Cross-domain AU Detection: Domains, Learning Approaches, and Measures

Itir Onal Ertugrul[1], Jeffrey F. Cohn[2], László A. Jeni[1], Zheng Zhang[3], Lijun Yin[3] and Qiang Ji[4]
[1] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
[3] Department of Computer Science, State University of New York at Binghamton, USA
[4] Rensselaer Polytechnic Institute, Troy, NY, USA

***Abstract*— Facial action unit (AU) detectors have performed well when trained and tested within the same domain. Do AU detectors transfer to new domains in which they have not been trained? To answer this question, we review literature on cross-domain transfer and conduct experiments to address limitations of prior research. We evaluate both deep and shallow approaches to AU detection (CNN and SVM, respectively) in two large, well-annotated, publicly available databases, Expanded BP4D+ and GFT. The databases differ in observational scenarios, participant characteristics, range of head pose, video resolution, and AU base rates. For both approaches and databases, performance decreased with change in domain, often to below the threshold needed for behavioral research. Decreases were not uniform, however. They were more pronounced for GFT than for Expanded BP4D+ and for shallow relative to deep learning. These findings suggest that more varied domains and deep learning approaches may be better suited for promoting generalizability. Until further improvement is realized, caution is warranted when applying AU classifiers from one domain to another.**

I. INTRODUCTION

People communicate emotion, intentions, and physical states using facial expressions. Automatic detection of facial expressions is crucial in many areas: mental and physical health, education, and human-computer interaction among others. The most comprehensive method to annotate facial expression is the anatomically based Facial Action Coding System (FACS) [1], [2]. FACS action units (AU) alone or in combinations can describe nearly all possible facial expressions. Automatic detection of FACS action units has been an active area of research [3], [4], [5].

Studies typically evaluate performance of AU detection models by cross validating algorithms within independent partitions of the same domain. A domain may consist of one or more databases that are used in both training and testing. In this way, classifiers are evaluated by how well they generalize, or transfer, to unseen subsets of the domain in which they were trained. Cross-validation within domains protects against overfitting but cannot ensure generalizability to new domains.

In many applications we are interested in applying AU detectors to new domains. For instance, we might wish to apply a classifier trained in posed facial expressions of a single participant to spontaneous expressions of a group of participants. For domain transfer, differences between domains become relevant. Domains may differ in multiple ways. These may include context (e.g., participants alone or interacting with other participants), individual differences (e.g., gender, ethnicity, and age), orientation to camera, non-rigid head motion, lighting, video resolution, and base rates and intensity of specific action units (that is, how frequently and for how long they occur). All of these factors potentially influence AU detection.

To evaluate state of the art in domain transfer of AU detectors, we first review previous research. Our review identifies factors that leave in question the generalizability of AU detectors. These factors include lack of AU-specific findings, differences in data sampling and performance metrics, and relatively small numbers of subjects, which can attenuate performance, These factors are reviewed in Section II.

Taking these factors into account, we then investigate cross-domain generalizability using two large well-annotated databases that differ in context (inductions of varied emotions versus social interaction among previously unacquainted participants), individual differences among participants (e.g., sex, age, and ethnicity), orientation to the camera, non-rigid head motion, frequency and intensity with which various action units occur, and other factors. The databases are an expanded version of BP4D+ [6] and the Sayette Group Formation Task (i.e. GFT below) [7]. To explore whether models trained with one database generalize better than ones trained with another, we perform cross-domain experiments in both directions. To ensure that findings are not classifier specific, we use both deep and shallow approaches to AU detection. For the deep approach, we use a multi-label convolutional neural network; for the shallow approach we use the handcrafted features and support vector machine of Openface [8]. Openface is a state-of-the-art shallow approach that was trained to optimize AU detection performance. Because different test statistics quantify different aspects of performance, we report a variety of metrics. These include S score [9], [10], AUC, F1 (which is positive agreement when comparing two methods) and negative agreement (NA).

To summarize, we:

- Review the literature on cross-domain AU detection and identify current issues in inferring AU-specific transfer.
- Investigate cross-domain AU-specific generalizability in two large, well-annotated databases using both a deep

and a shallow approach to AU detection.
- Report a variety of metrics that quantify different aspects of performance.
- Compare AU-specific generalizability of different databases and of shallow and deep approaches to AU detection.
- Make available an expanded version of BP4D+ database (referred to as EB+ below). EB+ includes 2D video and frame-level AU annotation for 200 participants.
- Make code for the CNN publicly available.

II. RELATED WORK

Action unit detection has been studied extensively for nearly two decades [3], [5], [4]. Until recently, most approaches have used hand-crafted features. Examples include LBP [41], SIFT [42], [43], LGBP [44], HOG [45] and LBP-TOP [46]. With the emergence of deep learning, CNN methods have shown significant success for AU detection [47], [48]. Except for studies listed in Table 1, almost all work in AU detection has focused on within-domain performance. For many purposes, however, we wish to apply AU detectors learned in one domain to new domains. As in the related field of speech recognition, the impact of AU detection will be determined in large part by how well it can perform reliably when applied to new domains.

Table I summarizes studies that evaluate cross-domain AU detection. Some [36], [41] propose novel adaptation approaches. Most test domain transfer without adaptation. Comparisons among these studies with respect to generalizability of specific AU detectors is confounded by at least four factors.

One is the lack of AU specific cross-domain results. While many studies [23], [24], [25], [26], [27], [28], [29] report detailed within-domain results for each AU, AU-specific cross-domain results are seldom reported. Cross-domain results are limited to averages computed across all AUs. Measures aggregated across multiple AUs mask AU-specific findings.

Two, even when AU-specific results are reported, comparisons between studies are confounded by use of different performance metrics. Some studies [36], [35], [37], [39], [38], [40] use AU-specific frame-level F1s, others AUC [32], 2AFC [31], [33], or accuracy [31], [34]. These measures are not interchangeable. Lack of standard metrics also undermines comparisons of studies that report only average performance across multiple AUs. Some report precision [23], [24] while others report recall [23], [24] or Hamming loss [27], [28]. Without fungible metrics, results between studies lack comparability.

Three, comparisons between studies often are confounded by differences in the numbers of subjects, sequences, or frames sampled within common domains. Differences in the sampling of frames are common. For instance, two studies [30], [37] used CK to train classifiers and MMI to test them but used different numbers of subjects (11 [30] and 70 [37], respectively) from MMI. Similarly, three studies [31], [38], [39] used the same 41 subjects in BP4D to train their model and the same 27 subjects of DISFA to test it, but they used different frames. The number of frames for testing in DISFA was 4845 [31], 130K [38] and 64K [39]. These confound comparisons between studies.

And four, classifiers often are trained on relatively small databases, which impairs generalizability. Within-database results can be low when the number of subjects is insufficient [49]. The same is likely true with respect to generalizability across domains. To make strong inferences about generalizability, relatively large numbers of subjects are necessary in the training. Moreover, some databases may yield greater genralizability than others. At minimum generalizability should be compared for at least two databases.

III. METHOD

To investigate AU-specific cross-domain transfer, we use both deep and shallow approaches in two databases that represent different domains. The deep approach is a CNN architecture [50]; the shallow approach is a support vector machine (SVM) with hand-crafted features. One database is an extended version of BP4D+ [6]. The other is GFT [7]. As noted above, they differ in context (emotion induction by an experimenter versus a group formation task of multiple participants), individual differences among participants, non-rigid head motion, video resolution, composition of the FACS coding teams, and other factors. Both databases are well annotated and relatively large although not same (200 participants in EB+ and 150 in GFT). To ensure comparability between deep and shallow approaches, the same video frames and train and test assignments were used for both.

For the CNN, we report both within- and cross domain AU-specific results for both databases. For the shallow approach (Openface), we report cross-domain results to GFT but not to EB+. Because the release version of Openface was trained in part on BP4D, domain transfer to EB+ would be confounded by domain contamination. Preprocessing steps and AU detection methods of both the CNN and Openface are described below.

A. Deep Approach: Convolutional Neural Network

1) Face tracking and registration: Video was tracked and normalized using ZFace [51], 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 200×200 pixel image of faces with 80 pixels IOD.

2) Video-specific normalization: Because videos of multiple people are used to train and test the models, individual differences in appearance could influence the models. To reduce variation introduced by person specific appearance and highlight variation in facial expression, we subtract the mean frame of each video from all frames of that video. Considering that variation caused by change in pose is eliminated in the registration step, static pixels that do

TABLE I: Studies reporting cross-database AU detection results. $D_1 \rightarrow D_2$ denotes that models are trained on domain D_1 and tested with domain D_2. The column titled AU specific represents whether the study reports AU specific cross-domain performance (Yes) or average performance (No). Used evaluation metrics include 2AFC, AUC, F1, Classification Rate (CR), Recall (RC), Precision (PR), Hamming Loss, Average positive recognition rate (APRR), Average false-alarm rate (AFAR). Used databases include Cohn-Kanade (CK) [11], Extended Cohn-Kanade (CK+) [12], BP4D [13], UNBC Shoulder-Pain Archive (SP) [14], MMI [15], DISFA [16], SEMAINE (SEM) [17], SAL [18], GFT [7], RU-FACS [19], GEMEP-FERA (G-FERA) [20], ISL [21]. For more comprehensive review, see [22].

Study	Databases	Number of subjects	Number of sequences (s) / frames (f)	Number of AUs	AUs	AU specific	Metrics
[23]	MMI→CK SAL→SEM	MMI (10) SAL (10),	MMI (264 s), CK+ (55 s) SAL (35 s), SEM (10 s)	15	Avg of 15 AUs (not specified)	No	2AFC, F1 CR, RC, PR
[24]	MMI→CK	MMI (15)	MMI (264 s) CK (143 s)	18	1, 2, 4, 5, 6, 7, 9 10, 11, 12, 14, 15, 17 20, 24, 25, 27, 45	No	F1, CR, RC, PR
[25]	CK→G-FERA G-FERA→CK	CK (>100)	CK (8000 f) G-FERA (5000 f)	8	1, 2, 4, 6, 7, 12, 15, 17	No	F1
[26]	CK+, G-FERA, SP, DISFA (Train on one, test on the rest)	CK+ (123), G-FERA (7), SP (25), DISFA (27)	CK+ (593 s, 593*4 f), G-FERA (87 s), SP (200 s)	14	1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 17, 20, 25, 26	No	AUC, F1
[27], [28]	CK+→ISL	ISL (7)	ISL (7*19 s), CK+ (327 s, 327 * 2 f)	13	1, 2, 4, 5, 6, 7, 9 12, 17, 23, 24, 25, 27	No	Hamming L., F1
[29]	CK+→SEM	CK+ (210)	CK+ (593 s, 593 f)	10	1, 2, 4, 5, 6, 7, 12, 17, 25, 26	No	F1
[30]	CK→MMI	MMI (11), CK (>100)	MMI (54 s)	13	1, 2, 4, 5, 6, 7 9, 12, 15, 17, 23, 25, 27	Yes	APRR, AFAR
[31]	BP4D→CK+, BP4D→DISFA, DISFA→CK+, DISFA→BP4D	CK+ (123), DISFA (27), BP4D (41)	CK+ (582 s), DISFA (4845 f)	10	1, 2, 4, 5, 6, 9 12, 15, 17, 20	Yes	ACC, 2AFC
[32]	CK+→SP	CK+ (123), SP (25)	CK+ (593 s, 593 f), SP (48,398 f)	6	4, 6, 7, 9, 10, 43	Yes	AUC
[33]	CK+→G-FERA, G-FERA→CK+	CK+ (123), G-FERA (10)	CK+ (593 s CK+ ,593 f)	17	1, 2, 4, 5, 6, 7, 9 11, 12, 15, 17, 20 23, 24, 25, 26, 27	Yes	2AFC
[34]	DISFA→G-FERA, G-FERA→DISFA	DISFA (27), G-FERA (7)	DISFA (32 s, 32*4000 f), G-FERA (87 s)	8	1, 2, 4, 6, 12, 17, 25, 26	Yes	CR (Per seq)
[35]	BP4D→GFT GFT→BP4D	BP4D (41), GFT (50)	BP4D (328 s, 146,847 f), GFT (254,451 f)	12	1, 2, 4, 6, 7, 10, 12, 14, 15 17, 23, 24	Yes	F1
[36]	RU-FACS→G-FERA, GFT→RU-FACS	RU-FACS (34), G-FERA (7), GFT (42)	G-FERA (87 s), Ru-FACS (29 s, 180K f) , GFT (∼302K f)	8	1, 2, 4, 6, 12, 14, 15, 17	Yes	AUC, F1
[37]	MMI→CK , CK→MMI	MMI (70)	MMI (244 s), CK (153 s)	16	1, 2, 4, 5, 6, 7, 9, 10, 12, 15, 20, 24, 25, 26, 27, 45	Yes	F1
[38]	SEM, BP4D, DISFA (Train on one, test on the rest)	BP4D (41), SEM (31), DISFA (27)	BP4D (150K f), SEM (93K f), DISFA (130K f),	8	2, 12, 17 (all) 25 (DISFA→SEM), 1, 4, 6, 15 (DISFA→BP4D)	Yes	F1
[39]	BP4D→DISFA	BP4D (41) DISFA (27)	BP4D (328 s, 328*300 f) DISFA (27 * 2400 f)	8	1, 2, 4, 6, 9, 12, 25, 26	Yes	AUC, F1
[40]	BP4D→DISFA DISFA→BP4D	DISFA (27) BP4D (41)	Varies between (10 - 500 f)	7	1, 2, 4, 6, 12, 15, 17	Yes	AUC, F1

not change with the expression will be black after mean image subtraction. We thus eliminate regions that do not vary greatly so that our models will focus only on the dynamic parts of the frames, which are changing with the expression.

3) AU Detection: We trained a convolutional neural network (CNN) containing three convolutional layers and two fully connected layers (see Figure 1). Frames obtained after video-specific normalization are converted into grayscale

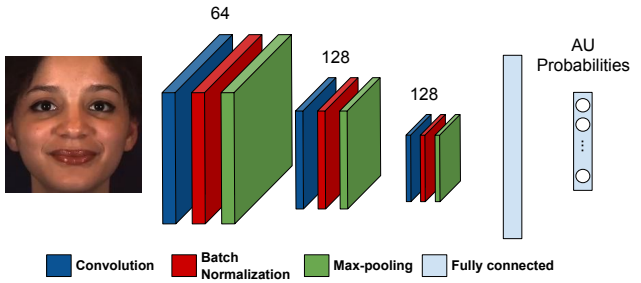

Fig. 1: Overview of the deep network used for within-domain and cross-domain experiments.

images and fed as inputs to the network. We employ 64, 128, and 128 filters of 5×5 pixels in three convolutional layers with a stride of 2, 1 and 1, respectively. After convolution, rectified linear unit (ReLU) is applied to the output of the convolutional layers in order to add non-linearity to the model. We apply batch normalization to the outputs of all convolutional layers. The network contains three max-pooling layers that are applied after batch normalization. We apply max-pooling with a 2×2 window such that the output of max-pooling layer is downsampled with a factor of 2. Output of the last maxpooling layer is connected to the fully connected layer of size 400. Finally, the output of first fully connected layer is connected to the final layer having $N = 12$ neurons. A sigmoid activation function is used at the output of final dense layer for non-linearity[1].

Because we perform multi-label AU detection, we use binary cross-entropy loss as follows:

$$L = \sum_{n=1}^{N} \left[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n) \right]. \quad (1)$$

Values obtained at the output neurons are between [0,1], corresponding to the probability of 12 AUs. During test time, we assign the positive AU occurrence label to the instances with probability above a threshold, which is optimized on training set.

B. Shallow Approach: Openface

OpenFace 2.0 [8] uses Convolutional Experts Constrained Local Model (CE-CLM) [52] for facial landmark detection and tracking. It employs HOG features extracted from similarity aligned 112×112 pixel face images and facial shape features for AU detection. It performs person-specific normalization, in which the median frame of a video is subtracted from all frames of the video, and prediction correction. Output of AU detection module of Openface is 0/1 label for absence/presence of each AU in each frame.

IV. Experiments

A. Databases

We performed experiments with two large spontaneous, well-annotated databases that differ in multiple ways. GFT [7] involves social interaction among groups of three previously unacquainted young adults. A third of the groups are drinking an alcoholic beverage; a third a placebo beverage; and a third fruit juice. Alcohol and placebo effects are common and have been reported previously [53], [54], [55]. The other database (EB+) is a series of emotion inductions of a single participant by an experimenter, which elicits more intense action units with different rates of occurrence. BP4D+ is reported in [6]. The databases differ as well in participant characteristics, range of head pose, non-rigid head motion, illumination, and video resolution.

Both databases were manually annotated by different teams of highly qualified FACS coders. We included 12 AUs that occurred in more than 3% of the frames in both databases. That is, AU 1, AU 2, AU 4, AU 6, AU 7, AU 10, AU 12, AU 14, AU 15, AU 17, AU 23, and AU 24. Because Openface does not output occurrence for AU 24, results for AU 24 are reported for the CNN only.

Expanded BP4D+ (EB+)[2] is a manually FACS annotated database of spontaneous behavior. Video is 2D with resolution of 1040×1392. Average video duration is around 44 seconds. Well-designed tasks (e.g. interviews, physical activities) initiated by an experimenter are used to elicit varied emotions. Face orientation is nearly frontal and out-of-plane head rotation is not common. It contains videos from a total of 200 subjects (140 subjects from BP4D+ [6], 60 additional subjects) associated with 5 to 8 tasks. We use a total of 1261 number of videos having a total of 395K frames. Positive samples are defined as the ones with intensities equal to or higher than B-level, and the remaining ones are negative samples.

GFT[3] [7] is a manually FACS annotated database of spontaneous behavior in 150 young adults in three-person groups. Behavior is unscripted and each video is approximately 2min in duration (approximately 517k frames in all). Moderate out-of-plane head motion is frequent and occlusion is common, making AU detection more challenging. Positive samples are defined as ones with intensities equal to or higher than B-level, and the remaining ones are negative samples.

B. Settings

Database splits We perform both within-domain and cross-domain experiments. In within-domain experiments, 5-fold cross validation is used. For EB+, each fold consists of 160 subjects for training and tuning and 40 subjects for testing. In GFT, each fold consists of 120 subjects for training and tuning and 30 subjects for testing. In cross-domain experiments, data from all subjects in the source domain is used for training; and data from all subjects in the other domain is used for testing.

Evaluation metrics Different metrics capture different properties about the AU detection performance. Choices of one or another metric depend on a number of factors, including preferences of investigators, purposes of the task,

[1] http://www.jeffcohn.net/resources/AFAR/

[2] http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html

[3] https://osf.io/7wcyz/

the nature of the data, etc. Following Girard and colleagues [7], we report a variety of metrics: S score (free-margin kappa), area under ROC curve (AUC), F1 and negative agreement (NA).

F1 is the most commonly used metric in AU detection literature. It is the harmonic mean of precision (P) and recall (R) $\frac{2RP}{R+P}$ which is also equivalent to positive agreement (PA) $\frac{2tp}{2tp+fp+fn}$ when only two methods are compared (e.g., CNN and manual AU coding). F1 can tell the performance on correct predictions on positive samples.

Negative agreement (NA) is the complement of F1 and is equal to $\frac{2tn}{2tn+fp+fn}$. It evaluates the solution by the harmonic agreement of samples not including AUs.

Area under the Receiver Operating Characteristics Curve (AUC) is equal to the probability that a classifier will rank a randomly chosen frame in which AU is present higher than a randomly chosen one in which AU is absent. Therefore, this measure shows the success of classifier to rank frames with and without AU. AUC was proven to be better than the accuracy metrics for evaluating classifier performance [56].

S score or free-marginal kappa coefficient is computed as $\frac{2tp+2tn}{tp+fp+fn+tn}$ [7]. It provides an overall, chance-adjusted summary statistic. It is equal to the ratio of observed nonchance-agreement to possible nonchance-agreement and it estimates chance agreement by assuming that each category is equally likely to be chosen at random.

Many of the AUs occur infrequently (i.e., have low base rates). S score and AUC are robust to imbalanced data while F1 and NA are not [57], which should be taken into account when evaluating results for AUs occur infrequently.

Network and training settings We trained CNNs with batches of 100 samples. We chose stochastic gradient descent optimizer with a learning rate of 1e-3 and a momentum of 0.9 for better generalizability to unseen domains. Our implementation is based on the PyTorch and we performed all experiments on NVidia 1080ti GPU.

C. Results

We report first within-domain CNN results for EB+ and GFT. Second, we compare within- and between domain CNN results in both databases. We then compare cross-domain results between CNN and Openface, which affords a comparison between a deep (CNN) and shallow (Openface) approach. For each set of comparisons we controlled for Type I error using Bonferroni correction. With experiment-wise error of 0.05 and $2 * 12 = 24$ comparisons in each set, a p of 0.002 is the critical value for significance.

1) Within-domain results: Table IIa and Table IIc show within-domain AU-specific results obtained by the CNN deep approach for EB+ and GFT databases, respectively. In the next to last row are reported averages across the 12 AUs outputted by CNN. For comparability with Openface, which outputs one fewer AU (AU 24), the last row shows the average of the 11 AUs that are common to both approaches.

Imbalanced classes are evident in both databases (see base rate (BR) columns in Table II. In EB+, seven of 12 AUs occur in fewer than 15 percent of frames. In GFT, five of 12 AUs occur in fewer than 15 percent of frames. This level of skew means fewer positive examples available for training and testing and decreases the range of F1 scores in particular [57]. Average F1 scores in both databases are in the moderate range. For AUs that occur in more than 15 percent of the frames, F1 scores are far better (0.75 to 0.88 in EB+ and 0.75 to 0.80 in GFT) and are higher in EB+ than in GFT.

AUC scores are consistently high in EB+ and moderate to high in GFT. The same pattern as found for F1 is found for AUC. AUC is higher for AUs that occur in more than 15 percent of the frames. The effect of base rate is likely due to the greater challenge of learning AUs that occur less frequently.

S scores (free-margin kappa) range from moderate to high in both databases. Although S scores show a less consistent relation to base rate, they show the same difference between databases. Results for EB+ generally are higher than those for GFT. Overall, most but not all S scores are within the range that is acceptable for observational research in psychology where kappa scores of 0.7 are expected. These findings are consistent with the hypothesis that AUs can be reliably detected within the same domains in which they were trained. They also suggest that some databases are easier or more difficult for AU detection. Sources of variation need to be better understood.

Within-domain results are better for EB+ than for GFT. Significance results in Table III reveal that, within-domain results on EB+ database is significantly better for 8 of the 12 AUs when S scores are compared and for 7 of the 12 AUs when AUC values are compared. For AU 6 and AU 12, within-domain results of both databases are similarly good.

2) Within-domain Cross-domain comparison: A critical question is whether AU-detectors generalize to new domains. Table IIb and Table IId reports AU specific cross-domain results for CNN. They report GFT to EB+ and EB+ to GFT.

When we compare within-domain and cross-domain results, we observe a decrease in average cross-domain results for both domains. Average AUC and F1 values are 0.864 and 0.631 for within EB+ (see Table IIa) while they are 0.719 and 0.458 for GFT $\rightarrow$ EB+ cross-domain (see Table IIb). Therefore, we observe decrease of 0.145 and 0.173 for AUC and F1, respectively. For each individual AU, there is a degradation in S score, AUC, F1 and NA values. Therefore, GFT does not generalize well to EB+ database. The highest F1 and AUC values for GFT $\rightarrow$ EB+ are obtained with AU 10 and AU 12, whose base rates are high.

On the contrary, average AUC and F1 values are 0.789 and 0.481 for within GFT (see Table IIc) while they are 0.736 and 0.463 for EB+ $\rightarrow$ GFT cross-domain (see Table IId). We observe a decrease of 0.053 and 0.018 for AUC and F1, respectively. S score and AUC values of EB+ $\rightarrow$ GFT are worse than within GFT results for all individual AUs. Although a decrease is observed for each individual AU, it is rather slight, meaning that testing GFT with a model

TABLE II: Within-domain and cross-domain AU detection results.

(a) Within-domain: EB+

-	BR	S	AUC	F1	NA
AU1	0.09	0.787	0.811	0.468	0.941
AU2	0.07	0.856	0.816	0.437	0.961
AU4	0.07	0.873	0.879	0.526	0.966
AU6	0.43	0.685	0.925	0.821	0.859
AU7	0.63	0.646	0.894	0.864	0.748
AU10	0.59	0.713	0.926	0.881	0.820
AU12	0.53	0.736	0.945	0.876	0.858
AU14	0.42	0.566	0.853	0.749	0.809
AU15	0.10	0.776	0.808	0.408	0.938
AU17	0.14	0.643	0.791	0.344	0.897
AU23	0.14	0.722	0.852	0.569	0.917
AU24	0.03	0.943	0.895	0.245	0.986
Average 12 AUs	**0.27**	**0.745**	**0.866**	**0.599**	**0.892**
Average 11 AUs	**0.29**	**0.727**	**0.864**	**0.631**	**0.883**

(b) Cross-domain: GFT $\rightarrow$ EB+

-	S	AUC	F1	NA
AU1	0.743	0.722	0.312	0.929
AU2	0.844	0.671	0.224	0.959
AU4	0.855	0.755	0.204	0.962
AU6	0.369	0.747	0.577	0.749
AU7	0.269	0.690	0.678	0.578
AU10	0.469	0.811	0.767	0.692
AU12	0.532	0.847	0.757	0.774
AU14	0.235	0.707	0.631	0.602
AU15	0.651	0.656	0.268	0.901
AU17	0.377	0.645	0.302	0.799
AU23	0.254	0.659	0.320	0.743
AU24	0.734	0.724	0.135	0.928
Average 12 AUs	**0.528**	**0.720**	**0.431**	**0.801**
Average 11 AUs	**0.509**	**0.719**	**0.458**	**0.790**

(c) Within-domain: GFT

-	BR	S	AUC	F1	NA
AU1	0.09	0.827	0.828	0.437	0.953
AU2	0.12	0.770	0.814	0.449	0.935
AU4	0.04	0.928	0.748	0.198	0.982
AU6	0.33	0.679	0.907	0.746	0.882
AU7	0.42	0.525	0.843	0.721	0.791
AU10	0.41	0.621	0.886	0.765	0.840
AU12	0.33	0.744	0.933	0.798	0.905
AU14	0.43	0.249	0.662	0.500	0.691
AU15	0.18	0.580	0.698	0.339	0.875
AU17	0.17	0.639	0.675	0.170	0.898
AU23	0.12	0.737	0.688	0.168	0.928
AU24	0.07	0.853	0.811	0.129	0.962
Average 12 AUs	**0.22**	**0.679**	**0.791**	**0.452**	**0.887**
Average 11 AUs	**0.24**	**0.664**	**0.789**	**0.481**	**0.880**

(d) Cross-domain: EB+ $\rightarrow$ GFT

-	S	AUC	F1	NA
AU1	0.741	0.729	0.258	0.929
AU2	0.597	0.720	0.338	0.881
AU4	0.817	0.710	0.180	0.952
AU6	0.562	0.852	0.688	0.832
AU7	0.251	0.791	0.666	0.573
AU10	0.490	0.850	0.728	0.759
AU12	0.541	0.866	0.703	0.813
AU14	0.083	0.651	0.621	0.420
AU15	0.314	0.622	0.324	0.770
AU17	0.219	0.646	0.334	0.724
AU23	0.669	0.660	0.248	0.907
AU24	0.533	0.725	0.231	0.862
Average 12 AUs	**0.485**	**0.735**	**0.443**	**0.785**
Average 11 AUs	**0.480**	**0.736**	**0.463**	**0.778**

(e) Cross-domain: Openface GFT

-	S	AUC	F1	NA
AU1	0.658	0.701	0.373	0.901
AU2	0.579	0.689	0.386	0.873
AU4	0.636	0.565	0.102	0.899
AU6	0.489	0.761	0.676	0.789
AU7	0.306	0.645	0.589	0.699
AU10	0.510	0.769	0.738	0.770
AU12	0.472	0.779	0.694	0.768
AU14	0.040	0.565	0.610	0.376
AU15	0.412	0.584	0.323	0.812
AU17	0.408	0.610	0.346	0.809
AU23	0.305	0.519	0.196	0.778
AU24	-	-	-	-
Average 12 AUs	**-**	**-**	**-**	**-**
Average 11 AUs	**0.438**	**0.653**	**0.458**	**0.770**

trained on EB+ can give good results. The model trained on EB+ generalizes well to GFT database.

Significance results in Table III reveal that, except for AU4, within-domain results are significantly better than cross-domain results for GFT when S score is used and for EB+ when AUC is used.

Recall that, EB+ has videos of larger number of individuals, higher base rates of AUs, and contains nearly frontal faces, while GFT has larger variation due to moderate head pose, making AU detection a more challenging problem. We can interpret our results in a way that, if a model is trained with a domain having infrequent AUs, it is likely to have generalizability problems on even relatively less challenging domains. However, if the model is trained with a more balanced domain, it can generalize better to others, provided that other variations, such as pose are minimized in the preprocessing step.

3) Cross-domain comparison of deep and shallow approaches: We report cross-domain results with deep approach and Openface on GFT. Training set of current release of Openface contains BP4D, whose tasks, base rates of AUs, pose and illumination conditions are the same with EB+. Therefore, we do not report test results using Openface with EB+ since it would not correspond to a cross-domain experiment.

Since we report AU specific detection results and test both models on the same domain, we can directly compare AU detection results of deep and shallow approaches. By comparing Table IId with Table IIe we can infer that, deep model gives slightly better S score, F1 and NA on average, while average AUC of deep approach is much higher than Openface. When we analyze F1s for individual AUs, deep approach outperforms Openface in all AUs except for AU1, AU2, AU10 and AU17. AUC values of deep approach are significantly ($p < 0.05$) better than the ones obtained with shallow approach for all AUs except for AU1. *S* values of AUs obtained with deep approach are generally better and they are significantly better than shallow approach for AU 1, AU 4, AU 6, AU 14, and AU 23 (see Table III).

Notice that, if we would only report AUC values as in [32], we would say that following a deep approach generalizes better compared to a shallow one. On the other hand, if we would only report F1s as in [37], [38], [35], we would infer that deep and shallow approaches perform similar for cross-

TABLE III: Significance of differences between classifiers by t-test. * is $p < 0.05$, ** is $p < 0.01$, *** is $p < 0.001$. The latter are significant after correcting for multiple comparisons. For comparison between deep and shallow, deep was greater than shallow except for the shaded cells.

-	Within EB+ > Within GFT		Within GFT > Cross on GFT		Within EB+ > Cross on EB+		Deep > Shallow	
AU	S	AUC	S	AUC	S	AUC	S	AUC
1	n.s.	n.s.	***	***	***	***	***	n.s.
2	***	n.s.	***	***	*	***	n.s.	*
4	n.s.	***	***	n.s.	n.s.	***	***	***
6	n.s.	n.s.	***	***	***	***	**	***
7	***	**	***	***	***	***	n.s.	***
10	***	*	***	***	***	***	n.s.	***
12	n.s.	n.s.	***	***	***	***	n.s.	***
14	***	***	***	***	***	***	*	***
15	***	***	***	***	***	***	***	***
17	***	***	***	***	***	***	***	**
23	***	***	***	***	***	***	***	***
24	***	n.s.	***	***	***	***	-	-

domain experiments. With a comparison of only S score values, we would conclude that deep approach is slightly better. Since we report results with all the measures for both approaches, we can interpret that, deep approach ranks instances with AUs present or absent much better, both deep and shallow approaches perform similar on positive instances and when the effect of chance is discarded, deep approach performs slightly better.

V. DISCUSSION AND FUTURE WORK

We reviewed studies that report cross-domain results and identified major problems in comparing generalizability of different approaches. These are failure to report AU specific results, variability in the number of subjects or frames used to obtain test results, and variability in the measures used to quantify performance. To overcome these problems, we recommend that investigators use comparable subjects and frames and report AU specific results using multiple measures that quantify varied aspects of performance. We recommend S score, AUC, F1, and NA on all available frames of the domain. With these recommendations, within- and cross-domain results can be rigorously compared.

To address limitations of previous research in AU-specific domain transfer, we performed cross-domain experiments using both a deep and a shallow approach using two large, well-annotated databases, namely EB+ and GFT, that differ from each other in key respects. Additional databases were initially considered (Bosphorus, BP4D, DISFA, SEMAINE, FERA, UNBC and CK+), but all had been used in training OpenFace. To control for experiment-wise error in statistical tests, we used Bonferroni correction.

In both deep and shallow approaches, we sought to maximize generalizability. For instance, we used video-specific normalization to reduce individual differences in appearance. And in the deep approach we used stochastic gradient descent, which has been shown to provide better generalizability to unseen domains. Even with such efforts, our results reflect that AU detectors that perform well within the same domain perform less well on new domains. The decrease occurred for all of the AUs examined. In many cases performance decreased to below the threshold acceptable for behavioral research.

Commercial systems, including iMotions, Affectiva and Noldus, profess to recognize AU and holistic facial expressions. Considering the low cross-domain generalizability of the state-of-the-art, we urge caution in applying such systems to new domains. Use in new domains should first be validated on a subset of manually annotated video. If systems fail this validation step, re-training is recommended. This is not possible with current commercial systems but is an option with OpenFace and the CNN used here.

All machine learning methods, whether shallow or deep, implicitly assume that representations and classifiers are drawn from the same domains [58]. When this assumption is violated, additional learning is required. Domain adaptation approaches for AU detection would be indicated.

VI. ACKNOWLEDGMENTS

This research was supported in part by NIH awards NS100549 and MH096951 and NSF awards CNS-1629716, CNS-1629898, and CNS-1629856.

REFERENCES

[1] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system: Research nexus network research information," *Salt Lake City, UT*, 2002.

[2] J. F. Cohn and P. Ekman, "Measuring facial action," *The new handbook of methods in nonverbal behavior research*, pp. 9–64, 2005.

[3] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE TPAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[4] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, 2017.

[5] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE TPAMI*, vol. 38, no. 8, pp. 1548–1568, 2016.

[6] 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 *CVPR*, 2016, pp. 3438–3446.

[7] J. M. Girard, W.-S. Chu, L. A. Jeni, and J. F. Cohn, "Sayette group formation task (gft) spontaneous facial expression database," in *FG*. IEEE, 2017, pp. 581–588.

[8] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *FG*. IEEE, 2018, pp. 59–66.

[9] E. M. Bennett, R. Alpert, and A. Goldstein, "Communications through limited-response questioning," *Public Opinion Quarterly*, vol. 18, no. 3, pp. 303–308, 1954.

[10] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and psychological measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[11] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *FG*. IEEE, 2000, p. 46.

[12] 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*. IEEE, 2010, pp. 94–101.

[13] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[14] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *FG*. IEEE, 2011, pp. 57–64.

[15] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, p. 5.

[16] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[18] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *LREC Workshop on Corpora for Research on Emotion and Affect*. ELRA, 2008, pp. 1–4.

[19] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, J. R. Movellan *et al.*, "Automatic recognition of facial actions in spontaneous expressions." *Journal of multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[20] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," in *ACII*. Springer, 2007, pp. 476–487.

[21] "Isl facial expression databases." rensslear Polytechnic Institute, Undated.

[22] J. F. Cohn, I. Onal Ertugrul, W.-S. Chu, J. M. Girard, L. A. Jeni, and Z. Hammal, "Affective facial computing: Generalizability across domains," in *Multimodal Behavior Analysis in the Wild*. Elsevier, 2019, pp. 407–441.

[23] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling." *IEEE Trans. Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.

[24] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE TPAMI*, vol. 32, no. 11, pp. 1940–1954, 2010.

[25] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *IEEE Transactions on affective computing*, vol. 4, no. 2, pp. 127–141, 2013.

[26] A. Ruiz, J. Van de Weijer, and X. Binefa, "From emotions to action units with hidden and semi-hidden-task learning," in *ICCV*, 2015, pp. 3703–3711.

[27] S. Wang, Q. Gan, and Q. Ji, "Expression-assisted facial action unit recognition under incomplete au annotation," *Pattern Recognition*, vol. 61, pp. 78–91, 2017.

[28] J. Wang, S. Wang, and Q. Ji, "Facial action unit classification with hidden knowledge under incomplete annotation," in *ICMR*. ACM, 2015, pp. 75–82.

[29] Y. Wu and Q. Ji, "Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection," in *CVPR*, 2016, pp. 3400–3408.

[30] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE TPAMI*, vol. 29, no. 10, 2007.

[31] 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*. IEEE, 2015, pp. 609–615.

[32] A. Mohammadian, H. Aghaeinia, F. Towhidkhah *et al.*, "Subject adaptation using selective style transfer mapping for detection of facial action units," *Expert Systems With Applications*, vol. 56, pp. 282–290, 2016.

[33] T. Gehrig and H. K. Ekenel, "Facial action unit detection using kernel partial least squares," in *ICCVW*. IEEE, 2011, pp. 2092–2099.

[34] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random field models for facial expression analysis," *Image and Vision Computing*, vol. 2, no. 4, 2016.

[35] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *FG*. IEEE, 2017, pp. 25–32.

[36] ——, "Selective transfer machine for personalized facial expression analysis," *TPAMI*, vol. 39, no. 3, pp. 529–545, 2017.

[37] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 1, pp. 28–43, 2012.

[38] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *FG*, vol. 6. IEEE, 2015, pp. 1–6.

[39] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *CVPR*, 2016, pp. 3391–3399.

[40] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic, "Gaussian process domain experts for modeling of facial affect," *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4697–4711, 2017.

[41] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1964–1970, 2013.

[42] T. Simon, M. H. Nguyen, F. De La Torre, and J. F. Cohn, "Action unit detection with segment-based svms," in *CVPR*. IEEE, 2010, pp. 2737–2744.

[43] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events," in *CVPR*. IEEE, 2010, pp. 2574–2581.

[44] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[45] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1006–1016, 2012.

[46] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *ACII*. IEEE, 2013, pp. 356–361.

[47] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," *arXiv preprint arXiv:1803.05588*, 2018.

[48] Z. Hammal, W.-S. Chu, J. F. Cohn, C. Heike, and M. L. Speltz, "Automatic action unit detection in infants using convolutional neural network," in *ACII*. IEEE, 2017, pp. 216–221.

[49] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre, "How much training data for facial action unit detection?" in *FG*, vol. 1. IEEE, 2015, pp. 1–8.

[50] J. F. Cohn, L. A. Jeni, I. Onal Ertugrul, D. Malone, M. S. Okun, D. Borton, and W. K. Goodman, "Automated affect detection in deep brain stimulation for obsessive-compulsive disorder: A pilot study," in *ICMI*. ACM, 2018.

[51] 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.

[52] A. Zadeh, Y. C. Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection." in *ICCVW*, 2017, pp. 2519–2528.

[53] C. E. Fairbairn, M. A. Sayette, J. M. Levine, J. F. Cohn, and K. G. Creswell, "The effects of alcohol on the emotional displays of whites in interracial groups." *Emotion*, vol. 13, no. 3, p. 468, 2013.

[54] C. E. Fairbairn, M. A. Sayette, O. O. Aalen, and A. Frigessi, "Alcohol and emotional contagion: An examination of the spreading of smiles in male and female drinking groups," *Clinical Psychological Science*, vol. 3, no. 5, pp. 686–701, 2015.

[55] M. A. Sayette, K. G. Creswell, J. D. Dimoff, C. E. Fairbairn, J. F. Cohn, B. W. Heckman, T. R. Kirchner, J. M. Levine, and R. L. Moreland, "Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding," *Psychological science*, vol. 23, no. 8, pp. 869–878, 2012.

[56] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[57] 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.

[58] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on Knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.