AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild

Ali Mollahosseini, *Student Member, IEEE,* Behzad Hasani, *Student Member, IEEE,* and Mohammad H. Mahoor, *Senior Member, IEEE*

Abstract—Automated affective computing in the wild setting is a challenging problem in computer vision. Existing annotated databases of facial expressions in the wild are small and mostly cover discrete emotions (aka the categorical model). There are very limited annotated facial databases for affective computing in the continuous dimensional model (e.g., valence and arousal). To meet this need, we collected, annotated, and prepared for public distribution a new database of facial emotions in the wild (called *AffectNet*). *AffectNet* contains more than 1,000,000 facial images from the Internet by querying three major search engines using 1250 emotion related keywords in six different languages. About half of the retrieved images were manually annotated for the presence of seven discrete facial expressions and the intensity of valence and arousal. *AffectNet* is by far the largest database of facial expression, valence, and arousal in the wild enabling research in automated facial expression recognition in two different emotion models. Two baseline deep neural networks are used to classify images in the categorical model and predict the intensity of valence and arousal. Various evaluation metrics show that our deep neural network baselines can perform better than conventional machine learning methods and off-the-shelf facial expression recognition systems.

Index Terms—Affective computing in the wild, facial expressions, continuous dimensional space, valence, arousal.

1 Introduction

Affect is a psychological term used to describe the outward expression of emotion and feelings. Affective computing seeks to develop systems and devices that can recognize, interpret, and simulate human affects through various channels such as face, voice, and biological signals [1]. Face and facial expressions are undoubtedly one of the most important nonverbal channels used by the human being to convey internal emotion.

There have been tremendous efforts to develop reliable automated Facial Expression Recognition (FER) systems for use in affect-aware machines and devices. Such systems can understand human emotion and interact with users more naturally. However, current systems have yet to reach the full emotional and social capabilities necessary for building rich and robust Human Machine Interaction (HMI). This is mainly due to the fact that HMI systems need to interact with humans in an uncontrolled environment (aka wild setting) where the scene lighting, camera view, image resolution, background, users head pose, gender, and ethnicity can vary significantly. More importantly, the data that drives the development of affective computing systems and particularly FER systems lack sufficient variations and annotated samples that can be used in building such systems.

There are several models in the literature to quantify affective facial behaviors: 1) categorical model, where the emotion/affect is chosen from a list of affective-related categories such as six basic emotions defined by Ekman *et al.* [2], 2) dimensional model, where a value is chosen over a continuous emotional scale, such as valence and arousal [3] and 3) Facial Action Coding System (FACS) model, where all possible facial actions are described in terms of Action Units (AUs) [4]. FACS model explains facial movements and does not describe the affective state directly. There are several methods to convert AUs to affect space (e.g., EMFACS [5] states that the occurrence of AU6 and AU12 is a sign of happiness). In the categorical model, mixed emotions cannot adequately be transcribed into a limited set of words. Some researchers tried to define multiple distinct compound emotion categories (e.g., happily surprised, sadly fearful) [6] to overcome this limitation. However, still the set is limited, and the intensity of the emotion cannot be defined in the categorical model. In contrast, the dimensional model of affect can distinguish between subtly different displays of affect and encode small changes in the intensity of each emotion on a continuous scale, such as valence and arousal. Valence refers to how positive or negative an event is, and arousal reflects whether an event is exciting/agitating or calm/soothing [3]. Figure 1 shows samples of facial expressions represented in the 2D space of valence and arousal. As it is shown, there are several different kinds of affect and small changes in the same emotion that cannot be easily mapped into a limited set of terms existing in the categorical model.

The dimensional model of affect covers both intensity and different emotion categories in the continuous domain. Nevertheless, there are relatively fewer studies on developing automated algorithms in measuring affect using the continuous dimensional model (e.g., valence and arousal). One of the main reasons is that creating a large database to cover the entire continuous space of valence and arousal is expensive and there are very limited annotated face databases in the continuous domain. This paper contributes to the field of affective computing by providing a large annotated face database of the dimensional as well as the

• *Authors are with the Department of Electrical and Computer Engineering, University of Denver, Denver, CO, 80210.*
E-mail: amollah, bhasani, mmahoor@du.edu

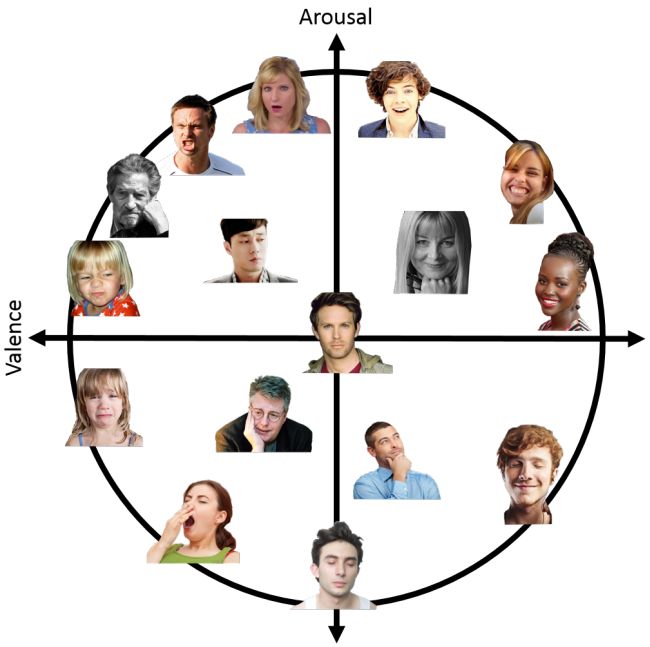


Fig. 1. Sample images in Valence Arousal circumplex

categorical models of affect.

The majority of the techniques for automated affective computing and FER are based on supervised machine learning methodologies. These systems require annotated image samples for training. Researchers have created databases of human actors/subjects portraying basic emotions [7], [8], [9], [10], [11]. Most of these databases mainly contain posed expressions acquired in a controlled lab environment. However, studies show that posed expressions can be different from unposed expressions in configuration, intensity, and timing [12], [13]. Some researchers captured unposed facial behavior while the subject is watching a short video [14], [15], engaged in laboratory-based emotion inducing tasks [16], or interacted with a computer-mediated tutoring system [17]. Although a large number of frames can be obtained by these approaches, the diversity of these databases is limited due to the number of subjects, head position, and environmental conditions.

Recently, databases of facial expression and affect in the wild received much attention. These databases are either captured from movies or the Internet, and annotated with categorical model [18], [19], [20], dimensional model [21], and FACS model [22]. However, they only cover one model of affect, have a limited number of subjects, or contain few samples of certain emotions such as disgust. Therefore, a large database, with a large amount of subject variations in the wild condition that covers multiple models of affect (especially the dimensional model) is a need.

To address this need, we created a database of facial **Affect** from the Inter**Net** (called *AffectNet*) by querying different search engines (Google, Bing, and Yahoo) using 1250 emotion related tags in six different languages (English, Spanish, Portuguese, German, Arabic, and Farsi). *AffectNet* contains more than one million images with faces and extracted facial landmark points. Twelve human experts manually annotated 450,000 of these images in both categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face. Figure 1 shows sample images from *AffectNet* and their valence and arousal annotations.

To calculate the agreement level between the human labelers, 36,000 images were annotated by two human labelers. *AffectNet* is by far the largest database of facial affect in still images which covers both categorical and dimensional models. The cropped region of the facial images, the facial landmark points, and the affect labels will be publicly available to the research community[1]. Considering the lack of in-the-wild large facial expressions datasets and more specifically annotated face datasets in the continuous domain of valence and arousal, *AffectNet* is a great resource which will enable further progress in developing automated methods for facial behavior computing in both the categorical and continuous dimensional spaces.

The rest of this paper is organized as follows. Section 2 reviews the existing databases and state-of-the-art methods for facial expression recognition with emphasis on the dimensional model and in the wild setting databases. Section 3 explains the process of collecting *AffectNet* images from the Internet and annotating the categorical and dimensional models. Section 4 presents two different baselines for automatic recognition of categorical emotions and prediction of dimensional valence and arousal in the continuous space using *AffecNet* images. Finally Section 5 concludes the paper.

2 Related Work

2.1 Existing databases

Early databases of facial expressions such as JAFFE [7], Cohn-Kanade [8], [9], MMI [10], and MultiPie [11] were captured in a lab-controlled environment where the subjects portrayed different facial expressions. This approach resulted in a clean and high-quality database of posed facial expressions. However, posed expressions may differ from daily life unposed (aka spontaneous) facial expressions. Thus, capturing spontaneous expression became a trend in the affective computing community. Examples of these environments are recording the responses of participants' faces while watching a stimuli (e.g., DISFA [14], AM-FED [15]) or performing laboratory-based emotion inducing tasks (e.g., Belfast [16]). These databases often capture multi-modal affects such as voice, biological signals, etc. and usually a series of frames are captured that enable researchers to work on temporal and dynamic aspects of expressions. However, the diversity of these databases is limited due to the number of subjects, head pose variation, and environmental conditions.

Hence there is a demand to develop systems that are based on natural, unposed facial expressions. To address this demand, recently researchers paid attention to databases in the wild. Dhall *et al.* [18] released *Acted Facial Expressions in the Wild (AFEW)* from 54 movies by a recommender system based on subtitles. The video clips were annotated with six basic expressions plus neutral. AFEW

1. Interested researcher can download a copy of AffectNet from: http://mohammadmahoor.com/databases-codes/

TABLE 1
The Summary and Characteristics of Reviewed Databases in Affect Recognition

Database	Database information	# of Subjects	Condition	Affect Modeling
CK+ [9]	- Frontal and 30 degree images	- 123	- Controlled - Posed	- 30 AUs - 7 emotion categories
MultiPie [11]	- Around 750,000 images - Under multiple viewpoints and illuminations	- 337	- Controlled - Posed	- 7 emotion categories
MMI [10]	- Subjects portrayed 79 series of facial expressions - Image sequence of frontal and side view are captured	- 25	- Controlled - Posed & Spontaneous	- 31 AUs - Six basic expression
DISFA [14]	- Video of subjects while watching a four minutes video - Clip are recorded by a stereo camera	- 27	- Controlled - Spontaneous	- 12 AUs
SALDB [23], [24]	- SAL - Audiovisual (facial expression,shoulder, audiocues) - 20 facial feature points, 5 shoulder points for video	- 4	- Controlled - Spontaneous	- Valence - Quantized [23] - Continuous [24]
RELOCA [25]	- Multi-modal audio, video, ECG and EDA	- 46	- Controlled - Spontaneous	- Valence and arousal (continuous) - Self assessment
AM-FED [15]	- 242 facial videos	- 242	- Spontaneous	- 14 AUs
DEAP [26]	- 40 one-minute long videos shown to subjects - EEG signals recorded	- 32	- Controlled - Spontaneous	- Valence and arousal (continuous) - Self assessment
AFEW [18]	- Videos	- 330	- Wild	- 7 emotion categories
FER-2013 [19]	- Images queried from web	- ∼35,887	- Wild	- 7 emotion categories
EmotioNet [22]	- Images queried from web - 100,000 images annotated manually - 900,000 images annotated automatically	- ∼100,000	- Wild	- 12 AUs annotated - 23 emotion categories based on AUs
Aff-Wild [21]	- 500 videos from YouTube	- 500	- Wild	- Valence and arousal (continuous)
FER-Wild [20]	- 24,000 images from web	- ∼24,000	- Wild	- 7 emotion categories
AffectNet **(This work)**	**- 1,000,000 images with facial landmarks** **- 450,000 images annotated manually**	**- ∼450,000**	**- Wild**	**- 8 emotion categories** **- Valence and arousal (continuous)**

contains 330 subjects aged 1-77 years and addresses the issue of temporal facial expressions in the wild. A static subset (SFEW [27]) is created by selecting some frames of AFEW. SFEW covers unconstrained facial expressions, different head poses, age range, occlusions, and close to real world illuminations. However, it contains only 700 images, and there are only 95 subjects in the database.

The *Facial Expression Recognition 2013 (FER-2013)* database was introduced in the ICML 2013 Challenges in Representation Learning [19]. The database was created using the Google image search API that matched a set of 184 emotion-related keywords to capture the six basic expressions as well as the neutral expression. Images were resized to 48x48 pixels and converted to grayscale. Human labelers rejected incorrectly labeled images, corrected the cropping if necessary, and filtered out some duplicate images. The resulting database contains 35,887 images most of which are in the wild settings. FER-2013 is currently the biggest publicly available facial expression database in the wild settings, enabling many researchers to train machine learning methods such as Deep Neural Networks (DNNs) where large amounts of data are needed. In FER-2013, the faces are not registered, a small number of images portray disgust (547 images), and unfortunately most of facial landmark detectors fail to extract facial landmarks at this resolution and quality. In addition, only the categorical model of affect is provided with FER-2013.

The *Affectiva-MIT Facial Expression Dataset (AM-FED)* database [15] contains 242 facial videos (160K frames) of people watching Super Bowl commercials using their webcam. The recording conditions were arbitrary with different illumination and contrast. The database was annotated frame-by-frame for the presence of 14 FACS action units, head movements, and automatically detected landmark points. AM-FED is a great resource to learn AUs in the wild. However, there is not a huge variance in head pose (limited profiles), and there are only a few subjects in the database.

The FER-Wild [20] database contains 24,000 images that are obtained by querying emotion-related terms from three search engines. The OpenCV face recognition was used to detect faces in the images, and 66 landmark points were found using Active Appearance Model (AAM) [28] and a face alignment algorithm via regression local binary features [29], [30]. Two human labelers annotated the images into six basic expressions and neutral. Comparing with FER-2013, FER-Wild images have a higher resolution with facial landmark points necessary to register the images. However, still a few samples portray some expressions such as disgust and fear and only the categorical model of affect is provided with FER-Wild.

The *EmotioNet* [22] consists of one million images of facial expressions downloaded from the Internet by selecting all the words derived from the word "feeling" in WordNet [31]. Face detector [32] was used to detect faces in these images and the authors visually inspected the resultant images. These images were then automatically annotated with AUs and AU intensities by an approach based on Kernel Subclass Discriminant Analysis (KSDA) [33]. The KSDA-based approach was trained with Gabor features centered on facial landmark with a Radial Basis Function (RBF) kernel. Images were labeled as one of the 23 (basic or compound) emotion categories defined in [6] based on

AUs. For example, if an image has been annotated as having AUs 1, 2, 12 and 25, it is labeled as happily surprised. A total of 100,000 images (10% of the database) were manually annotated with AUs by experienced coders. The proposed AU detection approach was trained on CK+ [9], DISFA [14], and CFEE [34] databases, and the accuracy of the automated annotated AUs was reported about 80% on the manually annotated set. EmotioNet is a novel resource of FACS model in the wild with a large amount of subject variation. However, it lacks the dimensional model of affect, and the emotion categories are defined based on annotated AUs and not manually labeled.

On the other hand, some researchers developed databases of the dimensional model in the continuous domain. These databases, however, are limited since the annotation of continuous dimensions is more expensive and necessitate trained annotators. Examples of these databases are *Belfast* [16], *RECOLA* [25], *Affectiva-MIT Facial Expression Dataset (AM-FED)* [15], and recently published *Aff-Wild Database* [21] which is the only database of dimensional model in the wild.

The Belfast database [16] contains recordings (5s to 60s in length) of mild to moderate emotional responses of 60 participants to a series of laboratory-based emotion inducing tasks (e.g., surprise response by setting off a loud noise when the participant is asked to find something in a black box). The recordings were labeled by information on self-report of emotion, the gender of the participant/experimenter, and the valence in the continuous domain. The arousal dimension was not annotated in Belfast database. While the portrayed emotions are natural and spontaneous, the tasks have taken place in a relatively artificial setting of a laboratory where there was a control on lighting conditions, head poses, etc.

The *Database for Emotion Analysis using Physiological Signals (DEAP)* [26] consists of spontaneous reactions of 32 participants in response to one-minute long music video clip. The EEG, peripheral physiological signals, and frontal face videos of participants were recorded, and the participants rated each video in terms of valence, arousal, like/dislike, dominance, and familiarity. Correlations between the EEG signal frequencies and the participants ratings were investigated, and three different modalities, i.e., EEG signals, peripheral physiological signals, and multimedia features on video clips (such as lighting key, color variance, etc.) were used for binary classification of low/high arousal, valence, and liking. DEAP is a great database to study the relation of biological signals and dimensional affect, however, it has only a few subjects and the videos are captured in lab controlled settings.

The RECOLA benchmark [25] contains videos of 23 dyadic teams (46 participants) that participated in a video conference completing a task which required collaboration. Different multi-modal data of the first five minutes of interaction, i.e., audio, video, ECG and EDA) were recorded continuously and synchronously. Six annotators measured arousal and valence. The participants reported their arousal and valence through the Self-Assessment Manikin (SAM) [35] questionnaire before and after the task. RECOLA is a great database of the dimensional model with multiple cues and modalities, however, it contains only 46 subjects and the videos were captured in the lab controlled settings.

Audio-Visual Emotion recognition Challenge (AVEC) series of competitions [36], [37], [38], [39], [40], [41] provided a benchmark of automatic audio, video and audio-visual emotion analysis in continuous affect recognition. AVEC 2011, 2012, 2013, and 2014 used videos from the SEMAINE [42] database videos. Each video is annotated by a single rater for every dimension using a two-axis joystick. AVEC 2015 and 2016 used the RECOLA benchmark in their competitions. Various continuous affect recognition dimensions were explored in each challenge year such as valence, arousal, expectation, power, and dominance, where the prediction of valence and arousal are studied in all challenges.

The *Aff-Wild* Database [21] is by far the largest database for measuring continuous affect in the valence-arousal space "in-the-wild". More than 500 videos from YouTube were collected. Subjects in the videos displayed a number of spontaneous emotions while watching a particular video, performing an activity, and reacting to a practical joke. The videos have been annotated frame-by-frame by three human raters, utilizing a joystick-based tool to rate valence and arousal. *Aff-Wild* is a great database of dimensional modeling in the wild that considers the temporal changes of the affect, however, it has a small subject variance, i.e., it only contains 500 subjects.

Table 1 summarizes the characteristics of the reviewed databases in all three models of affect, i.e., categorical model, dimensional model, and Facial Action Coding System (FACS).

2.2 Evaluation Metrics

There are various evaluation metrics in the literature to measure the reliability of annotation and automated affective computing systems. Accuracy, F1-score [49], Cohens kappa [50], Krippendorfs Alpha [51], ICC [52], area under the ROC curve (AUC), and area under Precision-Recall curve (AUC-PR) [53] are well-defined widely used metrics for evaluation of the categorical and FACS-based models. Since, the dimensional model of affect is usually evaluated in a continuous domain, different evaluation metrics are necessary. In the following, we review several metrics that are used in the literature for evaluation of dimensional model.

Root Mean Square Error (RMSE) is the most common evaluation metric in a continuous domain which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2} \quad (1)$$

where $\hat{\theta}_i$ and θ_i are the prediction and the ground truth of i^{th} sample, and n is the number of samples in the evaluation set. RMSE-based evaluation can heavily weigh the outliers [54], and it is not able to provide the covariance of prediction and ground-truth to show how they change with respect to each other. Pearsons correlation coefficient

TABLE 2
State-of-the-art Algorithms and Their Performance on the Databases Listed in Table 1.

Work	Database	Method	Results
Mollahosseini *et al.* [43]	CK+ MultiPie	- Inception based Convolutional Neural Network (CNN) - Subject-independent and cross-database experiments	- 93.2% accuracy on CK+ - 94.7% accuracy on MultiPie
Shan *et al.* [44]	MMI	- Different SVM kernels trained with LBP features - Subject-independent and cross-database experiments	- 86.9% accuracy on MMI
Zhang *et al.* [45]	DISFA	- $l_p norm$ multi-task multiple kernel learning - learning shared kernels from a given set of base kernels	- 0.70 F1-score on DISFA - 0.93 recognition rate on DISFA
Nicolaou *et al.* [24]	SALDB	- Bidirectional LSTM - Trained on multiple engineered features extracted from audio, facial geometry , and shoulder	- Leave-one-sequence-out - BLSTM-NN outperform SVR - Valence (RMSE=0.15 and CC=0.796) - Arousal (RMSE=0.21 and CC=0.642)
He *et al.* [46]	RECOLA	- Multiple stack of bidirectional LSTM (DBLSTM-RNN) - Trained on engineered features extracted from audio (LLDs), video (LPQ-TOP), 52 ECG features, and 22 EDA features	- Winner of AVEC 2015 challenge - Valence (RMSE=0.104 and CC=0.616) - Arousal (RMSE=0.121 and CC=0.753)
McDuff *et al.* [15]	AM-FED	- HOG features extracted - SVM with RBF kernel	- AUC 0.90, 0.72 and 0.70 for smile, AU2 and AU4 respectively
Koelstra *et al.* [26]	DEAP	- Gaussian naive Bayes classifier - EEG, physiological signals, and multimedia features - Binary classification of low/high arousal, valence, and liking	- 0.39 F1-score on Arousal - 0.37 F1-score on Valence - 0.40 F1-score on Liking
Fan *et al.* [47]	AFEW	- Trained on both video and audio. - VGG network are followed by LSTMs and combined with 3D convolution	- Winner of EmotiW 2016 challenge - 56.16% accuracy on AFEW
Tang *et al.* [48]	FER-2013	- CNN with linear one-vs-all SVM at the top	- Winner of the FER challenge - 71.2% accuracy on test set
Benitez-Quiroz *et al.* [22]	EmotioNet	- New face feature extraction method using Gabor filters - KSDA classification - Subject-independent and cross-database experiments	- ∼80% AU detection on EmotioNet
Mollahosseini *et al.* [20]	FER-Wild	- Trained on AlexNet - Noise estimation methods used	- 82.12% accuracy on FER-Wild

is therefore proposed in some literature [24], [36], [37] to overcome this limitation:

$$CC = \frac{COV\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \quad (2)$$

Concordance Correlation Coefficient (CCC) is another metric [40], [41] which combines the Pearsons correlation coefficient (CC) with the square difference between the means of two compared time series:

$$\rho_c = \frac{2\rho\sigma_{\hat{\theta}}\sigma_{\theta}}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2} \quad (3)$$

where ρ is the Pearson correlation coefficient (CC) between two time-series (e.g., prediction and ground-truth), $\sigma_{\hat{\theta}}^2$ and σ_{θ}^2 are the variance of each time series, and $\mu_{\hat{\theta}}$ and μ_{θ} are the mean value of each. Unlike CC, the predictions that are well correlated with the ground-truth but shifted in value are penalized in proportion to the deviation in CCC.

The value of valence and arousal are [-1,+1] and their signs are essential in many emotion-prediction applications. For example, if the ground-truth valence is +0.3, prediction of +0.7 is far better than prediction of -0.1, since +0.7 indicates a positive emotion similar to the ground-truth (despite both predictions have the same RMSE). Sign Agreement Metric (SAGR) is another metric that is proposed in [24] to evaluate the performance of a valence and arousal prediction system. SAGR is defined as:

$$SAGR = \frac{1}{n}\sum_{i=1}^{n}\delta(sign(\hat{\theta}_i), sign(\theta_i)) \quad (4)$$

where δ is the Kronecker delta function, defined as:

$$\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \quad (5)$$

The above discussed metrics are used to evaluate the categorical and dimensional baselines on *AffectNet* in Sec. 4.

2.3 Existing Algorithms

Affective computing is now a well-established field, and there are many algorithms and databases for developing automated affect perception systems. Since it is not possible to include all those great works, we only give a brief overview and cover the state-of-the-art methods that are applied on the databases explained in Sec. 2.1.

Conventional algorithms of affective computing from faces use hand-crafted features such as pixel intensities [55], Gabor filters [56], Local Binary Patterns (LBP) [44], and Histogram of Oriented Gradients (HOG) [14]. These hand-crafted features often lack enough generalizability in the wild settings where there is a high variation in scene lighting, camera view, image resolution, background, subjects head pose and ethnicity.

An alternative approach is to use Deep Neural Networks (DNN) to learn the most appropriate feature abstractions directly from the data and handle the limitations of hand-crafted features. DNNs have been a recent successful approach in visual object recognition [57], human pose estimation [58], face verification [59] and many more. This success is mainly due to the availability of computing power and existing big databases that allow DNNs to extract highly discriminative features from the data samples. There have been enormous attempts on using DNNs in automated facial expression recognition and affective computing [20], [43], [46], [47], [48] that are especially very successful in the wild settings.

Table 2 shows a list of the state-of-the-art algorithms and their performance on the databases listed in Table 1.

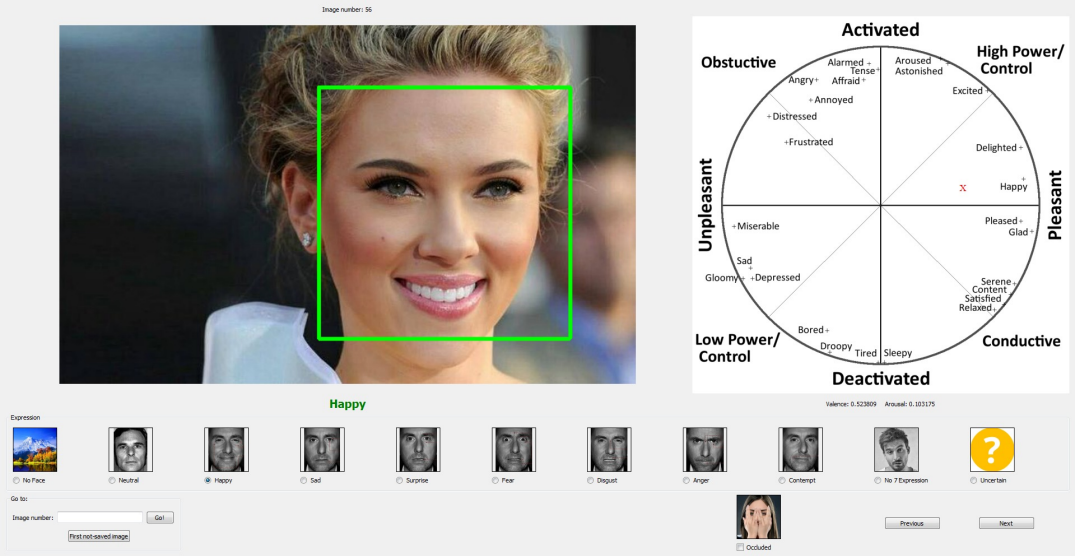


Fig. 2. A screen-shot of the software application used to annotate categorical and dimensional (valence and arousal) models of affect and the osculation tag if existing. Only one detected face in each image is annotated (shown in the green bounding box).

As shown in the table, the majority of these approaches have used DNNs to learn a better representation of affect, especially in the wild settings. Even some of the approaches, such as the winner of the AVEC 2015 challenge [46], trained a DNN with hand-crafted features and still could improve the prediction accuracy.

3 AffectNet

AffectNet (**Affect** from the Inter**Net**) is the largest database of the categorical and dimensional models of affect in the wild (as shown in Table 1). The database is created by querying emotion related keywords from three search engines and annotated by expert human labelers. In this section, the process of querying the Internet, processing facial images and extracting facial landmarks, and annotating facial expression, valence, and arousal of affect are discussed.

3.1 Facial Images from the Web

Emotion-related keywords were combined with words related to gender, age, or ethnicity, to obtain nearly 362 strings in the English language such as "joyful girl", "blissful Spanish man", "furious young lady", "astonished senior". These keywords are then translated into five other languages: Spanish, Portuguese, German, Arabic and Farsi. The direct translation of queries in English to other languages did not accurately result in the intended emotions since each language and culture has differing words and expressions for different emotions. Therefore, the list of English queries was provided to native non-English speakers who were proficient in English, and they created a list of queries for each emotion in their native language and inspected the quality of the results visually. The criteria for high-quality queries were those that returned a high percentage of human faces showing the intended queried emotions rather than drawings, graphics, or non-human objects. A total of 1250 search queries were compiled and used to crawl the search engines in our database. Since a high percentage of results returned by our query terms already contained neutral facial images, no individual query was performed to obtain additional neutral face.

Three search engines (Google, Bing, and Yahoo) were queried with these 1250 emotion related tags. Other search engines such as Baidu and Yandex were considered. However, they either did not produce a large number of facial images with intended expressions or they did not have available APIs for automatically querying and pulling image URLs into the database. Additionally, queries were combined with negative terms (e.g., "drawing", "cartoon", "animation", "birthday", etc.) to avoid non-human objects as much as possible. Furthermore, since the images of stock photo websites are posed unnaturally and contain watermarks mostly, a list of popular stock photo websites was compiled and the results returned from the stock photo websites were filtered out.

A total of ∼1,800,000 distinct URLs returned for each query were stored in the database. The OpenCV face recognition was used to obtain bounding boxes around each face. A face alignment algorithm via regression local binary features [29], [30] was used to extract 66 facial landmark points. The facial landmark localization technique was trained using the annotations provided from the 300W competition [60]. More than 1M images containing at least one face with extracted facial landmark points were kept for further processing.

The average image resolution of faces in AffectNet are 425×425 with STD of 349×349 pixels. We used Microsoft cognitive face API to extract these facial attributes on 50,000 randomly selected images from the database. According to MS face API, 49% of the faces are men. The average estimated age of the faces is 33.01 years with the standard deviation of 16.96 years. In particular, 10.85, 3.9, 30.19, 26.86, 14.46, and 13.75 percent of the faces are in age ranges [0, 10), [10, 20), [20, 30), [30, 40), [40, 50) and [50, -), respectively. MS face API detected forehead, mouth, and eye occlusions in 4.5, 1.08, and 0.49 percent of the images, respectively. Also, 9.63% of the faces wear glasses, 51.07 and 41.4% of the faces have eye and lip make-ups, respectively. In terms of head

pose, the average estimated pitch, yaw, roll are 0.0,-0.7, and -1.19 degrees, respectively.

3.2 Annotation

Crowd-sourcing services like Amazon Mechanical Turk are fast, cheap and easy approaches for labeling large databases. The quality of labels obtained from crowd-sourcing services, however, varies considerably among the annotators. Due to these issues and the fact that annotating the valence and arousal requires a deep understanding of the concept, we avoided crowd-sourcing facilities and instead hired 12 full-time and part-time annotators at the University of Denver to label the database. A total of 450,000 images were given to these expert annotators to label the face in the images into both discrete categorical and continuous dimensional (valence and arousal) models. Due to time and budget constraints each image was annotated by one annotator.

A software application was developed to annotate the categorical and dimensional (valence and arousal) models of affect. Figure 2 shows a screen-shot of the annotation application. A comprehensive tutorial including the definition of the categorical and dimensional models of affect with some examples of each category, valence and arousal was given to the annotators. Three training sessions were provided to each annotator, in which the annotator labeled the emotion category, valence and arousal of 200 images and the results were reviewed with the annotators. Necessary feedback was given on both the categorical and dimensional labels. In addition, the annotators tagged the images that have any occlusion on the face. The occlusion criterion was defined as if any part of the face was not visible. If the person in the images wore glasses, but the eyes were visible without any shadow, it was not considered as occlusion.

3.2.1 Categorical Model Annotation

Eleven discrete categories were defined in the categorical model of *AffectNet* as: Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face. The *None* ("None of the eight emotions") category is the type of expression/emotions (such as sleepy, bored, tired, seducing, confuse, shame, focused, etc.) that could not be assigned by annotators to any of the six basic emotions, contempt or neutral. However, valence and arousal could be assigned to these images. The *Non-face* category was defined as images that: 1) Do not contain a face in the image; 2) Contain a watermark on the face; 3) The face detection algorithm fails and the bounding box is not around the face; 4) The face is a drawing, animation, or painted; and 5) The face is distorted beyond a natural or normal shape, even if an expression could be inferred. If the annotators were uncertain about any of the facial expressions, images were tagged as *uncertain*. When an image was annotated as *Non-face* or *uncertain*, valence and arousal were not assigned to the image.

The annotators were instructed to select the proper expression category of the face, where the intensity is not important as long as the face depicts the intended emotion. Table 3 shows the number of images in each category. Table 4 indicates the percentage of annotated categories for queried emotion terms. As shown, the happy emotion had the highest hit-rate (48%), and the rest of the emotions had hit-rates less than 20%. About 15% of all query results were in the *No-Face* category, as many images from the web contain watermarks, drawings, etc. About 15% of all queried emotions resulted in neutral faces. Among other expressions, disgust, fear, and contempt had the lowest hit-rate with only 2.7%, 4%, and 2.4% hit-rates, respectively. As one can see, the majority of the returned images from the search engines were happy or neutral faces. The authors believe that this is because people tend to publish their images with positive expressions rather than negative expressions. Figure 3 shows a sample image in each category and its intended queries (in parentheses).

TABLE 3
Number of Annotated Images in Each Category

Expression	Number
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
Contempt	5,135
None	35,322
Uncertain	13,163
Non-Face	88,895

TABLE 4
Percentage of Annotated Categories for Queried Emotion Terms (%)

		Query Expression						
		HA	SA	SU	FE	DI	AN	CO
Annotated Expression	NE*	17.3	16.3	13.9	17.8	17.8	16.1	20.1
	HA	**48.9**	**27.2**	**30.4**	**28.6**	**33**	**29.5**	**30.1**
	SA	2.6	15.7	4.8	5.8	4.5	5.4	4.6
	SU	2.7	3.1	16	4.4	3.6	3.4	4.1
	FE	0.7	1.2	4.2	4	1.5	1.4	1.3
	DI	0.6	0.7	0.7	0.9	2.7	1.1	1
	AN	2.8	4.5	3.8	5.6	6	12.2	6.1
	CO	1.3	0.9	0.4	1.1	1.1	1.2	2.4
	NO	5.4	8.7	4.8	8.1	8.8	9.3	11.2
	UN	1.3	3.1	4.3	3.1	4.1	3.7	2.7
	NF	16.3	18.6	16.7	20.6	16.9	16.8	16.3

* NE, HA, SA, SU, FE, DI, AN, CO, NO, UN , and NF stand for Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face categories, respectively.

3.2.2 Dimensional (Valence & Arousal) Annotation

The definition of valence and arousal dimensions was adapted from [3] and was given to annotators in our tutorial as: "Valence refers to how positive or negative an event is, and arousal reflects whether an event is exciting/agitating or calm/soothing". A sample circumplex with estimated positions of several expressions, borrowed from [61], was provided in the tutorial as a reference for the annotators. The provided circumplex in the tutorial contained more than 34 complex emotions categories such as suspicious, insulted, impressed, etc., and used to train annotators. The annotators were instructed to consider the intensity of valence and arousal during the annotation. During the annotation process, the annotators were supervised closely and constant necessary feedback was provided when they were uncertain about some images.

Fig. 3. Samples of queried images from the web and their annotated tags. The queried expression is written in parentheses.

To model the dimensional affect of valence and arousal, a 2D Cartesian coordinate system was used where the x-axis and y-axis represent the valence and arousal, respectively. Similar to Russell's circumplex space model [3], our annotation software did not allow the value of valence and arousal outside of the circumplex. This allows us to convert the Cartesian coordinates to polar coordinates with $0 \leq r \leq 1$ and $0 \leq \theta < 360$. The annotation software showed the value of valence and arousal to the annotators when they selected a point in the circumplex. This helped the annotators to pick more precise locations of valence and arousal with a higher confidence.

A predefined estimated region of valence and arousal was defined for each categorical emotion in the annotation software (e.g., for happy emotion the valence is in (0.0, 1.0], and the arousal is in [-0.2, 0.5]). If the annotators select a value of valence and arousal outside of the selected emotion's region, the software indicates a warning message. The annotators were able to proceed, and they were instructed to do so, if they were confident about the value of valence and arousal. The images with the warning messages were marked in the database, for further review by the authors. This helped to avoid mistakes in the annotation of the dimensional model of affect.

Figure 4 shows the histogram (number of samples in each range/area) of annotated images in a 2D Cartesian coordinate system. As illustrated, there are more samples in the center and the right middle (positive valence and small positive arousal) of the circumplex, which confirms the higher number of Neutral and Happy images in the database compared to other categories in the categorical model. [2]

2. A numerical representation of annotated images in each range/area of valence and arousal is provided in the Appendix.

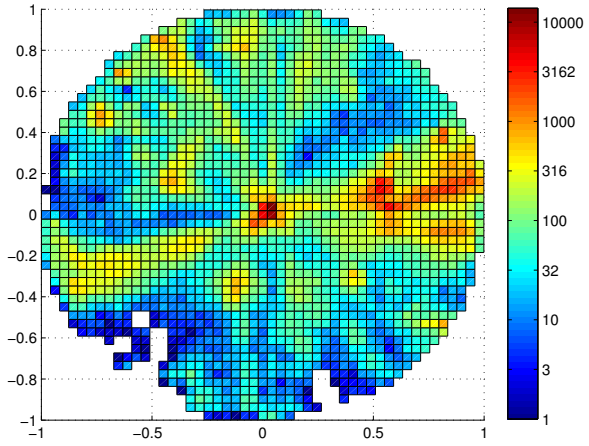


Fig. 4. Histogram (number of frames in each range/area) of valence and arousal annotations (Best viewed in color).

TABLE 5
Annotators' Agreement in Dimensional Model of Affect

	Same Category		All	
	Valence	Arousal	Valence	Arousal
RMSE	0.190	0.261	0.340	0.362
CORR	0.951	0.766	0.823	0.567
SAGR	0.906	0.709	0.815	0.667
CCC	0.951	0.746	0.821	0.551

3.3 Annotation Agreement

In order to measure the agreement between the annotators, 36,000 images were annotated by two annotators. The annotations were performed fully blind and independently, i.e., the annotators were not aware of the intended query or other annotator's response. The results showed that the annotators agreed on 60.7% of the images. Table 6 shows the agreement between two annotators for different categories. As it is shown, the annotators highly agreed on the *Happy* and *No Face* categories, and the highest disagreement occurred in the *None* category. Visually inspecting some of the images in the *None* category, the authors believe that the images in this category contain very subtle emotions and they can be easily confused with other categories (the last two example of Fig. 3 show images in the *None* category).

Table 5 shows various evaluation metrics between the two annotators in the continuous dimensional model of affect. These metrics are defined in Sec. 2.2. We calculated these metrics in two scenarios: 1) the annotators agreed on the category of the image; 2) on all images that are annotated by two annotators. As Table 5 shows, when the annotators agreed on the category of the image, the annotations have a high correlation and sign agreement (SAGR). According to Table 6, this occurred on only 60.7% images. However, there is less correlation and SAGR on overall images, since the annotators had a different perception of emotions expressed in the images. It can also be seen that the annotators agreed on valence more than arousal. The authors believe that this is because the perception of valence (how positive or negative the emotion is) is easier and less subjective than arousal (how excited or calm the subject is) especially in still images. Comparing the metrics in the existing dimensional

TABLE 6
Agreement Between Two Annotators in Categorical Model of Affect (%)

	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	None	Uncertain	Non-Face
Neutral	**50.8**	7.0	9.1	2.8	1.1	1.0	4.8	5.3	11.1	1.9	5.1
Happy	6.3	**79.6**	0.6	1.7	0.3	0.4	0.5	3.0	4.6	1.0	2.2
Sad	11.8	0.9	**69.7**	1.2	3.4	1.3	4.0	0.3	3.5	1.2	2.6
Surprise	2.0	3.8	1.6	**66.5**	14.0	0.8	1.9	0.6	4.2	1.9	2.7
Fear	3.1	1.5	3.8	15.3	**61.1**	2.5	7.2	0.0	1.9	0.4	3.3
Disgust	1.5	0.8	3.6	1.2	3.5	**67.6**	13.1	1.7	2.7	2.3	2.1
Anger	8.1	1.2	7.5	1.7	2.9	4.4	**62.3**	1.3	5.5	1.9	3.3
Contempt	10.2	7.5	2.1	0.5	0.5	4.4	2.1	**66.9**	3.7	1.5	0.6
None	**22.6**	12.0	14.5	8.0	6.0	2.3	16.9	1.3	9.6	4.3	2.6
Uncertain	13.5	12.1	7.8	7.3	4.0	4.5	6.2	2.6	12.3	**20.6**	8.9
Non-Face	3.7	3.8	1.7	1.1	0.9	0.4	1.7	0.4	1.2	1.4	**83.9**

databases (shown in Table 2) with the agreement of human labelers on *AffectNet*, suggest that *AffectNet* is a very challenging database and even human annotations have more RMSE than automated methods on existing databases.

4 Baseline

In this section, two baselines are proposed to classify images in the categorical model and predict the value of valence and arousal in the continuous domain of dimensional model. Since deep Convolutional Neural Networks (CNNs) have been a successful approach to learn appropriate feature abstractions directly from the image and there are many samples in *AffectNet* necessary to train CNNs, we proposed two simple CNN baselines for both categorical and dimensional models. We also compared the proposed baselines with conventional approaches (Support Vector Machines [62] and Support Vector Regressions [63]) learned from hand-crafted features (HOG). In the following sections, we first introduce our training, validation and test sets, and then show the performance of each proposed baselines.

4.1 Test, Validation, and Training Sets

Test set: The subset of the annotated images that are annotated by two annotators is reserved for the test set. To determine the value of valence and arousal in the test set, since there are two responses for one image in the continuous domain, one of the annotations is picked randomly. To select the category of image in the categorical model, if there was a disagreement, a favor was given to the intended query, i.e., if one of the annotators labeled the image as the intended query, the image was labeled with the intended query in the test set. This happened in 29.5% of the images with disagreement between the annotators. On the rest of the images with disagreement, one of the annotations was assigned to the image randomly. Since the test set is a random sampling of all images, it is heavily imbalanced. In other words, there are more than 11,000 images with happy expression while it contains only 1,000 images with contemptuous expression.

Validation set: Five hundred samples of each category is selected randomly as a validation set. The validation set is used for hyper-parameter tuning, and since it is balanced, there is no need for any skew normalization.

Training set: The rest of images are considered as training examples. The training examples, as shown in Table 3, are heavily imbalanced.

4.2 Categorical Model Baseline

Facial expression data is usually highly skewed. This form of imbalance is commonly referred to as *intrinsic* variation, i.e., it is a direct result of the nature of expressions in the real world. This happens in both the categorical and dimensional models of affect. For instance, Caridakis *et al.* [64] reported that a bias toward quadrant 1 (positive arousal, positive valence) exists in the SAL database. The problem of learning from imbalanced data sets has two challenges. First, training data with an imbalanced distribution often causes learning algorithms to perform poorly on the minority class [65]. Second, the imbalance in the test/validation data distribution can affect the performance metrics dramatically. Jeni *et al.* [53] studied the influence of skew on imbalanced validation set. The study showed that with exception of area under the ROC curve (AUC), all other studied evaluation metrics, i.e., Accuracy, F1-score, Cohens kappa [50], Krippendorfs Alpha [51], and area under Precision-Recall curve (AUC-PR) are affected by skewed distributions dramatically. While AUC is unaffected by skew, precision-recall curves suggested that AUC may mask poor performance. To avoid or minimize skew-biased estimates of performance, the study suggested to report both skew-normalized scores and the original evaluation.

We used AlexNet [57] architecture as our deep CNN baseline. AlexNet consists of five convolution layers, followed by max-pooling and normalization layers, and three fully-connected layers. To train our baseline with an imbalanced training set, four approaches are studied in this paper as *Imbalanced learning*, *Down-Sampling*, *Up-Sampling*, and *Weighted-Loss*. The imbalanced learning approach was trained with the imbalanced training set without any change in the skew of the dataset. To train the down-sampling approach, we selected a maximum of 15,000 samples from each class. Since there are less than 15,000 samples for some classes such as Disgust, Contempt, and Fear, the resulting training set is semi-balanced. To train the up-sampling approach, we heavily up-sampled the under-represented classes by replicating their samples so that all classes had the same number of samples as the class with maximum samples, i.e., Happy class.

The weighted-loss approach weighted the loss function for each of the classes by their relative proportion in the training dataset. In other words, the loss function heavily penalizes the networks for misclassifying examples from under-represented classes, while penalizing networks less

TABLE 7
F1-Scores of four different approaches of training AlexNet

	Imbalanced				Down-Sampling				Up-Sampling				Weighted-Loss			
	Top-1		Top-2		Top-1		Top-2		Top-1		Top-2		Top-1		Top-2	
	Orig*	Norm*	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm
Neutral	0.63	0.49	0.82	0.66	0.58	0.49	0.78	0.70	0.61	0.50	0.81	0.64	0.57	0.52	0.81	0.77
Happy	0.88	0.65	0.95	0.80	0.85	0.68	0.92	0.85	0.85	0.71	0.95	0.80	0.82	0.73	0.92	0.88
Sad	0.63	0.60	0.84	0.81	0.64	0.60	0.81	0.78	0.6	0.57	0.81	0.77	0.63	0.61	0.83	0.81
Surprise	0.61	0.64	0.84	0.86	0.53	0.63	0.75	0.83	0.57	0.66	0.80	0.81	0.51	0.63	0.77	0.86
Fear	0.52	0.54	0.78	0.79	0.54	0.57	0.80	0.82	0.56	0.58	0.75	0.76	0.56	0.66	0.79	0.86
Disgust	0.52	0.55	0.76	0.78	0.53	0.64	0.74	0.81	0.53	0.59	0.70	0.72	0.48	0.66	0.69	0.83
Anger	0.65	0.59	0.83	0.80	0.62	0.60	0.79	0.78	0.63	0.59	0.81	0.77	0.60	0.60	0.81	0.81
Contempt	0.08	0.08	0.49	0.49	0.22	0.32	0.60	0.70	0.15	0.18	0.42	0.42	0.27	0.59	0.58	0.79

*Orig and Norm stand for **Orig**inal and skew-**Norma**lized, respectively.

for misclassifying examples from well-represented classes. The entropy loss formulation for a training example (X, l) is defined as:

$$E = -\sum_{i=1}^{K} H_{l,i} log(\hat{p_i}) \tag{6}$$

where $H_{l,i}$ denotes row l penalization factor of class i, K is the number of classes, and $\hat{p_i}$ is the predictive softmax with values $[0, 1]$ indicating the predicted probability of each class as:

$$\hat{p_i} = \frac{exp(x_i)}{\sum_{j=1}^{K} exp(x_j)} \tag{7}$$

Equation (6) can be re-written as:

$$\begin{aligned} E = & -\sum_i H_{l,i} log(\frac{exp(x_i)}{\sum_j exp(x_j)}) \\ & = -\sum_i H_{l,i} x_i + \sum_i H_{l,i} log(\sum_j exp(x_j)) \\ & = log(\sum_j exp(x_j)) \sum_i H_{l,i} - \sum_i H_{l,i} x_i \end{aligned} \tag{8}$$

The derivate with respect to the prediction x_k is:

$$\begin{aligned} \frac{\partial E}{\partial x_k} &= \frac{\partial}{\partial x_k}[log(\sum_j exp(x_j)) \sum_i H_{l,i}] - \frac{\partial}{\partial x_k}[\sum_i H_{l,i} x_i] \\ &= (\sum_i H_{l,i}) \frac{1}{\sum_j exp(x_j)} \frac{\partial}{\partial x_k} \sum_j exp(x_j) - H_{l,k} \\ &= (\sum_i H_{l,i}) \frac{exp(x_k)}{\sum_j exp(x_j)} - H_{l,k} \\ &= (\sum_i H_{l,i}) \hat{p_k} - H_{l,k} \end{aligned} \tag{9}$$

When $H = I$, the identity, the proposed weighted-loss approach gives the traditional cross-entropy loss function. We used the implemented Infogain loss in Caffe [66] for this purpose. For simplicity, we used a diagonal matrix defined as:

$$H_{ij} = \begin{cases} \frac{f_i}{f_{min}}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where f_i is the number of samples of the i^{th} class and f_{min} is the number of samples in the most under-represented class, i.e., Disgust class in this situation.

Before training the network, the faces were cropped and resized to 256×256 pixels. No facial registration was performed at this baseline. To augment the data, five crops of 224×224 and their horizontal flips were extracted from the four corners and the center of the image at random during the training phase. The networks were trained for 20 epochs using a batch size of 256. The base learning rate was set to 0.01, and decreased step-wise by a factor of 0.1 every 10,000 iterations. We used a momentum of 0.9.

Table 7 shows the top-1 and top-2 F1-Scores for the imbalanced learning, down-sampling, up-sampling, and weighted-loss approaches on the test set. Since the test set is imbalanced, both the skew-normalized and the original scores are reported. The skew normalization is performed by random under-sampling of the classes in the test set. This process is repeated 200 times, and the skew-normalized score is the average of the score on multiple trials. As it is shown, the weighted-loss approach performed better than other approaches in the skew-normalized fashion. The improvement is significant in under-represented classes, i.e., Contempt, Fear, and Disgust. The imbalanced approach performed worst in the Contempt and Disgust categories since there were a few training samples of these classes compared with other classes. The up-sampling approach also did not classify the Contempt and Disgust categories well, since the training samples of these classes were heavily up-sampled (almost 20 times), and the network was over-fitted to these samples. Hence the network lost its generalization and performed poorly on these classes of the test set.

The confusion matrix of the weighted-loss approaches is shown in Table 8. The weighted-loss approach classified the samples of Contempt and Disgust categories with an acceptable accuracy but did not perform well in Happy and Neutral. This is because the network was not penalized enough for misclassifying examples from these classes. We believe that a better formulation of the weight matrix H based on the number of samples in the mini-batches or other data-driven approaches can improve the recognition of well-represented classes.

Table 9 shows accuracy, F1-score, Cohens kappa, Krippendorfs Alpha, area under the ROC curve (AUC), and area under the Precision-Recall curve (AUC-PR) on the test sets. Except for the accuracy, all the metrics are calculated in a binary-class manner where the positive class contains the samples labeled by the given category, and the negative class contains the rest. The reported result in Table 9 is the average of these metrics over eight classes. The accuracy is defined in a multi-class manner in which the number of correct predictions is divided by the total number of samples in the test set. The skew-normalization is performed by

TABLE 8
Confusion Matrix of Weighted-Loss Approach on the Test Set

		Predicted							
		NE	**HA**	**SA**	**SU**	**FE**	**DI**	**AN**	**CO**
Actual	**NE**	**53.3**	2.8	9.8	8.7	1.7	2.5	10.4	10.9
	HA	4.5	**72.8**	1.1	6.0	0.6	1.7	1.0	12.2
	SA	13.0	1.3	**61.7**	3.6	5.8	4.4	9.2	1.2
	SU	3.4	1.2	1.7	**69.9**	18.9	1.7	2.8	0.5
	FE	1.5	1.5	4.6	13.5	**70.4**	4.2	4.3	0.2
	DI	2.0	2.2	5.8	3.3	6.2	**68.6**	10.6	1.3
	AN	6.2	1.2	5.0	3.2	5.8	11.1	**65.8**	1.9
	CO	16.2	13.1	3.5	3.1	0.5	4.3	5.7	**53.8**

balancing the distribution of classes in the test set using random under-sampling and averaging over 200 trials. Since the validation set is balanced, there is no need for skew-normalization.

We compared the performance of CNN baseline with a Support Vector Machine (SVM) [62]. To train SVM, the faces in the images were cropped and resized to 256×256 pixels. HOG [67] features were extracted with the cell size of 8. We applied PCA retaining 95% of the variance to reduce the HOG features dimensionality from 36,864 to 6,697 features. We used a linear kernel SVM in Liblinear package [68] (which is optimized for large-scale linear classification and regression). Table 9 shows the evaluation metrics of SVM. Reported AUC and AUCPR values for SVM are calculated using the LibLinear's resulting decision values. We calculated the scores of predictions using a posterior-probability transformation sigmoid function. Comparing the performance of SVM with the CNN baselines on AffectNet, indicates that CNN models perform better than conventional SVM and HOG features in all metrics.

We also compared the baseline with an available off-the-shelf expression recognition system (Microsoft Cognitive Services emotion API [69]). The MS cognitive system had an excellent performance on Neutral and Happy categories with an accuracy of 0.94 and 0.85, respectively. However, it performed poorly on other classes with an accuracy of 0.25, 0.27 and 0.04 in the Fear, Disgust and Contempt categories. Table 9 shows the evaluation metrics on the MS cognitive system. Comparing the performance of the MS cognitive with the simple baselines on AffectNet indicates that *AffectNet* is a challenging database and a great resource to further improve the performance of facial expression recognition systems.

Figure 5 shows nine samples of randomly selected misclassified images of the weighted-loss approach and their corresponding ground-truth. As the figure shows, it is really difficult to assign some of the emotions to a single category. Some of the faces have partial similarities in facial features to the misclassified images, such as nose wrinkled in disgust, or eyebrows raised in surprise. This emphasizes the fact that classifying facial expressions in the wild is a challenging task and, as mentioned before, even human annotators agreed on only 60.7% of the images.

4.3 Dimensional Model (Valence and Arousal) Baseline

Predicting dimensional model in the continuous domain is a real-valued regression problem. We used AlexNet [57] architecture as our deep CNN baseline to predict the value

Fig. 5. Samples of miss-classified images. Their corresponding ground-truth is given in parentheses.

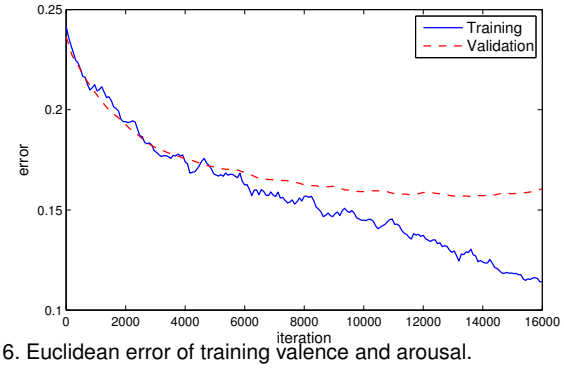


Fig. 6. Euclidean error of training valence and arousal.

of valence and arousal. Particularly, two separate AlexNets were trained where the last fully-connected layer was replaced with a linear regression layer containing only one neuron. The output of the neuron predicted the value of valence/arousal in continuous domain [-1,1]. A Euclidean (L2) loss was used to measure the distance between the predicted value ($\hat{y}_n$) and actual value of valence/arousal (y_n) as:

$$E = \frac{1}{2N}\sum_{n=1}^{N} ||\hat{y}_n - y_n||_2^2 \qquad (11)$$

The faces were cropped and resized to 256×256 pixels. The base learning rate was fixed and set to 0.001 during the training process. We used a momentum of 0.9. Training was continued until a plateau was reached in the Euclidean error of the validation set (approximately 16 epochs with a mini-batch size of 256). Figure 6 shows the value of training and validation losses over 16K iterations (about 16 epochs).

We also compared Support Vector Regression (SVR) [63] with our DNN baseline for predicting valence and arousal in *AffectNet*. In our experiments, first, the faces in the images were cropped and resized to 256×256 pixels. Histogram of Oriented Gradient (HOG) [67] features were extracted with the cell size of 8. Afterward, we applied PCA retaining 95%

TABLE 9
Evaluation Metrics and Comparison of CNN baselines, SVM and MS Cognitive on Categorical Model of Affect.

	CNN Baselines								SVM		MS Cognitive	
	Imbalanced		Down-Sampling		Up-Sampling		Weighted-Loss					
	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm
Accuracy	0.72	0.54	0.68	0.58	0.68	0.57	0.64	0.63	0.60	0.37	0.68	0.48
F_1-Score	0.57	0.52	0.56	0.57	0.56	0.55	0.55	0.62	0.37	0.31	0.51	0.45
Kappa	0.53	0.46	0.51	0.51	0.52	0.49	0.5	0.57	0.32	0.25	0.46	0.40
Alpha	0.52	0.45	0.51	0.51	0.51	0.48	0.5	0.57	0.31	0.22	0.46	0.37
AUC	0.85	0.80	0.82	0.85	0.82	0.84	0.86	0.86	0.77	0.70	0.83	0.77
AUCPR	0.56	0.55	0.54	0.57	0.55	0.56	0.58	0.64	0.39	0.37	0.52	0.50

TABLE 10
Baselines' Performances of Predicting Valence and Arousal on Test Set

	CNN (AlexNet)		SVR	
	Valence	Arousal	Valence	Arousal
RMSE	0.394	0.402	0.494	0.400
CORR	0.602	0.539	0.429	0.360
SAGR	0.728	0.670	0.619	0.748
CCC	0.541	0.450	0.340	0.199

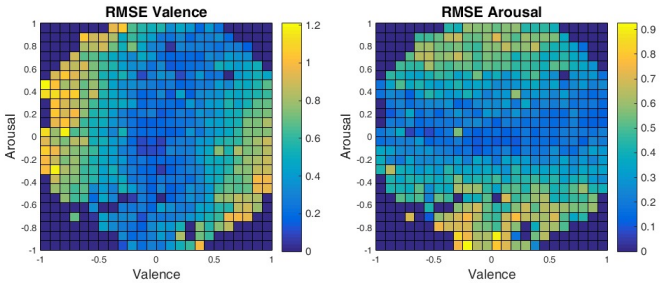


Fig. 7. RMSE of predicted valence and arousal using AlexNet and Euclidean (L2) loss (Best viewed in color).

of the variance of these features to reduce the dimensionality. Two separate SVRs were trained to predict the value of valence and arousal. Liblinear [68] package was used to implement SVR baseline.

Table 10 shows the performances of the proposed baseline and SVR on the test set. As shown, the CNN baseline can predict the value of valence and arousal better than SVR. This is because the high variety of samples in *AffectNet* allows the CNN to extract more discriminative features than hand-crafted HOG, and therefore it learned a better representation of dimensional affect.

The RMSE of CNN baseline (AlexNet) between the predicted valence and arousal and the ground-truth are shown in Fig. 7. As illustrated, the CNN baseline has a lower error rate in the center of circumplex. In particular, predicting low-valence mid-arousal and low-arousal mid-valence areas were more challenging. These areas correspond to the expressions of contempt, bored, and sleepy. It should be mentioned that predicting valence and arousal in the wild is a challenging task, and as discussed in Sec. 3.3, the disagreement between two human annotators has RMSE=0.367 and RMSE=0.481 for valence and arousal, respectively.

5 Conclusion

The analysis of human facial behavior is a very complex and challenging problem. The majority of the techniques for automated facial affect analysis are mainly based on machine learning methodologies, and their performance highly depends on the amount and diversity of annotated training samples. Recently, databases of facial expression and affect in the wild received much attention. However, existing databases of facial affect in the wild only cover one model of affect, have a limited number of subjects, or contain few samples of certain emotions.

The Internet is a vast source of facial images, most of which are captured in uncontrolled conditions. These images are often taken in the wild under natural conditions. In this paper, we introduced a new publicly available database of a facial **Affect** from the Inter**Net** (called *AffectNet*) by querying different search engines using emotion related tags in six different languages. *AffectNet* contains more than 1M images with faces and extracted landmark points. Twelve human experts manually annotated 450,000 of these images in both the categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face.

The agreement level of human labelers on a subset of *AffectNet* showed that expression recognition and predicting valence and arousal in the wild is a challenging task. The two annotators agreed on 60.7% of the category of facial expressions, and there was a large disagreement on the value of valence and arousal (RMSE=0.34 and 0.36) between the two annotators.

Two simple deep neural network baselines were examined to classify the facial expression images and predict the value of valence and arousal in the continuous domain of dimensional model. Evaluation metrics showed that simple deep neural network baselines trained on *AffectNet* can perform better than conventional machine learning methods and available off-the-shelf expression recognition systems. *AffectNet* is by far the largest database of facial expression, valence and arousal in the wild, enabling further progress in the automatic understanding of facial behavior in both categorical and continuous dimensional space. The interested investigators can study categorical and dimensional models in the same corpus, and possibly co-train them to improve the performance of their affective computing systems. It is highly anticipated that the availability of this database for the research community, along with the recent advances in deep neural networks, can improve the performance of automated affective computing systems in recognizing facial expressions and predicting valence and arousal.

Acknowledgments

This work is partially supported by the NSF grants IIS-1111568 and CNS-1427872. We gratefully acknowledge the

support of NVIDIA Corporation with the donation of the Tesla K40 GPUs used for this research.

REFERENCES

[1] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2005, pp. 981–995. 1

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971. 1

[3] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980. 1, 7, 8

[4] P. Ekman and W. V. Friesen, "Facial action coding system," 1977. 1

[5] W. V. Friesen and P. Ekman, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, p. 36, 1983. 1

[6] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014. 1, 3

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205. 2

[8] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, 2001. 2

[9] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101. 2, 3, 4

[10] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 5–pp. 2, 3

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010. 2, 3

[12] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 02, pp. 121–132, 2004. 2

[13] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 2006, pp. 162–170. 2

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 151–160, 2013. 2, 3, 4, 5

[15] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888. 2, 3, 4, 5

[16] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012. 2, 4

[17] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration." in *EDM*, 2013, pp. 43–50. 2

[18] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 509–516. 2, 3

[19] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015. 2, 3

[20] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 2, 3, 5

[21] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. A. Nicolaou, and G. Zhao, "Facial affect "in-the-wild": A survey and a new database," in *International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Affect "in-the-wild" Workshop*, June 2016. 2, 3, 4

[22] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16), Las Vegas, NV, USA*, 2016. 2, 3, 5

[23] M. A. Nicolaou, H. Gunes, and M. Pantic, "Audio-visual classification and fusion of spontaneous affective data in likelihood space," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3695–3699. 3

[24] ——, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011. 3, 5

[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8. 3, 4

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012. 3, 4, 5

[27] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2106–2112. 3

[28] A. Mollahosseini and M. H. Mahoor, "Bidirectional warping of active appearance model," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, 2013, pp. 875–880. 3

[29] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692. 3, 6

[30] L. Yu, "face-alignment-in-3000fps," https://github.com/yulequan/face-alignment-in-3000fps, 2016. 3, 6

[31] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995. 3

[32] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004. 3

[33] D. You, O. C. Hamsici, and A. M. Martinez, "Kernel optimization in discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 631–638, 2011. 3

[34] 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 *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 57–64. 4

[35] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994. 4

[36] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 415–424. 4, 5

[37] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456. 4, 5

[38] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10. 4

[39] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10. 4

[40] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and

workshop," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1335–1336. 4, 5

[41] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016-depression, mood, and emotion recognition workshop and challenge," *arXiv preprint arXiv:1605.01600*, 2016. 4, 5

[42] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012. 4

[43] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016. 5

[44] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009. 5

[45] X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using {l} _ {p}-norm mkl multiclass-svm," *Machine Vision and Applications*, pp. 1–17, 2015. 5

[46] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80. 5, 6

[47] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450. 5

[48] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013. 5

[49] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2006, pp. 1015–1021. 4

[50] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960. 4, 9

[51] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970. 4, 9

[52] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979. 4

[53] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 245–251. 4, 9

[54] S. Bermejo and J. Cabestany, "Oriented principal component analysis for large margin classifiers," *Neural Networks*, vol. 14, no. 10, pp. 1447–1461, 2001. 4

[55] M. Mohammadi, E. Fatemizadeh, and M. H. Mahoor, "Pca-based dictionary building for accurate facial expression recognition via sparse representation," *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1082–1092, 2014. 5

[56] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, 2002. 5

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 5, 9, 11

[58] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1653–1660. 5

[59] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708. 5

[60] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016. 6

[61] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 116–123, 2013. 7

[62] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 9, 11

[63] A. Smola and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997. 9, 11

[64] G. Caridakis, K. Karpouzis, and S. Kollias, "User and context adaptive neural networks for emotion recognition," *Neurocomputing*, vol. 71, no. 13, pp. 2553–2562, 2008. 9

[65] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009. 9

[66] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678. 10

[67] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893. 11

[68] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008. 11, 12

[69] "Microsoft cognitive services - emotion api," https://www.microsoft.com/cognitive-services/en-us/emotion-api, (Accessed on 12/01/2016). 11

Ali Mollahosseini received the BSc degree in computer software engineering from the Iran University of Science and Technology, Iran, in 2006, and the MSc degree in computer engineering - artificial intelligence from AmirKabir University, Iran, in 2010. He is currently working toward the Ph.D. degree and is a graduate research assistant in the Department of Electrical and Computer Engineering at the University of Denver. His research interests include deep neural networks for the analysis of facial expression, developing humanoid social robots and computer vision.

Behzad Hasani received the BSc degree in computer hardware engineering from Khaje Nasir Toosi University of Technology, Tehran, Iran, in 2013, and the MSc degree in computer engineering - artificial intelligence from Iran University of Science and Technology, Tehran, Iran, in 2015. He is currently pursuing his Ph.D. degree in electrical & computer engineering and is a graduate research assistant in the Department of Electrical and Computer Engineering at the University of Denver. His research interests include Computer Vision, Machine Learning, and Deep Neural Networks, especially on facial expression analysis.

Mohammad H. Mahoor received the BS degree in electronics from the Abadan Institute of Technology, Iran, in 1996, the MS degree in biomedical engineering from the Sharif University of Technology, Iran, in 1998, and the Ph.D. degree in electrical and computer engineering from the University of Miami, Florida, in 2007. He is an Associate Professor of Electrical and Computer Engineering at DU. He does research in the area of computer vision and machine learning including visual object recognition, object tracking and pose estimation, motion estimation, 3D reconstruction, and human-robot interaction (HRI) such as humanoid social robots for interaction and intervention with children with special needs (e.g., autism) and elderly with depression and dementia. He has received over $3M of research funding from state and federal agencies including the National Science Foundation. He is a Senior Member of IEEE and has published about 100 conference and journal papers.

APPENDIX A

TABLE 11
Samples of Annotated Categories for Queried Emotion Terms

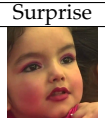
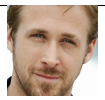
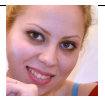
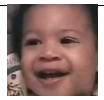
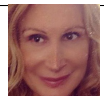
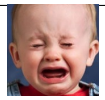
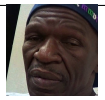
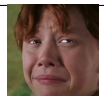
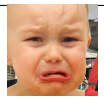
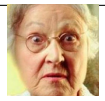
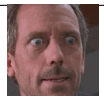
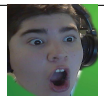
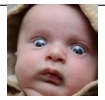
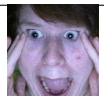
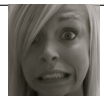
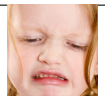
Annotated Expression	Queried Expression: Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
Neutral							
Happy							
Sad							
Surprise							
Fear							
Disgust							
Anger							
Contempt							
None							
Uncertain							
Non-Face							

TABLE 12
Samples of Annotated Images by Two Annotators (Randomly selected)

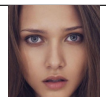
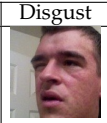
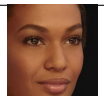
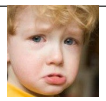
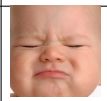
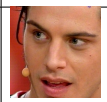
Annotator 2 \ Annotator 1	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	None	Uncertain	Non-Face
Neutral											
Happy											
Sad											
Surprise											
Fear								No Disagreement			
Disgust											
Anger											
Contempt											
None											
Uncertain											
Non-Face											

TABLE 13
Agreement percentage between Two Annotators in Categorical Model of Affect (%)

	A1*	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
A1	0.0**	69	70	68	0	0	0	0	0	0	0	0
A2	69	0	64.9	68.3	0	0	0	64.7	0	0	0	0
A3	70	64.9	0	70.6	67.4	69.9	63	62.3	0	48.1	0	0
A4	68	68.3	70.6	0	70.4	70.8	64.3	67.5	0	27.5	0	0
A5	0	0	67.4	70.4	0	70.6	0	0	0	0	0	0
A6	0	0	69.9	70.8	70.6	0	0	0	0	0	0	0
A7	0	0	63	64.3	0	0	0	0	0	75.8	0	0
A8	0	64.7	62.3	67.5	0	0	0	0	51.1	0	0	0
A9	0	0	0	0	0	0	0	51.1	0	0	54.4	0
A10	0	0	48.1	27.5	0	0	75.8	0	0	87.5	0	61.9
A11	0	0	0	0	0	0	0	0	54.4	0	0	0
A12	0	0	0	0	0	0	0	0	0	61.9	0	0

* A1 to A12 indicate Annotators 1 to 12
** Zero means that there were no common images between the two annotators

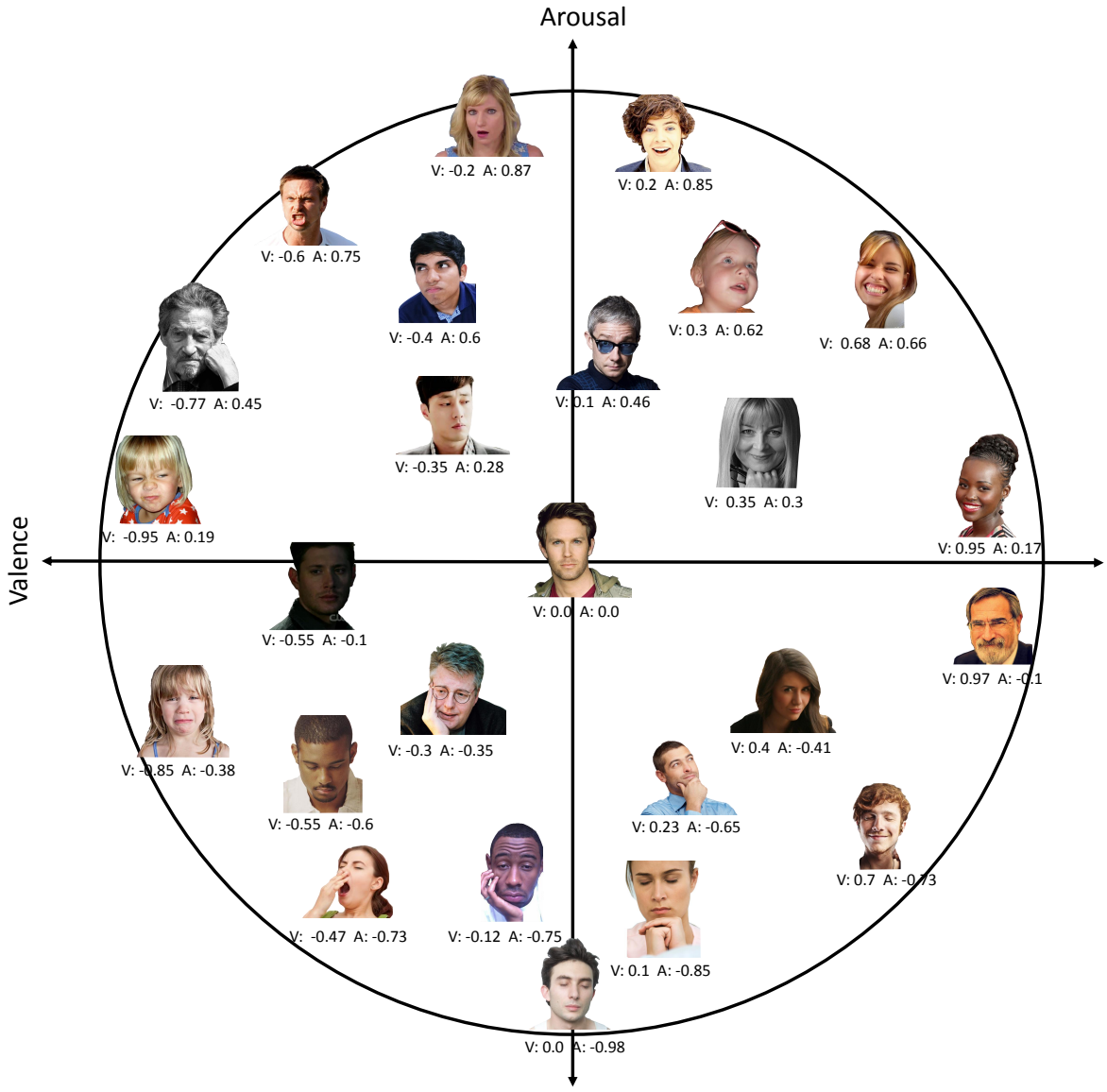


Fig. 8. Sample images in Valence Arousal circumplex with their corresponding Valence and Arousal values (V: Valence, A: Arousal).

TABLE 14
Number of annotated images in each range/area of valence and arousal

		Valence									
		[-1,-.8]	[-.8,-.6]	[-.6,-.4]	[-.4,-.2]	[-.2,0]	[0,.2]	[.2,.4]	[.4,.6]	[.6,.8]	[.8,1]
Arousal	[.8,1]	0	0	21	674	1021	521	60	57	0	0
	[.6,.8]	0	74	161	561	706	1006	432	738	530	0
	[.4,.6]	638	720	312	505	2689	1905	1228	992	3891	957
	[.2,.4]	6770	9283	3884	2473	5530	2296	3506	1824	2667	1125
	[0,.2]	3331	1286	2971	4854	14083	15300	4104	9998	13842	9884
	[-.2,0]	395	577	5422	3675	9024	23201	6237	42219	23281	21040
	[-.4,-.2]	787	1364	3700	6344	2804	1745	821	5241	10619	9934
	[-.6,-.4]	610	7800	2645	3571	2042	2517	1993	467	1271	921
	[-.8,-.6]	0	3537	8004	4374	5066	3379	4169	944	873	0
	[-1,-.8]	0	0	4123	1759	4836	1845	1672	739	0	0

TABLE 15
Evaluation Metrics and Comparison of CNN baselines, SVM and MS Cognitive on Categorical Model of Affect on the Validation Set.

	CNN Baselines				SVM	MS Cognitive
	Imbalanced	Down-Sampling	Up-Sampling	Weighted-Loss		
Accuracy	0.40	0.50	0.47	0.58	0.30	0.37
F_1-Score	0.34	0.49	0.44	0.58	0.24	0.33
Kappa	0.32	0.42	0.38	0.51	0.18	0.27
Alpha	0.39	0.42	0.37	0.51	0.13	0.23
AUCPR	0.42	0.48	0.44	0.56	0.30	0.38
AUC	0.74	0.47	0.75	0.82	0.68	0.70

TABLE 16
Baselines' Performances of Predicting Valence and Arousal on the Validation Set

	CNN (AlexNet)		SVR	
	Valence	Arousal	Valence	Arousal
RMSE	0.37	0.41	0.55	0.42
CORR	0.66	0.54	0.35	0.31
SAGR	0.74	0.65	0.57	0.68
CCC	0.60	0.34	0.30	0.18