compared a baseline HOG method, a standard multi-label CNN, and feature-based state-of-the-arts. The baseline HOG was used to train a linear SVM for each AU. This baseline has been shown to outperform other appearance descriptors (*i.e.*, Gabor/Daisy) [43]. Because HOG is unsupervised, for fairness, we evaluated a *cross-dataset* protocol that trained AlexNet on the other dataset, termed as ANet$^{\mathsf{T}}$. fc7 features extracted by ANet$^{\mathsf{T}}$ were then used in comparison with HOG descriptors. Linear SVMs served as the base classifier, which implicitly tells how separable each feature was, *i.e.*, higher classification rate suggests an easier linear separation, and validates that a good representation could reduce the burden of designing a sophisticated classifier. We evaluated ANet$^{\mathsf{T}}$ on a 3-fold protocol, while we expect similar results could be obtained using 10-fold.

Another alternative is our modified AlexNet (ANet), as mentioned in Sec. III-A, with slightly different architecture and loss function (multi-label cross-entropy instead of multi-class softmax). ANet stood for a standard multi-label CNN, a representative of *feature learning* methods. On the other hand, CPM [39] and JPML [41] are feature-based state-of-the-art methods that were reported on the two datasets. Both CPM and JPML used HOG features [39], [41]. They differ in attacking the AU detection problem from different perspectives. CPM is one candidate method of *personalization*, which aims at identifying reliable training samples for adapting a classifier that best separates samples of a test subject. On the other hand, JPML models *AU correlations*, and meanwhile considers patch learning to select important facial patches for specific AUs. We ran all experiments following protocols in Sec. IV-B.

Results and discussion: Tables I and II show F1 metrics reported on 12 AUs; "Avg" for the mean score of all AUs. According to the results, we discuss our findings in hope to answer three fundamental questions:

1) *Could we learn a representation that better generalizes across subjects or datasets for AU detection?* On both datasets, compared to HOG, ANet$^{\mathsf{T}}$ trained with a cross-dataset protocol on average yielded higher scores with a few exceptions. In addition, for both 3-fold and 10-fold protocols where ANet was trained on exclusive subjects, ANet consistently outperformed HOG over all AUs. These observations provide an encouraging evidence that the learned representation was transferable even when being tested across subjects and datasets, which also coincides with the findings in the image and video classification community [18], [29], [30]. On the other hand, as can be seen, ANet trained within datasets leads to higher scores than ANet$^{\mathsf{T}}$ trained across datasets. This is because of the dataset biases (*e.g.*, recording environment, subject background, etc.) that could cause distributional shifts in the feature space. In addition, due to the complexity of deep models, the performance gain of ANet trained on more data (10-fold) became larger than ANet trained on 3-fold, showing the generalizability of deep models increases with the growing number of training samples. Surprisingly, compared to HOG trained on 10-fold, ANet trained on 3-fold showed comparable scores, even with $\sim$30% fewer data than what HOG was used. All suggests that features less sensitive to the identity of subjects could improve AU detection performance.

2) *Could the learned temporal dependencies improve performance, and how?* The learned temporal dependencies was aggregated into the hybrid network denoted as "ours". On both 3-fold and 10-fold protocols, our hybrid network consistently outperformed ANet in all metrics. This improvement can be better told by comparing their F1-event scores. The proposed network used CNNs to extract spatial representations, stacked LSTMs to model temporal dependencies, and then performs a spatiotemporal fusion. From this view, predictions with fc7 features can be treated as a spacial case of ANet—a linear hyperplane with a portion of intermediate features. In general, adding temporal information helped predict AUs except for a few in GFT. A possible explanation is that in GFT, the head movement was more frequent and dramatic, and thus makes temporal modeling of AUs more difficult than moderate head movements in BP4D. In addition, adding temporal prediction into the fusion network attained an additional performance boost, leading to the highest F1 score on both datasets with either the 3-fold or the 10-fold protocols. This shows that the spatial and temporal cues are complementary, and thus is crucial to incorporate all of them into an AU detection system.

3) *Would jointly considering all issues in one framework improve AU detection?* This question aims to examine if the hybrid network would improve the performance of the methods that consider the aforementioned issues independently. To answer this question, we implemented CPM [39] as a personalization method that deals with representation issues, and JPML [41] as a multi-label learning method that deals with AU relations. Our modified ANet served as a feature learning method. All parameters settings were determined following the descriptions in the original papers. To draw a valid discussion, we fixed the exact subjects for all methods. Observing 3-fold on both datasets, the results are mixed. In

TABLE I

F1-FRAME ON GFT DATASET [4]

	3-fold protocol					cross	10-fold protocol				
AU	HOG	CPM	JPML	ANet	Ours	ANet$^{\intercal}$	HOG	CPM	JPML	ANet	Ours
1	12.1	30.7	17.5	**31.2**	29.9	9.9	30.3	29.9	28.5	57.5	**63.0**
2	13.7	30.5	20.9	29.2	25.7	10.8	25.6	25.7	25.5	61.4	**74.6**
4	5.5	–	3.2	**71.9**	68.9	45.4	–	–	–	**75.9**	68.5
6	30.6	61.3	**70.5**	64.5	67.3	46.2	66.2	67.3	73.1	61.6	**66.3**
7	26.4	70.3	65.5	67.1	**72.5**	51.5	70.9	72.5	70.2	**80.1**	74.5
10	38.4	65.9	**67.9**	42.6	67.0	23.5	65.5	67.0	67.1	54.5	**70.3**
12	35.2	74.0	74.2	73.1	**75.1**	55.2	74.2	75.1	78.3	**79.8**	78.2
14	55.8	**81.1**	52.4	69.1	80.7	62.8	79.6	80.7	61.4	**84.2**	80.4
15	9.5	25.5	20.3	27.9	**43.5**	14.2	34.1	43.5	28.0	40.3	**50.5**
17	31.3	44.1	48.3	**50.4**	49.1	34.2	49.2	49.1	42.4	61.6	**61.9**
23	19.5	19.9	31.8	34.8	**35.0**	21.8	28.3	35.0	29.6	47.0	**58.2**
24	12.9	27.2	28.5	**39.0**	31.9	18.9	31.9	31.6	28.0	**56.3**	50.8
Avg	24.2	48.2	41.8	50.0	**53.9**	32.9	50.5	52.4	48.4	63.4	**66.4**

TABLE II

F1-FRAME METRICS ON BP4D DATASET [40]

	3-fold protocol					cross	10-fold protocol				
AU	HOG	CPM	JPML	ANet	Ours	ANet$^{\intercal}$	HOG	CPM	JPML	ANet	Ours
1	21.1	**43.4**	32.6	40.3	31.4	32.7	46.0	46.6	33.9	54.7	**70.3**
2	20.8	**40.7**	25.6	39.0	31.1	26.0	38.5	38.7	36.2	56.9	**65.2**
4	29.7	43.3	37.4	41.7	**71.4**	29.0	48.5	46.5	42.2	**83.4**	83.1
6	42.4	59.2	42.3	62.8	**63.3**	61.9	67.0	68.4	62.9	94.3	**94.7**
7	42.5	61.3	50.5	54.2	**77.1**	59.4	72.2	73.8	69.9	93.0	**93.2**
10	50.3	62.1	72.2	**75.1**	45.0	67.4	72.7	74.1	72.5	98.9	**99.0**
12	52.5	68.5	74.1	78.1	**82.6**	76.2	83.6	84.6	72.0	94.4	**96.5**
14	35.2	52.5	65.7	44.7	**72.9**	47.1	59.9	62.2	62.6	82.9	**86.8**
15	21.5	**36.7**	38.1	32.9	34.0	21.7	41.1	44.3	38.2	55.4	**63.3**
17	30.7	**54.3**	40.0	47.3	53.9	47.1	55.6	57.5	46.5	81.1	**82.7**
23	20.3	**39.5**	30.4	27.3	38.6	21.6	40.8	41.7	38.3	63.7	**73.5**
24	23.0	37.8	42.3	**40.1**	37.0	31.3	42.1	39.7	41.5	74.3	**81.6**
Avg	32.5	50.0	45.9	48.6	**53.2**	43.4	55.7	56.5	51.4	77.8	**82.5**

GFT, ANet and JPML achieved 3 and 2 highest F1 scores; in BP4D, CPM and ANet reached 5 and 2 highest F1 scores. An explanation is because, although CNNs possess high degree of expressive power, the number training samples in 3-fold (33% left out for testing) were insufficient and might resulted in overfitting. In the 10-fold experiment, when training data was abundant, the improvements became clearer, as the parameters of the complex model can better fit our task. Overall, in most cases, our hybrid network outperformed alternative approaches by a significant margin, showing the benefits for considering all issues in one framework.

E. Visualization of learned AU models

To better understand and interpret the proposed network, we implement a gradient ascent approach [28], [38] to visualize each AU model. More formally, we solve for such input image $\mathcal{I}^\star$ by solving the optimization problem:

$$\mathcal{I}^\star = \arg\max_{\mathcal{I}} A_\ell(\mathcal{I}) - \Omega(\mathcal{I}), \tag{6}$$

where $A_\ell(\mathcal{I})$ is an activation function for the ℓ-th unit of the fc8 layer given an image $\mathcal{I}$, and $\Omega(\cdot)$ is a regularization function that penalizes $\mathcal{I}$ to enforce a natural image prior. In particular, we implemented $\Omega(\cdot)$ as a sequential operation of L_2 decay, clipping pixels with small norm, and Gaussian blur [38]. The optimization was done by iteratively updating a randomized and zero-centered image with the backprop gradient of $A_\ell(\mathcal{I})$. In other words, each pixel of $\mathcal{S}$ was renewed gradually to increase the activation of the ℓ-th AU. This process continued until 10,000 iterations.

Fig. 5 shows our visualizations of each AU model learned by the CNN architecture described in Sec. III-A. As can be seen, most models match the attributes described in FACS [10]. For instance, model AU12 (lip corner puller) exhibits a strong "$\smile$" shape to the mouth, overlapped with some vertical "stripes", implying the appearance of teeth is commonly seen in AU12. Model AU14 (dimpler) shows the dimple-like wrinkle beyond lip corners, which, compared to AU12, gives the lip corners a downward cast. Model AU15 (lip corner depressor) shows a clear "$\frown$" shape to the mouth, producing an angled-down shape at the corner. For upper face AUs, model AU6 (cheek raiser) captures deep texture of raised-up cheeks, narrowed eyes, as well as a slight "$\smile$" shape to the mouth, suggesting its frequent co-occurrence with AU12 in spontaneous smiles. Models AU1 and AU2 (inner/outer brow raiser) both capture the arched shapes to the eyebrows, horizontal wrinkles above eyebrows, as well as the widen eye cover that are stretched upwards. Model AU4 (brow lowerer) captures the vertical wrinkles between the eyebrows and narrowed eye cover that folds downwards.

Our visualizations suggest that the CNN was able to identify these important spatial cues to discriminate AUs, even though we did not ask the network to specifically learn these AU attributes. Furthermore, the global structure of a face was actually preserved throughout the network, despite that convolutional layers were designed for local abstraction (*e.g.*, corners and edges as shown in Fig. 2(d)). Lastly, the widespread agreements between the synthetic images and FACS [10] confirm that the learned representation is able to describe and reveal co-occurring attributes across multiple AUs. We believe such AU co-occurrence is captured due to the multi-label structure in the proposed network. This was not shown possible in standard hand-crafted features in AU detection (*e.g.*, shape [15], [22], HOG [39], [41], LBP [17], [33], or Gabor [33]). To the best of our knowledge, this is the first time to visualize how machines see facial AUs.

V. CONCLUSION AND FUTURE WORK

We have presented a hybrid network that jointly learns 3 factors for multi-label AU detection: *Spatial representation*, *temporal modeling*, and *AU correlation*. To the best of our knowledge, this is the first study that shows a possibility of laerning the three seemingly unrelated aspects within one framework. The hybrid network is motivated by existing progress on deep models, and takes advantage of spatial CNNs, temporal LSTMs, and their fusions to achieve multi-label AU detection. Experiments on two large spontaneous AU datasets demonstrate the performance over a standard CNN and feature-based state-of-the-art methods. In addition, our visualization of learned AU models showed, for the first time, how machines see each facial AU. Future work include deeper analysis of the hybrid network (especially the LSTM portion), incorporation of bi-directional LSTMs, training an entire network in one-go, and compare the proposed model between single-label and multi-label settings.

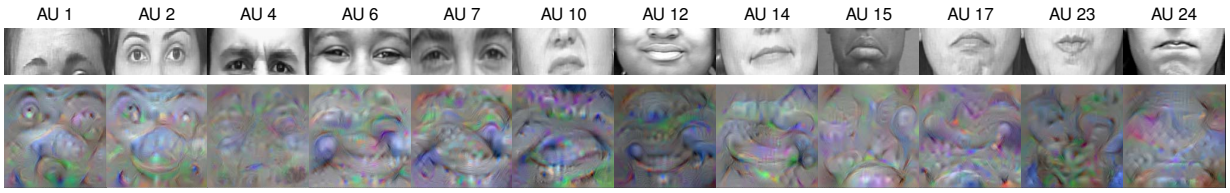

Fig. 5. Synthetically generated images to maximally activate individual AU neurons in the fc8 layer of CNN, trained on GFT [4], showing what each AU model "wants to see". The learned models show high agreement on attributes described in FACS [10]. (best view electronically)

Acknowledgment: This work was supported in part by US National Institutes of Health grants GM105004 and MH096951. The authors also thank NVIDIA for supporting this research with a Tesla K40c GPU, and Jiabei Zeng and Kaili Zhao for assisting partial experiments.

REFERENCES

[1] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *CVPR*, 2009.

[2] J. Chen, M. Kim, Y. Wang, and Q. Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *CVPR*, 2009.

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *CVPR*, 2013.

[4] J. F. Cohn and M. A. Sayette. Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. *Behavior Research Methods*, 42(4):1079–1086, 2010.

[5] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. IntraFace. In *Automatic Face and Gesture Recognition*, 2015.

[6] F. De la Torre and J. F. Cohn. Facial expression analysis. *Visual Analysis of Humans: Looking at People*, page 377, 2011.

[7] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang. Facial action unit event detection by cascade of tasks. In *IEEE Conference on International Conference on Computer Vision*, 2013.

[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[9] S. Du and A. M. Martinez. Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in clinical neuroscience*, 17(4):443, 2015.

[10] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[11] S. Eleftheriadis, O. Rudovic, and M. Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *ICCV*, 2015.

[12] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *ACII*, 2015.

[13] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013.

[14] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *AFGR*, 2015.

[15] S. Jaiswal and M. F. Valstar. Deep learning the dynamic appearance and shape of facial action units. 2016.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[17] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *AFGR*, 2011.

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[20] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. on Information Theory*, 37(1):145–151, 1991.

[21] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, 2014.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[23] A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *JMLR*, 13:1589–1608, 2012.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[25] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *TPAMI*, 37(5):944–958, 2015.

[26] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *ACM MM*, 2014.

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *TPAMI*, 37(6):1113–1133, 2015.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[29] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *CVPR*, 2015.

[31] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *TPAMI*, (10):1683–1699, 2007.

[32] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(4):966–979, 2012.

[33] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *ICCV*, 2013.

[34] A. Wong and M. You. Entropy and distance of random graphs with application to structural pattern recognition. *TPAMI*, (5):599–609, 1985.

[35] B. Wu, S. Lyu, B.-G. Hu, and Q. Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7):2279–2289, 2015.

[36] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*, 2015.

[37] S. Yang, O. Rudovic, V. Pavlovic, and M. Pantic. Personalized modeling of facial action unit intensity. In *Advances in Visual Computing*, pages 269–281. Springer, 2014.

[38] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. 2015.

[39] J. Zeng, W.-S. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. In *ICCV*, 2015.

[40] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *AFGR*, 2013.

[41] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *CVPR*, 2015.

[42] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.

[43] Y. Zhu, F. De la Torre, J. F. Cohn, and Y.-J. Zhang. Dynamic cascades with bidirectional bootstrapping for spontaneous facial action unit detection. *IEEE Trans. on Affective Computing*, 2:79–91, 2011.