Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

Article

Active AU Based Patch Weighting for Facial Expression Recognition

Weicheng Xie, Linlin Shen *, Meng Yang and Zhihui Lai

School of Computer Science & Software Engineering, Shenzhen University, Nanhai Ave 3688, Shenzhen, Guangdong 518060, China

* Correspondence: llshen@szu.edu.cn; Tel.: +86-755-8693-5089

Abstract: Facial expression has lots of applications in human-computer interaction. Although feature extraction and selection have been well studied, the specificity of each expression variation is not fully explored in state-of-the-art works. In this work, the problem of multiclass expression recognition is converted into triplet-wise expression recognition. For each expression triplet, a new feature optimization model based on Action Unit (AU) weighting and patch weight optimization is proposed to represent the specificity of the expression triplet. Sparse representation based approach is then proposed to detect the active AUs of testing sample for better generalization. The algorithm achieved competitive accuracies of 89.67% and 94.09% for Jaffe and CK+ databases, respectively. Better cross-database performance has also been observed.

Keywords: expression recognition; expression triplet; feature optimization; AU weighting; active AU detection

---

1. Introduction

With the help of facial expression recognition, human-computer interaction can automatically obtain the information of human face and infer the psychologic status of the user, which can be applied to driver monitoring, face paralysis expression recognition, intelligent access control and so on.

Recognition of six basic expressions like happy (Ha), angry (An), surprise (Su), fear (Fe), disgust (Di), sad (Sa) and the neutral (Ne) expression can be categorized into 3D-based and 2D-based approaches. The 3D-based expression recognition is a current research hot topic [1], which often employs the geometry features like differential curvature [2,3] based on an aligned face mesh [4]. The 2D-based approaches are currently prevailing due to the easy accessibility of the training samples. Facial action coding system (FACS) [5] is one of the important 2D approaches, i.e., the expressions are described and tracked according to some basic action units (AUs). The FACS was defined by Ekman [6] to reflect the deformation status of the corresponding facial part, which was developed based on a set of discrete emotions and initially applied to measure some specific facial muscle movements named as AUs. While AUs were often used as an intermediate step for recognizing the basic expressions, image-based feature representation is often considered for recognizing the expression directly. Dynamic recognition based on expression images is important in face animation and multimedia analysis [7–10]. As less information about the considered expression is available, static image based recognition is more challenging.

Deep learning with convolutional neural network (CNN) such as multiscale feature based CNN [11], hierarchical committee based CNN [12] and architecture improved CNN [13] has also been applied for static expression recognition. Pramerdorfer and Kampel [14] gave a detailed survey about these algorithms. Although it is experimentally verified in [15] that visually similar features to the facial AUs are obtained by the CNN-based algorithms, the weighting and optimization of these AU alike features in these algorithms are not fully studied. Meanwhile, CNN-based algorithms require large number of training parameters and high computational complexity [16,17]. Thus, our work focuses on exploring the feature optimization and active AU detection, which can be further integrated into the CNN to further improve the encoded features.

For static image encoding, many features have been proposed and evaluated. Examples are Gabor surface feature [18], Haar-like features [19], histograms of oriented gradients (HOG), local binary patterns (LBP) [20,21], radial encoding feature [22] and key point movement feature [23]. Combination of texture and geometric features was introduced [24] to solve the problem of minor expression deformation when wrinkle feature is unavailable [25]. In this work, the Gabor surface feature (GSF) proposed in [18] is employed for the feature representation. Gabor magnitude surface can reflect the differential geometry information even when the considered face is slightly deformed. Thus, it can discriminate different wrinkle textures on the expression face.

Based on the devised features, feature selection was often conducted for not only boosting the efficiency [26], but also improving the recognition accuracy and the generalization ability [27,28]. Feature selection can be conducted on the whole image, such as independent component analysis (ICA) [29], linear discriminant analysis (LDA) [30], rotational invariant dimensionality reduction [31], maximum margin projection [32] and supervised locally linear embedding [33]. For patch and landmark point-based feature representation, salient expression regions were manually located [34]. Automatic feature selection was usually realized with AdaBoost [19,20,26,35]. For better generalization ability, some feature selection algorithms attempted to obtain a relatively sparse number of features with the incorporation of optimization algorithms. Jia et al. [36] weighted the LBP feature with sparse representation, Zafeiriou and Pitas [37] proposed the sparse feature graph for the recognition. Feature selection has also been achieved by feature reduction or transformation, such as margin maximization [38], normalized cut [28], multitask joint sparse presentation [39] and locality preserving projection (LPP) [40]. An unified classification system [41] integrating feature selection and reduction was proposed based on boosted deep belief network (BDBN).

These algorithms devised and selected the same features for all the categories of expressions (such as six basic expressions), which may leave out the feature speciality and multi-scale property. Thus, pairwise expression features were proposed to improve the recognition performance. Kyperountas et al. [42] employed a pairwise expression recognition with a class separability measure by pairwise inter-class difference maximization and intra-class difference minimization. Happy and Routray [43] extracted pairwise appearance features with LDA for the recognition. Based on different pairwise expression features, feature selections were also different. Liu et al. [44] proposed feature disentangling machine (FDM) to learn different pairwise features. Besides of the common features for all the expressions, the specific features for each expression pair were also selected with multi-task learning in [45]. When most of these algorithms extract expression features from patches, they may not fully consider the causal relation between the patches since facial expressions are often demonstrated in the scale of facial parts involving multiple patches.

The AU-based features integrates the causal relation of patch features naturally since they reflect the deformation status of the facial parts. Tian et al. [5] used the classified AUs for the facial expression recognition, where the AUs were encoded by the geometric size and deformation. Tong et al. [46] constructed the Bayesian network of the causal relation of facial AU features with the corresponding conditional probability table. The expression appearance variance is represented with the assembly of AUs by deep network [47]. Zhao et al. [48] proposed component feature based on face block and weight assignment for expression recognition from near-infrared videos. The AU deformation intensities were estimated with regressors and used to train different classifiers for expression recognition [34]. Li et al. [49] introduced an unified probabilistic framework to simultaneously represent the facial AU evolvement, interactions and observations with different levels of feature representations and classification system.

However, current AU-based algorithms modeled the AU relations without considering the weights of patches contained in each AU. AU feature is suitable for encoding the macro-scale information of each expression since it integrates the large-scale information of face parts. However, they are not good for encoding the micro-scale feature since the combination space of AUs is limited when they are not carefully organized. Moreover, most of these algorithms learned the features from the training expression samples. However, the active feature implied in each testing expression sample is not fully exploited. Model learned from training samples might produce wrong classification when the testing sample is significantly different. However, only few works detect and use the active features of testing sample for recognition.

Although a lot of research works have been conducted in face expression recognition, the following problems are still to be addressed. First, current algorithms recognize the seven expressions entirely with an uniform feature weight, which may leave out the feature speciality and multi-scale property. In this work, the seven expression recognition is divided into multiple sub-problems with appropriate subsets, i.e., the expression triplet. Moreover, the weight vector w.r.t. each expression triplet are fine tuned individually to fully consider its specificity; Second, the patch-based and AU-based features are often encoded separately without considering the influence of their deficiencies. In this work, the weights of patches contained in each AU are finely optimized to represent characteristics of different expressions. In this way, the advantages of large-scale (AU-based) and small-scale (patch-wise) features are both explored; Third, few of current works make use of the specificity of each testing sample before recognition, wrong classification could occur when the testing sample is significantly different. Thus, this work exploits the active features of each testing expression.

The main novelties of this work are mainly on three aspects. First, a two-stage expression recognition using the idea of expression triplet weighting is introduced for representation of diversity among different expressions; Second, a new offline weight optimization for the patches contained in each AU is proposed to increase the discrimination abilities of both patch and AU features. Third, online detection of active AUs for each testing sample is proposed to fully exploit its specificity for feature encoding.

This paper is structured as follows. Section 2 gives a description about the proposed algorithm step by step. The experimental results of the proposed algorithm on public databases are presented in Section 3. Finally, discussions and some conclusions are addressed in Section 4.

2. The Proposed Algorithm

2.1. Framework of the Algorithm

The sketch of expression recognition is illustrated in Figure 1. In offline training, faces were divided into a number of non-overlapped patches, regions, and facial parts like eyes, nose and mouth, where Gabor surface features were extracted. AUs for each face part is defined as well. For each expression triplet combination, the weights of AUs and patches were optimized. In testing, the standard seven class based expression recognition was conducted at the first stage and the top three expression candidates were proposed. Based on the suggested expression triplet, the active AUs of the testing sample were detected and weighted using the learned weight vector. Weighted SVM was finally applied to give the expression label. The entire algorithm is then elaborated in the following sections.

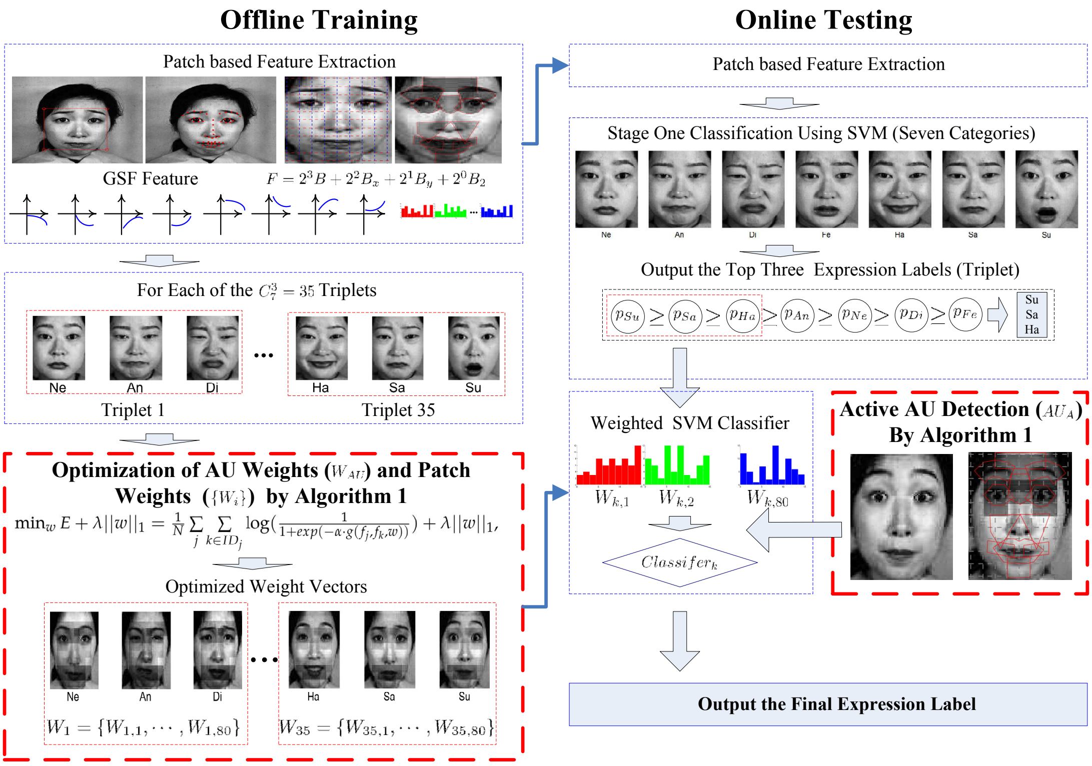

Figure 1. The framework of the proposed algorithm.

2.2. Region Definition and Feature Extraction

2.2.1. Definition of Patch, Region, Part and AU

Face image was first aligned with reference to the feature points located using the approach presented in [50,51], and then resized to 84×68. Only the central region with size 80×64 is cropped out for the following processing. Illumination was normalized by the method proposed in [52].

To extract features representing local face variations caused by expression, the face image was divided into 10×8 patches. Based on the 80 patches, 12 regions with relatively fixed shapes (RFSs) were defined to represent the expression sensitive face parts (PT). Figure 2c shows the 12 regions and Table 1 lists the involved landmarks for each region. When patches are used to represent local texture, regions encode the important variations correlated with expression changes. Figure 3 presents the involved regions used to define each of the seven face parts, i.e., eyes, brows, mouth, forehead, nose root, nasolabial regions and chin ($PT_1 - PT_7$).

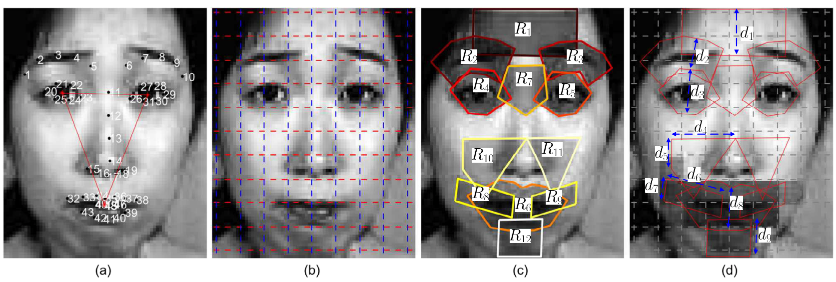

Figure 2. The patches and RFS regions. (**a**) The landmarks used for alignment; (**b**) The patches; (**c**) The RFS regions for defining face parts; (**d**) The sizes $d_1 - d_9$ of the RFS regions. The darker region on the mouth part denotes the patches having nonempty intersection with R_6, R_8, R_9 in (**c**), which presents an example of the relation between patch, region and part.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

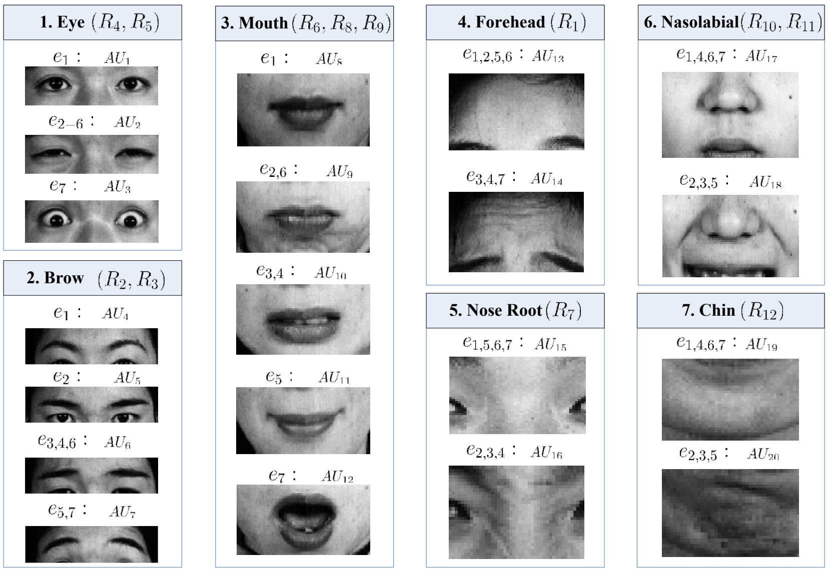

Figure 3. The AUs of seven facial parts with the corresponding RFSs and expression labels. The first part (eye, PT_1) consists of two regions R_4, R_5, which can be classified as $AU_1 - AU_3$ according to its eye status. The listed $e_1 - e_7$ are the expressions related with each defined AU, i.e., neutral, angry, disgust, fear, happy, sad and surprise.

Table 1. Feature point sequences for constructing the RFS regions in Figure 2c and the region sizes in Figure 2d. Take R_1 as an example, points $\{P_3, P_8\}$ define the length and width d_1 of R_1.

R_1 $\{P_3, P_8\}$, $d_1 = \frac{1}{2}\|\|P_3 - P_8\|\|_2$	R_2 $\{P_1 - P_5\}$, $d_2 = 2\|\|P_1 - P_2\|\|_2$	R_3 $\{P_6 - P_{10}\}$, The same as d_2	R_4 $\{P_{20} - P_{25}\}$, $d_3 = 2\|\|P_{20} - P_{21}\|\|_2$
R_5 $\{P_{26} - P_{31}\}$, The same as d_3.	R_6 $\{P_{32} - P_{43}\}$, $d_8 = 4\|\|P_{32} - P_{33}\|\|_2$	R_7 $\{P_{12}, P_{23}, P_5, P_6, P_{26}\}$	R_8 $\{P_{32}, P_{33}, P_{43}\}$, $d_6 = 2\|\|P_{32} - P_{43}\|\|_2$, $d_7 = \|\|P_{33} - P_{43}\|\|_2$
R_9 $\{P_{37}, P_{38}, P_{39}\}$, The same as d_6, d_7	R_{10} $\{P_{32}, P_{13}, P_{20}, P_{15}\}$, $d_4 = \|P^x_{20} - P^x_{13}\|$, $d_5 = \|P^y_{13} - P^y_{15}\|$	R_{11} $\{P_{38}, P_{13}, P_{29}, P_{15}\}$, The same as d_4, d_5	R_{12} $\{P_{39}, P_{43}\}$, $d_9 = \|\|P_{43} - P_{39}\|\|_2$

For encoding the face part variations related with expression [53], Action Units (AUs) have been widely used in literature. We defined 20 AUs to encode the seven expressions named as $e_1 - e_7$, i.e., neutral, angry, disgust, fear, happy, sad and surprise. Different from the AU labeling in [53], the part regions of each AU are manually labeled for training samples in this work. See Figure 3 for the definition of 20 AUs and their relationship with the 12 regions and the 7 parts. For example, AU_3 encodes wide opening eyes, which is usually correlated with e_7, i.e., surprise. As both AU_1 and AU_2 encode the status of the two eyes, they mainly consist of regions around eyes, i.e., R_4 and R_5. Figure 2d labels the involved patches of an example mouth part ($AU_8 - AU_{12}$) with dark grey, i.e., each AU of a part may involve multiple patches.

2.2.2. Feature Extraction

For each patch or RFS region, the Gabor filter [54] is employed for feature extraction, which is formulated as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at *Sensors* **2017**, *17*, 275; doi:10.3390/s17020275

$$\begin{cases} g(x_1,x_2) = \frac{k_s^2}{\sigma^2} exp[-\frac{k_s^2}{2\sigma^2}(x_1'^2 + x_2'^2)]exp(k_s x_1' i) \\ x_1' = \cos\alpha_a x_1 + \sin\alpha_a x_2, x_2' = -\sin\alpha_a x_1 + \cos\alpha_a x_2, \\ k_s = \frac{\pi}{\sqrt{2}^{2+s}}, s = 1,\cdots,n_s; \alpha_a = \frac{a\pi}{8}, a = 1,\cdots,n_a. \end{cases} \tag{1}$$

where $\sigma = 1.8\pi$, $n_s = 5, n_a = 8$, (k_s, α_a) define the amplitude and orientation of the central frequency.

To encode the texture magnitude map, the Gabor surface feature (GSF) [18] is employed since it can depict the curvature information of the wrinkles and direction of the expression texture. To extract GSF, the $n_s \times n_a$ Gabor magnitude images are first extracted, which are then encoded by local binary pattern (LBP) to reduce the feature sensitivity to misalignment. More precisely, feature GSF for the k-th patch p_k is formulated as follows

$$pf_{p_k} = 2^3 B + 2^2 B_x + 2^1 B_y + 2^0 B_2, \tag{2}$$

where B, B_x, B_y, B_2 are the binarizations of the Gabor magnitude image I, its first-order and second-order gradient pictures I_x, I_y, $I_{xx} + I_{yy}$ corresponding to the patch p_k. As an example, for each pixel (i, j) of p_k, its binary value is defined as

$$B_{i,j} = \begin{cases} 1 & \text{if } I_{i,j} \geq ThresMed_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where the threshold $ThresMed_{i,j}$ is the median of the pixel values of patch p_k. Thus, pf_{p_k} is the output feature map with value ranging from 0 to 15, which is further transformed to the histogram for feature representation. For each face patch or region, the corresponding feature GSF is then vectorized as a $16 \times n_s \times n_a$ dimension vector, where n_s, n_a are defined in Equation (1). Finally, the feature of the i-th expression sample is represented as

$$f_i = (pf_{p_1}, \cdots, pf_{p_n}), \tag{4}$$

where n is the number of patches or regions. For convenience of the following illustration, the feature of the i-th expression sample can be also grouped as $(f_i^{(1)}, \cdots, f_i^{(7)})$ according to the seven face parts presented in Figure 3.

2.3. Feature Optimization

Based on the feature representation, the weights of AU and patches for each expression triplet are optimized. First, the original AUs are weighted with conditional transition probability matrix, based on which, weight optimization is performed to weigh the patches involved in each AU in the second step. These two steps are conducted on the training samples and are offline. The third step is to select the active AUs for each testing sample. The entire procedure of the feature optimization is presented in Algorithm 1.

Algorithm 1 AU weighting, patch weight optimization and active AU detection.

1: **Offline Training:**
2: AU weighting (W_{AU}) using the conditional probability matrix presented in Section 2.3.1;
3: Patch-wise weight optimization of weight vectors ($\{W_i\}$) by multi-task sparse learning presented in Section 2.3.2;
4: **Online Testing:**
5: Active AU (AU_A) detection for testing samples by sparse representation presented in Section 2.3.3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at *Sensors* **2017**, *17*, 275; doi:10.3390/s17020275

2.3.1. AU Weighting

Motivated from the causal AU pair extraction with large transition probability by the Bayesian network (BN) [46], the representative abilities of all the AUs for each expression are weighted in this work to decrease the influence of weakly related AUs and provide a constraint for the following patch-wise weight optimization.

With the labeled AUs of all the training samples, the causal relation network between the AUs is obtained with the conditional probability matrix. The probability $p^e_{i|j}$ of the i-th AU conditioned on the j-th AU w.r.t. the e-th expression is defined as the product of the co-occurrence and co-absence probabilities as follows

$$p^e_{i|j} = op^e_{i|j} \cdot ap^e_{i|j}, \tag{5}$$

where the co-occurrence probability of i,j-th AUs conditioned on j-th AU is defined as the conditional probability

$$op^e_{i|j} = p(i \in AU^e | j \in AU^e), \tag{6}$$

where AU^e denotes all the action units of the e-th expression. And the co-absence probability of action units i,j conditioned on $j \notin AU^e$ is defined as the probability

$$ap^e_{i|j} = p(i \notin AU^e | j \notin AU^e), \tag{7}$$

Then the degree of causal relation of AU pair (i,j) is defined as follows

$$p^e_{i,j} = min(p^e_{i|j}, p^e_{j|i}). \tag{8}$$

The 'min' function of two conditional probabilities is adopted to avoid abnormal probability resulted from the imbalance of expressions related with each AU. For example (see Figure 3), AU_2 is related to a large number of expressions, which may result in significantly larger arrival transition probability.

For each expression, a relation probability matrix of dimension 20×20 is obtained with Equation (8). With the causal relation matrixes for all the expressions, the representative ability of AU for each expression can be found by simultaneously maximizing the representative ability for the considered expression and minimizing the representative abilities for the other expressions. That is, the representative ability of the AU pair (i,j) for the e-th expression is obtained as follows

$$RepAb^e_{i,j} = \begin{cases} 1 + \frac{1}{|\{l \neq e\}|}(p^e_{i,j} - \sum_{l \neq e} p^l_{i,j}) & \text{if } \{i,j\} \subset AU^e \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

where $|\{l \neq e\}|$ represents the number of elements in the set $\{l \neq e\}$. Finally, the representative ability of the i-th AU for the e-th expression is obtained as follows

$$RA^e_i = \begin{cases} \frac{1}{|\{j \neq i\}|} \sum_{j \neq i} RepAb^e_{i,j} & \text{if } i \in AU^e \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

For applying $\{RA^e_i, 1 \leq i \leq 20, 1 \leq e \leq 7\}$ for recognition, the maximum AU representative ability of each expression and face part is collected, which are denoted as RAP^e_i and presented as follows

$$RAP^e_i = \max_{j \in IA_i} RA^e_j \tag{11}$$

where the set IA_i denotes the AU indices corresponding to the i-th part. As the correspondence between AU and face part (PT), the maximum representative abilities of the parts are used to weigh the corresponding AUs, which are denoted as $W_{AU} = \{RAP^e_i, 1 \leq i \leq 7, 1 \leq e \leq 7\}$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

Due to the number limitation of AUs and the training samples, the representation space of AU is limited when they are simply organized. In order to expand the representation space of the AUs, the contribution of the patches contained in each AU is weighted by the following weight optimization model in this work.

2.3.2. Patch Weight Optimization

Based on the weights of AUs for each expression, a weight optimization model is proposed to weigh the patches of each AU for the considered expression triplet $G = \{e_1, e_2, e_3\}$.

The objective is to minimize a loss function with the weight sparseness and regularization constraints, which is presented as follows

$$\begin{cases} \min_w E + \lambda||w||_1 = \frac{1}{N}\sum_j \sum_{k\in ID_j} \log(\frac{1}{1+exp(-\alpha\cdot g(f_j,f_k,w))}) + \lambda||w||_1, \\ s.t. ||w_{PT_i}||_1 = RAP_i^G ||w||_1, i = 1,\cdots,7; ||w||_2 = 1. \end{cases} \tag{12}$$

where f_j, f_k are the features of the j-th and k-th training samples of the triplet G, whose patch feature is defined in Equation (2) and reduced to two dimensions using PCA and LDA [30]. $N = \sum_j |ID_j|$ and ID_j records all the training sample indices. Vector w records the weights of all the patches, w_{PT_i} records the weights of the patches related with the i-th part PT_i, parameter α is fixed to 1. RAP_i^G denotes the normalized representative abilities w.r.t. the considered expression triplet G, which is formulated as follows

$$RAP_i^G = \max_{1\le j\le 3} RAP_i^{e_j}, RAP_i^G \leftarrow \frac{RAP_i^G}{\sum_i RAP_i^G}. \tag{13}$$

The loss function $g(f_j, f_k, w)$ reflects the similarity loss of the feature vectors f_j, f_k with the weight vector w, which is constructed to minimize the intra-class variance and maximize the inter-class variance as follows

$$g(f_j, f_k, w) = \begin{cases} < f_j, f_k \cdot w > - < f_j, f_{j_0} \cdot w > & \text{if } \mathcal{L}(f_k) \neq \mathcal{L}(f_j) \\ < f_j, f_{j_0} \cdot w > - < f_j, f_k \cdot w > & \text{if } \mathcal{L}(f_k) = \mathcal{L}(f_j) \\ \text{with } j_0 = arg \max_{\{t:\mathcal{L}(f_t)=\mathcal{L}(f_j)\}} < f_j, f_t \cdot w > & \end{cases} \tag{14}$$

where $\mathcal{L}(f_k)$ is the expression label of the training feature vector f_k.

For solving the optimization problem (12), the gradient of the first term of the minimization objective function in Equation (12) is formulated as follows

$$\frac{\partial E}{\partial w} = \frac{\alpha}{N}\Sigma_j\Sigma_{k\in ID_j}\frac{e_{j,k}(w)-1}{e_{j,k}(w)}\cdot\frac{\partial g(f_j,f_k,w)}{\partial w}, \tag{15}$$

where $e_{j,k}(w) = 1 + exp(-\alpha\cdot g(f_j, f_k, w))$ and $\frac{\partial g(f_j,f_k,w)}{\partial w}$ is computed based on Equation (14) as follows

$$\frac{\partial g(f_j,f_k,w)}{\partial w} = \begin{cases} f_j\cdot f_k - f_j\cdot f_{j_0} & \text{if } \mathcal{L}(f_k) \neq \mathcal{L}(f_j) \\ f_j\cdot f_{j_0} - f_j\cdot f_k & \text{if } \mathcal{L}(f_k) = \mathcal{L}(f_j). \end{cases} \tag{16}$$

The sparseness term of model (12) adopts the L_1 norm, which is a special case of the L_1/L_2 mixed-norm employed in the work [45,55]. Thus, the optimization model (12) is solved with modified multi-task sparse learning algorithm employed in [45,55] with several differences presented as follows. The overall optimization algorithm is elaborated in Algorithm 2.

- The weight of each patch is initialized as a ratio of the corresponding AU representative ability as follows

$$w_{P_j,0} \leftarrow \frac{RAP_i^G}{|PT_i|}, w_0 \leftarrow \frac{w_0}{||w_0||_2}, \tag{17}$$

where i is the index of part including the j-th patch, $|PT_i|$ denotes the number of patches in the part PT_i, $w_{P_j,0}$ records the weights of the j-th patch P_j. The initialization procedure is presented in Step 3 of Algorithm 2;

- The weight vector w_{s+1} and auxiliary vector v_{s+1} in $s+1$-th iteration of Algorithm 2 are normalized to satisfy the constraint defined in Equation (12) as follows

$$w_{P_j,s+1} \leftarrow \frac{w_{P_j,s+1} RAP_i^G}{||w_{PT_i,s+1}||_1}, w_{s+1} \leftarrow \frac{w_{s+1}}{||w_{s+1}||_2}, \tag{18}$$

where i is the index of part including the j-th patch, w_{PT_i} denotes the weights of the part PT_i. The normalization is employed in Steps 11 and 19;
- Compared with [45], optimization model (12) is proposed by minimizing the feature similarity bias of different expression classes in Equation (14), which uses the information of mutual feature difference and contains more information than that of expression label matching in [45]. The corresponding objective function and the gradient vector are changed according to Equations (12) and (15), as revealed in Step 5.

Algorithm 2 The modified multi-task sparse optimization.

1: Obtain the feature vectors of the training and testing samples.
2: Initialization: the coefficients $\lambda = \eta = 1e^{-2}, S = 3e^1, \epsilon = 1e^{-10}, MaxNumF = 4$.
3: Initialize the weight vector w_0 as in Equation (17).
4: **for** $s = 0, \cdots, S$ **do**
5: Compute the objective value $FunV$ with Equation (12) and the gradient $\frac{\partial E}{\partial w_s}$ with Equation (15).
6: **if** $s \geq 2$ AND $FunV - pFunV \geq -\epsilon$ **then**
7: $NumF \leftarrow NumF + 1$.
8: **else**
9: $NumF \leftarrow 0$.
10: **end if**
11: Perform $w_{s+1} = v_s - \eta \frac{\partial E}{\partial w_s}$, normalize current weight vector w_{s+1} with Equation (18). Renew the i-th element of the weight vector w_{s+1} as follows.
12: **if** $|w_{i,s+1}| \geq \lambda\eta$ **then**
13: $w_{i,s+1} = (1 - \frac{\lambda\eta}{|w_{i,s+1}|})w_{i,s+1}$
14: **else**
15: $w_{i,s+1} = 0$
16: **end if**
17: $a_{s+1} = \frac{2}{s+3}, \delta_{s+1} = w_{s+1} - w_s, pFunV \leftarrow FunV$
18: $v_{s+1} = w_{s+1} + \frac{1-a_s}{a_s} a_{s+1}\delta_{s+1}$
19: Normalize weight vectors w_{s+1}, v_{s+1} with Equation (18).
20: **if** $||\delta_{s+1}||_2 \leq \epsilon$ OR $NumF > MaxNumF$ **then**
21: break
22: **end if**
23: **end for**

With the weight optimization model (12) and Algorithm 2, the weights of the patches for each expression triplet are obtained. The number of optimized weight vectors $\{W_i\}$ equals to $C_7^3 = 35$, i.e., the number of expression triplets.

2.3.3. Active AU Detection

Though each expression is related with several AUs, these AUs may not present at the same time. For example, while the AUs involved brows and eyes are present for the surprise expression shown in Figure 6e,j, the AU involved mouth was less active. In this case, error may occur if the features extracted from the AU involved mouth are included for expression recognition. To address this issue, we proposed a sparse representation based approach to identify the parts where the corresponding AUs are active for each testing sample before expression recognition.

For the k-th part of the i-th testing sample in Figure 3, the sparse representation is represented as follows

$$\min_{c^{(k)}} \frac{1}{2}||f_i^{(k)} - D^{(k)}c^{(k)}||_2^2 + \lambda||c^{(k)}||_1. \tag{19}$$

where $f_i^{(k)}$ is vectorized features of the i-th testing sample and $D^{(k)} = [f_{tr_1}^{(k)}, f_{tr_2}^{(k)}, \cdots, f_{tr_n}^{(k)}]$ are the patch features corresponding to the k-th face part (PT_k) of all the training samples of the candidate expression triplet and the neutral expression. Weight vector $c^{(k)}$ records the n-dimensional sparse representation coefficients, and λ is the regularization parameter set as $1e^{-3}$ in this work [56,57].

With the part-based sparse representation, the coefficients w.r.t. the AUs related with neutral and non-neutral expressions are obtained, where the AUs related with neutral expression (AU_{NE}) are $\{AU_1, AU_4, AU_8, AU_{13}, AU_{15}, AU_{17}, AU_{19}\}$ and the others are the AUs related with non-neutral expressions (AU_{EX}) as presented in Figure 3. More precisely, for the k-th part of the weight vector $c^{(k)}$, the corresponding weight components w.r.t. the AUs related with neutral and non-neutral expressions are obtained as follows

$$(c_{NE,j}^{(k)}, c_{EX,j}^{(k)}) = c_j^{(k)} \cdot (r_{NE,j}^{(k)}, r_{EX,j}^{(k)}). \tag{20}$$

where $r_{NE,j}^{(k)}, r_{EX,j}^{(k)}$ are the number ratios of the j-th feature element of the training samples w.r.t. the AUs of neutral and non-neutral expressions, respectively. That is, the weight vector $c^{(k)}$ is grouped into the sub-vectors $c_{NE}^{(k)}, c_{EX}^{(k)}$ with neutral and non-neutral expressions. Finally, the activeness of PT_k of the i-th testing sample is defined as follows

$$ActV_i^{(k)} = \frac{1}{n}\sum_{j=1}^{n} c_{EX,t_j}^{(k)} - c_{NE,t_j}^{(k)} \tag{21}$$

where t_j is the index of the weight with the the j-th largest value in the vector $c_{EX}^{(k)}$ or $c_{NE}^{(k)}$, $n = 10$ is the number of patches set to reduce the influence of abnormal weight components by sparse representation (19).

To judge whether PT_k or the corresponding AU is active or not, we treat each training sample f_i with non-neutral label as the testing sample and obtain its activeness value $TrActV_i^{(k)}$ of PT_k with Equation (21), where the part feature of $f_i^{(k)}$ is removed from the dictionary $D^{(k)}$ when obtaining the sparse coefficients $c^{(k)}$ in Equation (19). Finally, the considered part of testing image is decided to be active if $ActV_i^{(k)}$ is larger than the average of training activeness values as follows

$$fg_i^{(k)} = \begin{cases} 1 & \text{if } ActV_i^{(k)} \geq \frac{1}{n_{tr}}\sum_{j=1}^{n_{tr}} TrActV_j^{(k)} \\ 0 & otherwise. \end{cases} \tag{22}$$

where n_{tr} is the number of training samples with non-neutral labels. When the number of the selected active AUs for the i-th testing sample is less than two, which is likely to happen for the neutral expression sample, then the AUs with the top two largest activeness values $ActV_i^{(k)}$ presented in Equation (21) are determined to be active.

Figure 4 presents the sparse representation coefficients and the activeness values for an example surprise expression, where (a) and (b) present the sparse coefficients corresponding to the active 'Brow' and the non-active 'NoseRoot'. The training dictionary of the testing sample consists of 27 neutral expressions and 82 non-neutral expressions such as surprise, laugh and fear. Figures (a) and (b) show that the number of non-zero coefficients corresponding to non-neutral expression samples for active 'Brow' is significantly larger than that for non-active 'NoseRoot'. Figure 4c shows the activeness values of the seven parts of the same example, which clearly suggests that 'Brow' is the most active part and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

'NoseRoot' is the most non-active part. While 'Brow', 'Eye' and 'Forehead' are decided to be active and included for feature representation, non-active parts like 'Mouth', 'NoseRoot', 'Nasolabial' and 'Chin' regions will not involve in the following expression recognition.

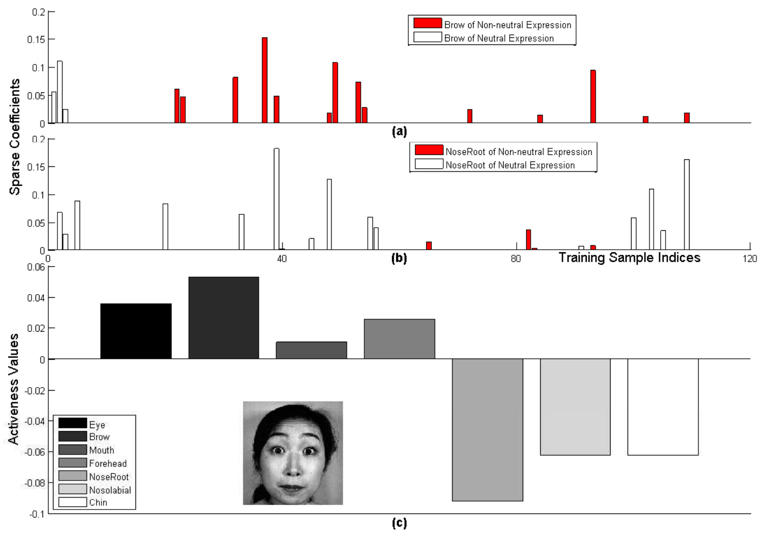

Figure 4. The sparse coefficients and the activeness values of an example surprise expression. (**a**,**b**) show the sparse representation coefficients of 'Brow' and 'NoseRoot'; (**c**) presents the example surprise expression and the activeness values of its seven parts.

After the active AU detection for each testing sample, the optimized patch weights for the corresponding candidate expression triplet with Algorithm 2 are used to weigh the selected active AUs (AU_A) and the involved patches for the following recognition.

2.4. Weighted SVM for Classification

After the feature weight optimization and AU activeness detection, support vector machine (SVM) with slightly modified kernel function is employed for the classification [58]. Rather than treat the feature weights as variables in SVM and obtain them with mutual information [59], the optimized feature weights learned in Section 2.3.2 are directly integrated with the patches involved with the detected active AUs in Section 2.3.3 for the recognition. That is, the new inner product $\langle f_i, f_j \rangle_w$ of two features f_i, f_j with weight vector w is defined as follows

$$\langle f_i, f_j \rangle_w = \langle f_i, w \cdot f_j \rangle. \tag{23}$$

where $\langle x, y \rangle = x^T y$ is the inner product of two vectors, $x \cdot y = (x_1 y_1, \cdots, x_n y_n)$ is the dot product of two vectors. Finally, with the new defined inner product (23) as the kernel function, SVM is used for the recognition.

3. Experimental Results

We perform the experiments using MATLAB 2014b on a PC with a 4GHZ core processor and 32 GB RAM. For the experimental testing, the Jaffe [60], Cohn-Kanade (CK+) [61] and SFEW2 [62] databases are employed for the performance and feature optimization study. Another three databases, i.e., TFEID [63], Yale-A database (YALE) [64] and EURECOM [65], are employed for the generalization testing. Among them, the database SFEW2 were collected in the wild and the faces were captured with un-controlled head poses and lighting conditions. Actually, the appearance of the same expression

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

is different from person to person, to guarantee that the image really represents a specific expression, these collected expressions are labelled by two independent labellers [62]. The remaining databases were videoed in the controlled lighting condition and the faces are all frontal, the corresponding participants were instructed by an experimenter to perform a series of facial displays for each expression [61].

The Jaffe database consists of 213 expression images of 10 Japanese female models, which can be categorized to six basic and the neutral expressions, i.e., angry (An), disgust (Di), fear (Fe), happy (Ha), sad (Sa) and surprise (Su). The CK+ database consists of 593 expression sequences from 123 subjects, where 327 sequences are labeled with one of seven expressions (angry, disgust, fear, happy, sad, surprise and contempt). Each sequence contains a set of captured frames when the subject changes his expression, 1033 expression images, i.e., the neutral and three non-neutral images sampled from each expression sequence are used for testing. The database SFEW2 is derived from the sub-challenge of static expression recognition in The Third Emotion Recognition in the Wild Challenge [62], which includes 958, 436 and 372 training, validation and testing samples of seven basic expressions. As the labels of testing set are not publicly available, the validation set was used in this paper for testing. The images were videoed in the un-controlled condition with different lightings, head poses, profiles, resolutions and face colors. Five landmark points were located with [50,66] for face alignment.

The Taiwanese Facial Expression Image Database (TFEID) database consists of 268 expression images from 40 subjects (20 females, 20 males), each of the subject presents six basic expressions and the neutral expression. The Yale expression database consists of 60 expression images from 15 subjects, each of the subject presents three basic expressions (happy, sad and surprise) and the neutral expression. The EURECOM Kinect Face Dataset (EURECOM) consists of 312 expression images from 52 subjects (14 females, 38 males), each subject presents two basic expressions (happy and surprise) and the neutral expression. The expression images are captured in two sessions at an interval of about two weeks. The six basic and the neutral examples of the six databases are demonstrated in Figure 5. For the following experiment, the person-independent strategy with ten-fold setting is employed for testing and comparison. More precisely, the considered database is divided into ten groups with approximately equal number of person IDs. While nine of them were used for training, the remaining group was used for testing. The process was randomly repeated ten times and the average accuracy is recorded as the final result.

Figure 5. Examples of six expression databases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

3.1. Number of Candidate Expressions Suggested by the First Stage Classifier

Take Jaffe database as example, some expressions in the dataset are quite difficult to discriminate, even for human eyes. For example, the angry and sad expressions shown in Figure 6a and d, f and g are very similar. It would be more plausible to develop a hierarchy system which could discriminate the easy categories at the first stage, and then differentiate the difficult categories at the second stage.

Figure 6. Expression examples with similar appearance from Jaffe database.

To decide the number of candidate expressions proposed by the first stage classifier, we show in Figure 7 the variation of accuracy with the value of k when top-k strategy is adopted for expression recognition. A classification is said to be correct if one of the top-k labels returned by the system matches the true label of the sample. The accuracy generally increases with the values of rank, k. While the accuracy of 91.5% was achieved for $k = 2$, the accuracy reached 96% for $k = 3$. To reach a trade-off between accuracy and efficiency, we set $k = 3$ for the first stage classification, i.e., the top three expression labels were assigned for the testing sample at the first stage. Based on the three candidates, the final label was given by a different model trained using finer features at the second stage.

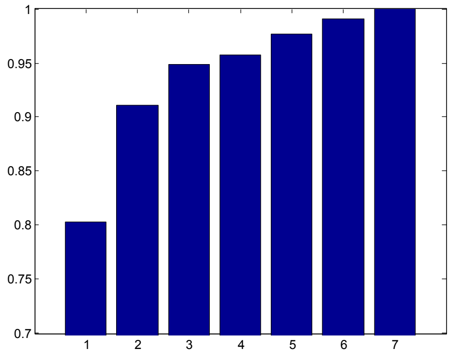

Figure 7. The variation of recognition rates with the value of ranks for Jaffe database.

3.2. Recognition Performance Analysis

To evaluate the effects of different models like AU weights, patch weight optimization and active AU detection, we tested performance of the recognition system with/without those models on Jaffe, CK+ and SFEW2 databases. For traditional one stage recognition, the features extracted from each patch were concatenated (see Equation (4)) and input to SVM for classification. The feature was further optimized using the proposed AU weight, patch weight optimization and active AU detection. The recognition performance of the system for different models are tabulated in Table 2. One can observe from the table that the proposed models significantly boost the performance. For example, when all of the three models were used, the recognition performance increased from 82.63% to 89.67%, from 89.06% to 94.09% and from 42.2% to 46.1%, for Jaffe, CK+ and SFEW2 datasets, respectively.

Table 2. The effects of different models to the overall recognition rate (%).

Database	Seven Expression	Triplet-Wise Based Two Stage Classification			
	Patch Feature	Patch Feature	Patch Feature+ AU Weight	Patch Feature+ AU Weight+ Patch Weight	Patch Feature+ AU Weight+ Patch Weight+ Active AU Detection
Jaffe	82.63	83.10	84.04	86.85	**89.67**
CK+	89.06	89.55	91.09	93.32	**94.09**
SFEW2	42.2	42.2	42.66	43.81	**46.1**

Figure 8 shows the top three most representative AUs of each expression, one can observe from the figure that the most representative parts of the surprise expression (g) are the brows, eyes and mouth. The most representative regions of the laugh expression (e) are the brow, mouth and nasolabial parts.

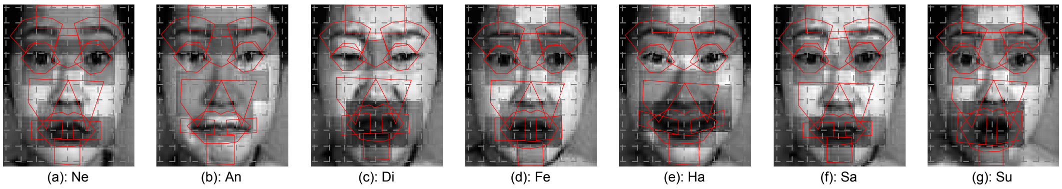

Figure 8. The top three most representative parts for each expression. The darker regions denote the larger representative abilities.

To analyze the performance of the patch weight optimization, Figure 9 depicts the optimized weight vectors of five expression triplets of the Jaffe database. It can be seen from the figure that the weights of the patches of each AU are further optimized. With the proposed weight optimization, the discrimination ability of the weighted patches for expressions with small variation is increased and performance improvements on the databases Jaffe and CK+ are observed in the fifth column of Table 2.

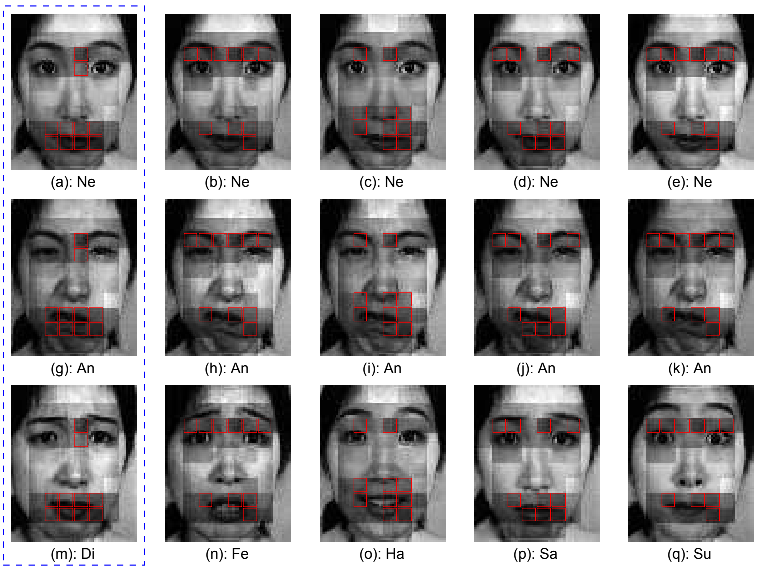

Figure 9. The visualization of optimized patch weights for five expression triplets. The darker the patch is, the larger is the weight. The patches with the top ten largest weights are labeled with red rectangles.

Due to the limited number of training samples, the weight optimization is not always beneficial to the recognition rate improvement. Figure 10 demonstrates the variation of the objective function

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

values in Equation (12) and the testing accuracy of an example expression triplet (angry, fear and sad) w.r.t. the number of iterations on the Jaffe database, when active AU detection was not applied. It can be seen that the recognition rate is not always increasing with the descendant of the objective function values due to the difference between the testing and training samples. Thus, active AUs could be detected to represent the specific features for each testing expression sample.

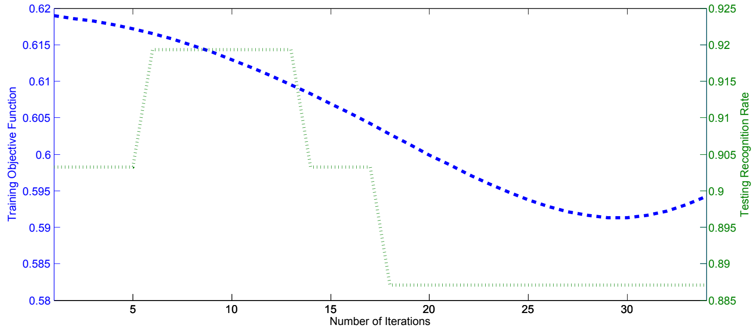

Figure 10. The evolutions of the objective function values and the testing recognition rates of expression triplet: fear, angry and sad w.r.t. the number of iterations.

To study the effect of the active AU detection for recognition, Figure 11 presents the top two active AUs of six example testing expressions, where (c) and (i) show that the brow and eye parts are more active than the other parts for the expression sample presented in Figure 6e. When these active parts are used for the feature encoding, the expression samples will be correctly recognized.

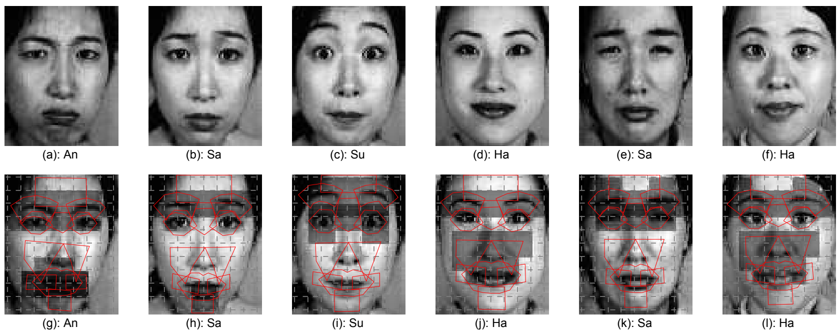

Figure 11. The top two active AUs of six example testing images.

To analyze the algorithm performance in overall, the confusion matrix of the final recognition results on the databases Jaffe and CK+ are presented in Tables 3 and 4, respectively. Both tables show that the angry, fear and sad expressions are relatively difficult to be correctly recognized. The difficulty is verified by expressions presented in Figure 6, where faces present similar features, not only in the appearance but also in the face part deformation. Table 4 suggests that the sad expression is mostly misclassified as the neutral expression (error rate 14.28%).

Table 3. Confusion matrix (%) of the proposed recognition algorithm on Jaffe database.

Expression	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**90**	3.33	0	0	0	0	6.67
An	0	**83.33**	6.67	0	0	10	0
Di	0	3.45	**93.1**	0	3.45	0	0
Fe	3.13	0	0	**87.5**	6.25	3.12	0
Ha	0	0	0	0	**100**	0	0
Sa	0	6.45	3.23	6.45	6.45	**77.42**	0
Su	0	0	0	0	3.33	0	**96.67**

Table 4. Confusion matrix (%) of the proposed recognition algorithm on CK+ database.

Expression	Ne	An	Di	Fe	Ha	Sa	Su
Ne	**94.34**	2.84	0	0	0.94	0.94	0.94
An	9.63	**85.93**	4.44	0	0	0	0
Di	0	0	**100**	0	0	0	0
Fe	2.67	0	0	**86.67**	5.33	1.33	4
Ha	0	0	0	0	**100**	0	0
Sa	14.28	1.2	0	0	0	**84.52**	0
Su	2.81	0	0	0	0	0	**97.19**

3.3. Feature Optimization Comparison

This section mainly compare the performance of the proposed weight optimization algorithm with other related algorithms, such as Adaboost [19,20,26,35], linear discriminant analysis (LDA) [30,43], Chi square statistic (CSS) [48], multi-task salient patch selection (MTSPS) [45] and the uniform weights (UWs) setting. For the Adaboost feature selection [35], the strong classifier of the final recognition is linearly composed of a number of patch-based weak classifiers. In the expression recognition [48], only the Chi square statistic for weight assignment is employed. In the feature selection [43], the patch saliency score is related with the classification accuracy of the training expression samples, where PCA and LDA are employed to reduce the feature dimension. The salient feature selection in [45] trains a set of active common and specific expression patches. The same strategy of the triplet-mode and the GSF feature is employed for a fair comparison. The recognition rates obtained by these algorithms on the databases Jaffe and CK+ are presented in Table 5.

Table 5. The recognition rates (%) of different feature selection algorithms on two databases.

Database	UWs (Uniform Weights)	Adaboost [35]	LDA [43]	CSS [48]	MTSPS [45]	Ours
Jaffe	83.10	81.22	82.63	83.57	85.45	**89.67**
CK+	89.55	88.58	90.71	90.22	92.45	**94.09**

Table 5 shows that the recognition rates of Adaboost, LDA are lower than that of the uniform weights UWs. CSS achieves slightly better performance than UWs on Jaffe database. In these models, the specificity of each expression and the causal relation information among AUs are not sufficiently exploited. To reduce the effects of personal ID information, the salient feature selection in [45] integrated the common and specific expression features and higher recognition rates are achieved.

Different from the other feature selection algorithms, the AU based feature optimization in the proposed algorithm weighs the AUs and the corresponding patches with conditional transient probability matrix. The discrimination information contained in both large-scale AUs and small-scale patches are considered. Moreover, active AUs of each testing expression sample are also detected for the feature encoding. The best recognition rates achieved in Table 5 justified the advantages of the proposed feature optimization.

3.4. Comparison with State of the Art

In this section, comparison of the overall recognition rates with a number of the state-of-the-art algorithms is conducted. To make the comparison fair, the competing algorithms were all tuned for the best performance. The comparison results on the databases Jaffe, CK+ and SFEW2 are demonstrated in Tables 6–8, respectively, where the algorithm description, the category, the number of subjects, testing protocol and the final recognition rates are considered.

Table 6. Comparison of different algorithms on Jaffe database.

Algorithm	Category	Subjects	Protocol	Recognition Rate (%)
Feature and Classifier Selection [42]	Traditional	10	10-fold	85.92
Radial Feature [22]	Traditional	10	10-fold	89.67
Supervised LLE [33]	Traditional	10	10-fold	86.75
Ours	Traditional	10	10-fold	89.67
Deep CNN [11]	Deep learning-based	10	10-fold	88.6
Deep Belief Network [41]	Deep learning-based	10	10-fold	**91.8**

Table 7. Comparison of different algorithms on CK+ database.

Algorithm	Category	Subjects	Protocol	Recognition Rate (%)
Maximum Margin Projection [32]	Traditional	100	5-fold	89.2
Feature Selection with GMM [67]	Traditional	97	10-fold	89.1
SVM (RBF) and Boosted-LBP [20]	Traditional	96	10-fold	91.4
Radial Feature [22]	Traditional	94	10-fold	91.51
Ours	Traditional	106	10-fold	**94.09**
AU Deep Network [47]	Deep learning-based	118	10-fold	92.05
Deep Neural Network [68]	Deep learning-based	106	5-fold	93.2

Table 8. The accuracy (%) of different algorithms on SFEW2 database.

Pyramid of Histogram of Gradients+Local Phase Quantisation+Non-linear SVM [62]	Hierarchical Committee CNN [12]	Multiple CNN [69]	Transfer Learning Based CNN [70]	Ours
35.93	**56.4**	56.19	48.5	46.1

For Jaffe database, our proposed algorithm achieves competitive recognition rate among all the algorithms in Table 6. The algorithm [41] using deep belief network yields the highest recognition rate of 91.8%. However, feature selection and classifier training are time-consuming and the process requires several days for each database. Rather than using a well-designed feature representation, the proposed algorithm achieves the best accuracy of 89.67% as the radial feature based algorithm [22] among the traditional algorithms. For the CK+ database, the proposed algorithm achieves the highest accuracy of 94.09%. As we are focused on seven class expression recognition, those work developed for six expressions like [21,37,41,43–45,49] are not included for comparison in this paper.

The feature and classifier adopted in the proposed algorithm are significantly different from the convolutional neural network (CNN)-based algorithms. In the following, the database (SFEW2) collected in the wild is taken to compare the overall performance between CNN and the proposed algorithms. As SFEW2 was used in Emotion Recognition in the Wild Challenge for performance evaluation, we directly take the accuracies of participants for comparison. All of the top three participants adopted CNN and their results are listed in Table 8, together with that of our approach.

While our approach achieves the top performance for CK+ database, CNN-based methods perform much better for the wild dataset, i.e., SFEW2. As CNN-based algorithms employ randomly cropped face regions for dataset augmentation, they are less sensitive to the face misalignment than the traditional algorithms. However, when large training dataset is not available and the images were mostly frontal faces, e.g., Jaffe and CK+, the traditional approaches could perform better than CNN based approaches. Furthermore, the network and parameters of CNN need to be finely tuned, which is much more time consuming than traditional algorithms.

Peer-reviewed version available at Sensors 2017, 17, 275; doi:10.3390/s17020275

3.5. Cross-Database Performance Study

To study the generalization ability of the proposed model, cross-database experiments are conducted and the corresponding recognition rates are presented in Table 9. In this testing, while one database is set as the training set, the other database is used as the testing set for evaluation.

Table 9. Comparison of cross-database recognition rates (%).

Algorithm	CK+ Training	Jaffe Training	CK+ and Jaffe Training		
	Jaffe Testing	CK+ Testing	TFEID	YALE	EURECOM
SVM and LBP [20]	41.3	-	-	-	-
Radial Feature [22]	**55.87**	**54.05**	61.94	60.66	-
LPP [40]	30.52	27.97	-	-	-
SR [36]	40.5	-	-	-	-
Ours	46.01	47.05	**78.73**	**63.33**	43.27

It can be seen from Table 9 that the radial feature encoding [22] with the probability projection achieves the highest accuracy when the databases Jaffe and CK+ are used for testing and training, respectively. The proposed algorithm achieves a competitive recognition rate of 46.01%, which is better than the recognition rate of 32.86% achieved by [22] when the employed probability projection is replaced with the Borda count strategy. When Jaffe is used as the training and CK+ is used for testing, the proposed algorithm also achieves competitive accuracy.

To further study the generalization ability of the proposed model, databases of CK+ and Jaffe are used as the training, while one of the other three databases is chosen for the testing. The accuracy is presented in the last three columns of Table 9, which shows that the proposed algorithm achieves much better recognition rate than the algorithm [22] on database TFEID and a competitive recognition rate on the database YALE.

4. Discussion and Conclusions

In this work, a two-stage expression recognition model based on triplet-wise feature optimization is proposed, the novelty of the this work is concentrated on three aspects. First, overall facial expression recognition is transformed into the triplet-wise mode to sufficiently exploit the specificity of each expression. Second, AU weighting and patch weight optimization are proposed for each expression triplet. Lastly, online detection of active AUs is proposed for each testing expression sample to reduce the influence of the non-active features in recognition. Experimental results and comparison with the related state-of-the-art algorithms verify the effectiveness and competitiveness of the proposed algorithm.

Although competitive results are obtained with the proposed model, there still leaves room for further improvement. First, feature optimization of more than two stages can be explored for the performance improvement. Second, more efficient features should be devised and integrated into the feature optimization model. Third, the cross-database recognition rates are still not high enough for the real application, which will be explored in our future work. Lastly, the ideas of AU weighting, feature sparseness optimization and active AU detection can be combined with CNN-based algorithms to improve the feature encoding based on face frontalization [71].

Acknowledgments: The authors thank the anonymous reviewers for their helpful comments and suggestions. The work was supported by Natural Science Foundation of China under grand No. 61672357, 61602315 and 61402289, the Science and Technology Innovation Commission of Shenzhen under Grant JCYJ20160422144110140, China Postdoctoral Science Foundation under grant No. 2015M572363, the Science Foundation of Guangdong Province under grant No. 2014A030313556 and 2014A030313558.

Author Contributions: Linlin Shen is the corresponding author who designed the algorithm and revised the paper. Weicheng Xie conceived, designed and performed the experiments, analyzed the data and wrote the paper.

Meng Yang and Zhihui Lai provided some suggestions and comments for the performance improvement of the recognition algorithm.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697.
2. Vezzetti, E.; Marcolin, F.; Fracastoro, G. 3D face recognition: An automatic strategy based on geometrical descriptors and landmarks. *Rob. Auton. Syst.* **2014**, *62*, 1768–1776.
3. Vezzetti, E.; Marcolin, F. 3D landmarking in multiexpression face analysis: A preliminary study on eyebrows and mouth. *Aesthet. Plast. Surg.* **2014**, *38*, 796–811.
4. Liu, J.; Zhang, Q.; Tang, C. Base mesh extraction for different 3D faces based on ellipse fitting. In Proceedings of the IEEE Advanced Information Technology, Electronic and Automation Control Conference, Chongqing, China, 19–20 December 2015; pp. 285–294.
5. Tian, Y.; Kanade, T.; Fcohn, J. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115.
6. Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*, 2nd ed.; Oxford Univ. Press: New York, NY, USA, 2005.
7. Valstar, M.; Pantic, M. Fully automatic facial action unit detection and temporal analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 17–22 June 2006; p. 149.
8. Elaiwat, S.; Bennamoun, M.; Boussaid, F. A spatio-temporal RBM-based model for facial expression recognition. *Pattern Recognit.* **2016**, *49*, 152–161.
9. Liu, M.; Shan, S.; Wang, R.; Chen, X. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1749–1756.
10. Rudovic, O.; Pavlovic, V.; Pantic, M. Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2634–2641.
11. Fasel, B. Multiscale facial expression recognition using convolutional neural networks. In Proceedings of the Third Indian Conference on Computer Vision, Graphics & Image Processing, Ahmadabad, India, 16–18 December 2002; p. 8123.
12. Kim, B.K.; Roh, J.; Dong, S.Y.; Lee, S.Y. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *J. Multimodal User Interfaces* **2016**, *10*, 173–189.
13. Burkert, P.; Trier, F.; Afzal, M.Z.; Dengel, A.; Liwicki, M. DeXpression: Deep convolutional neural network for expression recognition. *arXiv* **2016**, arXiv:1509.05371.
14. Pramerdorfer, C.; Kampel, M. Facial expression recognition using convolutional neural networks: State of the art. *arXiv* **2016**, arXiv:1612.02903.
15. Khorrami, P.; Paine, T.L.; Huang, T.S. Do deep neural networks learn facial action units when doing expression recognition? In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 11–12 December 2015; pp. 19–27.
16. Liu, B.; Wang, M.; Foroosh, H.; Tappen, M.; Penksy, M. Sparse convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 806–814.
17. Graham, B. Spatially-sparse convolutional neural networks. *Comput. Sci.* **2014**, *34*, 864–867.
18. Yan, K.; Chen, Y.; Zhang, D. Gabor surface feature for face recognition. In Proceedings of the First Asian Conference on Pattern Recognition, Beijing, China, 28–30 November 2011; pp. 288–292.
19. Whitehill, J.; Womlin, C. Haar features for FACS AU recognition. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 10–12 April 2006; pp. 97–101.
20. Shan, C.; Gong, S.; Wmcowan, P. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816.
21. Ouyang, Y.; Sang, N.; Huang, R. Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. *Neurocomputing* **2015**, *149*, 71–78.

22. Gu, W.; Xiang, C.; Vvenkatesh, Y.; Huang, D.; Lin, H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognit.* **2012**, *45*, 80–91.
23. Ghimire, D.; Lee, J. Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. *Sensors* **2013**, *13*, 7714–7734.
24. Kotsia, I.; Zafeiriou, S.; Pitas, I. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognit.* **2008**, *41*, 833–851.
25. Xie, W.; Shen, L.; Jiang, J. A novel transient wrinkle detection algorithm and its application for expression synthesis. *IEEE Trans. Multimedia* **2016**, doi:10.1109/TMM.2016.2614429.
26. Bartlett, M.; Littlewort, G.; Bartonfrank, M.; Lainscsek, C.; Fasel, I.; Rmovellan, J. Recognizing facial expression: Machine learning and application to spontaneous behavior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 568–573.
27. Hesse, N.; Gehrig, T.; Gao, H.; Kemalekenel, H. Multi-view facial expression recognition using local appearance features. In Proceedings of the International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012; pp. 3533–3536.
28. Sha, T.; Song, M.; Bu, J.; Chen, C.; Tao, D. Feature level analysis for 3D facial expression recognition. *Neurocomputing* **2011**, *74*, 2135–2141.
29. Liu, C.; Wechsler, H. Independent component analysis of Gabor features for face recognition. *IEEE Trans. Neural Netw.* **2003**, *14*, 919–928.
30. Jlyons, M.; Budynek, J.; Akamatsu, S. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1357–1362.
31. Lai, Z.; Yong, X.; Jian, Y.; Shen, L.; Zhang, D. Rotational invariant dimensionality reduction algorithms. *IEEE Trans. Cybern.* **2016**, doi:10.1109/TCYB.2016.2578642.
32. Nikitidis, S.; Tefas, A.; Pitas, I. Maximum margin projection subspace learning for visual data analysis. *IEEE Trans. Image Process.* **2014**, *23*, 4413–4425.
33. Liang, D.; Yang, J.; Zheng, Z.; Chang, Y. A facial expression recognition system based on supervised locally linear embedding. *Pattern Recognit. Lett.* **2005**, *26*, 2374–2389.
34. Zhang, Y.; Zhang, L.; Ahossain, M. Adaptive 3D facial action intensity estimation and emotion recognition. *Expert Syst. Appl.* **2015**, *42*, 1446–1464.
35. Silapachote, P.; Rkaruppiah, D.; Rhanson, A. Feature selection using Adaboost for face expression recognition. In Proceedings of the the Fourth IASTED International Conference on Visualization, Imaging, and Image Processing, Marbella, Spain, 6–8 September 2004; pp. 84–89.
36. Jia, Q.; Gao, X.; Guo, H.; Luo, Z.; Wang, Y. Multi-layer sparse representation for weighted LBP-patches based facial expression recognition. *Sensors* **2015**, *15*, 6719–6739.
37. Zafeiriou, S.; Pitas, I. Discriminant graph structures for facial expression recognition. *IEEE Trans. Multimedia* **2008**, *10*, 1528–1540.
38. Giladbachrach, R.; Navot, A.; Tishby, N. Margin based feature selection-theory and algorithms. In Proceedings of the International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; pp. 43–50.
39. Jia, S.; Hu, J.; Xie, Y.; Shen, L. Gabor cube selection based multitask joint sparse representation for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1–14.
40. Wang, L.; Wang, K.; Li, R. Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition. *IET Comput. Vis.* **2015**, *9*, 655–662.
41. Liu, P.; Han, S.; Meng, Z.; Tong, Y. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1805–1812.
42. Kyperountas, M.; Tefas, A.; Pitas, I. Salient feature and reliable classifier selection for facial expression classification. *Pattern Recognit.* **2010**, *43*, 972–986.
43. Happy, S.L.; Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **2015**, *6*, 1–12.
44. Liu, P.; Tianyizhou, J.; Wtsang, I.; Meng, Z.; Han, S.; Tong, Y. Feature disentangling machine-A novel approach of feature selection and disentangling in facial expression analysis. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 151–166.
45. Zhong, L.; Liu, Q.; Yang, P.; Huang, J.; Metaxas, D. Learning multiscale active facial patches for expression analysis. *IEEE Trans. Cybern.* **2015**, *45*, 2562–2569.
46. Tong, Y.; Liao, W.; Ji, Q. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1683–1699.

47. Liu, M.Y.; Li, S.X.; Shan, S.G.; Chen, X.L. AU-aware deep network for facial expression recognition. In Proceedings of the Tenth IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–6.
48. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikainen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619.
49. Li, Y.; Wang, S.; Zhao, Y.; Ji, Q. Simultaneous facial feature tracking and facial expression recognition. *IEEE Trans. Image Process.* **2013**, *22*, 2559–2573.
50. Sun, Y.; Wang, X.; Tang, X. Deep convolutional network cascade for facial point detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , Portland, OR, USA, 23–28 June 2013; pp. 3476–3483.
51. Tzimiropoulos, G. Project-out cascaded regression with an application to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3659–3667.
52. Tan, X.; Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **2010**, *19*, 1635–1650.
53. Ekman, P.; Friesen, W.V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: San Francisco, CA, USA, 1978.
54. Shen, L.; Bai, L. A review on Gabor wavelets for face recognition. *Pattern Anal. Appl.* **2006**, *9*, 273–292.
55. Chen, X.; Pan, W.; Kwok, J.T.; Carbonell, J.G. Accelerated gradient method for multi-task sparse learning problem. In Proceedings of the IEEE International Conference on Data Mining, Miami, FL, USA, 6–9 December 2009; pp. 746–751.
56. Lee, H.; Battle, A.; Raina, R.; Ng, A.Y. Efficient sparse coding algorithms. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 4–7 December 2006; pp. 801–808.
57. Yang, M.; Zhang, L.; Yang, J.; Zhang, D. Robust sparse coding for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 625–632.
58. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 389–396.
59. Xing, H.J.; Ha, M.H.; Tian, D.Z.; Hu, B.G. A novel support vector machine with its features weighted by mutual information. In Proceedings of the International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 315–320.
60. Jlyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wavelets. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
61. Kanade, T.; Fcohn, J.; Tian, Y. Comprehensive database for facial expression analysis. In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 26–30 March 2000; p. 46.
62. Dhall, A.; Ramana Murthy, O.V.; Goecke, R.; Joshi, J.; Gedeon, T. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In Proceedings of the ACM International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 423–426.
63. Chen, L.F.; Yen, Y.S. Brain Mapping Laboratory, Institute of Brain Science. Ph.D. Thesis, National Yang-Ming University, Taipei, Taiwan, 2007.
64. Nbelhumeur, P.; Phespanha, J.; Jkriegman, D. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 711–720.
65. Min, R.; Kose, N.; Dugelay, J. KinectFaceDB: A Kinect database for face recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2014**, *44*, 1534–1548.
66. Zhang, J.; Shan, S.; Kan, M.; Chen, X. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 1–16.
67. Elguebaly, T.; Bouguila, N. Simultaneous high-dimensional clustering and feature selection using asymmetric Gaussian mixture models. *Image Vis. Comput.* **2015**, *34*, 27–41.
68. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–9 March 2016; pp. 1–10.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 January 2017 doi:10.20944/preprints201701.0120.v1

Peer-reviewed version available at *Sensors* **2017**, *17*, 275; doi:10.3390/s17020275

69. Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the ACM International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 435–442.
70. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the ACM International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
71. Hassner, T.; Harel, S.; Paz, E.; Enbar, R. Effective face frontalization in unconstrained images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4295–4304.