EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild

C. Fabian Benitez-Quiroz*, Ramprakash Srinivasan*, Aleix M. Martinez
Dept. Electrical and Computer Engineering
The Ohio State University
*These authors contributed equally to this paper.

Abstract

Research in face perception and emotion theory requires very large annotated databases of images of facial expressions of emotion. Annotations should include Action Units (AUs) and their intensities as well as emotion category. This goal cannot be readily achieved manually. Herein, we present a novel computer vision algorithm to annotate a large database of one million images of facial expressions of emotion in the wild (i.e., face images downloaded from the Internet). First, we show that this newly proposed algorithm can recognize AUs and their intensities reliably across databases. To our knowledge, this is the first published algorithm to achieve highly-accurate results in the recognition of AUs and their intensities across multiple databases. Our algorithm also runs in real-time (>30 images/second), allowing it to work with large numbers of images and video sequences. Second, we use WordNet to download 1,000,000 images of facial expressions with associated emotion keywords from the Internet. These images are then automatically annotated with AUs, AU intensities and emotion categories by our algorithm. The result is a highly useful database that can be readily queried using semantic descriptions for applications in computer vision, affective computing, social and cognitive psychology and neuroscience; e.g., "show me all the images with happy faces" or "all images with AU 1 at intensity c."

1. Introduction

Basic research in face perception and emotion theory cannot be completed without large annotated databases of images and video sequences of facial expressions of emotion [7]. Some of the most useful and typically needed annotations are Action Units (AUs), AU intensities, and emotion categories [8]. While small and medium size databases can be manually annotated by expert coders over several months [11, 5], large databases cannot. For example, even if it were possible to annotate each face image very fast by an expert coder (say, 20 seconds/image)[1], it would take 5,556 hours to code a million images, which translates to 694 (8-hour) working days or 2.66 years of uninterrupted work.

This complexity can sometimes be managed, e.g., in image segmentation [18] and object categorization [17], because everyone knows how to do these annotations with minimal instructions and online tools (e.g., Amazon's Mechanical Turk) can be utilized to recruit large numbers of people. But AU coding requires specific expertise that takes months to learn and perfect and, hence, alternative solutions are needed. This is why recent years have seen a number of computer vision algorithms that provide fully- or semi-automatic means of AU annotation [20, 10, 22, 2, 26, 27, 6].

The major problem with existing algorithms is that they either do not recognize all the necessary AUs for all applications, do not specify AU intensity, are too computational demanding in space and/or time to work with large database, or are only tested within databases (i.e., even when multiple databases are used, training and testing is generally done within each database independently).

The present paper describes a new computer vision algorithm for the recognition of AUs typically seen in most applications, their intensities, and a large number (23) of basic and compound emotion categories *across databases*. Additionally, images are annotated semantically with 421 emotion keywords. (A list of these semantic labels is in the Supplementary Materials.)

Crucially, our algorithm is the first to provide reliable recognition of AUs and their intensities across databases and runs in real-time (>30 images/second). This allows us to automatically annotate a large database of a million facial expressions of emotion images "in the wild" in about 11 hours in a PC with a 2.8 GHz i7 core and 32 Gb of RAM.

The result is a database of facial expressions that can be readily queried by AU, AU intensity, emotion category, or

[1] Expert coders typically use video rather than still images. Coding in stills is generally done by comparing the images of an expressive face with the neutral face of the same individual.

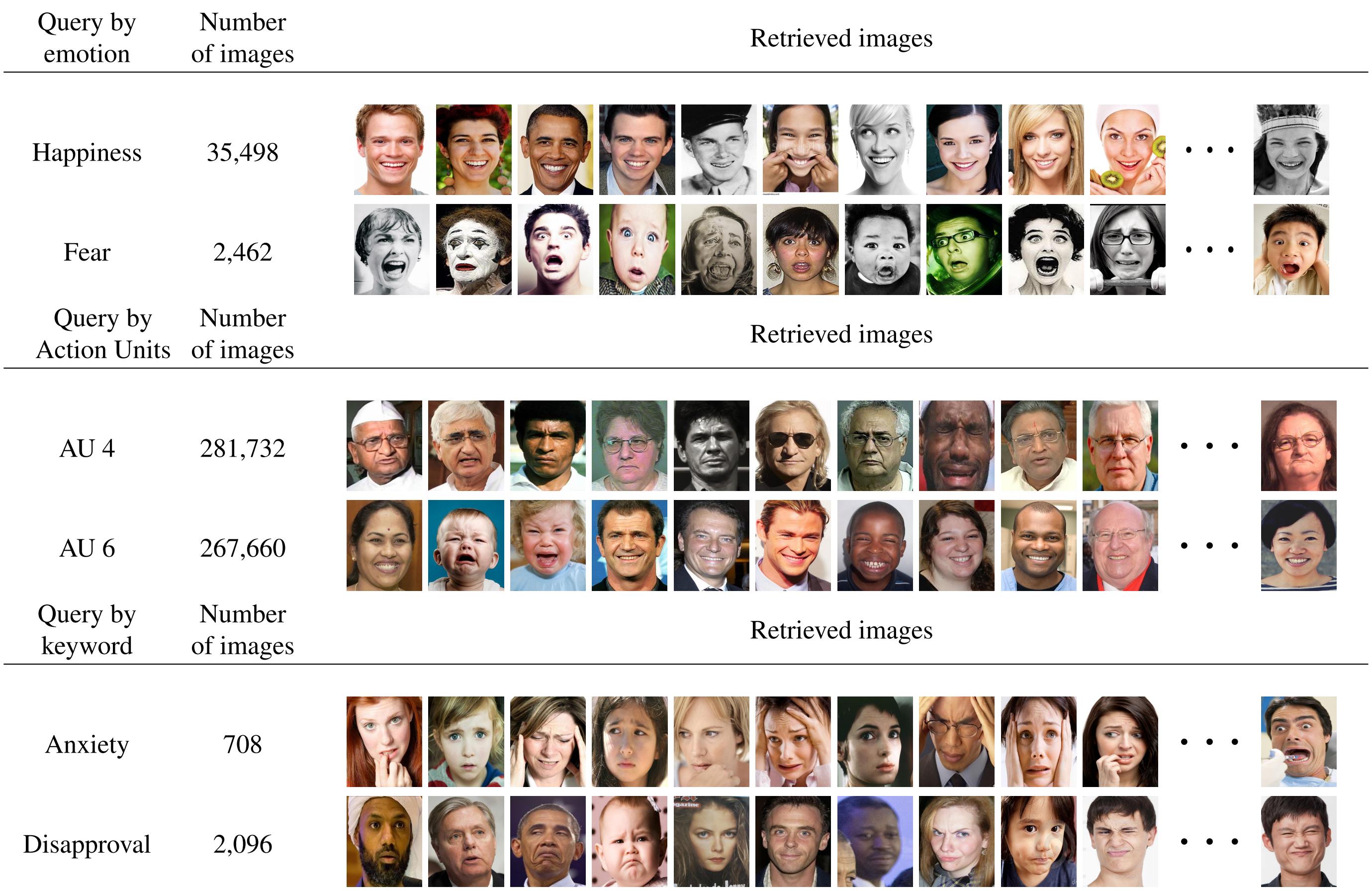


Figure 1: The computer vision algorithm described in the present work was used to automatically annotate emotion category and AU in a million face images in the wild. These images were downloaded using a variety of web search engines by selecting only images with faces and with associated emotion keywords in WordNet [15]. Shown above are three example queries. The top example is the results of two queries obtained when retrieving all images that have been identified as happy and fearful by our algorithm. Also shown is the number of images in our database of images in the wild that were annotated as either happy or fearful. The next example queries show the results of retrieving all images with AU 4 or 6 present, and images with the emotive keyword "anxiety" and "disaproval."

emotion keyword, Figure 1. Such a database will prove invaluable for the design of new computer vision algorithms as well as basic, translational and clinical studies in social and cognitive psychology, social and cognitive neuroscience, neuromarketing, and psychiatry, to name but a few.

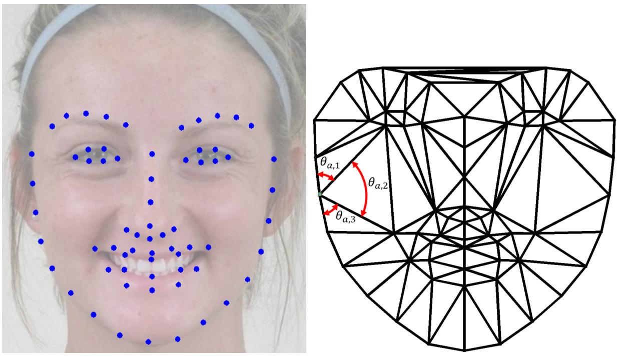
2. AU and Intensity Recognition

We derive a novel approach for the recognition of AUs. Our algorithm runs at over 30 images/second and is highly accurate even across databases. Note that, to date, most algorithms have only achieved good results within databases. *The major contributions of our proposed approach is that it achieves high recognition accuracies even across databases and runs in real time.* This is what allows us to automatically annotate a million images in the wild. We also categorize facial expressions within one of the twenty-three basic and compound emotion categories defined in [7]. Categorization of emotion is given by the detected AU pattern of activation. Not all images belong to one of these 23 categories. When this is the case, the image is only annotated with AUs, not emotion category. If an image does not have any AU active, it is classified as a neutral expression.

2.1. Face space

We start by defining the feature space employed to represent AUs in face images. Perception of faces, and facial expressions in particular, by humans is known to involve a combination of shape and shading analyses [19, 13].

Shape features thought to play a major role in the per-



(a) (b)

Figure 2: (a) Shown here are the normalized face landmarks $\hat{\mathbf{s}}_{ij}$ ($j = 1, \dots, 66$) used by the proposed algorithm. Fifteen of them correspond to anatomical landmarks (e.g., corners of the eyes, mouth and brows, tip of the nose, and chin). The others are pseudo-landmarks defined about the edge of the eyelids, mouth, brows, lips and jaw line as well as the midline of the nose going from the tip of the nose to the horizontal line given by the center of the two eyes. The number of pseudo-landmarks defining the contour of each facial component (e.g., brows) is constant. This guarantees equivalency of landmark position across people. (b) The Delaunay triangulation used by the algorithm derived in the present paper. The number of triangles in this configuration is 107. Also shown in the image are the angles of the vector $\boldsymbol{\theta}_a = (\theta_{a1}, \dots, \theta_{aq_a})^T$ (with $q_a = 3$), which define the angles of the triangles emanating from the normalized landmark $\hat{\mathbf{s}}_{ija}$.

ception of facial expressions of emotion are second-order statistics of facial landmarks (i.e., distances and angles between landmark points) [16]. These are sometimes called configural features, because they define the configuration of the face.

Let $\mathbf{s}_{ij} = \left(\mathbf{s}_{ij1}^T, \dots, \mathbf{s}_{ijp}^T\right)^T$ be the vector of landmark points in the j^{th} sample image ($j = 1, \dots, n_i$) of AU i, where $\mathbf{s}_{ijk} \in \mathbb{R}^2$ are the 2D image coordinates of the k^{th} landmark, and n_i is the number of sample images with AU i present. These face landmarks can be readily obtained with state-of-the-art computer vision algorithms. Specifically, we combine the algorithms defined in [24, 9] to automatically detect the 66 landmarks shown in Figure 2a. Thus, $\mathbf{s}_{ij} \in \mathbb{R}^{132}$.

All training images are then normalized to have the same inter-eye distance of τ pixels. Specifically, $\hat{\mathbf{s}}_{ij} = c\,\mathbf{s}_{ij}$, where $c = \tau / \|\mathbf{l} - \mathbf{r}\|_2$, $\mathbf{l}$ and $\mathbf{r}$ are the image coordinates of the center of the left and right eye, $\|.\|_2$ defines the 2-norm of a vector, $\hat{\mathbf{s}}_{ij} = \left(\hat{\mathbf{s}}_{ij1}^T, \dots, \hat{\mathbf{s}}_{ijp}^T\right)^T$ and we used $\tau = 300$. The location of the center of each eye can be readily computed as the geometric mid-point between the landmarks defining the two corners of the eye.

Now, define the shape feature vector of configural features as,

$$\mathbf{x}_{ij} = \left(d_{ij12}, \dots, d_{ijp-1\,p}, \boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_p^T\right)^T, \qquad (1)$$

where $d_{ijab} = \|\hat{\mathbf{s}}_{ija} - \hat{\mathbf{s}}_{ijb}\|_2$ are the Euclidean distances between normalized landmarks, $a = 1, \dots, p-1$, $b = a + 1, \dots, p$, and $\boldsymbol{\theta}_a = (\theta_{a1}, \dots, \theta_{aq_a})^T$ are the angles defined by each of the Delaunay triangles emanating from the normalized landmark $\hat{\mathbf{s}}_{ija}$, with q_a the number of Delaunay triangles originating at $\hat{\mathbf{s}}_{ija}$ and $\sum_{k=1}^{q_a} \theta_{ak} \leq 360^o$ (the equality holds for non-boundary landmark points). Specifically, we use the Delaunay triangulation of the face shown in Figure 2b. Note that since each triangle in this figure can be defined by three angles and we have 107 triangles, the total number of angles in our shape feature vector is 321. More generally, the shape feature vectors $\mathbf{x}_{ij} \in \mathbb{R}^{p(p-1)/2+3t}$, where p is the number of landmarks and t the number of triangles in the Delaunay triangulation. With $p = 66$ and $t = 107$, we have $\mathbf{x}_{ij} \in \mathbb{R}^{2,466}$.

Next, we use Gabor filters centered at each of the normalized landmark points $\hat{\mathbf{s}}_{ijk}$ to model shading changes due to the local deformation of the skin. When a facial muscle group deforms the skin of the face locally, the reflectance properties of the skin change (i.e., the skin's bidirectional reflectance distribution function is defined as a function of the skin's wrinkles because this changes the way light penetrates and travels between the epidermis and the dermis and may also vary their hemoglobin levels [1]) as well as the foreshortening of the light source as seen from a point on the surface of the skin.

Cells in early visual cortex in humans can be modelled using Gabor filters [4], and there is evidence that face perception uses this Gabor-like modeling to gain invariance to shading changes such as those seen when expressing emotions [3, 19, 23]. Formally, let

$$g\left(\hat{\mathbf{s}}_{ijk}; \lambda, \alpha, \phi, \gamma\right) = \exp\left(\frac{s_1^2 + \gamma^2 s_2^2}{2\sigma^2}\right) \cos\left(2\pi \frac{s_1}{\lambda} + \phi\right), \qquad (2)$$

with $\hat{\mathbf{s}}_{ijk} = \left(\hat{s}_{ijk1}, \hat{s}_{ijk2}\right)^T$, $s_1 = \hat{s}_{ijk1} \cos\alpha + \hat{s}_{ijk2} \sin\alpha$, $s_2 = -\hat{s}_{ijk1} \sin\alpha + \hat{s}_{ijk2} \cos\alpha$, λ the wavelength (i.e., number of cycles/pixel), α the orientation (i.e., the angle of the normal vector of the sinusoidal function), ϕ the phase (i.e., the offset of the sinusoidal function), γ the (spatial) aspect ratio, and σ the scale of the filter (i.e., the standard deviation of the Gaussian window).

We use a Gabor filter bank with o orientations, s spatial scales, and r phases. We set $\lambda = \{4, 4\sqrt{2}, 4 \times 2, 4(2\sqrt{2}), 4(2 \times 2)\} = \{4, 4\sqrt{2}, 8, 8\sqrt{2}, 16\}$ and $\gamma = 1$, since these values have been shown to be appropriate to represent facial expressions of emotion [7]. The values of

o, s and r are learned using cross-validation on the training set. This means, we use the following set of possible values $\alpha = \{4, 6, 8, 10\}$, $\sigma = \{\lambda/4, \lambda/2, 3\lambda/4, \lambda\}$ and $\phi = \{0, 1, 2\}$ and use 5-fold cross-validation on the training set to determine which set of parameters best discriminates each AU in our face space.

Formally, let $\mathbf{I}_{ij}$ be the j^{th} sample image with AU i present and define

$$\mathbf{g}_{ijk} = \left(g\left(\hat{\mathbf{s}}_{ijk}; \lambda_1, \alpha_1, \phi_1, \gamma\right) * I_{ij}, \ldots, \right. \\ \left. g\left(\hat{\mathbf{s}}_{ij1}; \lambda_5, \alpha_o, \phi_r, \gamma\right) * I_{ij}\right)^T, \tag{3}$$

as the feature vector of Gabor responses at the k^{th} landmark points, where $*$ defines the convolution of the filter $g(.)$ with the image $\mathbf{I}_{ij}$, and λ_k is the k^{th} element of the set λ defined above; the same applies to α_k and ϕ_k, but not to γ since this is always 1.

We can now define the feature vector of the Gabor responses on all landmark points for the j^{th} sample image with AU i active as

$$\mathbf{g}_{ij} = \left(\mathbf{g}_{ij1}^T, \ldots, \mathbf{g}_{ijp}^T\right)^T. \tag{4}$$

These feature vecotros define the shading information of the local patches around the landmarks of the face and their dimensionality is $\mathbf{g}_{ij} \in \mathbb{R}^{5 \times p \times o \times s \times r}$.

Finally, putting everything together, we obtained the following feature vectors defining the shape and shading changes of AU i in our face space,

$$\mathbf{z}_{ij} = \left(\mathbf{x}_{ij}^T, \mathbf{g}_{ij}^T\right)^T, \quad j = 1, \ldots, n_i. \tag{5}$$

2.2. Classification in face space

Let the training set of AU i be

$$\mathcal{D}_i = \{ \left(\mathbf{z}_{i1}, y_{i1}\right), \ldots, \left(\mathbf{z}_{in_i}, y_{in_i}\right), \\ \left(\mathbf{z}_{in_i+1}, y_{in_i+1}\right), \ldots, \left(\mathbf{z}_{i\,n_i+m_i}, y_{i\,n_i+m_i}\right)\}, \tag{6}$$

where $y_{ij} = 1$ for $j = 1, \ldots, n_i$, indicating that AU i is present in the image, $y_{ij} = 0$ for $j = n_i + 1, \ldots, n_i + m_i$, indicating that AU i is *not* present in the image, and m_i is the number of sample images that do *not* have AU i active.

The training set above is also ordered as follows. The set

$$\mathcal{D}_i(a) = \{\left(\mathbf{z}_{i1}, y_{i1}\right), \ldots, \left(\mathbf{z}_{i\,n_{ia}}, y_{i\,n_{ia}}\right)\} \tag{7}$$

includes the n_{ia} samples with AU i active at intensity *a* (that is the lowest intensity of activation of an AU), the set

$$\mathcal{D}_i(b) = \{ \left(\mathbf{z}_{i\,n_{ia}+1}, y_{i\,n_{ia}+1}\right), \ldots, \\ \left(\mathbf{z}_{i\,n_{ia}+n_{ib}}, y_{i\,n_{ia}+n_{ib}}\right)\} \tag{8}$$

are the n_{ib} samples with AU i active at intensity *b* (which is the second smallest intensity), the set

$$\mathcal{D}_i(c) = \{ \left(\mathbf{z}_{i\,n_{ia}+n_{ib}+1}, y_{i\,n_{ia}+n_{ib}+1}\right), \ldots, \\ \left(\mathbf{z}_{i\,n_{ia}+n_{ib}+n_{ic}}, y_{i\,n_{ia}+n_{ib}+n_{ic}}\right)\} \tag{9}$$

are the n_{ic} samples with AU i active at intensity *c* (which is the next intensity), and the set

$$\mathcal{D}_i(d) = \{ \left(\mathbf{z}_{i\,n_{ia}+n_{ib}+n_{ic}+1}, y_{i\,n_{ia}+n_{ib}+n_{ic}1}\right), \ldots, \\ \left(\mathbf{z}_{i\,n_{ia}+n_{ib}+n_{ic}+n_{id}}, y_{i\,n_{ia}+n_{ib}+n_{ic}+n_{id}}\right)\} \tag{10}$$

are the n_{id} samples with AU i active at intensity *d* (which is the highest intensity we have in the databases we used), and $n_{ia} + n_{ib} + n_{ic} + n_{id} = n_i$.

Recall that an AU can be active at five intensities, which are labeled *a*, *b*, *c*, *d*, and *e* [8]. In the databases we will use in this paper, there are no examples with intensity *e* and, hence, we only consider the four other intensities.

The four training sets defined above are subsets of $\mathcal{D}_i$ and are thus represented as different *subclasses* of the set of images with AU i active. This observation directly suggests the use of a subclass-based classifier. In particular, we use Kernel Subclass Discriminant Analysis (KSDA) [25] to derive our algorithm. The reason we chose KSDA is because it can uncover complex non-linear classification boundaries by optimizing the kernel matrix and number of subclasses, i.e., while other kernel methods use cross-validation on the training data to find an appropriate kernel mapping, KSDA optimizes a class discriminant criterion that is theoretically known to separate classes optimally wrt Bayes. This criterion is formally given by $Q_i(\varphi_i, h_{i1}, h_{i2}) = Q_{i1}(\varphi_i, h_{i1}, h_{i2}) Q_{i2}(\varphi_i, h_{i1}, h_{i2})$, with $Q_{i1}(\varphi_i, h_{i1}, h_{i2})$ responsible for maximizing homoscedasticity (i.e., since the goal of the kernel map is to find a kernel space $\mathcal{F}$ where the data is linearly separable, this means that the subclasses will need to be linearly separable in $\mathcal{F}$, which is the case when the class distributions share the same variance), and $Q_{i2}(\varphi_i, h_{i1}, h_{i2})$ maximizes the distance between all subclass means (i.e., which is used to find a Bayes classifier with smaller Bayes error[2]).

Thus, the first component of the KSDA criterion presented above is given by,

$$Q_{i1}(\varphi_i, h_{i1}, h_{i2}) = \frac{1}{h_{i1} h_{i2}} \sum_{c=1}^{h_{i1}} \sum_{d=h_{i1}}^{h_{i1}+h_{i2}} \frac{\operatorname{tr}\left(\mathbf{\Sigma}_{ic}^{\varphi_i} \mathbf{\Sigma}_{id}^{\varphi_i}\right)}{\operatorname{tr}\left(\mathbf{\Sigma}_{ic}^{\varphi_i^2}\right) \operatorname{tr}\left(\mathbf{\Sigma}_{id}^{\varphi_i^2}\right)}, \tag{11}$$

where $\mathbf{\Sigma}_{il}^{\varphi_i}$ is the subclass covariance matrix (i.e., the covariance matrix of the samples in subclass l) in the kernel space defined by the mapping function $\varphi_i(.) : \mathbb{R}^e \to \mathcal{F}$, h_{i1} is the number of subclasses representing AU i is present in the image, h_{i2} is the number of subclasses representing

[2]To see this recall that the Bayes classification boundary is given in a location of feature space where the probabilities of the two Normal distributions are identical (i.e., $p(\mathbf{z}|\mathcal{N}(\mu_1, \Sigma_1)) = p(\mathbf{z}|\mathcal{N}(\mu_2, \Sigma_2))$, where $\mathcal{N}(\mu_i, \Sigma_i)$ is a Normal distribution with mean μ_i and covariance matrix Σ_i. Separating the means of two Normal distributions decreases the value where this equality holds, i.e., the equality $p(\mathbf{x}|\mathcal{N}(\mu_1, \Sigma_1)) = p(\mathbf{x}|\mathcal{N}(\mu_2, \Sigma_2))$ is given at a probability values lower than before and, hence, the Bayes error is reduced.

AU i is *not* present in the image, and recall $e = 3t + p(p-1)/2 + 5 \times p \times o \times s \times r$ is the dimensionality of the feature vectors in the face space defined in Section 2.1.

The second component of the KSDA criterion is,

$$Q_{i2}(\varphi_i, h_{i1}, h_{i2}) = \sum_{c=1}^{h_{i1}} \sum_{d=h_{i1}+1}^{h_{i1}+h_{i2}} p_{ic}\, p_{id}\, \|\mu_{ic}^{\varphi_i} - \mu_{id}^{\varphi_i}\|_2^2, \tag{12}$$

where $p_{il} = n_l/n_i$ is the prior of subclass l in class i (i.e., the class defining AU i), n_l is the number of samples in subclass l, and $\mu_{il}^{\varphi_i}$ is the sample mean of subclass l in class i in the kernel space defined by the mapping function $\varphi_i(.)$.

Specifically, we define the mapping functions $\varphi_i(.)$ using the Radial Basis Function (RBF) kernel,

$$k(\mathbf{z}_{ij_1}, \mathbf{z}_{ij_2}) = \exp\left(-\frac{\|\mathbf{z}_{ij_1} - \mathbf{z}_{ij_2}\|_2^2}{\upsilon_i}\right), \tag{13}$$

where υ_i is the variance of the RBF, and $j_1, j_2 = 1, \dots, n_i + m_i$. Hence, our KSDA-based classifier is given by the solution to,

$$\upsilon_i^*, h_{i1}^*, h_{i2}^* = \arg \max_{\upsilon_i, h_{i1}, h_{i2}} Q_i(\upsilon_i, h_{i1}, h_{i2}). \tag{14}$$

Solving for (14) yields the model for AU i, Figure 3. To do this, we first divide the training set $\mathcal{D}_i$ into five subclasses. The first subclass (i.e., $l = 1$) includes the sample feature vectors that correspond to the images with AU i active at intensity a, that is, the $\mathcal{D}_i(a)$ defined in (7). The second subclass ($l = 2$) includes the sample subset (8). Similarly, the third and fourth subclass ($l = 2$, 3) include the sample subsets (9) and (10), respectively. Finally, the five subclass ($l = 5$) includes the sample feature vectors corresponding to the images with AU i *not* active, i.e.,

$$\mathcal{D}_i(\text{not active}) = \{ (\mathbf{z}_{i\,n_i+1}, y_{i\,n_i+1}), \dots, (\mathbf{z}_{i\,n_i+m_i}, y_{i\,n_i+m_i}) \}. \tag{15}$$

Thus, initially, the number of subclasses to define AU i active/inactive is five (i.e., $h_{i1} = 4$ and $h_{i2} = 1$).

Optimizing (14) may yield additional subclasses. To see this, note that the derived approach optimizes the parameter of the kernel map υ_i as well as the number of subclasses h_{i1} and h_{i2}. This means that our initial (five) subclasses can be further subdivided into additional subclasses. For example, when no kernel parameter υ_i can map the non-linearly separable samples in $D_i(a)$ into a space where these are linearly separable from the other subsets, $D_i(a)$ is further divided into two subsets $D_i(a) = \{D_i(a_1), D_i(a_2)\}$. This division is simply given by a nearest-neighbor clustering. Formally, let the sample $\mathbf{z}_{i\,j+1}$ be the nearest-neighbor to $\mathbf{z}_{ij}$, then the division of $D_i(a)$ is readily given by,

$$\mathcal{D}_i(a_1) = \{(\mathbf{z}_{i1}, y_{i1}), \dots, (\mathbf{z}_{i\,n_a/2}, y_{i\,n_a/2})\} \tag{16}$$
$$\mathcal{D}_i(a_2) = \{(\mathbf{z}_{i\,n_a/2+1}, y_{i\,n_a/2+1}), \dots, (\mathbf{z}_{i\,n_a}, y_{i\,n_a})\}.$$

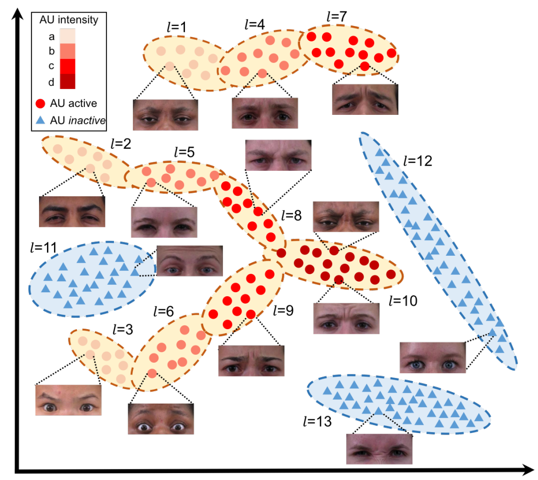


Figure 3: In the hypothetical model shown above, the sample images with AU 4 active are first divided into four subclasses, with each subclass including the samples of AU 4 at the same intensity of activation (*a*–*d*). Then, the derived KSDA-based approach uses (14) to further subdivide each subclass into additional subclasses to find the kernel mapping that (intrinsically) maps the data into a kernel space where the above Normal distributions can be separated linearly and are as far apart from each other as possible.

The same applies to $D_i(b)$, $D_i(c)$, $D_i(d)$ and $D_i(\text{not active})$. Thus, optimizing (14) can result in multiple subclasses to model the samples of each intensity of activation or non-activation of AU i, e.g., if subclass one ($l = 1$) defines the samples in $\mathcal{D}_i(a)$ and we wish to divide this into two subclasses (and currently $h_{i1} = 4$), then the first new two subclasses will be used to define the samples in $\mathcal{D}_i(a)$, with the fist subclass ($l = 1$) including the samples in $\mathcal{D}_i(a_1)$ and the second subclass ($l = 2$) those in $\mathcal{D}_i(a_2)$ (and h_{i1} will now be 5). Subsequent subclasses will define the samples in $D_i(b)$, $D_i(c)$, $D_i(d)$ and $D_i(\text{not active})$ as defined above. Thus, the order of the samples as given in D_i never changes with subclasses 1 through h_{i1} defining the sample feature vectors associated to the images with AU i active and subclasses $h_{i1} + 1$ through $h_{i1} + h_{i2}$ those representing the images with AU i not active. This end result is illustrated using a hypothetical example in Figure 3.

Then, every test image $\mathbf{I}_{test}$ can be readily classified as follows. First, its feature representation in face space $\mathbf{z}_{test}$ is computed as described in Section 2.1. Second, this vector is projected into the kernel space obtained above. Let us call this $\mathbf{z}_{test}^{\varphi}$. To determine if this image has AU i active, we find the nearest mean,

$$j^* = \arg \min_j \|\mathbf{z}_{test}^{\varphi_i} - \mu_{ij}^{\varphi_i}\|_2, \quad j = 1, \dots, h_{i1} + h_{i2}. \tag{17}$$

If $j^* \leq h_{i1}$, then $\mathbf{I}_{test}$ is labeled as having AU i active; otherwise, it is not.

The classification result in (17) also provides intensity recognition. If the samples represented by subclass l are a subset of those in $D_i(a)$, then the identified intensity is a. Similarly, if the samples of subclass l are a subset of those in $D_i(b)$, $D_i(c)$ or $D_i(d)$, then the intensity of AU i in the test image $\mathbf{I}_{test}$ is b, c and d, respectively. Of course, if $j^* > h_{i1}$, the images does not have AU i present and there is no intensity (or, one could say that the intensity is zero).

3. EmotioNet: Annotating a million face images in the wild

In the section to follow, we will present comparative quantitative results of the approach defined in Section 2. These results will show that the proposed algorithm can reliably recognize AUs and their intensities *across databases*. To our knowledge, this is the first published algorithm that can reliably recognize AUs *and* AU intensities across databases. This fact allows us to now define a fully automatic method to annotate AUs, AU intensities and emotion categories on a large number of images in "the wild" (i.e., images downloaded from the Internet). In this section we present the approach used to obtain and annotate this large database of facial expressions.

3.1. Selecting images

We are interested in face images with associated emotive keywords. To this end, we selected all the words derived from the word "feeling" in WordNet [15].

WordNet includes synonyms (i.e., words that have the same or nearly the same meaning), hyponyms (i.e., subordinate nouns or nouns of more specific meaning, which defines a hierarchy of relationships), troponymys (i.e., verbs of more specific meaning, which defines a hierarchy of verbs), and entailments (i.e., deductions or implications that follow logically from or are implied by another meaning – these define additional relationships between verbs).

We used these noun and verb relationships in WordNet to identify words of emotive value starting at the root word "feeling." This resulted in a list of 457 concepts that were then used to search for face images in a variety of popular web search engines, i.e., we used the words in these concepts as search keywords. Note that each concept includes a list of synonyms, i.e., each concept is defined as a list of one or more words with a common meaning. Example words in our set are: affect, emotion, anger, choler, ire, fury, madness, irritation, frustration, creeps, love, timidity, adoration, loyalty, etc. A complete list is provided in the Supplementary Materials.

While we only searched for face images, occasionally non-face image were obtained. To eliminate these, we checked for the presence of faces in all downloaded images with the standard face detector of [21]. If a face was not detected in an image by this algorithm, the image was eliminated. Visual inspection of the remaining images by the authors further identify a few additional images with no faces in them. These images were also eliminated. We also eliminated repeated and highly similar images. The end result was a dataset of about a million images.

3.2. Image annotation

To successfully automatically annotate AU and AU intensity in our set of a million face images in the wild, we used the following approach. First, we used three available databases with manually annotated AUs and AU intensities to train the classifiers defined in Section 2. These databases are: the shoulder pain database of [12], the Denver Intensity of Spontaneous Facial Action (DISFA) dataset of [14], and the database of compound facial expressions of emotion (CFEE) of [7]. We used these databases because they provide a large number of samples with accurate annotations of AUs an AU intensities. Training with these three datasets allows our algorithm to learn to recognize AUs and AU intensities under a large number of image conditions (e.g., each database includes images at different resolutions, orientations and lighting conditions). These datasets also include a variety of samples in both genders and most ethnicities and races (especially the database of [7]). The resulting trained system is then used to automatically annotate our one million images in the wild.

Images may also belong to one of the 23 basic or compound emotion categories defined in [7]. To produce a facial expression of one of these emotion categories, a person will need to activate the unique pattern of AUs listed in Table 1. Thus, annotating emotion category in an image is as simple as checking whether one of the unique AU activation patterns listed in each row in Table 1 is present in the image. For example, if an image has been annotated as having AUs 1, 2, 12 and 25 by our algorithm, we will also annotated it as expressing the emotion category happily surprised.

The images in our database can thus be searched by AU, AU intensity, basic and compound emotion category, and WordNet concept. Six examples are given in Figure 1. The first two examples in this figure show samples returned by our system when retrieving images classified as "happy" or "fearful." The two examples in the middle of the figure show sample images obtained when the query is AU 4 or 6. The final two examples in this figure illustrate the use of keyword searches using WordNet words, specifically, anxiety and disapproval.

4. Experimental Results

We provide extensive evaluations of the proposed approach. Our evaluation of the derived algorithm is divided

Category	AUs	Category	AUs
Happy	12, 25	Sadly disgusted	4, 10
Sad	4, 15	Fearfully angry	4, 20, 25
Fearful	1, 4, 20, 25	Fearfully surpd.	1, 2, 5, 20, 25
Angry	4, 7, 24	Fearfully disgd.	1, 4, 10, 20, 25
Surprised	1, 2, 25, 26	Angrily surprised	4, 25, 26
Disgusted	9, 10, 17	Disgd. surprised	1, 2, 5, 10
Happily sad	4, 6, 12, 25	Happily fearful	1, 2, 12, 25, 26
Happily surpd.	1, 2, 12, 25	Angrily disgusted	4, 10, 17
Happily disgd.	10, 12, 25	Awed	1, 2, 5, 25
Sadly fearful	1, 4, 15, 25	Appalled	4, 9, 10
Sadly angry	4, 7, 15	Hatred	4, 7, 10
Sadly surprised	1, 4, 25, 26	–	–

Table 1: Listed here are the prototypical AUs observed in each basic and compound emotion category.

into three sets of experiments. First, we present comparative results against the published literature using within-databases classification. This is needed because, to our knowledge, only one paper [20] has published results across databases. Second, we provide results across databases where we show that our ability to recognize AUs is comparable to that seen in within database recognition. And, third, we use the algorithm derived in this paper to automatically annotate a million facial expressions in the wild.

4.1. Within-database classification

We tested the algorithm derived in Section 2 on three standard databases: the extended Cohn-Kanade database (CK+) [11], the Denver Intensity of Spontaneous Facial Action (DISFA) dataset [14], and the shoulder pain database of [12].

In each database, we use 5-fold-cross validation to test how well the proposed algorithm performs. These databases include video sequences. Automatic recognition of AUs is done at each frame of the video sequence and the results compared with the provided ground-truth. To more accurately compare our results with state-of-the-art algorithms, we compute the F1 score, defined as, F1 score $= 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$, where Precision (also called positive predictive value) is the fraction of the automatic annotations of AU i that are correctly recognized (i.e., number of correct recognitions of AU i / number of images with detected AU i), and Recall (also called sensitivity) is the number of correct recognitions of AU i over the actual number of images with AU i.

Comparative results on the recognition of AUs in these three databases are given in Figure 4. This figure shows comparative results with the following algorithms: the Hierarchical-Restricted Boltzmann Machine (HRBM) algorithm of [22], the nonrigid registration with Free-Form Deformations (FFD) algorithm of [10], and the l_p-norm algorithm of [26]. Comparative results on the shoulder database

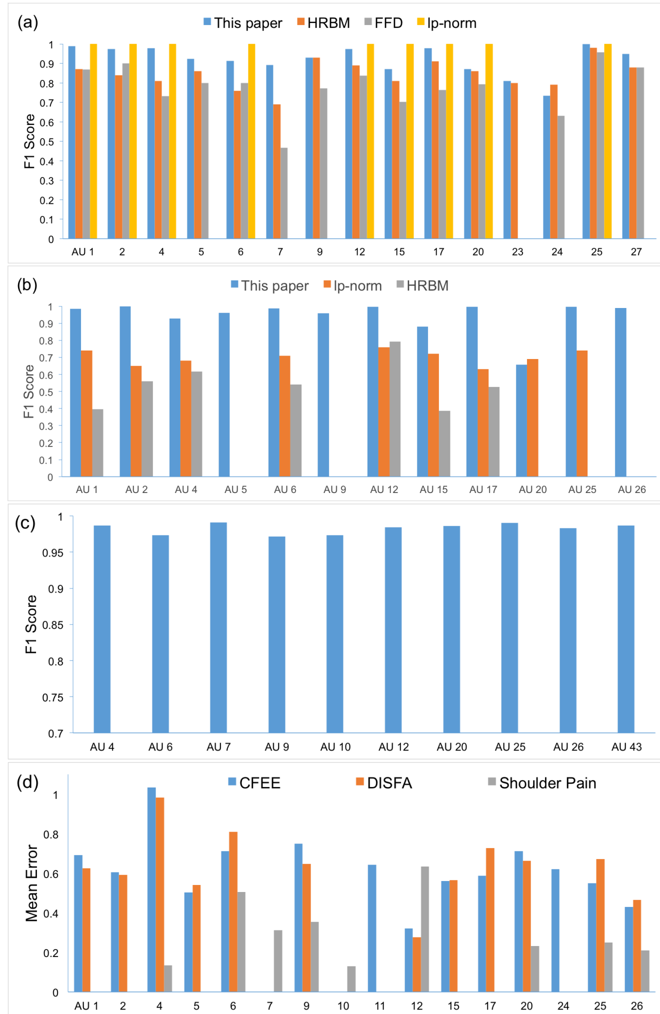


Figure 4: Cross-validation results within each database for the method derived in this paper and those in the literature. Results correspond to (a) CK+, (b) DISFA, and (c) shoulder pain databases. (d) Mean Error of intensity estimation of 16 AUs in three databases using our algorithm.

can be found in the Supplementary Materials. These were not included in this figure because the papers that report results on this database did not disclose F1 values. Comparative results based on receiver operating characteristic (ROC) curves are in the Supplementary Materials.

Next, we tested the accuracy of the proposed algorithm in estimating AU intensity. Here, we use three databases that include annotations of AU intensity: CK+ [11], DISFA [14], and CFEE [7]. To compute the accuracy of AU intensity estimation, we code the four levels of AU intensity *a*-*d* as 1-4 and use 0 to represent inactivity of the AU, then compute Mean Error $= n^{-1}\sum_{i=1}^{n}|\text{Estimated AU intensity} - \text{Actual AU intensity}|$, n the number of test images.

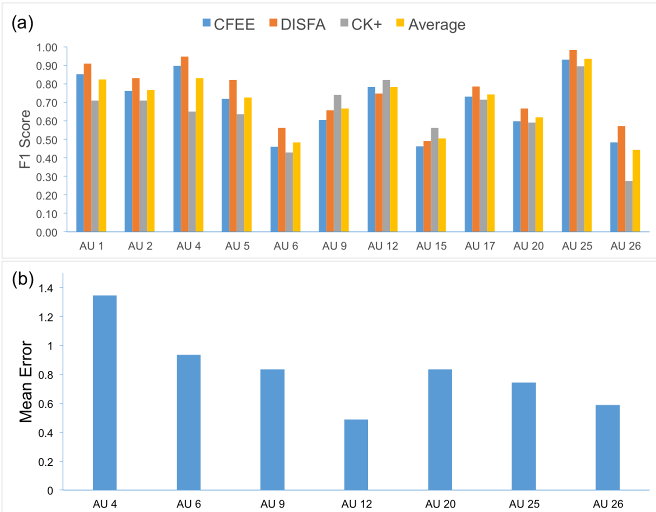


Figure 5: (a). Leave-one-database out experiments. In these experiments we used three databases (CFEE, DISFA, and CK+). Two of the databases are used for training, and the third for testing, The color of each bar indicates the database that was used for testing. Also shown are the average results of these three experiments. (b) Average intensity estimation across databases of the three possible leave-one out experiments.

Additional results (e.g., successful detection rates, ROCs) as well as additional comparisons to state-of-the-art methods are provided in the Supplementary Materials.

4.2. Across-database classification

As seen in the previous section, the proposed algorithm yields results superior to the state-of-the-art. In the present section, we show that the algorithm defined above can also recognize AUs accurately across databases. This means that we train our algorithm using data from several databases and test it on a separate (independent) database. This is an extremely challenging task due to the large variability of filming conditions employed in each database as well as the high variability in the subject population.

Specifically, we used three of the above-defined databases – CFEE, DISFA and CK+ – and run a leave-one-database out test. This means that we use two of these databases for training and one database for testing. Since there are three ways of leaving one database out, we test all three options. We report each of these results and their average in Figure 5a. Figure 5b shows the average Mean Error of estimating the AU intensity using this same leave-one-database out approach.

4.3. EmotioNet database

Finally, we provide an analysis of the used of the derived algorithm on our database of a million images of facial expressions described in Section 3. To estimate the accuracy of these automatic annotations, we proceeded as follows. First, the probability of correct annotation was obtained by computing the probability of the feature vector $\mathbf{z}^{\varphi}_{test}$ to belong to subclass j^* as given by (17). Recall that j^* specifies the subclass closest to $\mathbf{z}^{\varphi}_{test}$. If this subclass models samples of AU i active, then the face in $\mathbf{I}_{test}$ is assumed to have AU i active and the appropriate annotation is made. Now, note that since this subclass is defined as a Normal distribution, $\mathcal{N}\left(\Sigma_{ij^*}, \mu_{ij^*}\right)$, we can also compute the probability of $\mathbf{z}^{\varphi}_{test}$ belonging to it, i.e., $p\left(\mathbf{z}^{\varphi}_{test} | \mathcal{N}\left(\Sigma_{ij^*}, \mu_{ij^*}\right)\right)$. This allows us to sort the retrieved images as a function of their probability of being correctly labeled. Then, from this ordered set, we randomly selected $3,000$ images in the top $1/3$ of the list, $3,000$ in the middle $1/3$, and $3,000$ in the bottom $1/3$.

Only the top $1/3$ are listed as having AU i active, since these are the only images with a large probability $p\left(\mathbf{z}^{\varphi}_{test} | \mathcal{N}\left(\Sigma_{ij^*}, \mu_{ij^*}\right)\right)$. The number of true positives over the number of true plus false positives was then calculated in this set, yielding $\mathbf{80.9}\%$ in this group. Given the heterogeneity of the images in our database, this is considered a really good result. The other two groups (middle and bottom 1/3) also contain some instances of AU i but recognition there would only be 74.9% and 67.2%, respectively, which is clearly indicated by the low probability computed by our algorithm. These results thus provide a quantitative measure of reliability for the results retrieved using the system summarized in Figure 1.

5. Conclusions

We have presented a novel computer vision algorithm for the recognition of AUs and AU intensities in images of faces. Our main contributions are: 1. Our algorithm can reliably recognize AUs and AU intensities *across databases*, i.e., while other methods defined in the literature only report recognition accuracies within databases, we demonstrate that the algorithm derived in this paper can be trained using several databases to successfully recognize AUs *and* AU intensities on an independent database of images not used to train our classifiers. 2. We use this derived algorithm to automatically construct and annotate a large database of images of facial expressions of emotion. Images are annotated with AUs, AU intensities and emotion categories. The result is a database of a million images that can be readily queried by AU, AU intensity, emotion category and/or emotive keyword, Figure 1.

Acknowledgments. Supported by NIH grants R01-EY-020834 and R01-DC-014498 and a Google Faculty Research Award.

References

[1] E. Angelopoulo, R. Molana, and K. Daniilidis. Multispectral skin color modeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001*, volume 2, pages II–635, 2001. 3

[2] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3515–3522. IEEE, 2013. 1

[3] A. Cowen, S. Abdel-Ghaffar, and S. Bishop. Using structural and semantic voxel-wise encoding models to investigate face representation in human cortex. *Journal of vision*, 15(12):422–422, 2015. 3

[4] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal Optical Society of America A*, 2(7):1160–1169, 1985. 3

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, 2012. 1

[6] A. Dhall, O. Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proc. of the 17th ACM Intl. Conf. on Multimodal Interaction (ICMI 2015). ACM*, 2015. 1

[7] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 1, 2, 3, 6, 7

[8] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), 2nd Edition*. Oxford University Press, 2015. 1, 4

[9] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1867–1874. IEEE, 2014. 3

[10] S. Koelstra, M. Pantic, and I. Y. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 32(11):1940–1954, 2010. 1, 7

[11] 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), IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 1, 7

[12] 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*, pages 57–64. IEEE, 2011. 6, 7

[13] A. M. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *The Journal of Machine Learning Research*, 13(1):1589–1608, 2012. 2

[14] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2), April 2013. 6, 7

[15] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2, 6

[16] D. Neth and A. M. Martinez. Emotion perception in emotionless face images suggests a norm-based representation. *Journal of Vision*, 9(1), 2009. 3

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014. 1

[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 1

[19] R. Russell, I. Biederman, M. Nederhouser, and P. Sinha. The utility of surface reflectance for the recognition of upright and inverted faces. *Vision research*, 47(2):157–165, 2007. 2, 3

[20] T. Simon, M. H. Nguyen, F. De La Torre, and J. F. Cohn. Action unit detection with segment-based svms. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2737–2744. IEEE, 2010. 1, 7

[21] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 6

[22] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *Computer Vision (ICCV), IEEE International Conference on*, pages 3304–3311. IEEE, 2013. 1, 7

[23] L. Wiskott, J. Fellous, N. Kuiger, and C. Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 3

[24] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 3

[25] D. You, O. C. Hamsici, and A. M. Martinez. Kernel optimization in discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):631–638, 2011. 4

[26] X. Zhang, M. H. Mahoor, S. M. Mavadati, and J. F. Cohn. A l_p-norm mtmkl framework for simultaneous detection of multiple facial action units. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1104–1111. IEEE, 2014. 1, 7

[27] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2207–2216, 2015. 1

Supplementary Materials
EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild

C. Fabian Benitez-Quiroz*, Ramprakash Srinivasan*, Aleix M. Martinez
Dept. Electrical and Computer Engineering
The Ohio State University
*These authors contributed equally to this paper.

1. Extended Experimental Results

The within-database classification results on the shoulder database were compared to methods described in the literature which report results using ROC (Receiver Operating Characteristic) curves. ROC curves are used to visually and analytically evaluate the performance of binary classifiers. Recall that our classifiers are binary, i.e., AU present (active) in the image or not. ROC plots display the *true positive rate* against the *false positive rate*. The true positive rate is the sensitivity of the classifier, which we have preciously defined as Recall in the main paper. The false positive rate is the number of negative test samples classified as positive (i.e., the image does not include AU i but is classified as having AU i present) over the total number of false positives plus true negatives. Note that the derived algorithm only provides a result, but this can be plotted in ROC space and compared to state-of-the-art methods. Furthermore, since we run a five-fold cross validation, we actually have five results plus the mean reported in the main document. Thus, we can plot six results in ROC space. These results are in Figure S1. Figure S2 provides the same ROC plots for the DISFA database.

As mentioned above, our proposed approach does not yield an ROC curve but rather a set of points in ROC space. We can nevertheless estimate an ROC curve by changing the value of the prior of each AU i. In the results reported in the main paper, we assumed equal priors for AU i active and not active. Reducing the prior of AU i active will decrease the false detection rate, i.e., it is less likely to misclassify a face that does not have AU i active as such. Increasing the prior of AU i active will increase the true positive detection rate. This is *not* what our algorithm does, but it is a simple extension of what can be obtained in applications where the use of priors is needed. Figures S3 and S4 provide the ROC curves thus computed on two of the databases used in the main paper, shoulder pain and DISFA.

The plots in Figures S3 allow us to compute the area under the curve for the results of our algorithm on the shoulder pain database. These and comparative results against the algorithms of [5] and [11] are in Table S1. Once again, we see that the results obtained with the proposed algorithm are superior than those reported in the literature.

We also computed the results on a recent database of spontaneous facial expressions, AM-FED [7]. Our F1 scores where as follows: .93 (AU 2), .89 (AU 4), .94 (AU 5), .82 (AU 9), .92 (AU 12), .75 (AU 14), .82 (AU 15), .92 (AU 17), .90 (AU 18), .72 (AU 26).

2. EmotioNet: Facial Expressions of Emotion in the Wild

We collected one million images of facial expressions of emotion in the wild. Images were downloaded from several popular web search engines by using the emotive keywords defined as nodes of the word "feeling" in WordNet [8] and with the requirement that a face be present in the image. The number of concepts (i.e., words with the same meaning) given by WordNet was 421. These words are listed in Tables S2-S5.

This search yielded a large number of images. These images were further evaluated to guarantee they included a face. This was done in two stages. First, we used the face detector of [10] to detect faces in these images. Images where a face was not detected by this algorithm were discarded. Second, the resulting images were visually inspected by the authors. Images that did not have a face, had a drawing of a face or pornography were eliminated. The end result was a dataset of one million images. This set of images in the wild was the one used in the present work. The number of images in these categories varies from a low of 47 to a maximum of $6,300$, and more than $1,000$ categories have $> 1,000$ images. The average number of sample images/category is 600 (805 stdv).

As described in the main paper, images were automatically annotated by our algorithm. First, our algorithm anno-

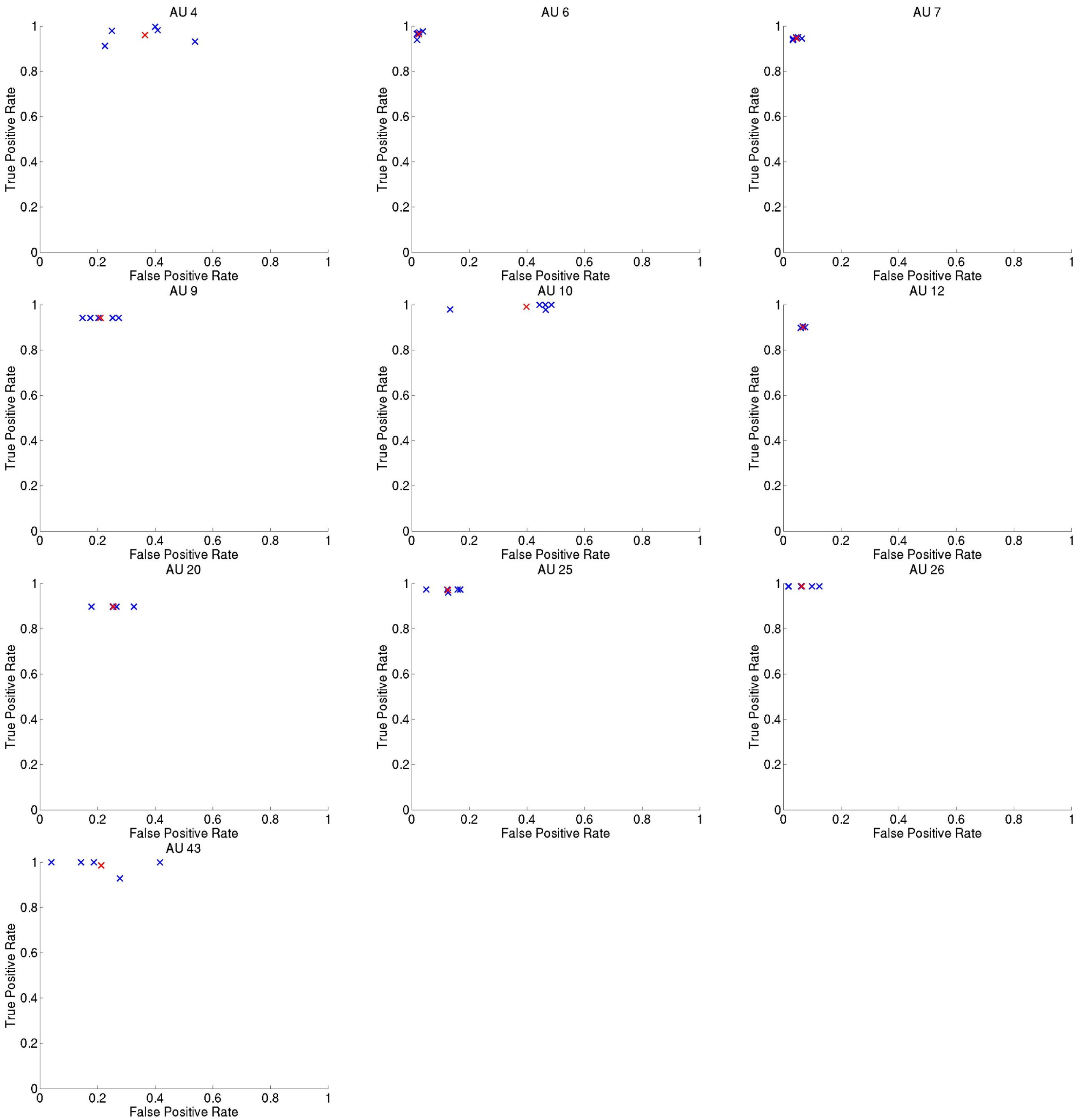


Figure S1: True positive rate against false positive rate of the proposed algorithm for each of the AUs automatically recognized in the images of the shoulder pain database. Shown in the figure are the five results of the five-fold cross-validation test (shown in blue) and the mean (shown in red).

tated AUs and AU intensities. The AUs we annotated were 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25 and 26, since these were the well represented ones in the databases used for training the system. Note that we need a set of accurately annotated AUs and AU intensities to be included during training.

Figure S5a shows the percentages of images in our database of facial expressions in the wild that where automatically annotated with AU i. For example, AU 1 was automatically annotated in over $200,000$ images.

Importantly, we manually FACS-coded 10% of this database. That is, a total of $100,000$ images were manually annotated with AUs by experienced coders in our lab-

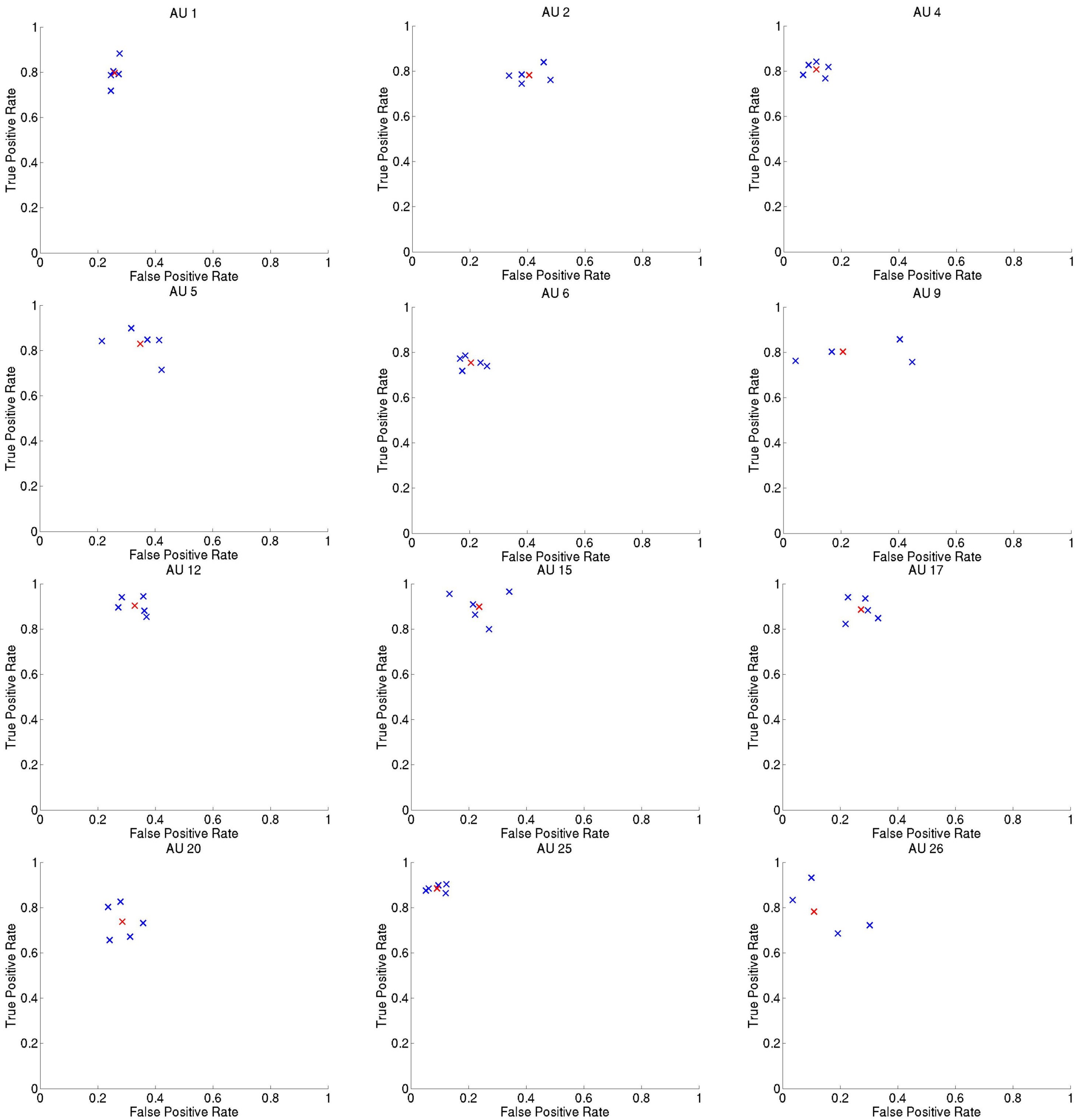


Figure S2: True positive rate against false positive rate of the proposed algorithm for each of the AUs automatically recognized in the images of the DISFA dataset.

AU	4	6	7	9	10	12	20	25	26	43
This paper	**82.45**	**93.48**	**88.57**	**92.56**	**86.15**	**98.54**	**91.13**	**81.46**	**87.19**	**95.47**
Lucey et al. [5]	53.7	86.2	70	79.8	75.4	85.6	66.8	73.3	52.3	90.9
Zafar et al. [11]	78.77	**91.2**		**92.1**						**96.53**

Table S1: Area under the curve for the results shown in Figure S3.

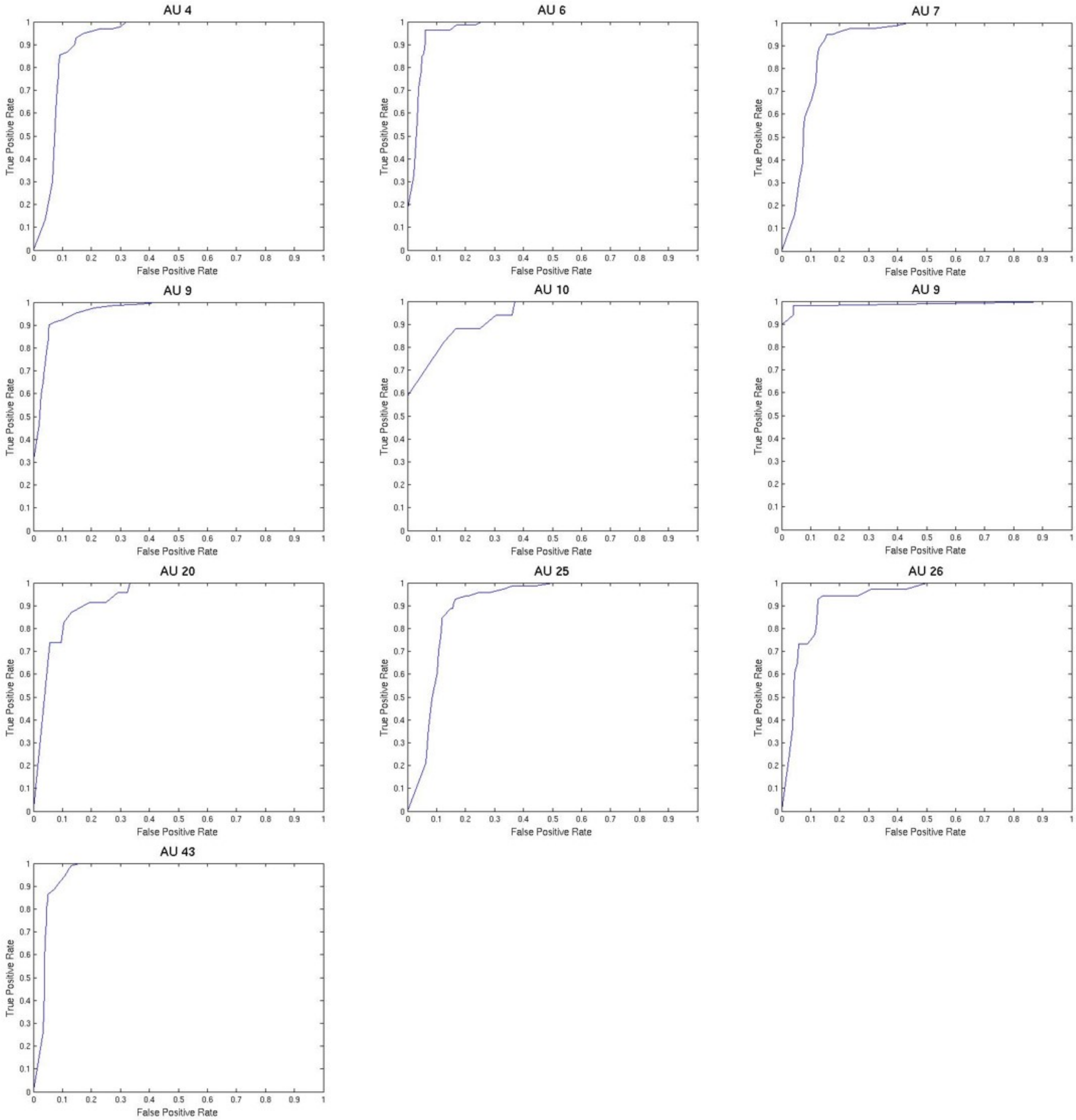


Figure S3: ROC curves for each AU on the shoulder pain database. ROC curves were computed by varying the value of the priors for AU i present and AU i not present.

oratory. This allowed us to estimate the AU detection accuracy of our algorithm, which was about 80%. Note this is extremely accurate given the heterogeneity of the images in the EmotioNet dataset. However, this number only considers correct true positive and true negatives, but does *not* include false negative. Additional work is needed to provide a full analysis of our proposed method on millions of images.

Once an image had been annotated with AUs and AU intensities, we used Table 1 to determine if the face in the image expressed one of the 23 basic or compound emotion categories described in [2, 3]. Note that a facial expression needs not belong to one of these categories. Only when the unique pattern of AU activation described in Ta-

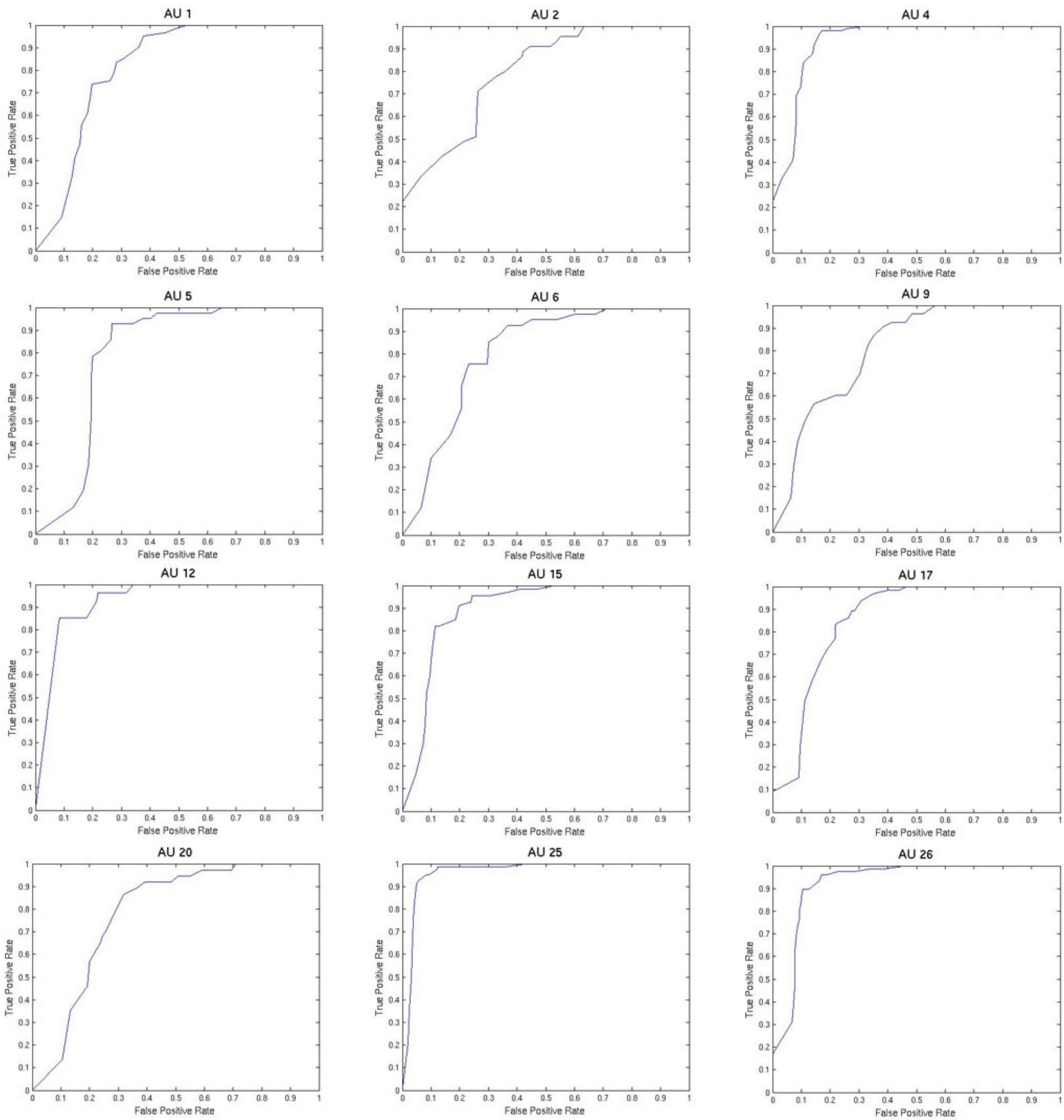


Figure S4: ROC curves for each AU on the DISFA database. ROC curves were computed by varying the value of the priors for AU i present and AU i not present.

ble 1 was present was the face classified as expressing one of these emotions. Figure S5b shows the percentage of images of each emotion category in our database. For example, over $78,000$ images include a facial expression of anger and about $76,000$ have an expression of sadly disgusted. Our algorithm has also been successfully used to detect the "not face" in images in the wild [1]. The "not face" is a grammmatical marker of negation and a facial expression of negation and disapproval.

The above two sections have shown additional quantitative results and analyses of the approach and database of facial expressions of emotion in the wild defined in the main paper. Figure S6 now shows qualitative examples of the images automatically annotated with AU 12 active (present).

3. Rank ordering AU classification

To retrieve images with AU i active, we rank-ordered images according to the posterior probability given by the logistic regression function in the face space of AU i. More formally, let $\mathbf{z}^{\varphi}$ be the sample feature vector of image $\mathbf{I}$ in the kernel space of AU i, then the posterior probability is given by,

$$\log\left(\frac{P(\text{AU } i \text{ active}|\mathbf{Z}=\mathbf{z}_{\varphi})}{P(\text{AU } i \text{ inactive}|\mathbf{Z}=\mathbf{z}^{\varphi})}\right)=\mathbf{b}_i+\mathbf{n}_i^T\mathbf{z}^{\varphi}, \quad \text{(S1)}$$

where $\mathbf{b}_i$ and $\mathbf{n}_i$ are the bias and normal of the hyperplane defining the classifier of AU i in kernel space. It is easy to show that (S1) above is equivalent to,

$$P(\text{AU } i \text{ active}|\mathbf{Z}=\mathbf{z}^{\varphi})=\frac{1}{1+e^{-\left(\mathbf{b}_i+\mathbf{n}_i^T\mathbf{z}^{\varphi}\right)}}. \quad \text{(S2)}$$

The parameters $\mathbf{b}_i$ and $\mathbf{n}_i$ are estimated with iterative re-weighted least squares on the training data $\mathcal{D}_i$ and by optimizing the following function,

$$(\mathbf{b}^*,\mathbf{n}_i^*)=\arg\min_{\mathbf{b},\mathbf{n}_i}\sum_{j=1}^{n_i+m_i}\left\{y_{ij}(\mathbf{b}_i+\mathbf{n}_i^T\mathbf{z}_i^{\varphi})-\right. \quad \text{(S3)}$$
$$\left.\log\left(1+e^{-\left(\mathbf{b}_i+\mathbf{n}_i^T\mathbf{z}^{\varphi}\right)}\right)\right\}.$$

The images we previously shown in Figure S6 are rank-ordered by (S3) such that the images in the first row have a greater posterior than those in the second row and, in general, the images of a top row have a larger posterior than those in its bottom row. Images in the same row have a similar posterior.

4. Ordinal Regression Metrics

The evaluation of AU intensities is a bit trickier than that of AU active/inactive because these are defined by ordinal variables. Unfortunately, evaluation of ordinal variables is a difficult problem. One popular solution is to use the Mean Zero-one Error (MZE), given by $n^{-1}\sum L(f(\mathbf{z}_i)\neq y_i)$, where n is the number of samples, $L(.)$ is an indicator function, $\mathbf{z}_i$ are the samples, y_i are the ordinal labels, and $f(.)$ is the function that estimates the ordinal variable y. Note that this metric does not take the ordinal nature of the labels y_i into account and thus misclassifying a sample $\mathbf{z}_i$ with ordinal value k by any other value but k is considered equally bad. This is not applicable to our case because misclassifying AU intensity by one ordinal step is better than misclassifying it by two which, in turn, is better than misclassifying it by three and so on.

Two other popular methods for evaluating one's estimates of ordinal variables are the Mean Absolute Error (MAE) and the Mean Square Error (MSE). Here, a function $g(.)$ is employed to assign real values to the ordinal categories, e.g., AU intensity $a=1$, $b=2$, $c=3$, $d=4$ and $e=5$. The error is then measured as $n^{-1}\sum|y_i-f(\mathbf{z}_i)|^b$, where y_i and $f(.)$ are now real numbers, and $b=1$ for MAE and $b=2$ for MSE. This is a popular option and was the one chosen to analyze the results in the main paper (with $b=1$).

The main problem with the aforementioned approach is that it assumes that the distance between any two ordinal values is the same, i.e., the distance between AU intensity a and b is the same as the distance between c and d. This is of course not necessarily true.

While the distance between any pair of AU intensities is difficult to define generally, its definition can be readily obtained in most applications. For example, in some applications, misclassifying intensity a as c is twice as bad as misclassifying a as b, and misclassifying intensity a as e is twice as bad as misclassifying a as c. This corresponds to a linear function and thus MSE or MAE are the most appropriate measurements. However, when misclassifying intensity a as c is only a little worse than misclassifying a as b, MAE and MSE need to be modified. This can be easily done by defining

$$\frac{1}{n}\sum_{i=1}^{n}|M(y_i,f(\mathbf{z}_i))|^b, \quad \text{(S4)}$$

where y_i and $\mathbf{z}_i$ now take values from the ordinal set $\{a,b,c,d,e\}$, $M(.,.)$ is a 5×5 matrix with each (p,q) entry specifying how bad our estimation of AU intensity is in our application. For example, we can define $M(.,.)$ as

	a	b	c	d	e
a	0	1	1.2	1.3	1.4
b	1	0	1	1.2	1.3
c	1.2	1	0	1	1.2
d	1.3	1.2	1	0	1
e	1.4	1.3	1.2	1	0

Using the above defined metric (and $b=1$) to calculate the AU intensity estimation errors of our derived algorithm across databases yields the following errors: .73 for AU 4, .62 for AU 6, .58 for AU 9, .46 for AU 12, .7 for AU 20, .43 for AU 25, and .49 for AU 26. These results would substitute those previously reported in Figure 5b and are based on the idea that misclassifying by one ordinal value is almost as bad as any other misclassification.

1	Feeling	62	Anxiousness, disquiet
2	Affect	63	Insecurity
3	Emotion	64	Disquietude, edginess, inquietude, uneasiness
4	Conditioned emotional response	65	Care,concern, fear
5	Anger, choler, ire	66	Willies
6	Fury, madness, rage	67	Sinking
7	Wrath	68	Misgiving, qualm, scruple
8	Lividity	69	Jitteriness, jumpiness, nervousness, restiveness
9	Enragement, infuriation	70	Angst
10	Offence, offense, umbrage	71	Joy, joyfulness, joyousness
11	Indignation, outrage	72	Elation, lightness
12	Dudgeon	73	Euphoria, euphory
13	Huffiness	74	Exultation, jubilance
14	Dander, hackles	75	Triumph
15	Irascibility, spleen	76	Excitement, exhilaration
16	Conniption, fit, scene, tantrum	77	Bang, boot, charge, flush, kick, rush, thrill
17	Annoyance,chafe, vexation	78	Intoxication
18	Irritation,pique, temper	79	Titillation
19	Frustration	80	Exuberance
20	Aggravation, exasperation	81	Love
21	Harassment, torment	82	Adoration, worship
22	Displeasure	83	Agape
23	Fear, fearfulness, fright	84	Crush, infatuation
24	Alarm, consternation, dismay	85	Amorousness, enamoredness
25	Creeps	86	Ardor, ardour
26	Chill, frisson, quiver, shiver, shudder, thrill, tingle	87	Devotedness, devotion
27	Horror	88	Benevolence
28	Hysteria	89	Beneficence
29	Affright, panic, terror	90	Heartstrings
30	Swivet	91	Caring, lovingness
31	Scare	92	Warmheartedness, warmth
32	Apprehension, apprehensiveness, dread	93	Hate, hatred
33	Trepidation	94	Loyalty
34	Boding, foreboding, premonition, presentiment	95	Abhorrence, abomination, detestation, execration, loathing, odium
35	Shadow	96	Misanthropy
36	Presage	97	Misogamy
37	Suspense	98	Misogynism, misogyny
38	Gloom, gloominess, somberness, somberness	99	Misology
39	Chill, pall	100	Misoneism
40	Timidity, timidness, timorousness	101	Murderousness
41	Shyness	102	Despising
42	Diffidence, self-distrust, self-doubt	103	Enmity, hostility
43	Hesitance, hesitancy	104	Animosity, animus
44	Unassertiveness	105	Antagonism
45	Intimidation	106	Aggression, aggressiveness
46	Awe, fear, reverence, veneration	107	Belligerence, belligerency
47	Anxiety	108	Warpath
48	Discomfiture, discomposure, disconcertion, disconcertment	109	Bitterness, gall, rancor, rancor, resentment
49	Trouble, worry	110	Huffishness, sulkiness
50	Grievance, grudge, score	111	Comfort
51	Enviousness, envy	112	Felicity, happiness
52	Covetousness	113	Beatification, beatitude, blessedness
53	Jealousy	114	Enlightenment, nirvana
54	Malevolence, malignity	115	Radiance
55	Maleficence	116	State
56	Malice, maliciousness, spite, spitefulness, venom	117	Unhappiness
57	Vengefulness, vindictiveness	118	Embitterment
58	Spirit	119	Sadness, sorrow, sorrowfulness
59	Embarrassment	120	Huffishness, sulkiness
60	Ecstasy, exaltation, rapture, raptus, transport	121	Bereavement, mourning
61	Gratification, satisfaction	122	Poignance, poignancy

Table S2: List of the WordNet concepts used as keywords to search images of faces in a variety of web search enginees.

123	Glow		184	Sex
124	Faintness		185	Pleasance, pleasure
125	Soul, soulfulness		186	Afterglow
126	Passion		187	Delectation, delight
127	Infatuation		188	Entrancement, ravishment
128	Abandon, wildness		189	Amusement
129	Ardor, ardor, fervency, fervor, fervor, fire		190	Schadenfreude
130	Zeal		191	Enjoyment
131	Storminess		192	Gusto, relish, zest, zestfulness
132	Sentiment		193	Pleasantness
133	Sentimentality		194	Comfort
134	Bathos, mawkishness		195	Consolation, solace, solacement
135	Complex		196	Alleviation, assuagement, relief
136	Ambivalence, ambivalency		197	Algolagnia, algophilia
137	Conflict		198	Sadism
138	Apathy		199	Sadomasochism
139	Emotionlessness, impassiveness, impassivity, indifference, phlegm, stolidity		200	Masochism
140	Languor,lassitude, listlessness		201	Pain, painfulness
141	Desire		202	Unpleasantness
142	Ambition,aspiration, dream		203	Hurt, suffering
143	Emulation		204	Agony, torment, torture
144	Nationalism		205	Throes
145	Bloodlust		206	Discomfort, irritation, soreness
146	Temptation		207	Distress, hurt, suffering
147	Craving		208	Anguish, torment, torture
148	Appetence, appetency, appetite		209	Self-torment, self-torture
149	Stomach		210	Tsoris
150	Addiction		211	Wound
151	Want, wish, wishing		212	Pang, stab, twinge
152	Velleity		213	Liking
153	Hungriness, longing, yearning		214	Leaning,propensity, tendency
154	Hankering, yen		215	Fancy, fondness, partiality
155	Pining		216	Captivation, enchantment, enthrallment, fascination
156	Lovesickness		217	Penchant,predilection, preference, taste
157	Wistfulness		218	Weakness
158	Nostalgia		219	Mysophilia
159	Homesickness		220	Inclination
160	Discontent,discontentment		221	Stomach
161	Disgruntlement		222	Undertow
162	Dysphoria		223	Friendliness
163	Dissatisfaction		224	Amicability, amicableness
164	Boredom ,ennui, tedium		225	Goodwill
165	Blahs		226	Brotherhood
166	Fatigue		227	Approval
167	Displeasure		228	Favor, favour
168	Disappointment, letdown		229	Approbation
169	Defeat, frustration		230	Admiration, esteem
170	Concupiscence, eros		231	Anglophilia
171	Love		232	Philhellenism
172	Aphrodisia		233	Philogyny
173	Passion		234	Dislike
174	Sensualism, sensuality, sensualness		235	Disinclination
175	Amativeness, amorousness, eroticism, erotism, sexiness		236	Anglophobia
176	Carnality, lasciviousness, lubricity, prurience, pruriency		237	Unfriendliness
177	Fetish		238	Alienation, disaffection, estrangement
178	Libido		239	Isolation
179	Lecherousness, lust, lustfulness		240	Antipathy, aversion, distaste
180	Nymphomania		241	Disapproval
181	Satyriasis		242	Contempt, despite, disdain, scorn
182	Itch, urge		243	Disgust
183	Caprice, impulse, whim		244	Abhorrence, abomination, detestation, execration, loathing, odium

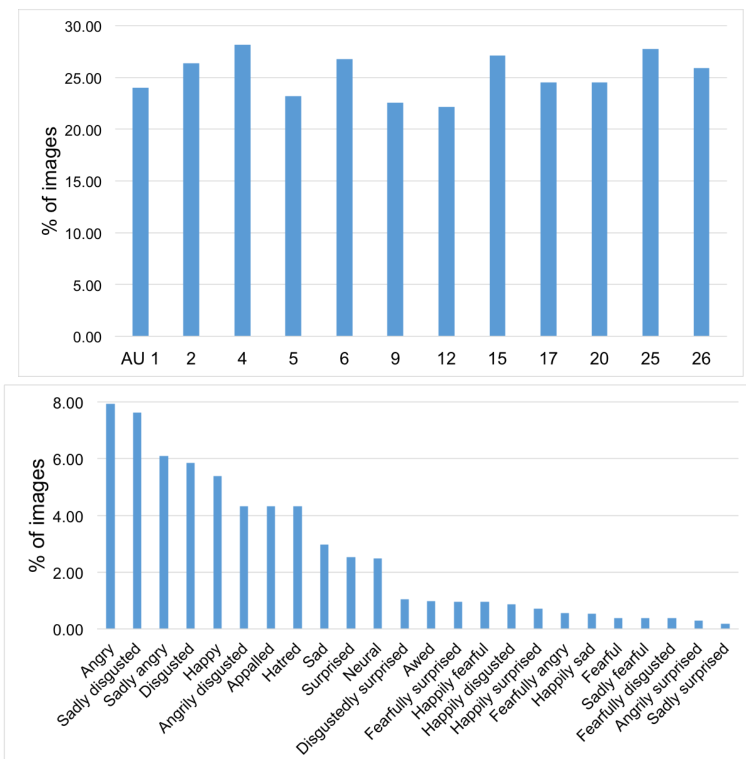
Table S3: Continues from Table S2.

245	Horror, repugnance, repulsion, revulsion	306	Sensation
246	Nausea	307	Tumult, turmoil
247	Creepy-crawlies	308	Calmness
248	Scunner	309	Placidity, placidness
249	Technophobia	310	Coolness, imperturbability
250	Antagonism	311	Dreaminess, languor
251	Gratitude	312	Bravery, fearlessness
252	Appreciativeness, gratefulness, thankfulness	313	Security
253	Ingratitude, ungratefulness	314	Confidence
254	Unconcern	315	Quietness, quietude, tranquility, tranquillity
255	Indifference	316	Ataraxis, heartsease, peace, peacefulness, repose, serenity
256	Aloofness, distance	317	Easiness,relaxation
257	Detachment, withdrawal	318	Happiness
258	Coldheartedness, hardheartedness, heartlessness	319	Bonheur
259	Cruelty, mercilessness, pitilessness, ruthlessness	320	Gladfulness, gladness, gladsomeness
260	Shame	321	Gaiety, merriment
261	Conscience	322	Glee, gleefulness, hilarity, mirth, mirthfulness
262	Self-disgust, self-hatred	323	Jocularity, jocundity
263	Embarrassment	324	Jolliness, jollity, joviality
264	Self-consciousness, uncomfortableness, uneasiness	325	Rejoicing
265	Shamefacedness,sheepishness	326	Belonging
266	Chagrin, humiliation, mortification	327	Comfortableness
267	Confusion, discombobulation	328	Closeness, intimacy
268	Abashment, bashfulness	329	Togetherness
269	Discomfiture, discomposure, disconcertion, disconcertment	330	Blitheness, cheerfulness
270	Pride, pridefulness	331	Buoyancy, perkiness
271	Dignity, self-regard, self-respect, self-worth	332	Carefreeness, insouciance, lightheartedness, lightsomeness
272	Self-esteem, self-pride	333	Contentment
273	Ego, egotism, self-importance	334	Satisfaction
274	Conceit, self-love, vanity	335	Pride
275	Humbleness, humility	336	Complacence, complacency, self-complacency, self-satisfaction
276	Meekness, submission	337	Smugness
277	Self-depreciation	338	Fulfillment, fulfilment
278	Amazement, astonishment	339	Gloat, gloating
279	Admiration,wonder, wonderment	340	Sadness, unhappiness
280	Awe	341	Dolefulness
281	Surprise	342	Heaviness
282	Stupefaction	243	Melancholy
283	Daze, shock, stupor	344	Gloom, gloominess, somberness, somberness
284	Devastation	345	Heavyheartedness
285	Expectation	346	Brooding, pensiveness
286	Anticipation, expectancy	247	Weltschmerz, world-weariness
287	Suspense	248	Misery
288	Fever	349	Desolation, forlornness, loneliness
289	Hope	350	Tearfulness, weepiness
290	Levity	351	Sorrow
291	Gaiety, playfulness	352	Brokenheartedness, grief, heartache, heartbreak
292	Gravity, solemnity	353	Dolor, dolour
293	Earnestness, seriousness, sincerity	354	Mournfulness, ruthfulness, sorrowfulness
294	Sensitiveness, sensitivity	355	Woe, woefulness
295	Sensibility	356	Plaintiveness
296	Insight, perceptiveness, perceptivity	357	Self-pity
297	Sensuousness	358	Regret, rue, ruefulness, sorrow
298	Feelings	359	Attrition, contriteness, contrition
299	Agitation	360	Compunction, remorse, self-reproach
300	Unrest	361	Guilt
301	Fidget, fidgetiness, restlessness	362	Penance, penitence, repentance
302	Impatience	363	Cheerlessness, uncheerfulness
303	Stewing	364	Joylessness
304	Stir	365	Depression
305	Electricity	366	Demoralization

Table S4: Continues from Tables S2-S3.

367	Helplessness	395	Jolliness, jollity, joviality
368	Despondence, despondency, disconsolateness, heartsickness	396	Distemper
369	Oppression, oppressiveness	397	Moodiness
370	Weight	398	Glumness, moroseness, sullenness
371	Dysphoria	399	Testiness, tetchiness, touchiness
372	Dejectedness, dispiritedness, downheartedness, low-spiritedness, lowness	400	Technophilia
373	Hope	401	Pet
374	Hopefulness	402	Sympathy
375	Encouragement	403	Concern
376	Optimis	404	Solicitousness, solicitude
377	Sanguineness, sanguinity	405	Softheartedness, tenderness
378	Despair	406	Kind-heartedness, kindheartedness
379	Hopelessness	407	Mellowness
380	Resignation, surrender	408	Exuberance
381	Defeatism	409	Compassion, compassionateness
382	Discouragement, disheartenment, dismay	410	Heartstrings
383	Intimidation	411	Tenderheartedness, tenderness
384	Pessimism	412	Ardor, ardour, elan, zeal
385	Cynicism	413	Mercifulness, mercy
386	Affection, affectionateness, fondness, heart, philia, tenderness warmheartedness, warmness	414	Choler, crossness, fretfulness, fussiness, irritability, peevishness, petulance
387	Attachment	415	Forgiveness
389	Protectiveness	416	Commiseration, pathos, pity, ruth
390	Regard, respect	417	Compatibility
391	Humor, mood, temper	418	Empathy
392	Peeve	419	Enthusiasm
393	Sulk, sulkiness	420	Gusto, relish, zest, zestfulness
394	Amiability	421	Avidity, avidness, eagerness, keenness

Table S5: Continues from Tables S2-S4.



(a)

(b)

Figure S5: (a) Percentage of images (y-axis) automatically annotated with AU i (x-axis). (b) Percentage of images (y-axis) automatically annotated with one of the 23 basic or compound emotion categories (x-axis) listed in Table 1.

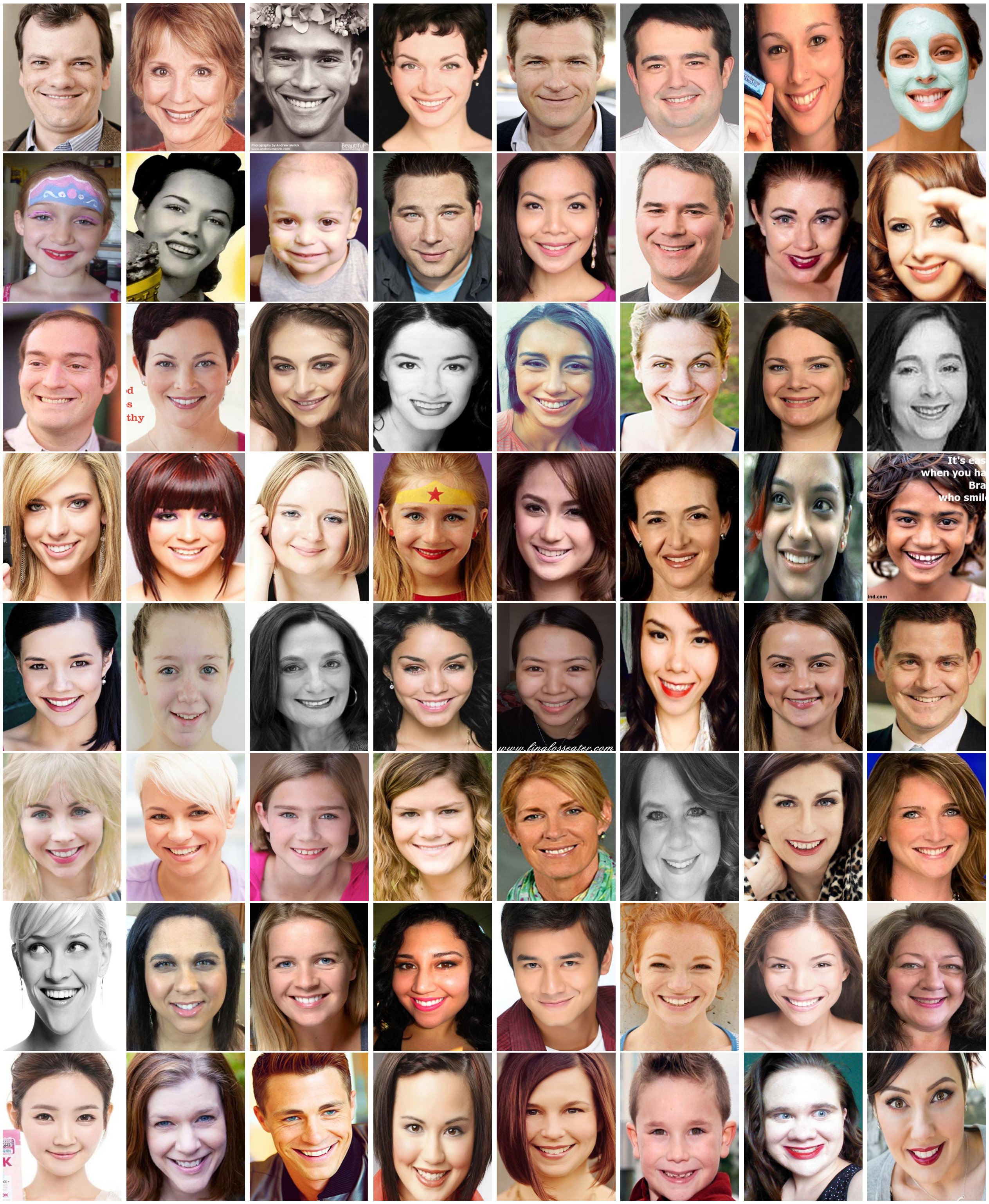


Figure S6: Sample images with AU 12 automatically annotated by our algorithm. The images are ranked according to the probability of AU 12 being active in the image.

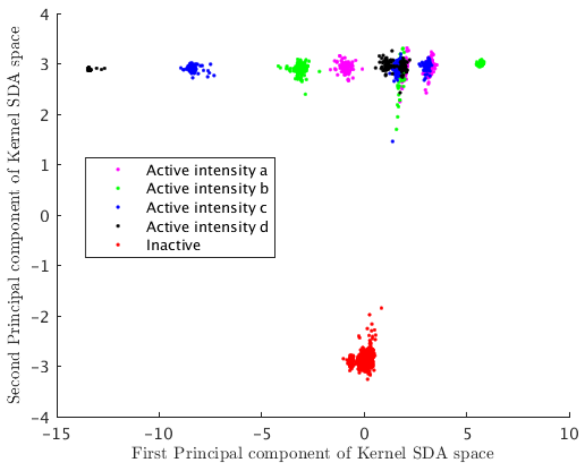


Figure S7: The first two KSDA components of the face space of an AU. Different colors correspond to distinct intensities of the AU. Note how some intensities are divided into subclasses, whereas others are not.

5. Subclass-based Representation of AUs

A key component of our algorithm is to assign the images with AU i active to distinct subclasses as a function of their intensity of activation. That is, images that show AU i active at intensity a are assigned to a subclass of class i, images showing AU i active at intensity b are assigned to a second subclass of class i, images showing AU i active at intensity c are assigned to a third subclass of class i, and images showing AU i active at intensity d are assigned to a fourth subclass of class i. This innovative approach is what allows us to simultaneously identify AUs and their intensities quickly and accurately in images.

This approach is related to the work of Subclass Discriminant Analysis (SDA) [12], which is a mathematical formulation specifically derived to identify the optimal number of subclasses to maximize spreadability of samples in different classes even when these are not defined by a Normal distribution. This is achieved by minimizing a criterion defined in [6], which guarantees Bayes optimality in this classification process under mild conditions.

The approach derived in the present paper is different in that we specify the initial subclass division, rather than using the Bayes criterion defined in [6]. Specifically, we derive a Kernel (SDA-inspired) algorithm to learn to simultaneously identify AUs and their intensities in images. This is done by first dividing the training data of each AU into five sets – one for each of the four intensities, $\mathcal{D}_i(a)$ to $\mathcal{D}_i(d)$, and another set to include the images that do not have that AU active $\mathcal{D}_i(\texttt{not active}) = \mathcal{D}_i - \cup_{j=a,b,c,d} \mathcal{D}_i(j)$. Thus, the initial number of subclasses for class AU i active is 4, i.e., $h_{i1} = 4$, and, the initial number of subclasses for AU i not active (i.e., not present) in the images is 1, i.e., $h_{i2} = 1$. This was illustrated in Figure 3 in the main paper. A 2D plot, for one of the AUs, with real data is now shown in Figure S7. Also, the sample images in each of these five sets are sorted using the nearest-neighbor algorithm of [12].

Next, we use the criterion derived in the main paper, $Q_i(\varphi_i, h_{i1}, h_{i2})$, to further optimize the number of classes and the parameters of the kernel mapping function. This criterion maximizes spherical-homoscedasticity in the RBF kernel space, which is known to minimize the Bayes classification error [4]. Note that, in the paper, we used the RBF kernel, but other options are possible, with each one yielding a different optimization function $Q(.)$.

6. Extended Discussion

The ability to automatically annotate facial action units in the wild in real time is likely to revolutionize research in the study or non-verbal communication and emotion theory. To date, most studies have focused on the analysis of data collected in the laboratory, even when this data corresponds to spontaneous facial expressions. Extending these studies to facial expressions in the wild is a necessary step.

The algorithm described in the present work achieves this goal, allowing researchers to analyze their data quickly and reliably. As a plus, our system is consistent with what is known about the visual perception of facial expressions of emotion by humans [3]. In fact, a recent result from our laboratory has identified a small region of interest in the human brain dedicated to the visual interpretation of facial actions [9]. The computer vision system defined in this paper could thus also help us advance our understanding of human perception.

References

[1] C. F. Benitez-Quiroz, R. B. Wilbur, and A. M. Martinez. The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150:77–84, 2016. 5

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

[3] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 4, 12

[4] O. C. Hamsici and A. M. Martinez. Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8(1583-1623):1–3, 2007. 12

[5] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(3):664–674, 2011. 1, 3

[6] A. M. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence, 27(12):1934–1944, 2005. 12

[7] 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*, pages 881–888, 2013. 1

[8] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[9] R. Srinivasan, J. D. Golomb, and A. M. Martine. A neural basis of facial action recognition in humans. *The Journal of Neuroscience*, 2016. 12

[10] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 1

[11] Z. Zafar and N. A. Khan. Pain intensity evaluation through facial action units. In *Pattern Recognition (ICPR), 22nd International Conference on*, pages 4696–4701. IEEE, 2014. 1, 3

[12] M. Zhu and A. M. Martinez. Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1274–1286, 2006. 12