IntraFace

Fernando De la Torre[†], Wen-Sheng Chu[†], Xuehan Xiong[†], Francisco Vicente[†], Xiaoyu Ding[†], Jeffrey Cohn[†‡]
[†]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
[‡]Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260

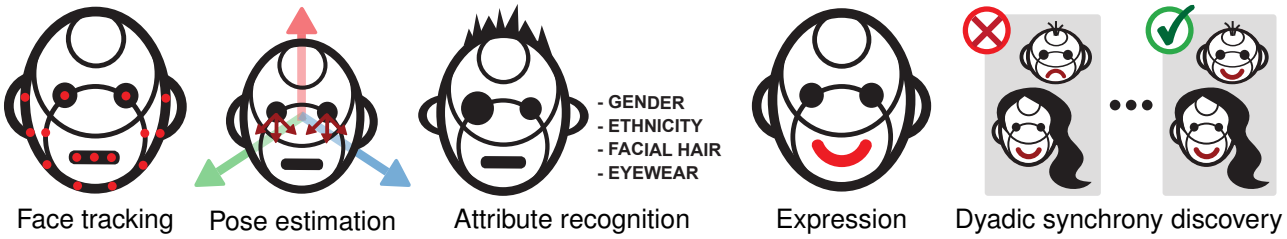


Fig. 1: An overview of the functionalities provided by **IntraFace** (**IF**)

***Abstract*— Within the last 20 years, there has been an increasing interest in the computer vision community in automated facial image analysis algorithms. This has been driven by applications in animation, market research, autonomous-driving, surveillance, and facial editing among others. To date, there exist several commercial packages for specific facial image analysis tasks such as facial expression recognition, facial attribute analysis or face tracking. However, free and easy-to-use software that incorporates all these functionalities is unavailable.**

This paper presents IntraFace (IF), a publicly-available software package for automated facial feature tracking, head pose estimation, facial attribute recognition, and facial expression analysis from video. In addition, IFincludes a newly develop technique for unsupervised synchrony detection to discover correlated facial behavior between two or more persons, a relatively unexplored problem in facial image analysis. In tests, IF achieved state-of-the-art results for emotion expression and action unit detection in three databases, FERA, CK+ and RU-FACS; measured audience reaction to a talk given by one of the authors; and discovered synchrony for smiling in videos of parent-infant interaction. IF is free of charge for academic use at http://www.humansensing.cs.cmu.edu/intraface/**.**

I. Introduction

Facial expression has been a focus of research in human behavior for over a century [12]. It is central to several leading theories of emotion [16], [32] and has been a focus of heated debates about issues in emotion science. In part because of its importance and potential uses, as well as its inherent challenges, automated facial expression analysis has been of keen interest in computer vision and machine learning. The last twenty years has witnessed dramatic advances in face detection, facial feature detection and tracking, face recognition, facial expression transfer, and facial attribute estimation. Emerging applications include surveillance [14], marketing [31], drowsy driver detection [29], parent-infant interaction [19], social robotics [3], telenursing [10], expression transfer for video gaming [23], animating avatars in multi-person games [24], interpersonal coordination [24], and subtle expression detection [20].

To meet the needs of these diverse applications, several consumer packages for facial image analysis have recently been introduced. Consumer software for facial expression analysis is available from companies such as Emotient[1] (previously CERT [27]), FaceReader[2], and NVSIO[3], among others. Commercial services for facial expression analysis are available as well, including Affectiva[4] and RealEyes[5]. These products and services can be difficult to use, publication of comparative results may be restricted, and the products prohibitively expensive for research applications. Furthermore, the code typically is closed; that is, users cannot modify it. For these and related reasons, it is typically difficult to compare and evaluate performance across different packages and services. To facilitate the use of facial image analysis software in the research community, we present **IntraFace** (**IF**), a publicly available software package that includes state-of-the-art algorithms for feature tracking, head pose estimation, facial attribute recognition, multi-person face detection and analysis, facial expression recognition, and facial synchrony detection from video or camera input. **IF** is available for non-commercial use without charge.

Figure 1 illustrates the functionalities provided by **IF**. Figure 2 illustrates a specific application of **IF** which is to measure audience demographics and reaction. In the example, **IF** detects and tracks multiple persons and reveals moments of attention and emotion reaction for realtime feedback to the speaker. When multiple faces are tracked as in this example, **IF** is able to detect facial synchrony as well. That is, it is able to find video segments that contain correlated facial behavior. For instance, in Figure 2, we would be interested in finding the moments when all members of the audience

[1] http://www.emotient.com/
[2] http://www.noldus.com/human-behavior-research/products/facereader
[3] http://www.nviso.ch/
[4] http://www.affdex.com/
[5] http://www.realeyes.me/

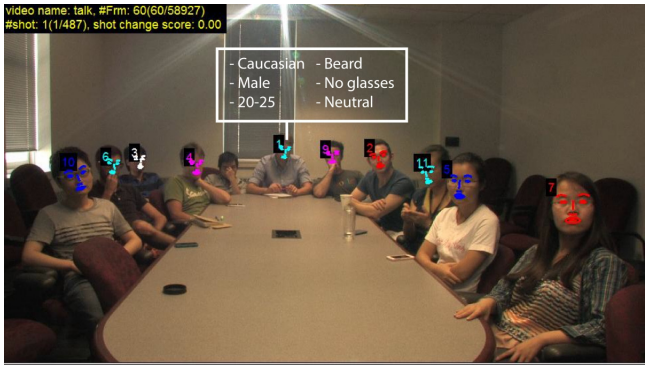


Fig. 2: Automatic output of **IntraFace** to measure audience reaction while attending a talk,"Common Sense for Research (and Life)," by one of the authors.

may be laughing, serious, or looking away.

Unsupervised or weakly-supervised discovery of synchrony from facial behavior has been a relatively unexplored problem in computer vision and facial image analysis. Because a naive exhaustive search approach to synchrony detection has a quadratic computational complexity with duration of the video, synchrony discovery has been impractical for other than specialized use. **IF** eliminates this limitation by using an efficient branch and bound (B&B) algorithm to tackle synchrony detection. **IF** can be applied to video of variable lengths for synchrony detection. The current implementation supports dyadic (two-person) synchrony detection (DSD). Future releases will extend synchrony detection to three or more persons.

In testing, **IF** achieves state-of-the-art results for facial expression detection on the FERA, CK+ and RU-FACS datasets in both within-dataset and cross-dataset scenarios. In addition to these results, we present two case studies for the use of **IF**: (1) CrowdCatch, an application to measure audience reaction. **IF** to provide a speakers with both on- and offline feedback; and (2) DSD, an application to detect synchronous facial behavior in videos of parent-infant interaction. For instance, we are able to detect the moments when the mother and the infant both smile.

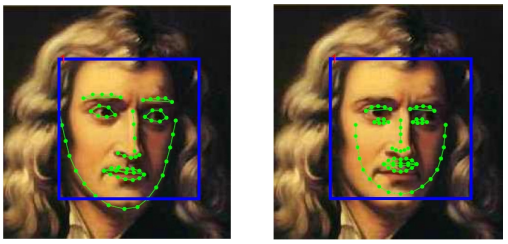
II. Facial Feature Tracking

This section describes our approach to facial feature detection and tracking in videos. For the notation convention, see the footnote[6].

A. Single face tracking

Facial feature detection and tracking in **IF** is implemented using the Supervised Descent Method (SDM) [35]. SDM is a supervised method that learns to optimize non-linear least squares problems. SDM learns generic descent maps in a supervised manner, and is able to overcome many drawbacks of second order optimization schemes, such as non-differentiability and expensive computation of the Jacobians and Hessians. Here, we give an overview of SDM in the context of facial feature detection and tracking. A detailed theoretical analysis and test results can be found in [36].

[6]Bold capital letters denote a matrix $\mathbf{X}$; bold lower-case letters denote a column vector $\mathbf{x}$. $\mathbf{x}_i$ represents the i^{th} column of the matrix $\mathbf{X}$. $\mathbf{X}_{ij}$ denotes the scalar in the i^{th} row and the j^{th} column of the matrix $\mathbf{X}$. All non-bold letters represent scalars. x_j denotes the scalar in the jth element of $\mathbf{x}$. $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix.



(a) $\mathbf{x}_*$ (b) $\mathbf{x}_0$

Fig. 3: (a) Manually labeled image with 66 landmarks. Blue outline indicates face detector. (b) Mean landmarks, $\mathbf{x}_0$, initialized using the face detector.

Given an image $\mathbf{d} \in \Re^{m\times 1}$ of m pixels, $\mathbf{d}(\mathbf{x}) \in \Re^{p\times 1}$ indexes p landmarks in the image. $\mathbf{h}$ is a non-linear feature extraction function (*e.g.*, HoG) and $\mathbf{h}(\mathbf{d}(\mathbf{x})) \in \Re^{128p\times 1}$ in the case of extracting HoG features. During training, we assume that the ground truth p landmarks (in our case $p = 49$) are known, and we will refer to them as $\mathbf{x}_*$ (see Figure 3a). Also, to mimic the testing scenario, we ran the face detector on the training images to provide an initial configuration of the landmarks ($\mathbf{x}_0$), which corresponds to an average shape (see Figure 3b). In this setting, face alignment can be framed as minimizing the following function over $\Delta\mathbf{x}$:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|\mathbf{h}(\mathbf{d}(\mathbf{x}_0 + \Delta\mathbf{x})) - \boldsymbol{\phi}_*\|_2^2, \qquad (1)$$

where $\boldsymbol{\phi}_* = \mathbf{h}(\mathbf{d}(\mathbf{x}_*))$ represents the HoG values at the manually labeled landmarks. In the training images, $\boldsymbol{\phi}_*$ and $\Delta\mathbf{x}$ are known.

Applying Newton's method to optimize Eq. (1) yields the following update

$$\Delta\mathbf{x}_k = -2\mathbf{H}^{-1}\mathbf{J}_\mathbf{h}^\top(\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_*), \qquad (2)$$

where $\mathbf{J}_h$ is the Jacobian matrix of $\mathbf{h}$ and $\mathbf{H}$ is the Hessian matrix of f evaluated at $\mathbf{x}_{k-1}$. However, two problems arise: first, HoG is a non-differentiable image operator, so in order to use Newton's method we have to perform expensive numerical methods to approximate the Jacobians and Hessians. Second, $\boldsymbol{\phi}_*$ is unknown in test time. To address these two issues, SDM rewrites Eq. (2) as a generic linear combination of feature vector $\boldsymbol{\phi}_{k-1}$ plus a bias term $\mathbf{b}_{k-1}$ that can be learned during training,

$$\Delta\mathbf{x}_k = \mathbf{R}_{k-1}\boldsymbol{\phi}_{k-1} + \mathbf{b}_{k-1}. \qquad (3)$$

Using training examples, SDM will learn a sequence of $\mathbf{R}_k, \mathbf{b}_k$ such that the succession of $\mathbf{x}_k$ converges to $\mathbf{x}_*$ for all images in the training set.

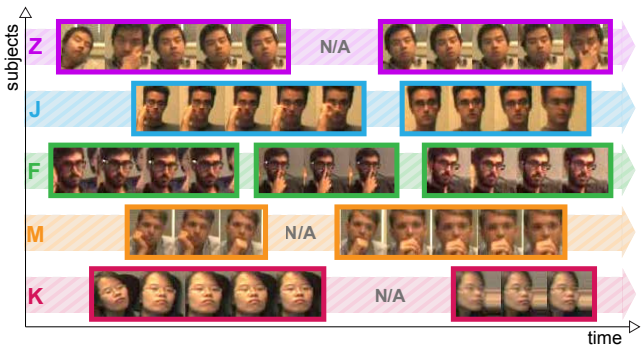


Fig. 4: An illustration of **IF** on multiple face tracking. Faces belong to the same person are identified and tracked across different scenes. "N/A" indicates the frames where the face is not present or the tracker fails to track the person.

B. Multiple face tracking and identification

A video is likely to contain multiple subjects and changes in scenes. Tracking and maintaining person identities across scenes is therefore a challenging task. To facilitate face tracking in this type of videos, **IF** provides a functionality of multiple face tracking and identification, which aims to track and tell whether the subject has been seen or not. First, we detect the differences in camera views and scene changes. Second, for each scene, we detect and track each face using our face tracker, and represent each track of a face as a *tracklet*. Finally, we associate person identities across different scenes.

Scene change detection: We detect scene changes by computing changes between consecutive frames. We first divide each frame into 32×32 blocks and compute the difference between edges and color histogram in corresponding blocks between two consecutive frames. If the differences are bigger than a threshold, we consider that the block belongs to different scenes. If more than 25% of the blocks are classified as belonging to different categories, we consider that the frames belong to different scenes.

Face association across different scenes: We detect and track each face within the same scene until the single face tracker is lost. Then, each set of tracked frames is grouped into a face tracklet (similar to [18]). Given several face tracklets from several scenes, we used a subspace distance measure to identify persons across tracklets. For each face tracklet i, we selected a predefined number of frames and vectorized them in a feature matrix $\mathbf{M}_i$. Each column of $\mathbf{M}_i$ contains HoG features at the landmarks for one frame. $\mathbf{U}_i$ contains the principal components of $\mathbf{M}_i$ that preserve 95% of the energy. The distance between the i^{th} and the j^{th} tracklet is defined as the average subspace distance between tracklets [33]:

$$d_{i,j} = \frac{||\mathbf{M}_j - \mathbf{U}_i\mathbf{U}_i^\top\mathbf{M}_j||_F^2}{||\mathbf{M}_j||_F^2} + \frac{||\mathbf{M}_i - \mathbf{U}_j\mathbf{U}_j^\top\mathbf{M}_i||_F^2}{||\mathbf{M}_i||_F^2}.$$

Two tracklets are associated to the same person if their average distance is lower than a threshold. Figure 4 illustrates an example where several subjects have been tracked and associated over time.

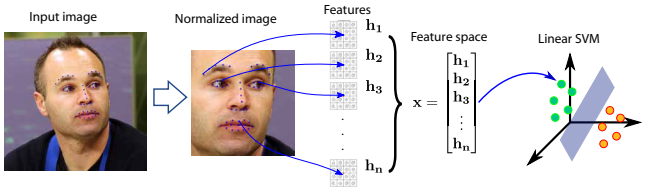


Fig. 5: Pipeline for facial attribute recognition.

TABLE I: Attribute Classification Results

Category	Attribute	$F1$ score
Ethnicity	East Asian	0.913
	Caucasian	0.930
	African-American	0.894
	Indian	0.934
Gender	Male/female	0.961
Facial hair	Mustache	0.942
	Beard	0.960
Eyewear	No glasses	0.970
	Eyeglasses	0.942
	Sunglasses	0.963

III. Facial Attribute Recognition

This section describes the facial attribute recognition algorithm used in **IF** to recognize ethnicity, gender, facial hair and eyewear on faces. These attributes can be classified efficiently using a linear SVM with HoG features [11]. Figure 5 illustrates our pipeline. First, facial landmarks are extracted using the SDM tracker, and then the image is normalized w.r.t. scale and rotation. Then HoG features are extracted at each landmark and concatenated in a vector $\mathbf{x}$. Finally, a linear SVM classifier is learned for each one of the attributes. We trained our attribute classifiers using the PubFig dataset [25], which contains approximately 60,000 labeled images. The training set is composed of 40,000 randomly selected samples, and the rest of the samples are used as test set. Table I shows the $F1$ scores obtained by our simple but effective attribute classifiers.

IV. Facial Expression Analysis and AU Detection

Psycholigists have classified human facial expression using two main categories: eight universal expressions [16] (see Figure 6) and a more anatomical descriptions called Action Units (AUs), defined by the Facial Action Coding System (FACS) [17] (see Figure 7). See [8], [13], [15], [30], [34] for a recent review of taxonomies and existing methods in facial expression analysis.

Automatic Facial Action unit detection (AFA) confronts a series of challenges. These include facial variations in pose, scale, illumination, occlusion, and individual differences in facial morphology (*e.g.*, heavy versus delicate brows, smooth versus deeply etched wrinkles) and behavior. To compensate for these variations, in **IF**, we adopted a transductive learning approach, Selective Transfer Machine (STM) [6],

Fig. 6: Basic emotions. From left to right: happiness, sadness, anger, fear, surprise, disgust, contempt, and embarrassment.

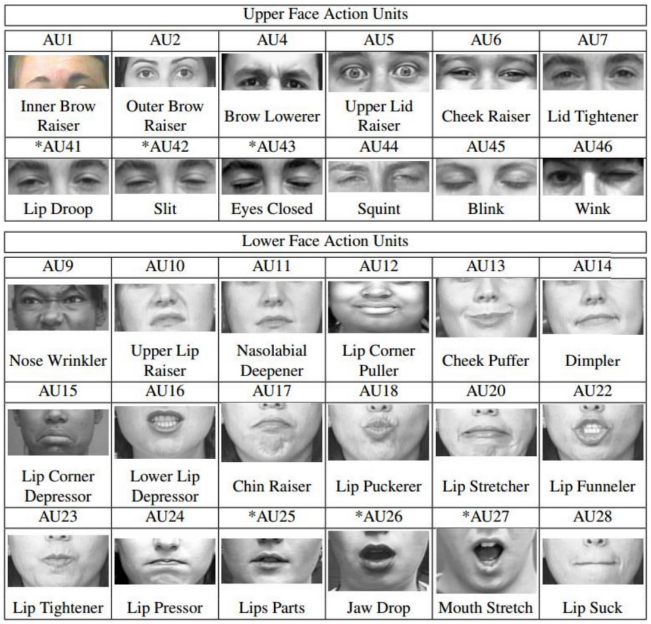


Fig. 7: Facial Action Units (AUs) of upper and lower face

to personalize a generic maximal-margin classifier. STM simultaneously learns a classifier while re-weighting the training samples that are most relevant to the test subject. In this paper, we further evaluate the method proposed in [6].

Denote the training set as $\mathcal{D}^{\text{tr}} = \{\mathbf{x}_i, y_i\}_{i=1}^{n_{\text{tr}}}$ that contains n_{tr} training images $\mathbf{x}_i$ and their labels $y_i \in \{+1, -1\}$, and $\mathbf{X}^{\text{tr}}, \mathbf{X}^{\text{te}}$ the sets of training and test images respectively. We adapt the STM formulation that minimizes the objective:

$$g(f, \mathbf{s}) = \min_{f,\mathbf{s}} R_f(\mathcal{D}^{\text{tr}}, \mathbf{s}) + \lambda \Omega_{\mathbf{s}}(\mathbf{X}^{\text{tr}}, \mathbf{X}^{\text{te}}), \qquad (4)$$

where $R_f(\mathcal{D}^{\text{tr}}, \mathbf{s})$ is the SVM empirical risk defined on the decision function f, and training set $\mathcal{D}^{\text{tr}}$ with each instance weighted by $\mathbf{s} \in \mathbb{R}^{n_{\text{tr}}}$. Each entry s_i corresponds to a positive weight for a training sample $\mathbf{x}_i$. $\Omega_{\mathbf{s}}(\mathbf{X}^{\text{tr}}, \mathbf{X}^{\text{te}})$ measures the mismatch between the training and the test distributions in a reproducing kernel Hilbert space $\mathcal{H}$ induced by some nonlinear feature mapping $\varphi(\cdot)$:

$$\Omega_{\mathbf{s}}(\mathbf{X}^{\text{tr}}, \mathbf{X}^{\text{te}}) = \left\| \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} s_i \varphi(\mathbf{x}_i^{\text{tr}}) - \frac{1}{n_{\text{te}}} \sum_{j=1}^{n_{\text{te}}} \varphi(\mathbf{x}_j^{\text{te}}) \right\|_{\mathcal{H}}^2 . \qquad (5)$$

The lower the value of $\Omega_{\mathbf{s}}$, the more similar the training and the test distributions are. $\lambda > 0$ is a tradeoff that balances the risk and the distribution mismatch. The goal of STM is to jointly optimize the decision function f as well as the selective coefficient $\mathbf{s}$, such that the resulting classifier can alleviate person-specific biases.

TABLE II: Emotion recognition on the CK+ dataset

	AUC				
Emotion	SVM	KMM	T-SVM	DA-SVM	**IF**
Anger	95.1	85.3	76.1	–	**96.4**
Contempt	**96.9**	94.5	88.8	–	**96.9**
Disgust	94.5	81.6	84.2	–	**96.0**
Fear	**96.6**	92.7	84.9	–	95.5
Happy	**99.4**	93.9	86.7	–	98.9
Sadness	**94.5**	76.0	78.7	–	93.3
Surprise	97.3	64.5	81.8	–	**97.6**
Avg	96.3	84.1	83.0	–	**96.4**

TABLE III: Emotion recognition on the FERA dataset

	AUC				
Emotion	SVM	KMM	T-SVM	DA-SVM	**IF**
Anger	31.1	66.5	70.4	**78.8**	78.6
Fear	31.9	81.4	64.5	83.9	**85.5**
Joy	90.2	33.5	78.9	71.1	**95.0**
Relief	20.4	74.8	76.8	87.9	**88.4**
Sadness	73.4	80.2	77.1	74.7	**84.8**
Avg	49.4	67.3	73.5	79.3	**86.5**

Given that the loss functions in $R_f(\mathcal{D}^{\text{tr}}, \mathbf{s})$ are convex sub-differential (*e.g.*, squared loss, logistic loss, Huber loss), STM in (4) becomes a standard bi-convex problem, and (4) can be simply solved by alternating between f and $\mathbf{s}$ using Newton's method or conjugate gradient. Once the optimization is done, the classification of test images is performed by applying the learned classifier f.

A. *Baseline for personalized facial expression recognition*

With **IntraFace**, we set up a baseline for facial expression recognition on three major benchmarks: CK+ [28], FERA Challenge [34] and RU-FACS [2]. We reported our results in terms of $F1$ score and Area Under the ROC Curve (AUC) on two tasks: basic emotion recognition and facial Action Unit (AU) detection. All experiments were conducted in a *cross-subject* scenario, which is also known as *leave-one-subject-out*, *i.e.*, training and test subjects were independent in all iterations. To carry out a credible experiment, we compared **IF** with generic methods and state-of-the-art transductive approaches, including a linear SVM [5], Kernel Mean Matching (KMM) [21], Transductive SVM (T-SVM) [9], and Domain Adaptation SVM (DA-SVM) [4]. The same features and training/test images were used for all methods. Below we describe each task in turn.

1) Basic emotion recognition: Similar to [34], we utilized all available frames for each algorithm. We only report AUC for this task because each video has only a single emotion label instead of frame-by-frame labeling, which makes $F1$ score meaningless. For CK+, 327 videos were given 7 basic and discrete emotions: *Anger*, *Contempt*, *Disgust*, *Fear*, *Happy*, *Sadness*, and *Surprise*. For FERA, 289 portrayals were asked to perform one of the five emotional states: *Anger*, *Fear*, *Joy*, *Sadness*, and *Relief*. We evaluated on the training set that includes 155 videos portrayed by 7 actors

TABLE IV: AU detection on the CK+ dataset

	AUC					$F1$ Score				
AU	SVM	KMM	T-SVM	DA-SVM	**IF**	SVM	KMM	T-SVM	DA-SVM	**IF**
1	79.8	68.9	69.9	72.6	**88.9**	61.1	44.9	56.8	57.7	**62.2**
2	**90.8**	73.5	69.3	71.0	87.5	73.5	50.8	59.8	64.3	**76.2**
4	74.8	62.2	63.4	69.9	**81.1**	62.7	52.3	51.9	57.7	**69.1**
6	89.7	87.7	60.5	**94.7**	94.0	75.5	70.1	47.8	68.2	**79.6**
7	82.1	68.2	55.7	61.4	**91.6**	59.6	47.0	43.8	53.1	**79.1**
12	88.1	89.5	76.0	**95.5**	92.8	76.7	74.5	59.6	59.0	**77.2**
15	93.5	66.8	49.9	94.1	**98.2**	75.3	44.4	40.4	76.9	**84.8**
17	90.3	66.6	73.1	94.7	**96.0**	76.0	53.2	61.7	81.4	**84.3**
Avg	86.1	72.9	64.7	81.7	**91.3**	70.0	54.7	52.7	64.8	**76.6**

TABLE VII: Cross-dataset AU detecttion: RU-FACS→FERA

	AUC					$F1$ Score				
AU	SVM	KMM	T-SVM	DA-SVM	**IF**	SVM	KMM	T-SVM	DA-SVM	**IF**
1	44.7	48.8	43.7	56.9	**63.2**	46.3	46.4	41.8	46.1	**50.4**
2	52.8	70.5	52.1	52.3	**74.0**	47.4	54.2	38.6	45.4	**54.6**
4	52.7	55.4	54.2	52.7	**58.6**	57.1	57.1	40.2	42.9	**57.4**
6	73.5	55.2	77.1	79.9	**83.4**	60.7	55.2	52.8	56.3	**72.7**
12	56.8	60.1	70.9	76.1	**78.1**	67.7	67.7	63.5	62.6	**71.5**
15	55.1	52.1	59.3	**60.2**	58.6	31.5	32.8	29.7	26.4	**41.1**
17	44.3	41.1	39.1	46.2	**52.7**	27.3	27.1	24.3	24.6	**31.4**
Avg	54.3	54.8	56.6	60.6	**66.9**	48.3	48.6	41.6	43.5	**54.2**

with 3∼5 instances of each emotion per actor.

Tables II and III show the results. In Table II, DA-SVM is omitted because it failed to converge due to insufficient testing data in CK+. One can see that a generic SVM performed fairly well, because in CK+ the positive (peak expressions) and negative samples (neutral faces) are relatively easy to separate. KMM and T-SVM performed sub-optimally, because they lack the refinement of instance weights, and thus are unable to correct badly estimated weights for learning the final classifier. This effect is particularly obvious when test data is insufficient as in this experiment. On the other hand, our method considers the labels for weight refinement and performed comparably.

The FERA dataset served as a more challenging benchmark for evaluating emotion recognition performance. See Table III for the results. Because each test video consists of tens of frames, DA-SVM was able to converge in most cases. The generic SVM performed poorly due to large variations in this dataset, such as head movements and spontaneous expressions. Without the ability to select meaningful training samples, the generic classifier suffered from the individual differences. Other cross-domain methods alleviated the person-specific biases and produced better results. Overall our method achieved the most satisfactory performance. Comparing Tables II and III, one can observe that when the data grows larger and more complex, the improvement of **IF** becomes more clear.

2) Facial AU detection: Tables IV∼VI show AUC and $F1$ scores on three datasets using **IF**. A linear SVM served as the baseline generic classifier. Note that KMM fails to perform better than the baseline on all datasets. An explanation is because KMM did not consider label information and produced less accurate sample weights. T-SVM performed similarly to SVM in FERA and RU-FACS, but worse than SVM in CK+. This may be because the samples in CK+ are more distinct than consecutive frames in FERA and RU-FACS. **IF** achieved 91% AUC, which is slightly better than the best published results (90.5% [1]), although the results are not directly comparable. Unlike **IF**, which used a penalized SVM, T-SVM did not consider re-weighting for training instances and used the losses from the training data. Hence it could not correct the weights for irrelevant samples, such as noise or outliers. On the other hand, DA-SVM extends T-SVM by progressively labeling test patterns and removing labeled training patterns. Not surprisingly, DA-SVM shows better performance than KMM and T-SVM because it used relevant samples for training resulting in a better personalized classifier. Similar to T-SVM, DA-SVM did not update the re-weightings using label information. Moreover, it does not always guarantee convergence to a correct solution. In our experiments, we faced the situation where DA-SVM failed to converge due to a large amount of samples lying within the margin bounds. In contrast, **IF** adopts a biconvex formulation, and is therefore guaranteed to converge to a critical point.

Our approach allows more than just subject adaptation. We performed an extra experiment to show that **IF** can be naturally extended for *cross-dataset* adaptation. Table VII shows the results of training on RU-FACS and testing on FERA. One can see that transductive approaches outperformed a generic SVM because a generic SVM does not model the biases between datasets. That is, under the cross-dataset scenario, the training and test distributions are likely different, which prevents SVM from transferring the knowledge from one dataset to another. In these cross-dataset experiments, we can clearly see the advantages of **IF** over a generic SVM classifier.

V. Dyadic Synchrony Discovery

Among possible elicitors of human emotion, social interactions may be the most powerful. In particular, dyadic interactions, such as coworkers, friends, romantic partners, family members and children, become one of the richest source of spontaneous emotion. See Figure 9 for an example of mother-infant interaction. **IF** provides a software to automatically discover correlated dyadic facial behaviors in the context of social interactions. We refer to this new problem as Dyadic Synchrony Discovery (DSD).

Problem formulation: Given a set of synchronized videos containing two interactive faces, we run **IF** and detect and track facial landmarks on the faces. The shape features are stacked into two matrices: $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$, where $\{\mathbf{a}_i\}_{i=1}^n$ and $\{\mathbf{b}_j\}_{j=1}^n$ contains vectorized shape features in our experiments. $\mathbf{A}[b_1, e_1] = [\mathbf{a}_{b_1}, \ldots, \mathbf{a}_{e_1}]$ denotes the subsequence of $\mathbf{A}$ that begins from frame b_1 and ends in frame e_1 (similarly for $\mathbf{B}[b_2, e_2]$). The problem of DSD consist on searching over all possible subsequences and find the one that maximize the correlation

TABLE V: AU detection on the FERA dataset

AU	AUC					$F1$ Score				
	SVM	KMM	T-SVM	DA-SVM	**IF**	SVM	KMM	T-SVM	DA-SVM	**IF**
1	71.5	43.3	72.2	83.3	**84.3**	56.5	48.5	60.3	59.1	**68.1**
2	73.9	51.0	74.3	**76.8**	73.3	56.9	50.2	58.5	57.1	**65.5**
4	58.5	53.5	42.8	**66.6**	60.0	43.5	39.8	36.9	**46.3**	43.3
6	80.4	60.2	81.1	**91.1**	87.7	63.7	58.7	63.8	**72.7**	71.6
7	66.9	59.4	70.8	**76.9**	75.4	63.1	63.5	63.7	**68.3**	66.2
12	77.7	58.8	74.8	74.5	**84.7**	79.1	68.4	77.6	75.5	**82.1**
15	55.5	58.7	67.2	67.5	**67.8**	33.4	35.2	35.2	**41.3**	39.3
17	59.8	51.8	63.8	**66.5**	63.3	32.0	27.8	36.2	**42.0**	35.9
Avg	68.0	54.6	68.4	**75.4**	74.5	53.5	49.0	54.0	57.8	**59.0**

TABLE VI: AU detection on the RU-FACS dataset

AU	AUC					$F1$ Score				
	SVM	KMM	T-SVM	DA-SVM	**IF**	SVM	KMM	T-SVM	DA-SVM	**IF**
1	72.0	74.0	72.0	77.0	**83.9**	40.8	37.7	37.4	35.5	**55.3**
2	66.6	58.6	71.1	76.5	**82.4**	35.7	32.2	36.2	34.1	**52.6**
4	74.8	62.2	50.0	76.4	**82.4**	25.2	14.5	11.2	**35.3**	30.4
6	89.1	88.8	61.6	60.3	**93.1**	58.3	39.2	33.1	42.9	**72.4**
12	86.7	87.0	86.7	84.4	**92.3**	61.9	63.0	62.6	71.4	**72.3**
14	71.8	67.8	74.4	70.4	**87.4**	31.3	25.8	25.8	40.9	**51.0**
15	72.5	68.8	73.5	58.1	**86.1**	32.3	29.5	32.3	34.9	**45.4**
17	78.5	76.7	79.5	75.7	**89.6**	39.5	35.6	44.0	46.5	**55.3**
Avg	76.5	72.3	71.1	72.3	**86.3**	40.6	37.3	40.6	42.7	**54.3**

Algorithm 1: Dyadic Temporal Commonality Discovery

input : A synchronized video pair $\mathbf{A}, \mathbf{B}$; minimal discovery length ℓ; commonality period T
output: Optimal intervals $\mathtt{r}^\star = [b_1, e_1, b_2, e_2]$

$L \leftarrow T + \ell$; // The largest possible searching period
$\mathtt{Q} \leftarrow$ empty priority queue; // Initialize priority queue $\mathtt{Q}$
for $t \leftarrow 0$ **to** $(n - L)$ **do**
  $\mathtt{R} \leftarrow [1+t, L+t, 1+t, L+t]$; // Diagonal regions
  $\mathtt{Q}.\mathtt{push}(\mathtt{bound}(\mathtt{R}), \mathtt{R})$; // Fill in $\mathtt{Q}$
end
$\mathtt{R} \leftarrow \mathtt{Q}.\mathtt{pop}()$; // Initialize $\mathtt{R}$
while $|\mathtt{R}| \neq 1$ **do**
  $\mathtt{R} \rightarrow \mathtt{R}' \cup \mathtt{R}''$; // Split into 2 disjoint sets
  $\mathtt{Q}.\mathtt{push}(\mathtt{bound}(\mathtt{R}'), \mathtt{R}')$; // Push $\mathtt{R}'$ and its bound
  $\mathtt{Q}.\mathtt{push}(\mathtt{bound}(\mathtt{R}''), \mathtt{R}'')$; // Push $\mathtt{R}''$ and its bound
  $\mathtt{R} \leftarrow \mathtt{Q}.\mathtt{pop}()$; // Pop top state from $\mathtt{Q}$
end
$\mathtt{r}^\star \leftarrow \mathtt{R}$; // Retrieve the optimal rectangle

of facial behavior. We formulate DSD as an integer programming over two intervals $[b_1, e_1] \subseteq [1, n]$ and $[b_2, e_2] \subseteq [1, n]$:

$$\max_{b_1,e_1,b_2,e_2} \quad \mathrm{corr}\left(\varphi_{\mathbf{A}[b_1,e_1]}, \varphi_{\mathbf{B}[b_2,e_2]}\right), \tag{6}$$
$$\text{s.t.} \quad \ell \leq e_i - b_i, \forall i \in \{1, 2\},$$
$$T \leq |b_1 - b_2|,$$

where $\varphi_{\mathbf{x}}$ is a feature mapping for a sequence $\mathbf{x}$, $\mathrm{corr}(\cdot, \cdot)$ is a correlation measurement between two feature vectors, ℓ controls the minimal length for each subsequence to avoid the trivial solution of both lengths being zero, and T is the size of temporal neighborhood where commonalities can occur. Note that, although the given video pair is synchronized, dyadic commonalities can appear in a slightly shifted time period, *e.g.*, the baby starts to smile after the mother smiles for a few seconds. The second constraint thus allows DSD to discover commonalities within a temporal window of T frames. A naive approach for solving (6) is to search over all possible locations for (b_1, e_1, b_2, e_2). However, it leads to an algorithm with computational complexity $\mathcal{O}(n^4)$, which is prohibitive for regular videos with thousands of frames.

Algorithm: Inspired by the Branch and Bound (B&B) approach for general temporal commonality discovery [7], we adapt a similar algorithm that discovers the global optimum of (6). Given the knowledge that dyadic commonalities only occur within a temporal neighborhood between two videos, we only need to search for a small number of regions in the temporal search space. Specifically, instead of exhaustively pruning the search space to a unique discovery (*e.g.*, [7], [26]), we constrain the space before the search begins. That is, denote $L = T + \ell$ as the largest possible period to search, we fill in the priority queue $\mathtt{Q}$ with the valid regions $\{[1+t, L+t, 1+t, L+t]\}_{t=0}^{n-L}$ and their associated bounds [7]. Because these regions lie sparsely along the diagonal in the 2-D search space, we are able to significantly reduce the searching from $\mathcal{O}(n^4)$ to $\mathcal{O}((n - T - \ell)T^2)$. $\mathtt{r}$ denotes a rectangle in the 2-D search space as a candidate solution (b_1, e_1, b_2, e_2). A rectangle set in the search space, *i.e.*, $\mathtt{R} = [B_1, E_1, B_2, E_2]$, indicates a set of parameter intervals, where $B_i = [b_i^{lo}, b_i^{hi}]$ and $E_i = [e_i^{lo}, e_i^{hi}], i \in \{1, 2\}$ are tuples of parameters ranging from frame lo to frame hi. Because correlation can be equivalently translated into distances under some constraints, *e.g.*, $\max_{\mathbf{x},\mathbf{y}} \mathbf{x}^\top \mathbf{y} \equiv \min_{\mathbf{x},\mathbf{y}} \|\mathbf{x} - \mathbf{y}\|_2^2$ given $\mathbf{x}$ and $\mathbf{y}$ are unitary, we follow a similar Branch-and-Bound (B&B) strategy as [7] to solve (6) with a guaranteed global optimum. Algorithm 1 summarizes the proposed DSD algorithm.

VI. Case Studies

This section describes two case studies using **IF**. The first one, CrowdCatch, shows how **IF** can be used to measure the emotional reaction of an audience. The second case study illustrates how to use the proposed DSD to find patterns of correlated facial behavior in videos of parent-infant interaction.

A. CrowdCatch

Public speaking, whether to groups small or large, is necessary in a wide range of occupations and social contexts. Speakers vary greatly in their preparation, prior experience, and skill. For speaker training, review, and monitoring of ongoing performance, real-time analysis of audience reaction would be of great benefit. With **IF** it becomes possible to automatically analyze emotion reactions from multiple listeners and provide immediate or time-delayed feedback to the speaker.

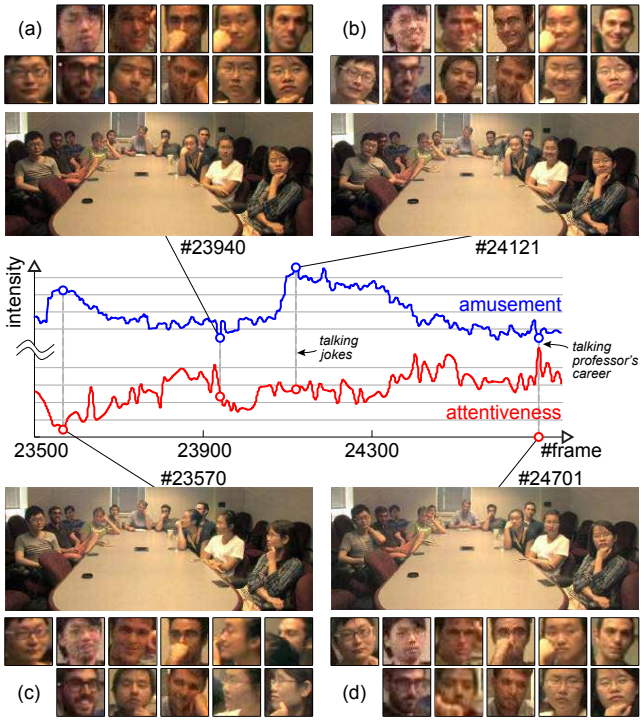


Fig. 8: An illustration of **IF** on capturing amusement and attentiveness during a 45-minute talk on "Common Sense for Research (and Life)". (a)/(b) and (c)/(d) show the lowest/highest amusement and attentiveness respectively. Check the text for the content about the talk.

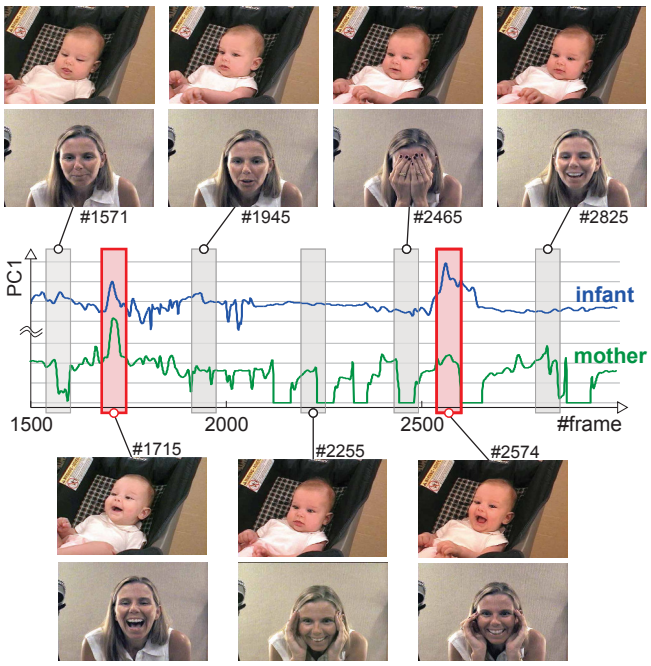


Fig. 9: An illustration of a mother-infant interaction during a Face-to-Face (FF) session. y-axis denotes the projected features onto the first principal component. Red framed rectangles indicate the discovered dyadic smiling faces; gray rectangles indicate other sampled frames.

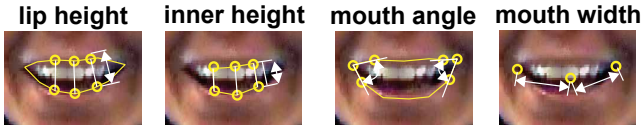


Fig. 10: An illustration of mouth features

In this section, we illustrate how **IF** can give presenters real-time feedback about audience attention and engagement. In the application, **IF** reports the percentage of audience members that are attentive to the speaker at any given time. This capability makes possible timely insights into the feelings of an audience and contingent corrections by the presenter in response to such feedback.

We recorded a 45-minute presentation, "Common Sense for Research (and Life)," by one of the authors to an audience of 12 students. The setup was illustrated above in Figure 2. Figure 8 illustrates the curves of attentiveness and amusement from **IF**. To compute dynamic variation in attention and amusement, continuous output of the pose and smile detectors was averaged within each video frame among all the participants. The large peak in Figure 8b corresponds to the moment of the following joke: "Artificial intelligence is no match for natural stupidity." Figure 8d, shows a dissociation between attention and positive affect. The speaker has just itemized a list of necessary but not necessarily enjoyable tasks that budding academics must accomplish. In the latter, the audience is attentive but perhaps less than enthusiastic about undertaking such tasks for themselves.

B. Parent-infant interaction

Emotional moments are memorable, especially those between parents and infants. Existing evidence suggests that infants show positive affect and well-coordinated contingent responsiveness with their parent(s). This experiment aims to discover these positive patterns in an unsupervised fashion.

We performed this experiment on a mother and infant dyad from the parent-infant dataset [22]. This dataset consists of face-to-face interactions of ethnically diverse six-month-old infants and their mother or father. All parents gave their informed consent to the procedures. Parent and infant were each recorded using hardware-synchronized cameras. The face-to-face interactions consisted of three contiguous episodes: three minutes of normal interaction (Face-to-Face: FF) where the parent plays with the infant as they might do at home, two minutes in which the parent remained oriented toward the infant but was unresponsive (Still-Face: SF), and three minutes in which the parent resumed normal play (Reunion: RE). For the case study, we selected hardware-synchronized video (5,684 frames) from the FF episode for one mother and her baby.

Similar to [7], we extracted shape features that corresponded to lip height, inner lips height, mouth width and mouth angle (see Figure 10 for an illustration). Figure 9 depicts synchronized facial behaviors in the synchronized videos. As shown in the figure, the red framed rectangles

indicate the discovered dyadic facial motions that correspond to joint smiles on both the mother and the infant. This is a moment of interest for parents during interactions with their infants and of interest to developmental psychologists that study such interactions. Recall that our algorithm is able to detect the start and end of the segments with maximum correlation using a user-specified time lag. In this case, $T = 30$, the infant's series is shifted by one second. The decision about time lag is informed by developmental literature on infant responsiveness.

VII. Conclusion

This paper presents **IntraFace**, a publicly available software package that integrates state-of-the-art algorithms for facial feature tracking, head pose estimation, facial attribute detection, synchrony detection, and facial expression analysis of multiple people in a video. This paper explores the new problem of synchrony detection—finding correlated facial behavior— and monitoring of audience attention and response to a speaker. In quantitative tests and in case studies, **IF** achieves state-of-the-art performance. **IF** is available at http://www.humansensing.cs.cmu.edu/intraface/ free of charge for non-commercial use to meet diverse applications in facial expression analysis.

VIII. Acknowledgments

Work on **IntraFace** began in 2008. Many people have contributed to the effort. We would like to thank in particular Ferran Altarriba, Marc Estruch, Santiago Ortega, Javier Lopez, Xavier Perez, Tomas Simon, and Zengyin Zhang for early contributions. **IF** has been supported in part by NIH grants MH096951 and GM105004, NSF grant RI-1116583, and FHWA grant DTFH61-14-C-00001.

References

[1] M. Bartlett, G. Littlewort, T. Wu, and J. Movellan. Computer expression recognition toolbox. In *Automatic Face & Gesture Recognition*, 2008.

[2] M. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[3] V. Bruce. What the human face tells the human mind: Some challenges for the robot-human interface. In *IEEE Int. Workshop on Robot and Human Communication*, 1992.

[4] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *PAMI*, 32(5):770–787, 2010.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

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

[7] W.-S. Chu, F. Zhou, and F. De la Torre. Unsupervised temporal commonality discovery. In *ECCV*. 2012.

[8] J. F. Cohn and F. De la Torre. *The Oxford Handbook of Affective Computing*, chapter Automated Face Analysis for Affective Computing. 2014.

[9] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive SVMs. *JMLR*, 7:1687–1712, 2006.

[10] Y. Dai, Y. Shibata, T. Ishii, K. Hashimoto, K. Katamachi, K. Noguchi, N. Kakizaki, and D. Ca. An associate memory model of facial expressions and its application in facial expression recognition of patients on bed. In *ICME*, pages 591 – 594, 2001.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[12] C. Darwin. *The expression of the emotions in man and animals.* New York: Oxford University., 1872/1998.

[13] F. De la Torre and J. F. Cohn. *Guide to Visual Analysis of Humans: Looking at People*, chapter Facial Expression Analysis. Springer, 2011.

[14] M. De la Torre, E. Granger, P. V. Radtke, R. Sabourin, and D. O. Gorodnichy. Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information Fusion*, 2014.

[15] X. Ding, W.-S. Chu, F. D. Torre, J. F. Cohn, and Q. Wang. Facial action unit event detection by cascade of tasks. In *ICCV*, 2013.

[16] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[17] P. Ekman, W. Friesen, and J. Hager. Facial action coding system: Research nexus. *Network Research Information, Salt Lake City, UT*, 2002.

[18] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing*, 27(5):545–559, Apr. 2009.

[19] E. E. Forbes, J. F. Cohn, N. B. Allen, and P. M. Lewinsohn. Infant affect during parent-infant interaction at 3 and 6 months: Differences between mothers and fathers and influence of parent history of depression. *Infancy*, 5:61–84, 2004.

[20] J. Gratch, L. Cheng, S. Marsella, and J. Boberg. Felt emotion and social context determine the intensity of smiles in a competitive video game. In *Automatic Face and Gesture Recognition*, 2013.

[21] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[22] Z. Hammal, J. F. Cohn, D. S. Messinger, W. I. Mattson, and M. H. Mahoor. Head movement dynamics during normal and perturbed parent-infant interaction. In *Affective Computing and Intelligent Interaction*, 2013.

[23] D. Huang and F. De la Torre. Bilinear kernel reduced rank regression for facial expression synthesis. In *ECCV*, 2010.

[24] E. Hutchings. Sony technology gives gaming avatars same facial expression as players, 2012.

[25] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[26] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[27] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition*, 2011.

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

[29] H. Matsuo and A. Khiat. Prediction of drowsy driving by monitoring driver's behavior. In *ICPR*, 2012.

[30] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[31] G. Shergill, H. Sarrafzadeh, O. Diegel, and A. Shekar. Computerized sales assistants: The application of computer technology to measure consumer interest;a conceptual framework. *Journal of Electronic Commerce Research*, 9(2):176–191, 2008.

[32] S. S. Tomkins. *Affect, imagery, consciousness.* New York: Springer., 1962.

[33] C. Vallespi and F. De la Torre. Automatic clustering of faces in meetings. In *ICIP*, 2006.

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

[35] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[36] X. Xiong and F. De la Torre. Supervised descent method for solving nonlinear least squares problems in computer vision. *arXiv preprint arXiv:1405.0601*, 2014.