FACIAL GRID TRANSFORMATION: A NOVEL FACE REGISTRATION APPROACH FOR IMPROVING FACIAL ACTION UNIT RECOGNITION

Shizhong Han, Zibo Meng, Ping Liu, and Yan Tong

Dept. of Computer Science and Engineering, University of South Carolina, Columbia, USA

ABSTRACT

Face registration is a major and critical step for face analysis. Existing facial activity recognition systems often employ coarse face alignment based on a few fiducial points such as eyes and extract features from equal-sized grid. Such extracted features are susceptible to variations in face pose, facial deformation, and person-specific geometry. In this work, we propose a novel face registration method named facial grid transformation to improve feature extraction for recognizing facial Action Units (AUs). Based on the transformed grid, novel grid edge features are developed to capture local facial motions related to AUs. Extensive experiments on two well-known AU-coded databases have demonstrated that the proposed method yields significant improvements over the methods based on equal-sized grid on both posed and more importantly, spontaneous facial displays. Furthermore, the proposed method also outperforms the state-of-the-art methods using either coarse alignment or mesh-based face registration.

***Index Terms*—** Facial action unit recognition, grid transformation, face registration.

1. INTRODUCTION

Facial activity is the most important and natural means for expressing human emotion and intention. Facial Action Coding System (FACS) [5] is the most comprehensive and commonly used system for facial activity analysis. Based on FACS, facial activity can be described by combinations of facial Action Units (AUs), each of which is anatomically related to the contraction of a specific set of facial muscles. An automatic system for facial AU recognition has emerging applications such as human behavior analysis, interactive games, computer-based learning, entertainment, telecommunication, and psychiatry, and thus has been extensively studied over the years as detailed in the survey papers [23, 21].

Generally, AU recognition is performed in three steps sequentially: face alignment/registration, feature extraction, and classification. The face is first detected and aligned to reduce variations in scale, rotation, and position, and more importantly, to establish the correspondence of major facial components such as eyes, nose, mouth across different face images. Alignment based on eyes is the most popular strategy since eyes are the most reliable facial components to be detected. Then, a set of features that are discriminative for the target AU are extracted and employed in classification.

Shape-based features, which purely describe positional changes and/or deformations of facial components, have been shown to be inferior to appearance-based features because the latter also capture transient facial appearance changes (e.g., wrinkles and bulges) and are less affected by errors in face alignment [1, 12, 21]. Most recently, histograms of Local Binary Patterns (LBP) [24, 22, 17, 21], Histograms of Oriented Gradients (HOG) [7, 4], and scale-invariant feature transform (SIFT) descriptors [7] have been widely adopted and shown promise in recognizing AUs. Assuming the face region is well aligned, the histogram-like features are often calculated from equal-sized facial grids, as shown in Fig. 1a. However, apparent misalignment can be observed and is primarily caused by variations in face pose and facial deformation (e.g., difference in the bottom of chin in Fig. 1a), as well as the diversity in human face geometry (e.g., difference in the nose tip in Fig. 1a). Since the face region is aligned based on eyes, the misalignment is more severe in the lower face.

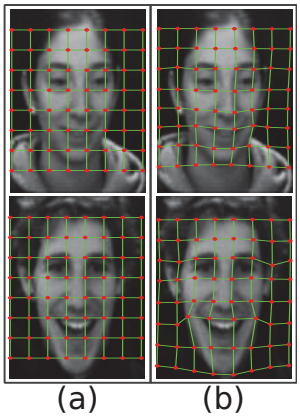

Fig. 1. Grid representation for AU recognition, where red dots denote grid vertices and features are extracted from each cell, respectively. a) Equal-sized grid adopted in current practice, where face region is aligned based on eyes. b) Transformed grid using the proposed method. Note improved correspondence for corners of eyebrows and the bottom of chin in (b).

To deal with this issue, we propose to *transform the facial grid*, so that contents enclosed in grid cells are semantically similar across different faces. Specifically, grid vertices denoted as red dots in Fig. 1 are transformed based on a set of facial landmarks, which are defined around the major facial components. As shown in Fig. 1b, the corners of eyebrows and the bottom of chin are well aligned for the two different subjects using the proposed method. Furthermore, the size and shape of the transformed grid capture the facial geometri-

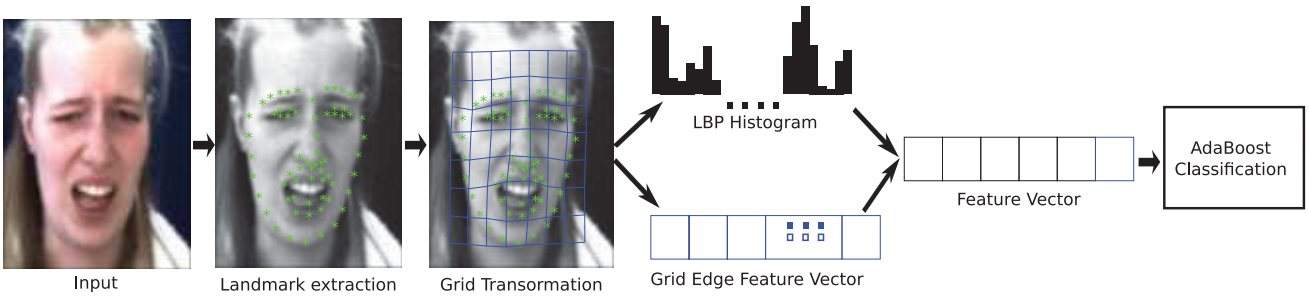

Fig. 2. The flowchart of the proposed method. Solid/hollow blue rectangles represent orientation/length grid edge features, respectively.

cal deformations caused by AUs. For example, elongated grid cells near the center of mouth imply AU26 (jaw drop). In addition, it is observed that the edges of the transformed grid often align with the contours of facial components (referring to the grid edges near the eyebrows, lip, and the bottom of chin in Fig. 1b). Thus, the changes in edge orientations characterize AU-related facial motions as well. Motivated by this, we further propose to *employ features extracted from grid edges* in addition to histogram-based appearance features. Fig. 2 gives the flowchart of the proposed AU recognition system.

Extensive experiments on Cohn-Kanade (CK) database [9] and the Denver Intensity of Spontaneous Facial Action (DISFA) database [16] have demonstrated that the proposed method yields significant improvement over the methods based on equal-sized grid on posed and more importantly, on natural and hence more challenging facial displays.

1.1. Related Work

Face registration is a crucial preprocessing step for facial AU recognition. Based on information employed in registration, methods can be categorized into coarse alignment and mesh-based registration. Coarse alignment [10, 19, 20, 22, 17, 6, 11, 8] takes a few fiducial points (e.g., eye centers, nose tip, and mouth center) to compute a global affine transformation to align the face region, so that the fiducial points are at the same locations across different face images. However, facial components far away from the utilized points cannot be aligned well. In addition, although these points are relatively reliable to detect, the coarse alignment is susceptible to detection error. In contrast, mesh-based registration [14, 2, 13, 18, 15, 16] warps the face into a common mean facial mesh based on a large set of facial landmarks. Then, shape-free texture features and/or "similarity normalized shape" [14] features are extracted for AU recognition. However, some regions such as forehead and cheek are difficult to align due to missing or ambiguous definition of facial landmarks.

Our work differs significantly from the mesh-based registration approaches in two aspects. First, variable-sized grid can cover the whole face region even forehead to fully capture discriminative appearance changes. Second, the novel grid edge features are capable of describing local facial deformations that are important to AU recognition.

2. METHODOLOGY

As shown in Fig. 2, our proposed AU recognition system consists of four major steps: *facial landmarks extraction*, *grid transformation*, *feature extraction*, and *AdaBoost classification*. In this section, we will discuss the proposed grid transformation and grid edge feature extraction method in details.

2.1. Grid Transformation based on Facial Landmarks

The main idea of the proposed method is to alleviate misalignment by adapting the standard equal-sized facial grid to a specific image. Specifically, the vertices of the standard grid will be transferred to a specific image according to their relationships to a set of control points. In this work, a set of facial landmarks estimated by Active Appearance Model (AAM) [3] and normalized based on eyes are employed.

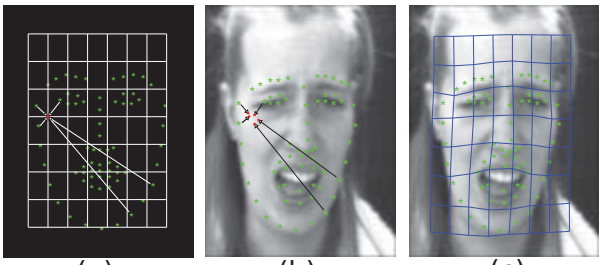

Fig. 3. An illustration of grid transformation. (a) Green stars denote a mean facial shape estimated from training images. Length of a white arrow represents distance between a landmark in the mean shape and a grid vertex in the standard grid. For clarity, only five arrows are displayed. (b) Red dots denote the predicted grid vertices using the five landmarks for a given face image. (c) The final transformed grid is shown in blue.

As shown in Fig. 3a, a mean facial shape denoted by $\bar{\mathbf{s}}$ (green stars) can be estimated from landmarks, given a set of training images. There are spatial relationships between $\bar{\mathbf{s}}$ and grid vertices in the standard grid denoted by $\bar{\mathbf{g}}$. Assuming that these spatial relationships denoted by $R(\bar{\mathbf{s}}, \bar{\mathbf{g}})$ can be transferred to a new face image, the transformed grid vertices (denoted by $\mathbf{g}$) can be calculated given facial landmarks detected on the current image ($\mathbf{s}$) as follows:

$$\mathbf{g} = f[R(\bar{\mathbf{s}}, \bar{\mathbf{g}}), \mathbf{s}] \quad (1)$$

where $f(\cdot)$ denotes a transformation function.

The relationship can be described as geometrical transformation such as pure translation, scaling, and rotation. In this work, we employ the simplest spatial relationship, i.e., pure translation, by assuming the difference between each pair of grid vertex and facial landmark is a constant. Thus, the relative positional difference between the i^{th} landmark and the j^{th} grid vertex is defined as

$$c_{i,j} = \bar{s}_i - \bar{g}_j \tag{2}$$

where $i = 1, \cdots, M$ (M is the number of landmarks) and $j = 1, \cdots, N$ (N is the number of grid vertices). Then, Eq. 1 can be simplified as:

$$g_{i,j} = (s_i - \bar{s}_i) + \bar{g}_j \tag{3}$$

Based on Eq. 3, each grid vertex has a predicted position $g_{i,j}$ using each landmark as shown in Fig. 3b. Then, the final predicted position for each grid vertex (g_j) is calculated as a weighted sum of all predictions as $g_j = \sum_{i=1}^{M} w_i * g_{i,j}$ (Fig. 3c). Since local facial deformation is desired, the closer the landmarks are to the grid vertex, the more important they are in determining g_j. Thus, the weight is defined to be inversely proportional to the distance between $\bar{g}_j$ and $\bar{s}_i$ as:

$$\hat{w}_i = \frac{1}{(|\bar{s}_i - \bar{g}_j|)^l} \tag{4}$$

The weights are further normalized as $w_i = \frac{\hat{w}_i}{\sum_{k=1}^{M} \hat{w}_k}$, where l is a control parameter. When l is big, g_j will more rely on the nearest landmarks. The impact of l on AU recognition will be analyzed in Section 3.1.

2.2. Grid Edge Feature Extraction

The grid vertices are estimated using facial landmarks, displacements of which characterize facial motions. Consequently, the size and shape of the transformed grid can be exploited in recognizing AUs. Specifically, features extracted from grid edges include the length and orientation of grid edges. For example, a moderately elongated grid edge e_2 in Fig. 4 implies the presence of AU26 (jaw drop); while an extremely elongated e_2 implies the presence of AU27 (mouth stretch). In another example, the orientation of grid edge e_1 represents strongly activated AU1 (inner brow raiser).

As shown in Fig. 4, the orientation feature θ_1 is calculated based on a local coordinate system centered at the closest grid vertex. For the length feature, in order to eliminate the influence of person-specific face geometry (width and height), vertical edges $\mathbf{e}_v$ and horizontal edges $\mathbf{e}_h$ will be normalized by dividing the mean lengths of the vertical and horizontal edges, respectively. The details of grid edge feature extraction are summarized in Algorithm 1.

In addition to grid edge features, histograms of LBP features are calculated based on the uniform pattern with 59 bins and from each quadrilateral cell in the transformed grid. Then, LBP histogram and grid edge features are employed as weak classifiers to train an AdaBoost classifier for each AU.

Algorithm 1 Grid edge feature extraction

Input: N 2-D grid vertices.

Output: Normalized length features and orientation features for K_v vertical and K_h horizontal edges.

for $i = 1$ to K_v **do**
  Get the vertical edge $\mathbf{e}_{v_i}$
  Calculate the length of edge $\hat{l}_{v_i} = \|\mathbf{e}_{v_i}\|_2$
  Calculate the orientation of the edge $\theta_{v_i} = \arctan(\frac{e_{v_{i,y}}}{e_{v_{i,x}}})$
end for
for $i = 1$ to K_v **do**
  Normalize the length of edge as $l_{v_i} = K_v \frac{\hat{l}_{v_i}}{\sum_{k=1}^{K_v} \hat{l}_{v_k}}$
end for
Length and orientation features for horizontal edges are calculated similarly.

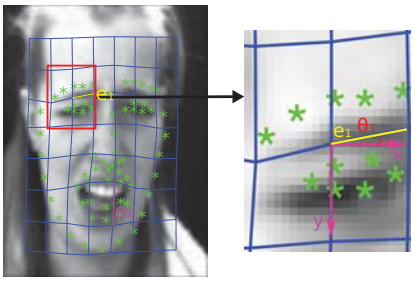

Fig. 4. Illustration of grid edge feature extraction. e_1 and e_2 are two grid edges. The figure on the right is a close look of the highlighted region in the left image. θ_1 represents the orientation feature of e_1.

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of proposed method, we have performed extensive experiments on two facial AU-coded image databases: the CK database [9] and the DISFA database [16]. For the standard grid, we empirically employed 7×7 cell size that is optimal for AU recognition. For grid transformation, facial landmarks are provided by the databases, which have been estimated using AAM fitting.

3.1. Performance Evaluation on CK Database

The CK database [9] contains 486 image sequences from 97 subjects and has been widely used for evaluating the performance of AU recognition. In the first experiment, the proposed method was evaluated on the CK database to demonstrate the generalization capability for a large population. Specifically, 14 AUs, which have been annotated frame-by-frame in the Enhanced CK database [20], are to be recognized. In our experiments, the CK database was divided into 8 subsets, where the subjects in any two of subsets were mutually exclusive. We then employed an 8-fold cross-validation on the CK database. The proposed method achieves an overall recognition rate of 94.96%, with a true positive rate (TPR) of 87.02% and a false alarm rate (FAR) of 4.03%, which are calculated from an average of the 8 runs.

We first compare the proposed method with three baseline methods: *Grid-eye* (standard grid normalized by eyes), *Grid-eye-mouth* (standard grid normalized by eyes and mouth), and

the proposed grid transformation without grid edge features. As shown in Fig. 5, the proposed method yields drastic improvement compared with *Grid-eye* (TPR 81.97% and FAR 5.72%) and *Grid-eye-mouth* (TPR 83.14% and FAR 5.47%). Compared with *Grid-eye*, the improvement is more significant on the 7 lower face AUs [1]: TPR increases from 79.41% (*Grid-eye*) to 87.88% (the proposed) and FAR decreases from 6.66% (*Grid-eye*) to 3.86% (the proposed).

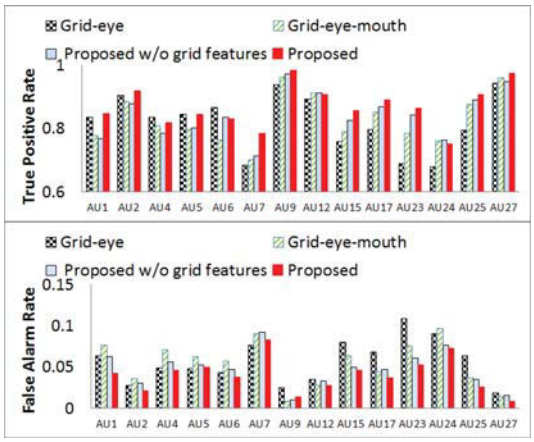

Fig. 5. Performance comparison on the CK database.

Although *Grid-eye-mouth* performs better than *Grid-eye* for lower face AUs, its performance of upper face AUs is inferior. Note that our proposed method outperforms them for both lower and upper face AUs, which demonstrates that grid transformation achieves a better face alignment in terms of improving AU recognition. Furthermore, the novel grid edge features have been demonstrated to be helpful in AU recognition. Compared with the method without grid features, the recognition performance is further improved: TPR increases from 85.38% (*Proposed w/o grid features*) to 87.02% (the proposed) and FAR decreases from 4.45% (*Proposed w/o grid features*) to 4.03% (the proposed).

The proposed method also outperforms the state-of-the-art approaches evaluated on the CK database. Lucey et al. [14] employed a mesh-based registration based on AAM and achieved an average recognition rate of 95.47% with an FAR 15.99% [2]. Note that they only employed the last frame of each sequence with AUs activated at higher intensity level; while our method was evaluated on the whole image sequences containing AUs at low intensity level, which are more difficult to recognize. Tong et al. [20] employed eye-based coarse alignment and achieved an average recognition rate of 93.33% with a TPR 86.3% by exploiting the relationships among AUs.

Varying the parameter l in Eq. 4 will affect grid transformation. [3] Here, we performed a study to find out if recognition performance is sensitive to the choice of l. In our experiments, when l ranges from 4 to 8, the average recognition rate remains almost the same (mean 94.9% with a standard deviation of 0.13%), which demonstrates that the proposed method is robust to l.

3.2. Performance Evaluation on DISFA Database

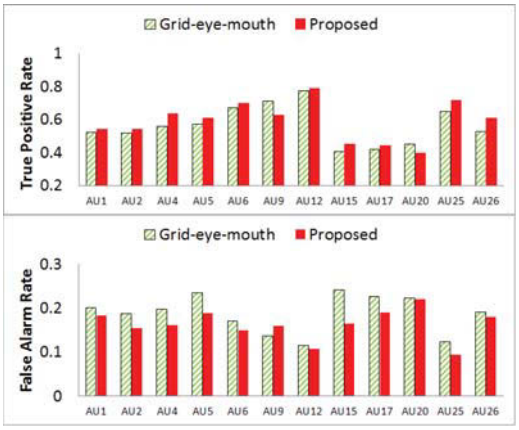

Fig. 6. Performance comparison on the DISFA database.

The DISFA database contains stereo videos of 27 subjects with different ethnicity [16]. Each subject watched a 4-minute video for eliciting emotion and showed spontaneous facial displays with natural head movements. There are a total of 130,814 images in this database. In our experiment, we use a 9-fold cross-validation strategy to evaluate the proposed method on spontaneous and thus more challenging data.

The proposed method achieved an average recognition rate of 81.17% on spontaneous DISFA database. As shown in Fig. 6, TPR increases from 56.43% (*Grid-eye-mouth*) to 58.77% and FAR decreases from 18.75% (*Grid-eye-mouth*) to 16.20%. The improvement is more impressive for some AUs that are difficult to recognize. For example, TPR of AU15 increases from 40.68% (*Grid-eye-mouth*) to 45.24% with a decrease in FAR from 24.19% (*Grid-eye-mouth*) to 16.37%.

4. CONCLUSION

In this paper, we propose a novel facial grid transformation method to improve feature extraction in facial AU recognition. Based on the transformed grid, novel grid edge features are developed to capture local facial motions related to AUs. Extensive experiments on two well-known AU-coded databases show that the proposed method yields significant improvements over the equal-sized grid based methods, especially for lower face AUs; and it also outperforms the state-of-the-art methods employing either coarse alignment or mesh-based face registration. In our future work, this method will be employed as a preprocessing step for other face-related problems such as facial expression recognition.

5. ACKNOWLEDGMENTS

This work was supported by National Science Foundation under CAREER Award IIS-1149787.

[1] AU12 (lip corner puller), AU15 (lip corner depressor), AU17 (chin raiser), AU23 (lip tightener), AU24 (lip presser), AU25 (lips part), and AU27 (mouth stretch) are the 7 lower face AUs recognized.

[2] We employed their published results [14] to calculate the average recognition rate for the 14 target AUs.

[3] In the previous experiments, l was set to 6 empirically.

6. REFERENCES

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

[2] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FG*, pages 915–920, 2011.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, 2001.

[4] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based HoG features. In *FG*, March 2011.

[5] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System: the Manual*. Research Nexus, Div., Network Information Research Corp., Salt Lake City, UT, 2002.

[6] T. Gehrig and H. Ekenel. Facial action unit detection using kernel partial least squares. In *ICCV Workshops*, pages 2092–2099, 2011.

[7] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang. Multi-view facial expression recognition. In *FG*, pages 1–6, 2008.

[8] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Trans. on Cybernetics*, 44(2):161–174, Feb 2014.

[9] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *FG*, pages 46–53, 2000.

[10] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li. Automated facial expression recognition based on facs action units. In *FG*, pages 390–395, April 1998.

[11] G. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *J. IVC*, 27(12):1741–1844, Nov. 2009.

[12] Patrick Lucey, Jeffrey Cohn, Simon Lucey, Iain Matthews, Sridha Sridharan, and Kenneth M Prkachin. Automatically detecting pain using facial actions. In *ACII*, pages 1–8, 2009.

[13] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. *IEEE T-SMC-B*, 41(3):664–674, 2011.

[14] S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In K. Kurihara, editor, *Face Recognition Book*. Pro Literatur Verlag, Mammendorf, Germany, April 2007.

[15] Mohammad H Mahoor, Mu Zhou, Kevin L Veon, Seyed Mohammad Mavadati, and Jeffrey F Cohn. Facial action unit recognition with sparse representation. In *FG*, pages 336–342, 2011.

[16] S. Mavadati, M. Mahoor, Kevin Bartlett, Philip Trinh, and J. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Trans. on Affective Computing*, 4(2):151–160.

[17] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In *FG Workshops*, pages 860 – 865, 2011.

[18] Tomas Simon, Minh Hoai Nguyen, Fernando De La Torre, and Jeffrey F Cohn. Action unit detection with segment-based svms. In *CVPR*, pages 2737–2744, 2010.

[19] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE T-PAMI*, 23(2):97–115, February 2001.

[20] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE T-PAMI*, 29(10):1683–1699, 2007.

[21] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analyis of the first facial expression recognition challenge. *IEEE T-SMC-B*, 42(4):966–979, 2012.

[22] Tingfan Wu, Nicholas J Butko, Paul Ruvolo, Jacob Whitehill, Marian Stewart Bartlett, and Javier R Movellan. Action unit recognition transfer across datasets. In *FG*, pages 889–896, 2011.

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE T-PAMI*, 31(1):39–58, Jan. 2009.

[24] G. Zhao and M. Pietiäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE T-PAMI*, 29(6):915–928, June 2007.