Discriminant Functional Learning of Color Features for the Recognition of Facial Action Units and their Intensities

C. Fabian Benitez-Quiroz, *Member, IEEE,* Ramprakash Srinivasan, *Student Member, IEEE,* and Aleix M. Martinez

Abstract—Color is a fundamental image feature of facial expressions. For example, when we furrow our eyebrows in anger, blood rushes in, turning some face areas red; or when one goes white in fear as a result of the drainage of blood from the face. Surprisingly, these image properties have not been exploited to recognize the facial action units (AUs) associated with these expressions. Herein, we present the first system to do recognition of AUs and their intensities using these functional color changes. These color features are shown to be robust to changes in identity, gender, race, ethnicity and skin color. Specifically, we identify the chromaticity changes defining the transition of an AU from inactive to active and use an innovative Gabor transform-based algorithm to gain invariance to the timing of these changes. Because these image changes are given by functions rather than vectors, we use a functional classifiers to identify the most discriminant color features of an AU and its intensities. We demonstrate that, using these discriminant color features, one can achieve results superior to those of the state-of-the-art. Finally, we define an algorithm that allows us to use the learned functional color representation in still images. This is done by learning the mapping between images and the identified functional color features in videos. Our algorithm works in realtime, i.e., >30 frames/second/CPU thread.

Index Terms—Facial expressions of emotion, face recognition, face perception, facial color, compound emotions, Gabor transform, color vision, time invariant, recognition in video, recognition in still images.

1 Introduction

The automatic recognition of facial Action Units (AUs) [1], [2] is a major problem in computer vision [3] with applications in engineering (e.g., advertising, robotics, artificial intelligence) [4], [5], [6], education [7], linguistics [8], psychology [9], [10], psychiatry [11], [12], cognitive science and neuroscience [13], [14], to name but a few.

Most past and current computer vision systems use spatio-temporal features (e.g., Gabor filters, high- and low-spatial filtering) [15], [16], shape [17], [18], shading [12], [19] and motion [20], [21] to identify AUs in images and video sequences.

Although color is clearly another important feature of facial expressions [22], *it is yet to be used as a feature for the recognition of AU activation*.

To clarify the importance of color, let us look at the example in Figure 1. As seen in this figure, when we contract and relax our facial muscles, the shading and color in our faces changes locally. For example, during a smile, the shading and color of the cheeks changes due to the use of AU 12 (lip corner puller). This contraction of facial muscles is known to change the brdf (bidirectional reflectance distribution function) of the face [23], yielding clearly visible image changes [24].

In this paper, we will exploit these color changes to detect AUs and their intensities. We will also demonstrate that these changes are consistent across identities, gender, race, ethnicity and skin color.

• *All authors are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43212.*

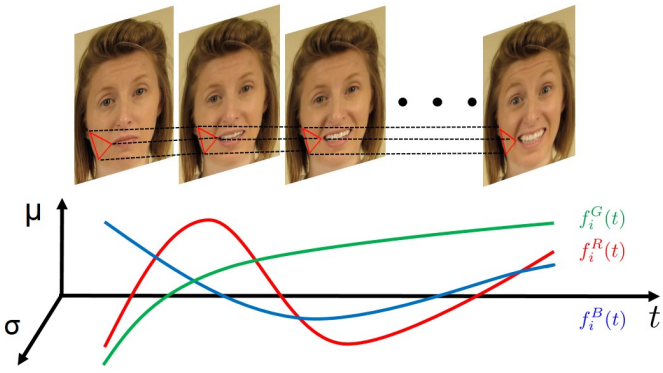


Fig. 1. *Top*: A few frames of a video sequences showing a facial expression of happily surprised. Note we have demarked a local region on that individual's right cheek with red lines. You may notice that the average and standard devision of the color of the pixels in this local region change over time. The value changes of the red, green and blue channels of the pixels in this local region are given in the *bottom* plot, with $f_j^R(t)$ showing the functional change in the red channel, $f_j^G(t)$ in the green, and $f_j^B(t)$ in the blue. Our contribution is to derive a method that can learn to identify when a facial action unit is active by exclusively using these color changes.

Note that we define color changes locally using a function $f_j(t) \in \mathbb{R}^6$, where $f_j(t) = \left(f_j^R(t), f_j^G(t), f_j^B(t)\right)^T$ describes the color changes in each of the three channels (R, G, B), $f_j^R(t), f_j^G(t), f_j^B(t) \in \mathbb{R}^2$, and j designates the j^{th} local region. Specifically, we use the local regions given by a set of automatically detected fiducial points [25], Figure 2. Aggregating these local functions, we obtain the global

color function $\mathbf{f}(.) = (f_1(t), \ldots, f_{107}(t))^T \in \mathbb{R}^{642}$, i.e., the function $f_j(t) \in \mathbb{R}^2$ of the 3 color channels in each of the 107 local regions, $j = 1, \ldots, 107$. The three channels are the red, green and blue (R, G, B) of the camera.

The color representation described in the preceding paragraph differs from previous shape and shading image descriptors in that its samples are given by functions defining color properties only, $\mathbf{f}_i(t)$, $i = 1, \ldots, n$, n the number of samples. (Note we have added a subscript i to our notation to identify the i^{th} sample feature function $\mathbf{f}_i(t)$.) This calls for the use of a discriminant approach that works with functions.

In order to work with these color functional changes, we derive an approach to represent them in DCT (Discrete Cosine Transform) space and use the Gabor transform to gain invariance to time. The use of the Gabor transform in our derivations is *key*, yielding a compact mathematical formulation for detecting the color changes of an AU regardless of when this occurs during a facial expression. That is, *the resulting algorithm is invariant to the duration, start and finish of the AU activation within a video of a facial expressions*.

Since these functions are defined in time, learning must be done over video sequences. But testing can be done in videos *and* still images. To use the learned functions in still images, we need to first find the color functional changes of an image. To do this, we use regression to learn the mapping between an image of a facial expression and the functional representation of the video of that same expression.

In summary, the present paper demonstrates, for the first time, that the use of these color descriptors yields classification accuracies superior to those reported in the literature. This shows that the contribution of some of these color features need to be uncorrelated to those of shading and shape features used previously.

The paper is organized as follows. Section 2 derives the color space used by our algorithm. Section 3 derives functional classifiers to identify where in the video sequence an AU is active/present. Section 4 defines a mapping from still images to the derived functional representation of color to allow recognition of AUs in images. Section 5 provides extensive experimental validation of the derived algorithm. We conclude in Section 6.

1.1 Related work

Despite many advances in object recognition, the automatic coding of AUs in videos and images remains an open problem in computer vision [3]. A 2017 challenge of AU annotations in images collected "in the wild" demonstrated that the problem is far from solved [26]. Even when large amounts of training data are available, deep nets have a hard time annotating AUs with good precision and recall [27], [28], [29]. In this paper, we tackle this problem by exploiting intrinsic functional changes of the color signal in facial expressions of emotion in video sequences. We then show how this can be readily extended to still images as well.

The Gabor transform is specifically suited for our problem, given its ability to identify a template in a function [30]. Or, alternatively, one could employ the Wavelet transform [31]. Here, we show how the Gabor transform can be used to find a template color function in a functional description of a video sequence without the need of a grid search. Similarly, color images have been used to identify optical flow [32] and other image features [33], but not AUs and the dynamic changes that define facial configurations. Nevertheless, color is known to play a major role in human vision [22]. Herein, we identify the discriminant color templates that specify AUs.

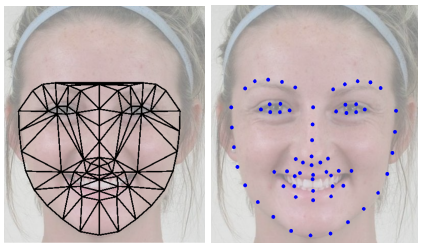


Fig. 2. The local regions of the face (left image) used by the derived algorithm. These local regions are obtained by Delaunay triangulation of the automatically detected fiducial points shown on the right image. These fiducial points, $\mathbf{s}_{ij}$ ($j = 1, \ldots, 66$), correspond to 15 anatomical landmarks (e.g., corners of the eyes, mouth and brows, tip of the nose, and chin) plus 51 pseudo-landmarks defined about the edge of the eyelids, brows, nose, lips and jaw line. The number of pseudo-landmarks defining the contour of each facial component (e.g., the brows) is constant as is their inter-landmark distance. This guarantees equivalency of landmark position across people. This triangulation yields 107 regions (patches).

2 Color Space

This section details the computations needed to construct the color feature space used by the proposed algorithm.

2.1 Local regions

We start with the i^{th} sample video sequence $V_i = \{\mathbf{I}_{i1}, \ldots, \mathbf{I}_{ir_i}\}$, where r_i is the number of frames and $\mathbf{I}_{ik} \in \mathbb{R}^{3qw}$ is the vectorized k^{th} color image of $q \times w$ RGB pixels. We now need to describe V_i as the sample function $\mathbf{f}_i(t)$ defined above, Figure 1.

To do this, we first identify a set of physical facial landmarks on the face and obtained the local regions using the algorithm of [25]. Formally, we define these landmark points in vector form as $\mathbf{s}_{ik} = (\mathbf{s}_{ik1}, \ldots, \mathbf{s}_{ik66})$, where i is the sample video index, k the frame number, and $\mathbf{s}_{ikl} \in \mathbb{R}^2$ are the 2D image coordinates of the l^{th} landmark, $l = 1, \ldots, 66$, Figure 2.

Next, let $D_{ij} = \{\mathbf{d}_{i1k}, \ldots, \mathbf{d}_{iPk}\}$ be the set of $P = 107$ image patches $\mathbf{d}_{ijk}$ obtained with the Delaunay triangulation shown in Figure 2 (left image), where $\mathbf{d}_{ijk} \in \mathbb{R}^{3q_{ij}}$ is the vector describing the j^{th} triangular local region of q_{ij} RGB pixels and, as above, i specifies the sample video number ($i = 1, \ldots, n$) and k the frame ($k = 1, \ldots, r_i$).

Note that the size (i.e., number of pixels, q_{ij}) of these local (triangular) regions not only varies across individuals but also within a video sequence of the same person. This is a result of the movement of the facial landmark points, a necessary process to produce a facial expression. This is evident in the images shown in Figures 1. Hence, we need to define a feature space that is invariant to the number of pixels in each of these local regions. We do this by

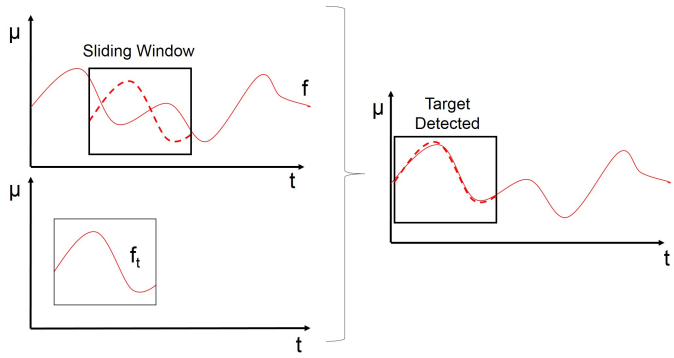


Fig. 3. The top left plot shows the function $f(.)$ of a video V. The bottom left plot is a template function $f_T(.)$ representing the color changes observed when people activate AU 1. Identifying this template $f_T(.)$ in $f(.)$ requires us to test all possible locations about time. This matching process in computationally expensive. The Gabor transform solves this complexity issue by identifying the location where the template function matches the color function without resorting to a sliding-window approach (right-most plot).

computing statistics on the color of the pixels in each local region as follows.

We compute the first and second (central) moments of the color of each local region,

$$\begin{aligned} \mu_{ijk} &= q_{ij}^{-1} \sum_{p=1}^{P} d_{ijkp} \\ \sigma_{ijk} &= \sqrt{q_{ij}^{-1} \sum_{p=1}^{P} (d_{ijkp} - \mu_{ijk})^2}, \end{aligned} \tag{1}$$

with $\mathbf{d}_{ijk} = (d_{ijk1}, \ldots, d_{ijkP})^T$ and μ_{ijk}, $\sigma_{ijk} \in \mathbb{R}^3$. The elements of σ_{ijk} are the mean and standard deviations of each individual color channel. We could compute additional moments, but this did not result in better classification accuracies in our experiments described below.

We can now construct the color feature vector of each local patch,

$$\mathbf{x}_{ij} = (\mu_{ij1}, \ldots, \mu_{ijr_i}, \sigma_{ij1}, \ldots, \sigma_{ijr_i})^T, \tag{2}$$

where, recall, i is the sample video index (V_i), j the local patch number and r_i the number of frames in this video sequence.

This feature representation defines the contribution of color in patch j. One can also include other proven features to increase the richness of this representation. For example, responses to filters or shape features. If these other features do not yield superior results to the representation in (2), then color does provide additional discriminant information beyond what has already been tried. In the experimental results, we show that these color features do indeed yield superior results to those of the state of the art. This demonstrates that color does provide supplementary discriminant features.

2.2 Invariant functional representation of color

We now derive an approach to define the above computed color information of equation (2) as a function invariant to time, i.e., our functional representation needs to be consistent regardless of where in the video sequence an AU becomes active.

This problem is illustrated in Figure 3. As made clear in this figure, we have the color function $f(.)$ that defines color variations of a video sequence V, and a template function $f_T(.)$ that models the color changes associated with the activation of an AU (i.e., from AU inactive to active). Our goal is to determine if $f_T(.)$ is in $f(.)$.

This problem can be readily solved by placing the template function $f_T(.)$ at each possible location in the time domain of $f(.)$. This is typically called a sliding-window approach, because it involves sliding the window left and right until all possible positions of $f_T(.)$ have been checked. Unfortunately, this is extremely time consuming.

To solve the problem of computational complexity defined in the preceding paragraph, we derive a matching method using the Gabor transform instead. The Gabor transform is specifically designed to determine the frequency and phase content of a local section of a function. This allows us to derive an algorithm to find the matching of $f_T(.)$ in $f(.)$ without having to resort to a sliding-window search. Let us define this process formally.

Without loss of generality let $f(t)$ be a function describing one of our color descriptors, e.g., the mean of the red channel in the j^{th} triangle of sample video i. Then, the Gabor transform of this function is,

$$G(t,f) = \int_{-\infty}^{\infty} f(\tau) g(\tau - t) e^{-2\pi \jmath \nu \tau} \, d\tau, \tag{3}$$

where $g(t)$ is a concave function [34] and $\jmath = \sqrt{-1}$. Herein, we use the pulse function,

$$g(t) = \begin{cases} 1, & 0 \leq t \leq L \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where L is a fixed time length.

Using (4) in (3) yields

$$\begin{aligned} G(t,f) &= \int_{t-L}^{t} f(\tau) e^{-2\pi \jmath \nu \tau} \, d\tau \\ &= e^{-2\pi \jmath \nu (t-L)} \int_{0}^{L} f(\tau + t - L) e^{-2\pi \jmath \nu \tau} \, d\tau. \end{aligned} \tag{5}$$

Note that (5) is the definition of a functional inner product in the span $[0, L]$ and, thus, $G(.,.)$ can also be written as follows,

$$G(t,f) = e^{-2\pi \jmath f (t-L)} \langle f(\tau + t - L), e^{-2\jmath \pi f \tau} \rangle, \tag{6}$$

where $\langle ., . \rangle$ is the functional inner product. It is important to point out that our definition of the Gabor transform in (6) is both continuous in time and frequency, in the noise-free case.

To compute the color descriptor of the i_1^{th} video, $f_{i_1}(t)$, we define all functions in a color spaces spanned by a set of b basis functions $\boldsymbol{\phi}(t) = \{\phi_0(t), ..., \phi_{b-1}(t)\}$, with $f_{i_1}(t) = \sum_{z=0}^{b-1} c_{i_1 z} \phi_z(t)$ and $\mathbf{c}_{i_1} = (c_{i_1 0}, \ldots, c_{i_1 b-1})^T$ the vector of coefficients. This allows us to compute the functional inner product of two color descriptors as,

$$\begin{aligned} \langle f_{i_1}(t), f_{i_2}(t) \rangle &= \sum_{\forall r,q} \int_0^L c_{i_1 r} \, \phi_{i_1}(t) \, c_{i_2 q} \, \phi_{i_2}(t) \, dt \\ &= \mathbf{c}_{i_1}^T \, \boldsymbol{\Phi}(t) \, \mathbf{c}_{i_2}, \end{aligned} \tag{7}$$

where $\mathbf{\Phi}$ is a $b \times b$ matrix with elements $\mathbf{\Phi}_{ij} = \langle \phi_i(t), \phi_j(t) \rangle$.

Our model assumes that statistical color properties change smoothly over time and that their effect in muscle activation has a maximum time span of L seconds. The basis functions that fit this description are the first several components of the real part of the Fourier series, i.e., normalized cosine basis.

Let the cosine bases be $\psi_z(t) = \cos(2\pi z t)$, $z = 0, \ldots, b-1$. The corresponding normalized bases are

$$\widehat{\psi}_z(t) = \frac{\psi_z(t)}{\sqrt{\langle \psi_z(t), \psi_z(t) \rangle}}. \tag{8}$$

We use this normalized basis set, because it allows us to have $\mathbf{\Phi} = \mathbf{Id}_b$, where $\mathbf{Id}_b$ denotes the $b \times b$ identity matrix, rather than an arbitrary positive definite matrix.

Importantly, the above derivations with the cosine bases, makes the frequency space implicitly discrete. This allows us to write our Gabor transform $\widetilde{G}(.,.)$ of color functions given in (6) as

$$\widetilde{G}(t, z) = \langle \tilde{f}_{i_1}(t), \widehat{\psi}_z(t) \rangle = c_{i_1 z}, \quad z = 0, \ldots, b-1, \tag{9}$$

where $\tilde{f}_{i_1}(t)$ is our computed function $f_{i_1}(t)$ in the interval $[t-L, t]$ and $c_{i_1 z}$ is the z^{th} coefficient.

The number of cosine basis functions b is determined by performing a grid-search between a minimum of 5 to a maximum of 20 basis functions. We pick the b that yield the best performance (as measured in section 5). It is crucial to note that since the above-derived approach *does not include the time domain*, we can always find these coefficients. This thus allows us to solve the matching of functions without resorting to the use of sliding windows.

In the next section we derive a functional classifier that exploits the advantages of this functional representation.

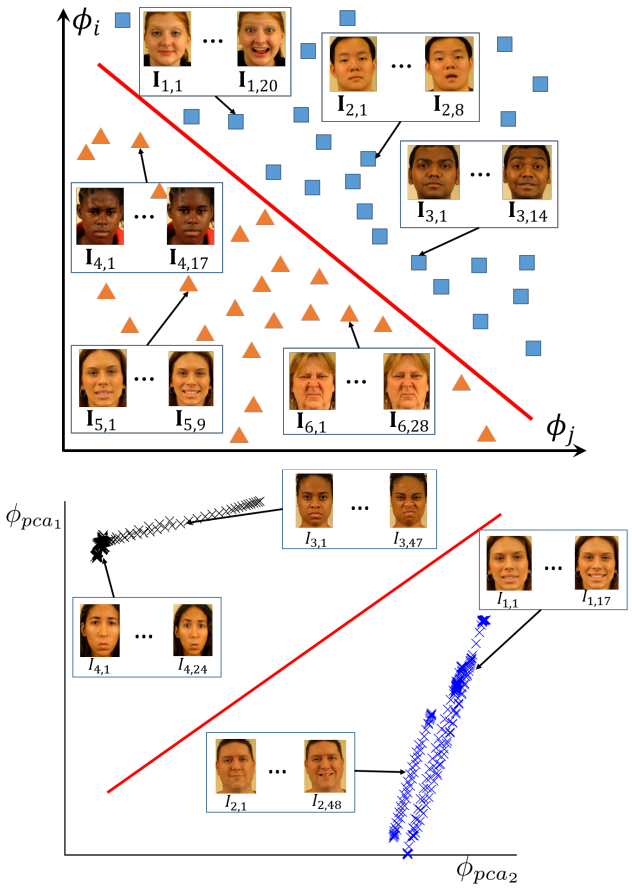


Fig. 4. Top: A schematic representation of positive samples (blue squares) and negative samples (orange triangles). Positive feature vectors correspond to videos with the activation of a specific AU. Negative sample videos do not have that AU present. Note that the sample videos need not be of the same length. Bottom: An example of a color functional space obtained with a SVM classifier for video sequences of facial expressions with AU 12 active/inactive.

3 Functional Classifier of Action Units

The key to our algorithm is to use the Gabor transform derived above to define a feature space invariant to the timing and duration of an AU. In the resulting space, we can employ any linear or non-linear classifier. Here, we report results on Support Vector Machines (SVM) and a Deep multilayer perceptron Network (DN).

3.1 Functional color space

As stated earlier, our feature representation is the collection of functions describing the mean and standard deviation of color information from distinct local patches, which requires simultaneous modeling of multiple functions. This is readily achieved in our formulation as follows.

We define a multidimensional function $\mathbf{\Gamma}_i(t) = \left(\gamma_i^1(t), \ldots, \gamma_i^g(t)\right)^T$, with each function $\gamma_i^e(t)$ the mean or standard deviation of a color channel in a given patch. Using the basis expansion approach described in Section 2.2, each $\gamma_i^e(t)$ is defined by a set of coefficients $\mathbf{c}_i{}^e$ and, thus, $\mathbf{\Gamma}_i(t)$ is given by:

$$\mathbf{c}_i{}^T = \left[(\mathbf{c}_i{}^1)^T, \ldots, (\mathbf{c}_i{}^g)^T \right]. \tag{10}$$

Using this notation, we can redefine the inner product for multidimensional functions. With our normalized Fourier cosine bases we get,

$$\langle \mathbf{\Gamma}_i(t), \mathbf{\Gamma}_j(t) \rangle = \sum_{e=1}^{g} \langle \gamma_i^e(t), \gamma_j^e(t) \rangle = \sum_{e=1}^{g} (\mathbf{c}_i^e)^T \mathbf{c}_j^e = \mathbf{c}_i{}^T \mathbf{c}_j. \tag{11}$$

We use a training set of video sequences to optimize each classifier. It is important to note that our approach is invariant to the length (i.e., number of frames) of a video, Figure 4. Hence, we do not require any alignment or cropping of the videos in our training or testing sets.

The approach derived above can readily extended to identify AU intensity. This is done using a multi-class classifier. In our experimental results, we trained our AU classifiers to detect each of the five intensities, a, b, c, d, and e [2] plus AU inactive (not present). This is a total of six classes.

Testing in videos is directly given by the equations derived above. But we can also use these learned functions to identify AUs in still images. The algorithm used to achieve this is presented in Section 4.

3.2 Support Vector Machines

The training set is $\{(\gamma_i(t), y_1), \ldots, (\gamma_n(t), y_n)\}$, where $\gamma_i(t) \in \mathcal{H}^v$, $\mathcal{H}^v$ is a Hilbert space of continuous functions with bounded derivatives up to order v, and $y_i \in \{-1, 1\}$ are their class labels, with $+1$ indicating that the AU is active and -1 inactive.

When the samples of distinct classes are linearly separable, the function $w(t)$ that maximizes class separability is given by

$$\begin{aligned} J(w(t), v, \boldsymbol{\xi}) \quad &= \quad \min_{w(t),v,\boldsymbol{\xi}} \left\{ \frac{1}{2} \langle w(t), w(t) \rangle + C \sum_{i=1}^{n} \xi_i \right\} \\ \text{subject to} \quad & \quad y_i \left(\langle w(t), \gamma_i(t) \rangle - v \right) \geq 1 - \xi_i, \\ & \quad \xi_i \geq 0, \end{aligned} \tag{12}$$

where v is the bias and, as above, $\langle \gamma_i(t), \gamma_j(t) \rangle = \int \gamma_i(t)\, \gamma_j(t)\, dt$ denotes the functional inner product, Figure 4, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$ are the slack variables, and $C > 0$ is a penalty value found using cross-validation [35].

Applied to our derived approach to model $\boldsymbol{\Gamma}_i$ using normalized cosine coefficients jointly with (11), transforms (12) to the following criterion

$$\begin{aligned} J(\mathbf{w}, v, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \min_{\mathbf{w},\xi,b,\mathbf{a}} \left\{ \frac{1}{2} \mathbf{w}^{\mathbf{T}} \mathbf{w} + C \sum_{i=1}^{n} \xi_i \right. \\ \left. - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\mathbf{w}^{\mathbf{T}} \mathbf{c_i} - v \right) - 1 + \xi_i \right) - \sum_{i=1}^{n} \theta_i \xi_i \right\}. \end{aligned} \tag{13}$$

where $C > 0$ is a penalty value found using cross-validation.

The bottom plot in Figure 4 shows the functional feature spaces of an actual AU classification – AU 12. Since one can only plot two-dimensional feature spaces, we projected the original color spaces onto the first two principal components of the data. This was done with Principal Components Analysis (PCA). The two resulting dimensions are labeled ϕ_{PCA_k}, $k = 1, 2$.

Once trained, this system can detect AU activation in video in real time, > 30 frames/second/CPU thread.

3.2.1 *Deep network approach using multilayer perceptron*

In the previous section, we used a SVM to define a linear classifier in Gabor-transform space. This formulation yields a linear classifier on the feature space of the $\mathbf{c}_i$. We will now use a deep network to identify non-linear classifiers in this color feature space.

We train a multilayer perceptron network (MPN) using the coefficients $\mathbf{c_i}$. This deep neural network is composed of 5 blocks of fully connected layers with batch normalization [36] and rectified linear units (ReLu) [37]. To effectively train the network, we used data augmentation by super-sampling the minority class (active), class weights and weight decay. A summary of the proposed architecture for *each* AU is in Table 1.

We train this neural network using gradient descent. The resulting algorithm works in real time, > 30 frames/second/CPU thread.

Layer type	Input size
Fully + batch normalization + ReLu	3,210
Fully + batch normalization + ReLu	1,056
Fully + batch normalization + ReLu	528
Fully + batch normalization + ReLu	132
Fully + batch normalization + ReLu	64
Fully + sigmoid	64

TABLE 1
Description of the deep network architecture used in this paper.

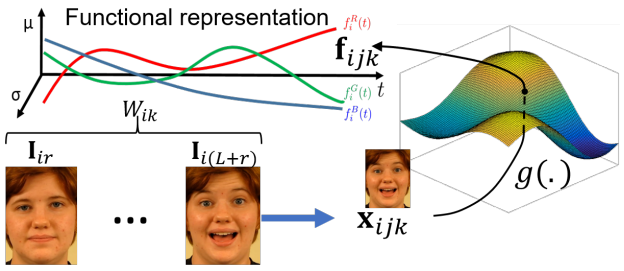


Fig. 5. Each video segment W_{ik} (shown on the bottom left) yields a feature representation $\mathbf{f}_{ijk}$ (top left), $j = 1, \ldots, 107$. We regress a function $g(.)$ to learn to map from the last image of W_{ik} to $\mathbf{f}_{ijk}$, $j = 1, \ldots, 107$ (right image).

4 AU DETECTION IN STILL IMAGES

People generally first observe dynamic facial expressions. Nonetheless, later, we have no problem recognizing facial expressions in still images. We derive an approach that allows our algorithm to recognize AUs in still images [38].

To be able to apply our algorithm to still images, we need a procedure that specifies the color functions $\mathbf{f}_i$ of image $\mathbf{I}_i$. That is, we need to define the mapping $h(\mathbf{I}_i) = \mathbf{f}_i$, Figure 5. And, recall, $\mathbf{f}_i$ is defined by its coefficients $\mathbf{c}_i^T$. These coefficents can be learned from training data using non-linear regression.

We start with a training set of m videos, $\{V_1, \ldots, V_m\}$. As above, $V_i = \{\mathbf{I}_{i1}, \ldots, \mathbf{I}_{ir_i}\}$. We consider every subset of consecutive frames of length L (with $L \leq r_i$), i.e., $W_{i1} = \{\mathbf{I}_{i1}, \ldots, \mathbf{I}_{iL}\}$, $W_{i2} = \{\mathbf{I}_{i2}, \ldots, \mathbf{I}_{i(L+1)}\}$, $\ldots$, $W_{i(r_i-L)} = \{\mathbf{I}_{i(r_i-L)}, \ldots, \mathbf{I}_{ir_i}\}$. This allows us to compute the color representations of all W_{ik} as described in Section 2.1. This yields $\mathbf{x}_{ik} = (\mathbf{x}_{i1k}, \ldots, \mathbf{x}_{i107k})^T$ for each W_{ik}, $k = 1, \ldots, r_i - L$. Following (2),

$$\mathbf{x}_{ijk} = (\mu_{ij1}, \ldots, \mu_{ijL}, \sigma_{ij1}, \ldots, \sigma_{ijL})^T, \tag{14}$$

where i and k specify the video W_{ik}, and j the patch, $j = 1, \ldots, 107$ (Figure 2).

Next we compute the functional color representations $\mathbf{f}_{ijk}$ of each W_{ik} for each of the patches, $j = 1, \ldots, 107$. This is done using the approach detailed in Section 2.2. That yields $\mathbf{f}_{ijk} = (c_{ijk1}, \ldots, c_{ijkQ})^T$, where c_{ijkq} is the q^{th} coefficient of the j patch in video W_{ij}.

Our training set is then given by the pairs $\{\mathbf{x}_{ijk}, \mathbf{f}_{ijk}\}$. This training set is used to regress the function $\mathbf{f}_{ijk} = h(\mathbf{x}_{ijk})$, Figure 5.

Specifically, let $\hat{\mathbf{I}}$ be a test image and $\hat{\mathbf{x}}_j$ its color representation in patch j. We use Kernel Ridge Regression to estimate the q^{th} coefficient of this test image as follows,

$$\hat{c}_{jq} = \mathcal{C}^T \left(\mathbf{K} + \lambda \mathbf{Id} \right)^{-1} \kappa(\hat{\mathbf{x}}_j), \tag{15}$$

where $\hat{\mathbf{x}}_j$ is the color feature vector of the j^{th} patch, $\mathcal{C} = \left(c_{1\,j\,1\,q}, \ldots, c_{m\,j\,(r_m-L)\,q}\right)^T$ is the vector of coefficients of the j^{th} patch in all training images, $\mathbf{K}$ is the Kernel matrix, $\mathbf{K}(i,j) = k(\mathbf{x}_{ijk}, \mathbf{x}_{\hat{i}j\hat{k}})$ (i and $\hat{i} = 1, \ldots, m$, k and $\hat{k} = 1, \ldots, r_i - L$), and $\kappa(\hat{\mathbf{x}}_j) = \left(k(\hat{\mathbf{x}}_j, \mathbf{x}_{1j1}), \ldots, k(\hat{\mathbf{x}}_j, \mathbf{x}_{mj(r_m-L)})\right)^T$. And, we use the Radial Basis Function kernel, $k(\mathbf{a}, \mathbf{b}; \eta) = \exp(-\eta\|\mathbf{a} - \mathbf{b}\|^2)$.

The parameters η and λ are selected to maximize accuracy and minimize model complexity. This is the same as optimizing the bias-variance tradeoff. We use the solution to the bias-variance problem presented in [39].

As shown above, we are ready to use the regressor on previously unseen test images. If $\hat{\mathbf{I}}$ is a previously unseen test image. Its functional representation is readily obtained as $\hat{\mathbf{c}} = h(\hat{\mathbf{x}})$, with $\hat{\mathbf{c}} = (c_{11}, \ldots, c_{107\,Q})^T$. This functional color representation can be directly used in the functional classifier derived above.

5 Experimental Results

The goal of this paper is to introduce a color feature space that can be efficiently and robustly used for the recognition of AUs. This section details experimental results of the theoretical work introduced above. We show that the proposed algorithm, which uses color features, performs better than state-of-the-art algorithms.

5.1 Comparative results

We provide comparative results against state-of-the-art algorithms on four publicly available datasets: Denver Intensity of Spontaneous Facial Action (DISFA) [40], Shoulder Pain (SP) [41], Binghamton-Pittsburgh 4D Spontaneous Expression Database (BP4D) [42], Affectiva-MIT Facial Expression Dataset (AM-FED) [43], and Compound Facial Expressions of Emotion (CFEE) [19]. AM-FED is a database of videos of facial expressions "in the wild," while DISFA, SP and BP4D are videos of spontaneous expressions collected in the lab. CFEE is a database of still images rather than video sequences.

In each database, we use subject independent 10-fold cross-validation, where all frames from a few subjects is held out from the training set and only used for testing. This ensures that subject specific patterns cannot be learned by the classifier. The results of the proposed algorithm are compared to the available ground-truth (manually annotated AUs). To more accurately compare our results with state-of-the-art algorithms, we compute the F_1 score, defined as, $F_1 = 2(\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall})$, 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 recognition of AU i over the actual number of images with AU i.

Comparative results on the first four datasets are given in Figure 6. Our results are compared against the methods of Emotionet [8], Hierarchical-Restricted Boltzmann Machine (HRBM) [44], Transferring Latent Task Structures (TLTS) [45], l_p-norm [46], Discriminant Label Embedding (DLE) [47], Cross-dataset Learning Support Vector Machine (CLM-SVM) [48], and Multi-Conditional Latent Variable Model (MC-LVM) [49]. We also make comparisons with our system annotating frames of the video sequences as still images, using the method described in section 4, and a version of our system that only computes these features in a single intensity channel (i.e., uses only grey-scale information). These results are labelled "This paper (still images)" and "This paper (grey-scale)".

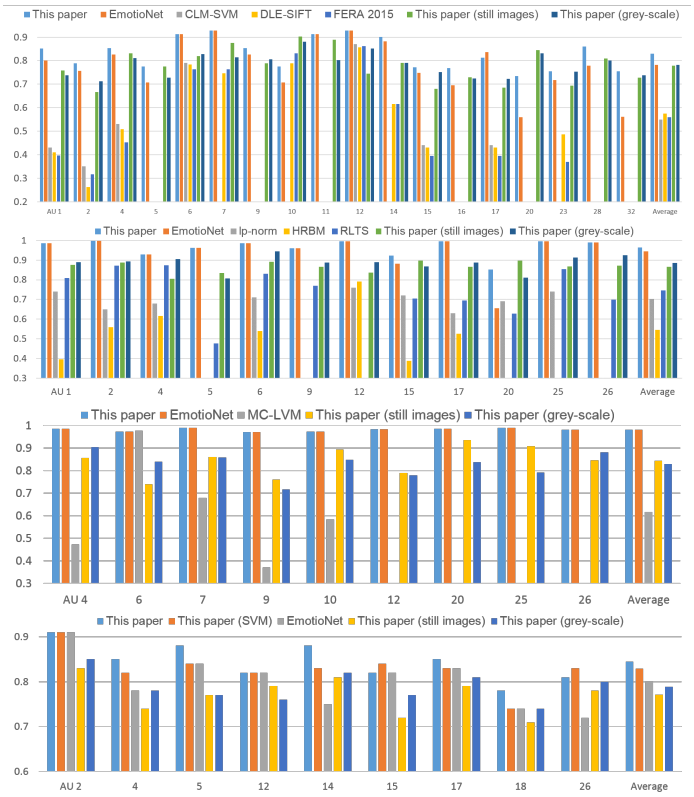


Fig. 6. F1 scores for the proposed approach and a variety of published algorithms. Note that not all published methods provide results for all AUs. This is why you see empty columns in the plots above. Average (Avg) is computed using the results of the available AUs in each algorithm. First plot: BP4D. Second plot: DISFA. Third plot: SP. Four plot: AM-FED.

As we can see in this figure, the herein derived color features achieve superior results to other feature representations previously used in the literature. Also, these results demonstrate that the proposed Gabor transform approach and functional classifier are efficient algorithms for the recognition of AUs in video. Further, it is apparent that both estimating the functional representation using the Gabor transform and using color based features is crucial to this system, as classifying only using still images or using only grey-scale images yields inferior results.

It is important to note that the results of the non-linear classifier (given by a deep network) are not significantly superior to those of a simple linear classifier. This is important, because it further demonstrates that the color features used herein efficiently seperate the feature vectors of different classes (i.e., AU active vs. inactive). This was previously illustrated in the bottom plot of Figure 4.

It is also important to note that our algorithm works

	DN-5 layer	DN-10 layer	SVM - RBF	SVM - polynomial	SVM - tanh
BP4D	0.912	0.911	0.859	0.853	0.847
DISFA	0.966	0.958	0.964	0.943	0.951
SP	0.928	0.927	0.925	0.928	0.925
AMFED	0.844	0.851	0.822	0.819	0.821

TABLE 2
Average F1 score for different classifiers

faster than real-time, >30 frames/second/CPU thread.

5.2 ROC curves

To further study the results on the proposed feature representation, we provide ROC (Receiver Operating Characteristic) curves in Figures 7-10. ROC plots display the *true positive rate* against the *false positive rate*. The true positive rate is the sensitivity of the classifier. The false positive rate is the number of negative test samples classified as positive over the total number of false positives plus true negatives.

ROC curves are computed as follows. The derivations of our approach have an equal priors assumption. That is, the probability of AU being active is the same as that of not being active. We can however vary the value of these priors. Reducing the prior of AU active will decrease the false detection rate, i.e., it is less likely to misclassify a face that does not have this AU active. Increasing the prior of AU active will increase the true positive detection rate. This is a simple extension of our algorithm that allows us to compute ROC curves.

The plots in Figures 7-10 allow us to compute the area under the curve for the results of our algorithm. We do this for all four datasets – BP4D, SP.0, DISFA and AM-FED. The results are in Table 3.

5.3 Invariance to skin color

One may wonder if the results reported above vary as a function of skin color/tone. Close analysis of our results shows that this is *not* the case, i.e., our feature representation is invariant to skin tone.

To demonstrate this, we divided our training samples into four groups as a function of skin color – from lighter to darker skin. We call these skin tonalities: levels 1, 2, 3 and 4. Level 1 represents the lightest tone and level 4 the darkest. The 10-fold cross-validation results using each of these four groups are shown in Figure 11.

A t-test showed no statistical difference in the results of Figure 11 across skin tones. The null hypothesis that these results are different was disproven: $p > .1$ in DISFA, $p > .8$ in BP4D, $p > .3$ in SP.

Figure 12 shows qualitative results on two videos of people of different ethnicity and skin color.

5.4 AU recognition in still images

In Section 4, we derived an approach that allows us to use our algorithm to detect the presence of AUs in still images; even thought the training was done using video sequences.

To achieve this, we trained the proposed regressor $h(.)$ using the three dataset of videos used in the preceding sections. Then, we test the trained system on the still images of the CFEE of [19]. This means that the functions of every test image $\hat{\mathbf{I}}$ are estimated as $\hat{\mathbf{f}} = h(\hat{\mathbf{x}})$.

Figure 13 provide comparative results against the EmotioNet algorithm of [25]. To our knowledge, this is the only other algorithm that has been applied to this dataset to date. It also provides comparative results of using only static color features for each image (i.e., without the proposed regressor)

We see that the results of the proposed algorithm are superior to those of previous approaches for AUs 1, 2, 4, 25 and 26. It is also clear that the regressor is a crucial component of the algorithm, as the results are inferior without it.

5.5 AU intensities

The recognition of intensity of activation of each AU is of high importance in most applications. This section demonstrates that our approach achieves intensity estimation errors that are smaller than those given by state-of-the-art algorithms.

Mean Absolute Error (MAE) is used to calculate the accuracy of the estimated intensities of AU activation. To do this each of the six levels of activation is given a numerical number. Specifically, AU not present (inactive) takes the values 0, intensity a the value 1, intensity b the value 2, intensity c the value 3, intensity d the value 4, and intensity e the value 5. The intensity estimate of the a^{th} samples with AU i active as given by an algorithm is u_{ia}. This estimates are compared to the ground-truth $\hat{u}_{ia}$,

$$MAE_i = \frac{1}{n_i}\sum_{a=1}^{n_i} \mid u_{ia} - \hat{u}_{ia} \mid \tag{16}$$

where n_i is the number of samples with AU i active.

Figure 14 provides comparative results for the recognition of AU intensity. This plot shows the results of the algorithm derived in this paper and those of Multi-Kernel Support Vector Machine (MK-SVM) [50], Context Sensitive Dynamic Ordinal Regression (CS-DOR) [51], Rotation Invariant Feature Regression (RIFR) [52], and EmotioNet [25].

6 Conclusions

The automatic recognition of facial action units and their intensities is a fundamental problem in computer visions with a large number of applications in the physical and biological sciences [3], [4], [38]. Recent computational models of the human visual system suggest that the recognition of facial expressions is based on the visual identification of these AUs, and a recent cognitive neuroscience experiment has identified a small brain region where the computations associated with this visual recognition likely take place [13].

AU	1	2	4	5	6	7	9	10	11	12	14
BP4D	.9656	.9736	.9808	.973	.9656	.9813	.9636	.9983	.9842	.9847	.9447
SP			.9935		.9931	.993	.9911	.9961		.9936	
DISFA	.9862	.9921	.9886	.9891	.9914		.9959			.9866	
AM-FED		.9913	.9858	.9927			.9873			.9873	.9947
AU	14	15	16	17	20	23	25	26	28	32	
BP4D	.9447	.9535	.9802	.9405	.9753	.9446			.9914	.9973	
SP					.9856		.9861	.9936			
DISFA		.9895		.985	.9884		.9956	.9859			
AM-FED	.9845		.9887				.9923				

TABLE 3
Area under the curve of the ROC curves shown in Figures 7-10.

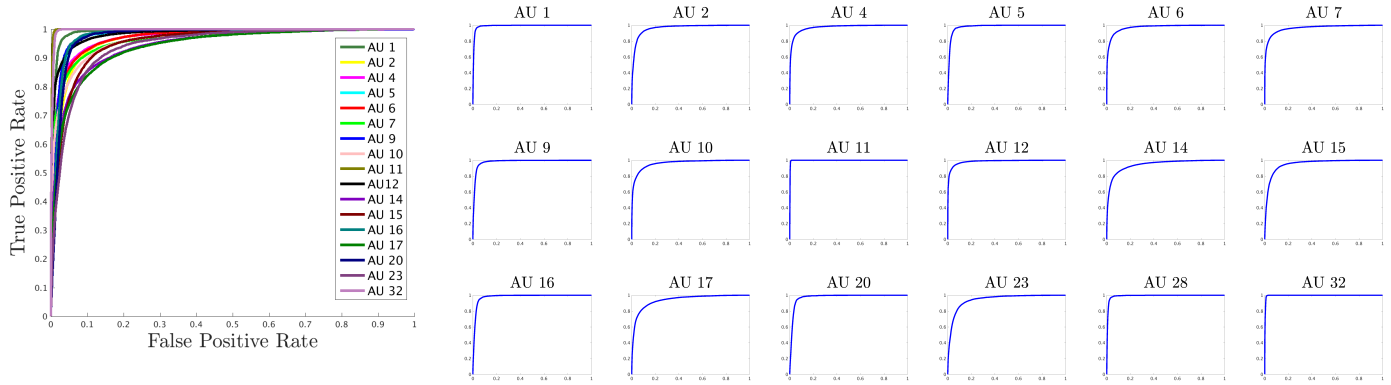


Fig. 7. ROC curves of the results on the BP4D dataset. The left image shows the ROC of all AUs combined. This shows the small variations between different AUs. The other plots show the ROCs of each AU. The area under the curve for each of these AUs are given in Table 3.

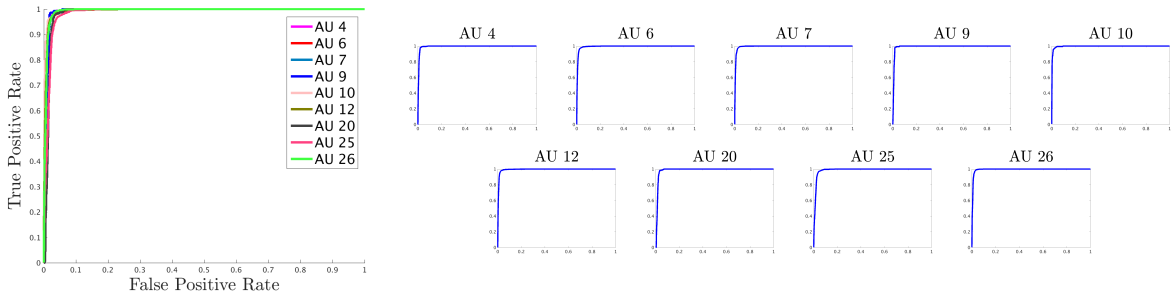


Fig. 8. ROC curves of our results on the Shoulder Pain dataset.

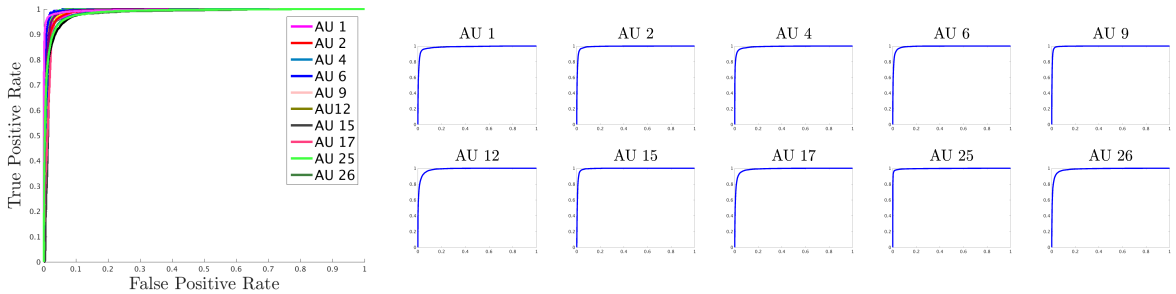


Fig. 9. ROC curves of our results on the DISFA dataset.

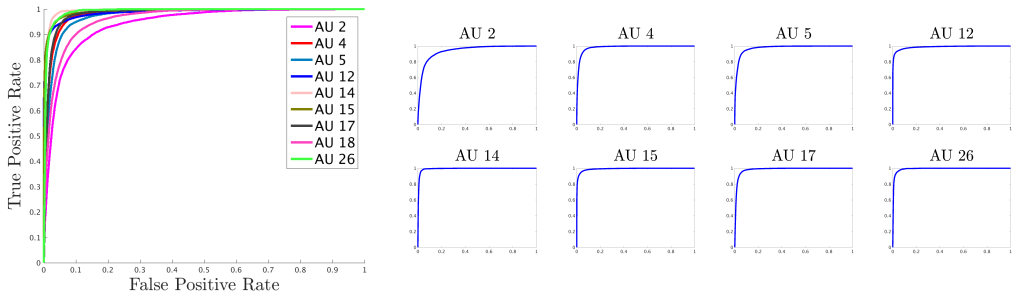


Fig. 10. ROC curves for the AM-FED dataset.

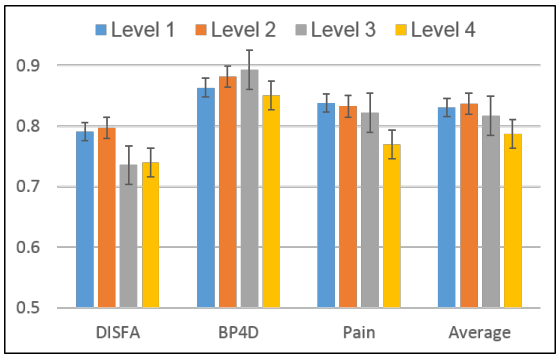


Fig. 11. Average and standard error bars of the F1-scores in each of the four skin tone levels across datasets. These results show no statistical difference between these four groups.

Previous approaches to this automatic recognition have exploited shading, shape, motion and spatio-temporal features [3], [12], [15], [16], [18], [19], [20], [21], [53], [54]. Remarkably absent from this list of features is color.

Nonetheless, faces are colorful. To produce a facial expression, one needs to move the facial muscles under our skin. These changes vary the brdf of the face and either increase or decrease the amount of blood or oxygenation in that local area. This yields clearly visible color changes that, to the authors knowledge, have not been exploited before.

The present work has derived the first comprehensive computer vision algorithm for the identification of AUs using color features.

We derived a functional representation of color and a highly innovative Gabor transform that are invariant to the timing and duration of these AU activations. We also define an approach that allows us to apply our trained functional color classifier to still test images. This was done by learning the mapping between the functional representation of color in video and images. Finally, we showed how these color changes can also be used to detect intensity of AU activation.

In summary, the present work reveals how facial color changes can be exploited to identify the presence of AUs in *videos and still images*. Skin color tone is shown to not have an effect on the efficacy of the derived algorithm.

Acknowledgments

This research was supported in part by the National Institutes of Health, grant R01-DC-014498, and the Human Frontier Science Program, grant RGP0036/2016. RS was partially supported by OSU's Center for Cognitive and Brain Sciences summer fellowship. We thank the reviewers for constructive feedback.

References

[1] P. Ekman and W. V. Friesen, *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.
[2] 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.
[3] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2016.
[4] J. F. Cohn and F. De la Torre, "Automated face analysis for affective computing," in *The Oxford Handbook of Affective Computing*, R. Calvo and S. D'Mello, Eds. Oxford University Press, USA, 2014, p. 131.
[5] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3d facial expression analysis in videos," in *International Workshop on Analysis and Modeling of Faces and Gestures*. Springer, 2005, pp. 293–307.
[6] X. Huang, A. Dhall, X. Liu, G. Zhao, J. Shi, R. Goecke, and M. Pietikainen, "Analyzing the affect of a group of people using multi-modal framework," *arXiv preprint arXiv:1610.03640*, 2016.
[7] A. Bellocchi, "Methods for sociological inquiry on emotion in educational settings," *Emotion Review*, vol. 7, pp. 151–156, 2015.
[8] C. F. Benitez-Quiroz, R. B. Wilbur, and A. M. Martinez, "The not face: A grammaticalization of facial expressions of emotion," *Cognition*, vol. 150, pp. 77–84, 2016.
[9] A. Todorov, C. Y. Olivola, R. Dotsch, and P. Mende-Siedlecki, "Social attributions from faces: Determinants, consequences, accuracy, and functional significance," *Psychology*, vol. 66, no. 1, p. 519, 2015.
[10] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre, "Spontaneous facial expression in unscripted social interactions can be measured automatically," *Behavior research methods*, vol. 47, no. 4, pp. 1136–1147, 2015.
[11] S. Cassidy, P. Mitchell, P. Chapman, and D. Ropar, "Processing of spontaneous emotional responses in adolescents and adults with autism spectrum disorders: Effect of stimulus type," *Autism Research*, vol. 8, no. 5, pp. 534–544, 2015.
[12] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.
[13] R. Srinivasan, J. Golomb, and A. M. Martinez, "A neural basis of facial action recognition in humans," *The Journal of Neurosceince*, 2016.
[14] A. E. Skerry and R. Saxe, "Neural representations of emotion are organized around abstract event features," *Current Biology*, vol. 25, no. 15, pp. 1945–1954, 2015.

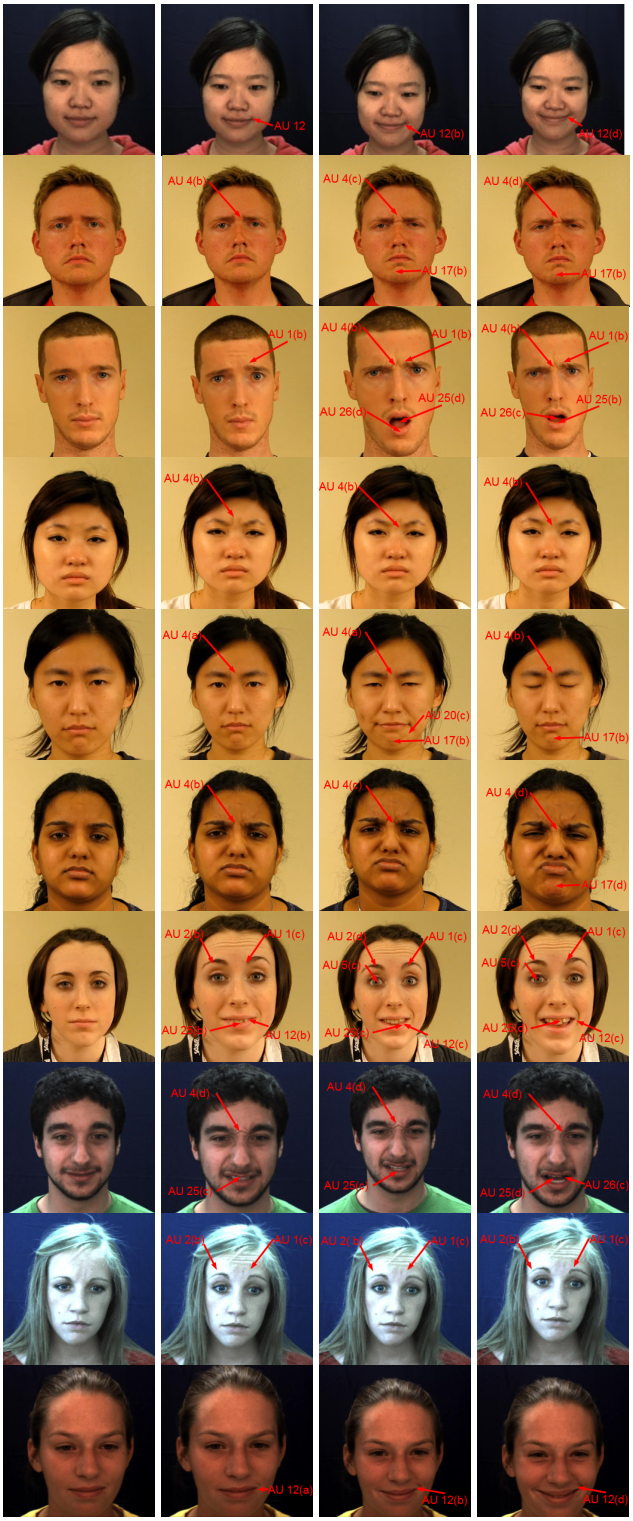


Fig. 12. Qualitative results. These correspond to the automatic annotation of AUs and their intensities. Intensities are in parentheses.

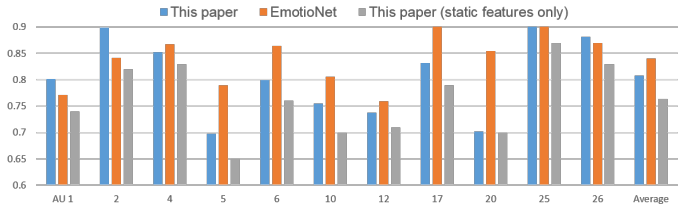


Fig. 13. Comparative F1 scores on CFEE. Our algorithms was trained using the videos of BP4D, DISFA and shoulder pain and tested on the still images of CFEE.

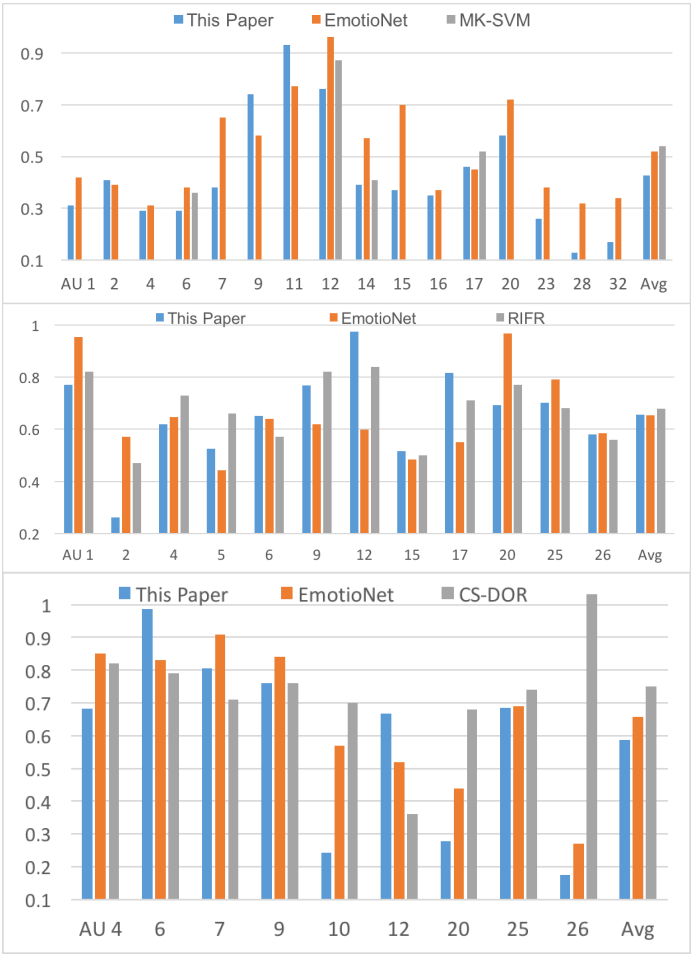


Fig. 14. Mean Absolute Error (MAE) for recognition of AU intensity on a variety of algorithms. Top plot: BP4D. Middle plot: DISFA. Bottom plot: shoulder pain.

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

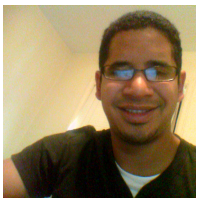
[16] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 229–234.

[17] 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*, vol. 13, no. 1, pp. 1589–1608, 2012.

[18] S. Jaiswal and M. F. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[19] 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.
[20] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 10, pp. 1683–1699, 2007.
[21] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image and Vision Computing*, vol. 31, no. 2, pp. 175–185, 2013.
[22] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Facial color is an efficient mechanism to visually transmit emotion," *Proceedings of the National Academy of Sciences*, vol. 115, no. 14, pp. 3581–3586, 2018.
[23] 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*, vol. 2, 2001, pp. II–635.
[24] M. A. Changizi, Q. Zhang, and S. Shimojo, "Bare skin, blood and the evolution of primate colour vision," *Biology Letters*, vol. 2, no. 2, pp. 217–221, 2006.
[25] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of half a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 2016.
[26] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.
[27] C. F. Benitez-Quiroz, Y. Wang, and A. M. Martinez, "Recognition of action units in the wild with deep nets and a new global-local loss," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3990–3999.
[28] C. A. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," *arXiv preprint arXiv:1803.05873*, 2018.
[29] J. C. Batista, V. Albiero, O. R. Bellon, and L. Silva, "Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 866–871.
[30] S. Blanco, C. D'Attellis, S. Isaacson, O. Rosso, and R. Sirne, "Time-frequency analysis of electroencephalogram series. ii. gabor and wavelet transforms," *Physical Review E*, vol. 54, no. 6, p. 6661, 1996.
[31] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on image processing*, vol. 1, no. 2, pp. 205–220, 1992.
[32] R. Niese, A. Al-Hamadi, A. Farag, H. Neumann, and B. Michaelis, "Facial expression recognition based on geometric and optical flow features in colour image sequences," *IET computer vision*, vol. 6, no. 2, pp. 79–89, 2012.
[33] S. M. Lajevardi and H. R. Wu, "Facial expression recognition in perceptual color space," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3721–3733, 2012.
[34] R. Carmona, W.-L. Hwang, and B. Torresani, *Practical Time-Frequency Analysis: Gabor and Wavelet Transforms, with an Implementation in S*. Academic Press, 1998, vol. 9.
[35] V. Vapnik, *The nature of statistical learning theory (2nd edition)*. Springer Science & Business Media, 2000.
[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift." in *ICML*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456.
[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
[38] A. M. Martinez, "Computational models of face perception," *Current Directions in Psychological Science*, 2017.
[39] D. You, C. F. Benitez-Quiroz, and A. M. Martinez, "Multiobjective optimization for model selection in kernel methods in regression," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 10, pp. 1879–1893, 2014.
[40] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "Disfa: A spontaneous facial action intensity database," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, April 2013.
[41] 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.
[42] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.
[43] 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.
[44] Z. Wang, Y. Li, S. Wang, and Q. Ji, "Capturing global semantic relationships for facial action unit recognition," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3304–3311.
[45] T. Almaev, B. Martinez, and M. Valstar, "Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3774–3782.
[46] 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 *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 1104–1111.
[47] A. Yüce, H. Gao, and J.-P. Thiran, "Discriminant multi-label manifold embedding for facial action unit detection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–6.
[48] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–6.
[49] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Multi-conditional latent variable model for joint facial action unit detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3792–3800.
[50] J. Nicolle, K. Bailly, and M. Chetouani, "Facial action unit intensity prediction via hard multi-task metric learning for kernel regression," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–6.
[51] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 944–958, 2015.
[52] D. Bingöl, T. Celik, C. W. Omlin, and H. B. Vadapalli, "Facial action unit intensity estimation using rotation invariant features and regression analysis," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 1381–1385.
[53] A. M. Martinez, "Matching expression variant faces," *Vision research*, vol. 43, no. 9, pp. 1047–1060, 2003.
[54] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 775–788, 2016.



C. Fabian Benitez-Quiroz (S'02-M'03-S'06) received the B.S. degree in electrical engineering from Pontificia Universidad Javeriana, Cali, Colombia, and the M.S. degree in electrical engineering from the University of Puerto Rico, Mayaguez, Puerto Rico, in 2004 and 2008, and a Ph.D. in Electrical and Computer Engineering from The Ohio State University (OSU) in 2015. He is currently a Postdoctoral researcher in the Computational Biology and Cognitive Science Lab at OSU. His current research interests include the analysis of facial expressions in the wild, functional data analysis, deformable shape detection, face perception and deep learning.

Ramprakash Srinivasan received the B.S. degree with honors in Electrical and Electronics Engineering from Anna University, Chennai, India, in 2013. He is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at The Ohio State University. His research interests include computer vision, machine learning and cognitive science. He is a student member of IEEE.

Aleix M. Martinez is a Professor in the Department of Electrical and Computer Engineering, The Ohio State University (OSU), where he is the founder and director of the Computational Biology and Cognitive Science Lab. He is also affiliated with the Department of Biomedical Engineering and to the Center for Cognitive and Brain Sciences, where he is a member of the executive committee. He has served as an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Affective Computing, Image and Vision Computing, and Computer Vision and Image Understanding. He has been an area chair for many top conferences and was Program Chair for CVPR, 2014. He is also a member of NIH's Cognition and Perception study section. He is most known for being the first to define many problems and solutions in face recognition (e.g., recognition under occlusions, expression, imprecise landmark detection), discriminant analysis (e.g., Bayes optimal solutions, subclass-approaches, optimal kernels), structure from motion (e.g., using kernel mappings to better model non-rigid deformations, noise invariance), and, most recently, demonstrating the existence of a much larger set of cross-cultural facial expressions of emotion than previously known (i.e., compound expressions of emotion) and the transmission of emotion through changes in facial color.