Crossing Domains for AU Coding: Perspectives, 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. How well do AU detectors transfer to domains in which they have not been trained? We review literature on cross-domain transfer and conduct experiments to address limitations of prior research. We evaluate generalizability in four publicly available databases. EB+ (an expanded version of BP4D+), Sayette GFT, DISFA and UNBC Shoulder Pain (SP). The databases differ in observational scenarios, context, participant diversity, range of head pose, video resolution, and AU base rates. In most cases performance decreased with change in domain, often to below the threshold needed for behavioral research. However, exceptions were noted. Deep and shallow approaches generally performed similarly and average results were slightly better for deep model compared to shallow one. Occlusion sensitivity maps revealed that local specificity was greater for AU detection within than cross domains. The findings suggest that more varied domains and deep learning approaches may be better suited for generalizability and suggest the need for more attention to characteristics that vary between domains. Until further improvement is realized, caution is warranted when applying AU classifiers from one domain to another.

Index Terms—Cross-domain generalizability, facial action unit detection, transfer learning.

1 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. We distinguish between cross-domain generalizability and fine-tuning and identify 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 2.

Taking these factors into account, we then investigate cross-domain generalizability using four databases that differ in context, 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. These databases are an expanded version of BP4D+ [6] (EB+), the Sayette Group Formation Task (i.e. GFT below) [7], DISFA [8] and UNBC Shoulder Pain Archive (SP) [9]. Two large well-annotated databases EB+ and GFT are used to train separate models and all of the four databases are used to test these models.

EB+ database involves inductions of varied emotions of a participant interacting with an experimenter while GFT involves social interaction among previously unacquainted participants. Context in DISFA, unlike EB+ and GFT, is non-social. No one other than the participant is present. In SP, although an experimenter is present, they remain passively

in the background. The stimulus (ice bath) is non-social.

AU base rates are higher in EB+ and GFT compared to DISFA and SP but different from each other. In both DISFA and SP the distribution of AUs and the correlation among them present particular challenges. In DISFA, the base rate of most AUs is low and limited to what occurs in a film-watching paradigm. In SP, AUs are limited mostly to those associated with pain; and the correlation among AUs differs markedly from that of the other databases. In SP, the appearance of AU 12 (oblique lip corner pull), for instance, typically is modified by pain-related actions (e.g., the upward pull of AU 9 or 10 and lateral stretch of AU 20). Such co-articulation effects along with the other differences present multiple challenges to generalizability.

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 hand-crafted features and support vector machine of Openface [10]. 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 [11], [12], AUC, F1 (which is positive agreement when comparing two methods) and negative agreement (NA).

In a further experiment, we investigate similarities and differences in the salience of facial regions when generalizing between domains. In order to understand and interpret at which facial regions classifiers look to detect specific AUs, we generate occlusion sensitivity maps. We compare occlusion sensitivity maps for within- and cross domain AU detection. The findings reveal important differences between within- and cross domain AU detectors.

An earlier version of this paper appeared as [13]. This version differs in multiple respects. It expands the literature review, investigates cross-domain generalizability using additional databases, increases the scope of experiments, and explores and visualizes AU-specific significant regions. The paper is organized as follows: Section 2 reviews the limitations in evaluating cross-domain generalizability and distinguishes between cross-domain generalizability and what is referred to as fine-tuning. Section 3 presents the deep and shallow approaches to investigate AU-specific cross-domain transfer. Section 4 evaluates cross-domain generalizability on two large and two smaller domains with a variety of metrics that quantify different aspects of performance and visualizes occlusion sensitivity maps. Section 5 discusses our findings and suggests future directions.

2 RELATED WORK

2.1 Cross-domain studies

Action unit detection has been studied extensively for nearly two decades [3], [4], [5]. Until recently, most approaches have used hand-crafted features. Examples include LBP [42], SIFT [43], [44], LGBP [45], HOG [46] and LBP-TOP [47]. With the emergence of deep learning, CNN methods have shown significant success for AU detection [48], [49]. 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 1 summarizes studies that evaluate cross-domain AU detection. Some [37], [42] propose novel adaptation approaches. Most test domain transfer without adaptation. Jiang et al. [24] explicitly analyzed temporal dynamics of facial actions using dynamic appearance descriptor Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP). Koelstra et al. [25] proposed a method based on nonrigid registration using free-form deformations to model dynamics of facial texture in near-frontal-view face image sequences for automatic AU detection. Li et al. [26] proposed a knowledge-driven model for AU recognition, which does not use training data and is learned from the generic domain knowledge that governs AU behaviors. a number of studies [27], [28], [29] designed AU detection methods which benefit from facial expression labels when AU annotations are limited. Wu et al. [30] proposed the Constrained Joint Cascade Regression Framework for simultaneous AU detection and facial landmark detection. Tong et al. [31] employed a dynamic Bayesian network (DBN) that systematically accounts for the relationships among AUs and their temporal evolutions for AU recognition. Among the studies personalizing generic classifiers, Mohammadian et al. [33] adapted the system to a new person using Selective Style Transfer Mapping and Chu et al. [37] used Selective Transfer Machine (STM) which attenuates person-specific mismatches. Gehrig et al. [34] used kernel partial least square regression for multi-label AU detection. Walecki et al. [35] proposed a variable-state Conditional Random Field model for dynamic facial expression recognition and AU detection. Valstar et al. [38] applied a combination of GentleBoost, support vector machines, and hidden Markov models to encode AUs and their temporal activation models. Baltrusaitis et al. [39] applied person-specific neutral expression normalisation, used hand-crafted appearance and geometric features to train SVM with multiple databases. Eleftheriadis et al. [41] performed domain adaptation using domain-specific Gaussian process experts for AU detection. Among the deep approaches Ghosh et al. [32] trained a multi-label convolutional neural network approach to learn a shared representation between multiple AUs directly. Chu et al. [36] used CNN to learn spatial features and LSTM to learn temporal dynamics for AU detection. Zhao et al. [40] combined deep region learning with multi-label classification for AU detection. 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 [24], [25], [26], [27], [28], [29], [30] 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], [37], [38], [39], [40], [41] use AU-specific frame-level F1s, others AUC [33], 2AFC

TABLE 1: 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) [14], Extended Cohn-Kanade (CK+) [15], BP4D [16], UNBC Shoulder-Pain Archive (SP) [9], MMI [17], DISFA [8], SEMAINE (SEM) [18], SAL [19], GFT [7], RU-FACS [20], GEMEP-FERA (G-FERA) [21], ISL [22]. For more comprehensive review, see [23].

Study	Databases	Number of subjects	Number of sequences (s) / frames (f)	Number of AUs	AUs	AU specific	Metrics
[24]	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
[25]	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
[26]	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
[27]	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
[28], [29]	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
[30]	CK+→SEM	CK+ (210)	CK+ (593 s, 593 f)	10	1, 2, 4, 5, 6, 7, 12, 17, 25, 26	No	F1
[31]	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
[32]	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
[33]	CK+→SP	CK+ (123), SP (25)	CK+ (593 s, 593 f), SP (48,398 f)	6	4, 6, 7, 9, 10, 43	Yes	AUC
[34]	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
[35]	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)
[36]	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
[37]	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
[38]	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
[39]	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
[40]	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
[41]	BP4D→DISFA DISFA→BP4D	DISFA (27) BP4D (41)	Varies between (10 - 500 f)	7	1, 2, 4, 6, 12, 15, 17	Yes	AUC, F1

[32], [34], or accuracy [32], [35]. 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 [24], [25] while others report recall [24], [25] or Hamming loss [28], [29]. 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 [31], [38] used CK to train classifiers and

MMI to test them but used different numbers of subjects (11 [31] and 70 [38], respectively) from MMI. Similarly, three studies [32], [39], [40] 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 [32], 130K [39] and 64K [40]. 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 [50]. 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.

2.2 Studies performing fine-tuning on the new domain

Training a deep network on one domain and then fine-tuning it in another domain has gained attention for AU detection. Several studies [40], [48], [51], [52], [53] have pre-trained models in BP4D and then fine-tuned them in DISFA. The studies typically select the top-performing CNN in BP4D to obtain face representations, fix convolutional layers, and retrain fully connected layers in DISFA. Face representations are learned on the source domain and unmodified in the new domain. The mappings of facial representations to AUs is fine-tuned using the new domain.

To accomplish fine-tuning, however, it is necessary to have and use AU labels in the new domain, which violates the independence of source and target domains. This assumption is necessary to evaluate domain transfer. Several studies have reported what they describe as cross-domain findings [40], [52] when, in fact, AU detectors contaminated by fine-tuning. Database independence is a necessary assumption to evaluate domain generalizability.

In DISFA AUs are labelled on a 0-5 intensity scale, where zero denotes that the AU did not occur, and 1-5 denotes that the AU occurred at one of five intensity levels (A or 1 =trace, E or 5 =maximum intensity). Because DISFA was scored only for intensity, it is necessary to apply a threshold with which to define occurrence.

Table 2 shows thresholding strategies of studies that have reported cross-domain or fine-tuning results on DISFA. While the cross-domain studies define AU occurrence as A level or higher, the ones reporting fine-tuning results define them variously from A level to D level. While a case may be made for defining occurrence at either A level or B level (per the FACS manual), not all studies have followed this practice. Some use a threshold of C or even D level. While this practice may make AUs easier to detect (subtle or even not so subtle ones may be ignored), comparability with cross-domain findings and with the FACS manual are lost. With the exception of the first BP4D [16], almost all widely used databases use a threshold of B level to define occurrence.

3 METHOD

To compare AU-specific within- and cross-domain transfer, we use both deep and shallow approaches in two databases

TABLE 2: Thresholding strategies of studies reporting cross-domain and fine-tuned AU detection results on DISFA.

Study	Thresholding	Cross-domain / Finetuned
[35]	A-level	Cross-domain
[39]	A-level	Cross-domain
[41]	A-level	Cross-domain
[32]	A-level	Cross-domain
[27]	Not reported	Cross-domain
[40]	C-level	Fine-tuned
[51]	B-level	Fine-tuned
[52]	A-level	Fine-tuned
[53]	D-level	Fine-tuned
[48]	B-level	Fine-tuned

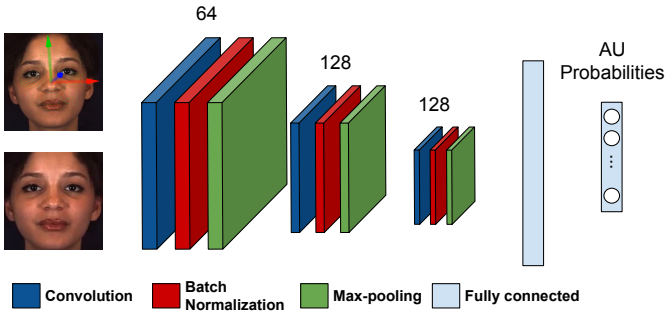


Fig. 1: Overview of the deep network used for within-domain and cross-domain experiments.

that represent different domains. The deep approach is a CNN architecture [54]; the shallow approach is a support vector machine (SVM) with hand-crafted features. One database used for training is an expanded version of BP4D+ [6] (which we refer to as EB+). 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 (200 participants and 395K frames in EB+ and 150 participants and 517K frames 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.

3.1 Deep Approach: Convolutional Neural Network

3.1.1 Face tracking and registration

Video was tracked and normalized using ZFace [55], 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.

3.1.2 Video-specific normalization

Faces vary markedly in geometry and appearance among people. Such differences are a potential source of error when models that have been trained on diverse faces are applied to those of faces that differ from them in face shape and appearance. To control for these individual differences in facial morphology, person-specific normalization has been proposed [39] and found to contribute to improved AU detection. Following this previous work, we included person-specific normalization in our pipeline. For each video frame, we subtracted the mean face shape and appearance of the video from which it came.

3.1.3 AU Detection

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

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

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. For within-domain experiments, threshold is 0.5. For cross-domain experiments, we optimize the threshold on the source domain using 5-fold cross validation. We identified the optimal threshold as the one giving the maximum performance averaged over all folds and use that optimal threshold while testing the model on the target domain.

3.2 Shallow Approach: Openface

OpenFace [10] is a state-of-the-art tool using a shallow approach for facial action unit detection. It uses Convolutional Experts Constrained Local Model (CE-CLM) [56] 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. It uses a linear kernel SVM for AU detection. Output of AU detection module of Openface is $0/1$ label for absence/presence of each AU in each frame.

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

4 Experiments

4.1 Databases

We performed experiments with four well-annotated databases, namely EB+, GFT, DISFA and SP that differ in size, context, resolution, pose, and participant diversity among other factors (see Fig. 2). While EB+ and GFT databases are large in terms of the number of subjects, DISFA and SP are much smaller. For that reason, we used DISFA and SP databases only to test the classifiers to obtain cross-domain results.

We considered using additional databases but they were either publicly unavailable or unsuited for our study. RU-FACS is not publicly available. In CK, CK+, MMI, GEMEP-FERA only 1-3 frames are coded for each participant. Manual annotations are obtained for a total of a few hundreds of frames, which is not sufficient for training a deep model. In addition, CK, GEMEP-FERA and subsets of CK+ and MMI databases are posed, in which facial behavior was deliberate. SEMAINE has labels for only three AUs in our set. Since the remaining annotations are lacking and we use multi-label classification to make use of AU correlations, we did not include SEMAINE in our experiments.

GFT [7] involves social interaction among 50 groups of three previously unacquainted young adults (150 participants in all). A third of the groups drink an alcoholic beverage; a third a placebo beverage they believe to contain alcohol; and a third fruit juice. Alcohol effects are common and have been reported previously [57], [58], [59]. EB+ is a series of emotion inductions or tasks of a single participant interacting with an experimenter, which elicits more intense action units with different rates of occurrence. BP4D+ is reported in [6]. DISFA involves participants alone in non-social context while watching videos to elicit spontaneous expressions. SP contains a series of active and passive range-of-motion tests performed by participants that suffer from shoulder pain. The focus is on the pain induction (movement of the affected shoulder) rather than social interaction. The databases differ as well in participant diversity, number of AUs that occur and their frequency and co-occurrence, range of head pose, non-rigid head motion, illumination, and video resolution.

EB+ and GFT were manually annotated by different teams of highly qualified, certified FACS coders from the same lab at different times. Occurrence was defined as B level or higher. 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. DISFA and SP were coded by different teams of certified FACS coders. Occurrence in both DISFA and SP indicates intensity of B or higher. DISFA and SP have annotations for a smaller set of AUs. For DISFA, the set

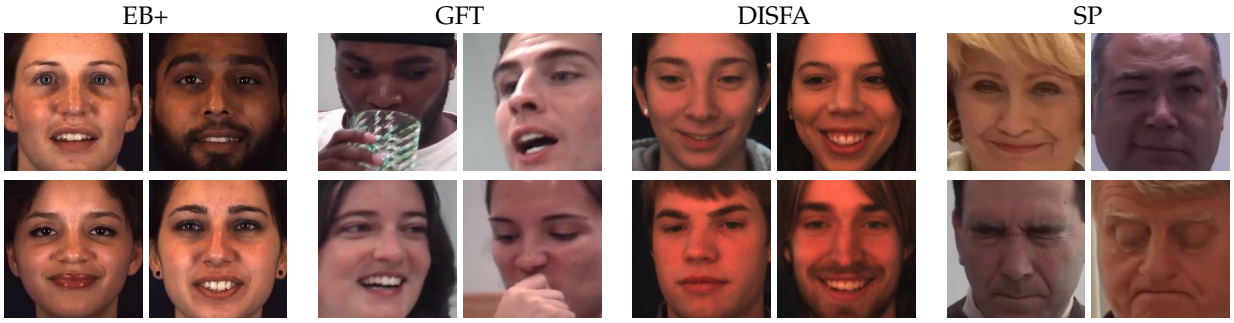


Fig. 2: Sample frames from different domains.

includes AU1, AU2, AU4, AU6 and AU12 while for SP, we have annotations for AU4, AU6, AU7, AU10 and AU12.

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, while the average annotated video duration is 13 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 limited. 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 1216 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). Video resolution is 720×480. 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.

DISFA [8] is a database of spontaneous behavior in 27 adults (12 women, 15 men). It is manually annotated for AU intensity from 0 to E-level. Participants watched a video clip consisting of 9 segments intended to elicit a range of facial expressions of emotion. Video resolution is 1024×768. Face orientation is nearly frontal. For each participant, 4845 video frames were recorded. We use a total of 130K frames in our experiments. Positive samples are defined as the ones with intensities equal to or higher than B-level, and the remaining ones are negative samples.

UNBC Shoulder Pain Archive (SP) [9] is a manually FACS annotated database. It contains 200 videos of 25 different patients having shoulder pain. Videos were recorded while the patients performed different types of arm movements. We use a total of 48K frames in our experiments. Positive samples are defined as the ones with intensities equal to or higher than B-level, and the remaining ones are negative samples.

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

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

4.2 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, 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 [60].

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

Fig. 3: Overview of the databases used in our experiments. Databases used for training (EB+ and GFT) are denoted with circles having thicker boundaries. Source and target of an arrow denote the database used for training and test, respectively. Self loops represent within-domain experiments. Each database and its properties are denoted with a specific color. Graphs in the corners show the respective base rates of AUs.

F1 and NA are not [61], 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.

4.3 AU Detection Results

We first report within-domain and cross-domain AU detection results for the deep models trained on EB+ and GFT, separately. Databases used as training or test set in within-domain and cross-domain experiments can be seen in Fig. 3. Then, we compare cross-domain results between CNN and Openface, which affords a comparison between a deep (CNN) and shallow (Openface) approach. For the comparisons between within-domain & cross-domain and deep & shallow, we performed significance tests in given Table 6. 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.

4.3.1 Within-domain and cross-domain results of deep model trained on EB+

Table 3 shows AU-specific results obtained by the deep model trained on EB+ database. While Table 3a shows within-domain results, Table 3b, Table 3c, and Table 3d show cross-domain results on GFT, DISFA, and SP databases, respectively. Contrary to the studies performing fine-tuning on the new domain, we directly test the models on new domains to infer generalizability.

Imbalanced classes are evident in all of the databases. 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. In DISFA and SP, none of the AUs occur in more 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 [61].

Average F1 score in EB+ is in the moderate range. Differences in the individual AU performances are present, which are related to base rates of AUs. For AUs that occur in more than 15 percent of the frames, F1 scores are far better (0.75 to 0.88). 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. 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.

A critical question is whether AU-detectors generalize to new domains. When we compare within-domain results in Table 3a and cross-domain results on GFT in Table 3b, we observe a decrease in average cross-domain results. Average AUC and F1 values are 0.729 and 0.599 for within EB+ (see Table 3a) while they are 0.658 and 0.443 for cross-domain results on GFT (see Table 3b). Therefore, we observe decrease of 0.071 and 0.156 for AUC and F1, respectively. For each individual AU, there is a degradation in S score, AUC, F1 and NA values. Significance results in Table 6 shows that within-domain results are significantly better for all AUs than cross-domain results on GFT.

When we compare within-domain results in Table 3a with cross-domain results on DISFA in Table 3c, we observe an increase in S, AUC, F1 and NA for AU1, AU2 and AU4. For these AUs, the classifier generalizes better to a new domain. A reason for better cross-domain results on DISFA may be that variation is limited in terms of pose, illumination and ethnicity. i) Limited variation (mostly Caucasian and young adults) in DISFA compared to larger variation in terms of ethnicity and age in EB+ ii) higher base rate of AU4 (0.15) compared to the base rate of EB+ (0.07) and iii) lack of pose in DISFA may be the potential reasons for better generalization. On the other hand, for AU6 and AU12 we observe a decrease in S, AUC and F1 in cross-domain results. The difference is only significant for AU12 (see Table 6). Average within-domain F1 value computed over AU1, AU2, AU4, AU6 and AU12 is 0.626, which is 11 percent higher than average cross-domain F1 value on DISFA. Cross-domain results on DISFA are generally better when NA is used. Similar to F1, NA is affected by the skew in the data and base rates of all AUs are very low for DISFA. In other words, negative samples greatly outnumber the positive ones and such imbalance may lead to larger NA.

Cross-domain results on SP in Table 3d are much smaller than within-domain results in Table 3a. For all AUs, we observe a decrease in AUC and F1 values. Only for AU1 and AU2, cross-domain results are better for S score and NA.

TABLE 3: Within-domain and cross-domain AU detection results (EB+).

(a) Within-domain: EB+

-	Base rate	S	AUC	F1	NA
AU1	0.09	0.787	0.670	0.468	0.941
AU2	0.07	0.856	0.659	0.437	0.961
AU4	0.07	0.873	0.690	0.526	0.966
AU6	0.43	0.685	0.839	0.821	0.859
AU7	0.63	0.646	0.811	0.864	0.748
AU10	0.59	0.713	0.846	0.881	0.820
AU12	0.53	0.736	0.867	0.876	0.858
AU14	0.42	0.566	0.779	0.749	0.809
AU15	0.10	0.776	0.656	0.408	0.938
AU17	0.14	0.643	0.601	0.344	0.897
AU23	0.14	0.722	0.736	0.569	0.917
AU24	0.03	0.943	0.595	0.245	0.986
Average 12 AUs	**0.27**	**0.745**	**0.729**	**0.599**	**0.892**

(b) Cross-domain: EB+ → GFT

-	Base rate	S	AUC	F1	NA
AU1	0.09	0.741	0.588	0.258	0.929
AU2	0.12	0.597	0.640	0.338	0.881
AU4	0.04	0.817	0.607	0.180	0.952
AU6	0.33	0.562	0.769	0.688	0.832
AU7	0.42	0.251	0.661	0.666	0.573
AU10	0.41	0.490	0.760	0.728	0.759
AU12	0.33	0.541	0.784	0.703	0.813
AU14	0.43	0.083	0.584	0.621	0.420
AU15	0.18	0.314	0.582	0.324	0.770
AU17	0.17	0.219	0.601	0.334	0.724
AU23	0.12	0.669	0.671	0.248	0.907
AU24	0.07	0.533	0.648	0.231	0.862
Average 12 AUs	**0.22**	**0.485**	**0.658**	**0.443**	**0.785**

(c) Cross-domain: EB+ →DISFA

-	Base rate	S	AUC	F1	NA
AU1	0.05	0.916	0.770	0.571	0.978
AU2	0.04	0.905	0.759	0.499	0.975
AU4	0.15	0.800	0.743	0.612	0.943
AU6	0.08	0.660	0.801	0.416	0.901
AU12	0.12	0.375	0.807	0.444	0.783
Average	**0.14**	**0.731**	**0.776**	**0.508**	**0.916**

(d) Cross-domain: EB+ → SP

-	Base rate	S	AUC	F1	NA
AU4	0.02	0.920	0.616	0.222	0.980
AU6	0.11	0.700	0.633	0.350	0.915
AU7	0.07	0.290	0.599	0.176	0.774
AU10	0.01	0.622	0.801	0.083	0.895
AU12	0.14	0.594	0.668	0.406	0.878
Average	**0.07**	**0.625**	**0.663**	**0.247**	**0.888**

TABLE 4: Within-domain and cross-domain AU detection results (GFT).

(a) Within-domain: GFT

-	Base rate	S	AUC	F1	NA
AU1	0.09	0.827	0.672	0.437	0.953
AU2	0.12	0.770	0.677	0.449	0.935
AU4	0.04	0.928	0.560	0.198	0.982
AU6	0.33	0.679	0.810	0.746	0.882
AU7	0.42	0.525	0.762	0.721	0.791
AU10	0.41	0.621	0.803	0.765	0.840
AU12	0.33	0.744	0.849	0.798	0.905
AU14	0.43	0.249	0.602	0.500	0.691
AU15	0.18	0.580	0.602	0.339	0.875
AU17	0.17	0.639	0.537	0.170	0.898
AU23	0.12	0.737	0.543	0.168	0.928
AU24	0.07	0.853	0.535	0.129	0.962
Average 12 AUs	**0.22**	**0.679**	**0.663**	**0.452**	**0.887**

(b) Cross-domain: GFT → EB+

-	Base rate	S	AUC	F1	NA
AU1	0.09	0.743	0.630	0.312	0.929
AU2	0.07	0.844	0.573	0.224	0.959
AU4	0.07	0.855	0.559	0.204	0.962
AU6	0.43	0.369	0.662	0.577	0.749
AU7	0.63	0.269	0.645	0.678	0.578
AU10	0.59	0.469	0.733	0.767	0.692
AU12	0.53	0.532	0.771	0.757	0.774
AU14	0.42	0.235	0.638	0.631	0.602
AU15	0.10	0.651	0.599	0.268	0.901
AU17	0.14	0.377	0.621	0.302	0.799
AU23	0.14	0.254	0.616	0.320	0.743
AU24	0.03	0.734	0.639	0.135	0.928
Average 12 AUs	**0.27**	**0.528**	**0.637**	**0.431**	**0.801**

(c) Cross-domain: GFT → DISFA

-	Base rate	S	AUC	F1	NA
AU1	0.05	0.818	0.733	0.370	0.951
AU2	0.04	0.832	0.762	0.379	0.955
AU4	0.15	0.794	0.701	0.553	0.942
AU6	0.08	0.688	0.867	0.475	0.908
AU12	0.12	0.698	0.902	0.624	0.906
Average	**0.14**	**0.766**	**0.793**	**0.480**	**0.932**

(d) Cross-domain: GFT → SP

-	Base rate	S	AUC	F1	NA
AU4	0.02	0.904	0.545	0.099	0.975
AU6	0.11	0.418	0.566	0.209	0.822
AU7	0.07	-0.328	0.590	0.151	0.454
AU10	0.01	0.113	0.425	0.016	0.714
AU12	0.14	0.679	0.534	0.164	0.911
Average	**0.07**	**0.357**	**0.529**	**0.128**	**0.775**

TABLE 5: Deep and shallow cross-domain results on GFT

(a) Deep model

-	S	AUC	F1	NA
AU1	0.741	0.588	0.258	0.929
AU2	0.597	0.640	0.338	0.881
AU4	0.817	0.607	0.180	0.952
AU6	0.562	0.769	0.688	0.832
AU7	0.251	0.661	0.666	0.573
AU10	0.490	0.760	0.728	0.759
AU12	0.541	0.784	0.703	0.813
AU14	0.083	0.584	0.621	0.420
AU15	0.314	0.582	0.324	0.770
AU17	0.219	0.601	0.334	0.724
AU23	0.669	0.671	0.248	0.907
Average 11 AUs	**0.480**	**0.659**	**0.463**	**0.778**

(b) Shallow model

-	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
Average 11 AUs	**0.438**	**0.653**	**0.458**	**0.770**

TABLE 6: 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. n.s. denotes not significant. For the shaded cell cross-domain results are greater than within-domain or shallow is greater than deep.

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

On average, cross-domain results (0.663 AUC and 0.247 F1) are much worse than average within-domain results (0.810 AUC and 0.793 F1).

4.3.2 Within-domain and cross-domain results of deep model trained on GFT

Table 4 shows AU-specific results obtained by the deep model trained on GFT. While Table 4a shows within-domain results, Table 4b, Table 4c, and Table 4d show cross-domain results on EB+, DISFA, and SP, respectively.

Within-domain average F1 score is in the moderate range. For AUs that occur in more than 20 percent of the frames, F1 scores and AUC values are much better. When we compare within-domain results on EB+ in Table 3a and within-domain results on GFT in Table 4a, we observe that F1 scores and AUCs are higher in EB+ than in GFT. Similar to the model trained on EB+, S scores obtained with the model trained on GFT range from moderate to high. Although S scores show a less consistent relation to base rate, they show the same difference between EB+ and GFT. Results for GFT in Table 4a are generally worse than those for EB+ in Table 3a. Significance results in Table 6 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. These findings suggest that EB+ is an easier database compared to GFT for AU detection. Sources of variation need to be better understood.

After analyzing within-domain results, we compare whether AU-detector trained on GFT generalizes to other domains. We observe a decrease in average cross-domain results for EB+ in Table 4b. Although the decrease in average F1 score and AUC is slight, it is large in S scores and NA values. Significance results in Table 6 show that, within-domain results on GFT 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. When S scores are compared, cross-domain results are significantly better than within-domain results for AU2 and AU15. When AUC values are compared, for AU14, AU17 and AU23 cross-domain results are significantly better.

When we compare within-domain results in Table 4a with cross-domain results on DISFA in Table 4c, we observe an increase in F1 only for AU4. For the remaining AUs, F1 values are lower and average F1 is 0.48, which is 4 percent lower than the average F1 score obtained over AU1, AU2, AU4, AU6 and AU12. On the other hand, cross-domain AUC values are high and are larger than within-domain AUC values for all AUs. Recall that, cross-domain AUC results from EB+ $\rightarrow$ DISFA in Table 3c are also larger than within-domain AUC results for most of the AUs. These findings suggest that classifiers trained on other domains

may generalize well to DISFA.

On the other hand, a comparison of within-domain results in Table 4a and cross-domain results on SP in Table 4d yields a large decrease in S, AUC and F1 values for all AUs. Decrease in average AUC and F1 computed over AU4, AU6, AU7, AU10 and AU12 are 0.227 and 0.517, respectively. For all AUs except for AU4, within-domain results are significantly better than cross-domain results when both S score and AUC are used. These results show that neither the model trained on EB+ nor the one trained on GFT generalizes well to SP database.

When we compare how the model trained on EB+ and the one trained on GFT generalizes on DISFA and SP, we observe that F1 results are consistently better and AUC results are generally better for the model trained on EB+ (Table 3c and Table 3d) than the one trained on GFT (Table 4c and Table 4d). CNNs trained on EB+ generalize better to unseen domains rather than the ones trained on GFT.

While in previous analyses we compare results within the same domain to results on other domains, it is important also to consider the expected best result for the new domain. When we look at average within-domain results on GFT in Table 4a, we observe that average F1 score is 0.452, which can be considered as the upper limit expected for GFT. When we test GFT with the model trained on EB+ in Table 3b, we observe an average F1 score of 0.443, which is very close to the expected best F1 score for GFT. Therefore, when tested on GFT, performance of the model trained on EB+ is nearly as good as the model trained within GFT. These results suggest that, in addition to the differences in domains, difficulty of domains (within-domain performance which may be considered as the expected upper limit for a domain) may also be another factor for the degradation in the performance in cross-domain experiments. If the target domain has low within-domain performance (as in GFT), decrease in the cross-domain performance from source domain to target (from EB+ to GFT) domain is likely.

4.3.3 Cross-domain comparison of deep and shallow models

We report cross-domain results with deep and shallow approaches 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. Similarly, as the training set of Openface includes DISFA and SP databases, we do not report cross-domain results on these databases.

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 5a with Table 5b we can infer that, deep model gives slightly better S score, F1, AUC and NA on average. When we analyze F1s for individual AUs, deep approach outperforms shallow one 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 AU4, AU6 and AU23. For the AUs with high baserates, both deep and shallow approaches perform similarly. 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 6). Notice that, these conclusions are drawn when the models are trained with large databases. Results may be different when small databases are used.

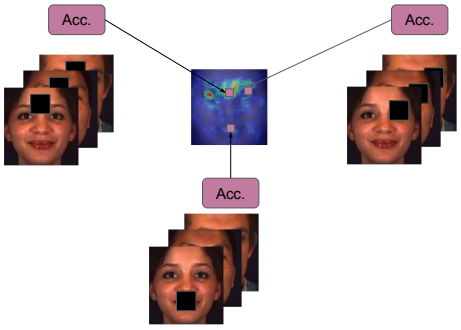


Fig. 4: Overview of generating occlusion sensitivity maps from accuracy values of occluded images. Note that patches are slid over the images after video-specific normalization.

If we would only report AUC values as in [33], or F1s as in [36], [38], [39], we would infer that deep and shallow approaches perform similar for cross-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 similar to shallow approach, both deep and shallow approaches perform similar on positive instances and when the effect of chance is discarded, deep approach performs slightly better.

4.4 Visualizing AU-specific significant regions

A natural question is whether the classifier looks at expected regions to detect the related action units (e.g., nasal root for AU 4) in within-domain and cross-domain experiments. Given the co-occurring nature of AUs in spontaneous behavior, important regions may not be trivial. To answer this question we systematically occlude different portions of the input and monitor the output of the classifier. Note that, our goal is not to understand whether our model is robust to occlusion. Instead, we use occlusion as a way to infer the significant facial regions the classifier is looking at to detect specific AUs.

In order to interpret the facial regions that cause the largest decrease in the accuracy when occluded, we generate occlusion sensitivity maps [62], [63] and visualize them for different AUs. For each AU, we randomly select 1000 images that contain the specified AU and are classified correctly by the model (true positives). We define a patch having size 15×15 whose pixel values are 0 (having black color). We first overlay the patch onto top-left corners of the 1000 input images. We test these occluded images with the same model and obtain accuracy value for the top-left position of the patch. We write the obtained accuracy value to the center pixel of the patch in the occlusion sensitivity map as shown in Fig. 4. We slide the patch over the image of size 200×200 with a stride 2 and repeat the same steps. In the end, we obtain accuracy values for 92×92 different positions of the patch. After an interpolation step, the resulting grids of

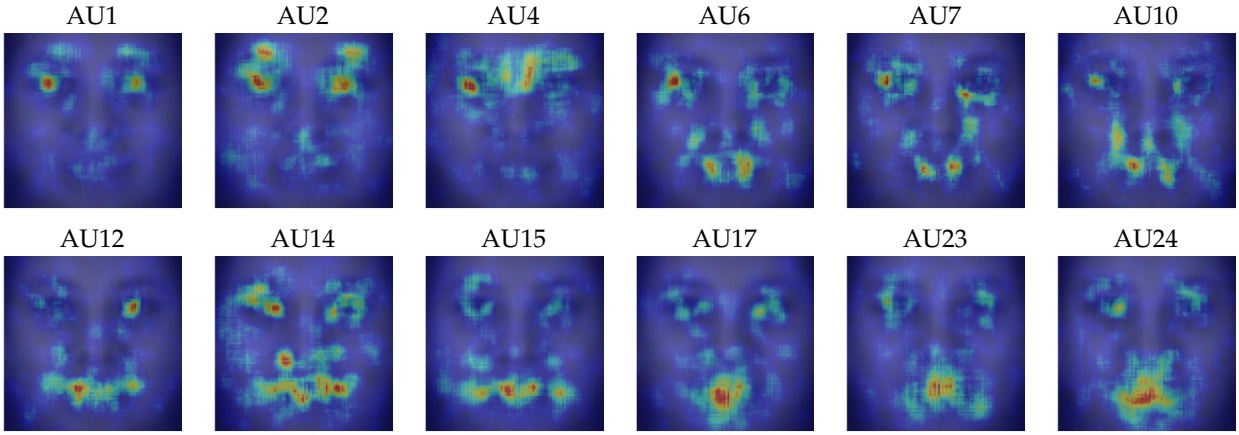


Fig. 5: Occlusion sensitivity maps (Within-domain: EB+).

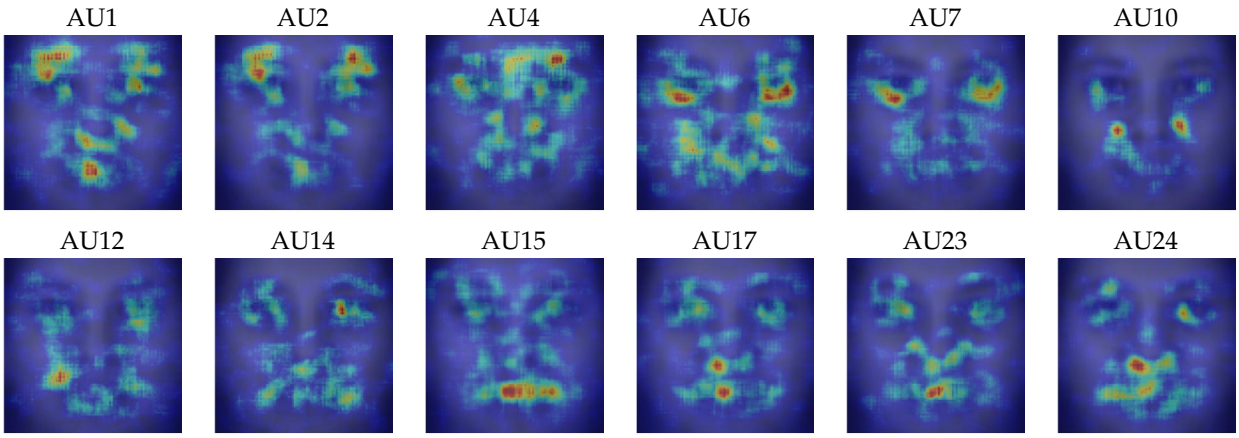


Fig. 6: Occlusion sensitivity maps (Cross-domain: GFT →EB+).

accuracy values yield occlusion sensitivity maps. We show occlusion sensitivity maps of within-domain experiment on EB+ in Fig. 5 and cross-domain experiment (GFT →EB+) in Fig. 6. In occlusion sensitivity maps, darker red colors represent the lowest accuracy of correctly estimating occluded positive samples while darker blue colors represent the parts, whose occlusion do not affect the accuracy a lot. Significant regions for each AU are the ones colored with red, whose occlusion by a patch leads to a great decrease in the accuracy. We use these maps to understand 1) if our models look to anticipated facial regions to when detecting AUs; 2) if our models learn the co-occurrence relationship between AUs; and 3) how these maps differ for within- and cross-domain comparisons.

In Fig. 5, for most of the AUs, the model learns where to look at the input to detect the specific AU correctly. For AU1, AU2 the most important regions are around eyes, eyebrows and forehead while for AU4 inner brow regions and eyes are significant. For AU12, AU14 and AU15, the classifier mainly looks at a long and narrow region around mouth and lip corners, while for AU17, AU23 and AU24 the significant regions are more local around mouth and chin. For AU6 and AU7 in addition to the eye region, the classifier also looks at mouth region due to the co-occurring nature of AUs.

When we compare occlusion sensitivity maps of cross-domain (Fig. 6) and within-domain (Fig. 5) experiments, we observe that significant regions in maps of within-domain experiments are more local while they are more distributed in the maps of cross-domain experiments. Since moderate-to-large head pose is present in the training domain GFT, and due to the domain differences between EB+ and GFT, the model looks at larger regions to detect specific AUs as expected. For example, the model trained on EB+ looks mainly around eye and eyebrow regions to detect AU1 when it is tested with frames from the same domain. On the other hand, the model trained on GFT uses the information around mouth patch in addition to eye and eyebrow regions to detect AU1 when it is tested with frames from EB+.

5 DISCUSSION AND FUTURE WORK

The future impact of AU detectors hinges on their ability to generalize from domains in which they have been trained to ones in which they have not. How well they generalize until now is an open question. We found that relevant studies failed to report AU-specific results; frustrated comparison with previous work by using different numbers of subjects or frames; lacked comparisons with one or more approaches; and failed to report sufficient test statistics to quantify different aspects of performance.

Without comparability across methods, domains, and test statistics, inferences about generalizabilty remain limited.

From our review, 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. Other investigators may recommend additional metrics. With these recommendations, within- and cross-domain results can be rigorously compared within and between AU-detection approaches.

In line with these recommendations, 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 the shallow approach (OpenFace) we used. 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. 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. In many cases performance decreased to below the threshold acceptable for behavioral research.

Within-domain results on EB+ are better than those on GFT. Because EB+ has higher AU base rates, nearly frontal faces, and higher resolution, which may better capture subtle details, it may be an easier database to detect AUs. Larger pose variation, lower AU base rates and the lower resolution of GFT may make AU detection more difficult.

Models trained on EB+ or GFT performed more poorly when applied to new domains. Reasons may include differences between domains in AU base rates, demographics (age, gender, and ethnicity), camera view (frontal or out-of-plane), video resolution, illumination, and context (induced emotion, physical pain, or group formation).

While cross-domain attenuation was common, generalization to some domains was better than that to others. Classifiers trained on EB+ or GFT generalized better to DISFA than to SP. Unlike EB+ and GFT, SP included tasks for physical pain and the correlation among AUs differs from that of the other databases. Also, participants in SP are older (average age = 49), and video resolution and AU base rates are all lower than in other domains. Participants in DISFA are young adults more similar in age to the subjects in EB+ (average age = 20.5) and in GFT (average age = 22.3), and DISFA has much higher resolution than SP. These differences in age range, AU base rates and video resolution may be key to the lower generalizability to SP.

We explored cross-domain generalizability using four well-annotated databases that differ in context, variation in head pose, illumination, age, gender, ethnicity, AU base rates, and resolution. Since these aspects could not be varied systematically with the given databases, we could only infer the reasons for decrease in the cross-domain performance. Such a systematic evaluation would require posed data, representative sampling of different racial groups, and a wide range of ages, illumination, and other parameters. These video then would need well-annotated AU annotation, ideally from multiple teams so that annotator variability might be considered. Future work could focus on collecting such databases and then systematically varying these aspects to see their effect on the detection performance of specific AUs.

We used CNN architecture given in Fig. 1 in our experiments since it has been shown to perform well and provide meaningful results in a cross-domain setting [54]. One of the limitations of our work is, we do not know whether a deeper CNN would perform better. However, in a recent work, Niinuma et al. [64] used a deeper network (VGG16) with a range of parameter settings. They reported cross-domain results on DISFA and for some AUs our cross-domain results are better. They also observe a decrease in cross-domain setting compared to within-domain results. We do not believe that results will be fundamentally different for other architectures. Yet, this remains to be tested.

We included as many databases as possible, used a range of test statistics and both deep and shallow approaches to answer the question about domain transfer. Given the lack of AU specific results and heterogeneity of metrics and database splits in the existing work, we cannot compare our results with the state-of-the-art. For example, we have reported test results on GFT for all of the 150 subjects and Chu et al. [36] reported results on GFT for 50 subjects. Moreover, we cannot compare our results on DISFA with the results reported in i) [40] since they finetuned their models on DISFA, ii) [39] since they use A-level thresholding contrary to recent common practice of B-level thresholding, and iii) [41] since they subsampled frames and did not use all frames to test their models. Therefore, we cannot answer how well our model generalizes compared to the state-of-the-art approaches, which is a limitation of our work.

Commercial systems, including iMotions, Affectiva and Noldus, profess to recognize AU and facial expressions. Considering the relatively 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 [65]. When this assumption is violated, additional learning is required. Domain adaptation approaches for AU detection would be indicated.

6 ACKNOWLEDGMENTS

This research was supported in part by NIH awards NS100549 and MH096951 and NSF awards IIS 1418026, 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] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[9] 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.

[10] 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.

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

[12] 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.

[13] I. Onal Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Cross-domain au detection: Domains, learning approaches, and measures," in *FG*. IEEE, 2019.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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 Trans. on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[19] 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.

[20] 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.

[21] 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.

[22] Q. Ji, "ISL facial expression databases," Rensslear Polytechnic Institute, https://www.ecse.rpi.edu/~cvrl/database/database.html.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

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

[46] 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 Trans. on Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 1006–1016, 2012.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: Deep nets with enhancing and cropping for facial action unit detection," *IEEE TPAMI*, vol. 40, no. 11, pp. 2583–2596, 2018.

[52] Z. Zhang, S. Zhai, and L. Yin, "Identity-based adversarial training of deep cnns for facial action unit recognition," in *BMVC*, 2018.

[53] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *ECCV*, 2018, pp. 298–313.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE TKDE*, vol. 17, no. 3, pp. 299–310, 2005.

[61] 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.

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

[63] I. Onal Ertugrul, L. A. Jeni, and J. F. Cohn, "Facscaps: Pose-independent facial action coding with capsules," in *CVPRW*, 2018, pp. 2130–2139.

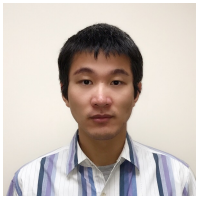
[64] K. Niinuma, L. A. Jeni, I. Onal Ertugrul, and J. F. Cohn, "Cross-domain au detection: Domains, learning approaches, and measures," in *BMVC*, 2019.

[65] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE TKDE*, vol. 22, no. 10, pp. 1345–1359, 2010.

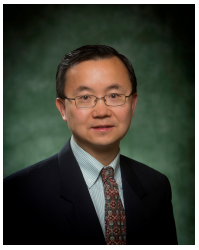
Itir Onal Ertugrul is a postdoctoral researcher at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. She was a postdoctoral researcher at Affect Analysis Group, University of Pittsburgh, PA, USA in 2017. She received her B.Sc., M.Sc. and Ph.D. degrees in computer engineering from Middle East Technical University, Ankara, Turkey, in 2011, 2013 and 2017, respectively. Her research interests include affective computing, facial action detection and understanding human behavior.

Jeffrey F. Cohn is Professor of Psychology, Psychiatry, and Intelligent Systems at the University of Pittsburgh and Adjunct Faculty at the Robotics Institute, Carnegie Mellon University. Dr. Cohn has led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis of facial expression and prosody and applied those tools to research in human emotion, interpersonal processes, social development, and psychopathology. He co-developed the influential Cohn-Kanade, Multi-PIE, Pain Archive, BP4D, and BP4D+ databases, co-edited special issues on facial expression analysis, and chaired international conferences in automatic face and gesture recognition, multimodal interaction, and affective computing.

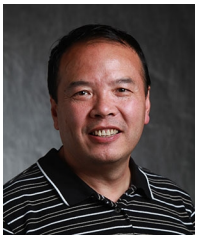
László A. Jeni is Systems Faculty in the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA, USA. He specializes in computer vision and computational behavior science, specifically in areas of modelling, understanding, and synthesizing human behavior using diverse sensors. He received his MSc degree in Computer Science from Eötvös Loránd University, Hungary, and his Ph.D. in 2012 from the University of Tokyo, Japan. To his credit he has over 50 peer reviewed publications. His honors include Best Paper Awards at the IEEE Conference on Human System Interaction (HSI'2011) for work on validating observations of human activity and at the Conference on Automatic Face and Gesture Recognition (FG'2015) for work on dense 3D face alignment from 2D video. He has organized computational face challenges and workshops at ECCV, ICCV, and FG.



Zheng Zhang received his B.Sc. from the department of Mathematics and Applied Mathematics, Tianjin University, China in 2012. Subsequently, he received his M.S. degree in Computer Science from the State University of New York at Binghamton in 2014. He is now a Ph.D. candidate in the same insitution at Graphics and Image Computing Lab. His research interests cover multimodal affective computing, domain adaptation and geometric deep learning.



Lijun Yin is a Professor of Computer Science, Director of Center for Imaging, Acoustics, and Perception Science at Binghamton University, Director of Graphics and Image Computing Laboratory, and Co-director of Seymour Kunis Media Core, T. J Watson School of Engineering and Applied Science at the State University of New York at Binghamton. He received Ph.D. of computer science from the University of Alberta, Canada and Master of Electrical Engineering from Shanghai Jiao Tong University in China. Dr. Yins research focuses on computer vision, graphics, HCI, and multimedia, specifically on face and gesture modeling, analysis, recognition, animation, and expression understanding. His research has been funded by the NSF, AFRL/AFOSR, NYSTAR, and SUNY Health Network of Excellence. Dr. Yin received the prestigious Lois B. DeFleur Faculty Prize for Academic Achievement Award (2019), James Watson Investigator Award of NYSTAR (2006), and SUNY Chancellor's Award for Excellence in Scholarship & Creative Activities (2014). He holds 11 US patents, and released four 2D/3D/4D facial expression databases to public, and has published over 130 papers in technical conferences and journals. Dr. Yin served as a program co-chair of FG 2013 and FG2018. He is currently serving on editorial board of IVC and PRL.



Qiang Ji received his Ph.D degree in Electrical Engineering from the University of Washington. He is currently a Professor with the Department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute (RPI). From 2009 to 2010, he served as a program director at the National Science Foundation (NSF), Arlington, VA, USA, where he managed NSFs computer vision and machine learning programs. He also held teaching and research positions with the Beckman Institute at University of Illinois at Urbana-Champaign, Urbana, IL, USA; the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA, USA; the Dept. of Computer Science at University of Nevada, Reno, Nevada, USA; and the Air Force Research Laboratory, Rome, NY, USA. Prof. Ji currently serves as the director of the Intelligent Systems Laboratory (ISL) at RPI. Prof. Ji's research interests are in computer vision, probabilistic graphical models, machine learning, and their applications in various fields. He has published over 300 papers in peer-reviewed journals and conferences, and has received multiple awards for his work. Prof. Ji is has served as an editor on several related IEEE and international journals and as a general chair, program chair, technical area chair, and program committee member for numerous international conferences/workshops. Prof. Ji is a fellow of the IEEE and the IAPR.