Algorithm Modeling of Smile Intensity Estimation Using a Spatial Attention Convolutional Neural Network

Zhenzhen Luo¹, Zongyu Sha¹, Chen Jin² and Ning Xu³*

¹School of Software, Jiangxi Normal University, Jiangxi, China
²School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
³School of Information Engineering, Wuhan University of Technology, Wuhan, China

*Corresponding author email: xuning@whut.edu.cn

Abstract. Smile intensity estimation is a challenging task as it required subtle feature extraction, self-adapted weighted model and classifier, complexity of the problem domains, and problems on fine-grained image recognition are some of the issues related to intensity estimation. In this study, we designed a self-weighted deep convolutional neural network architecture for smiles intensity estimation using graphics processing unit. In the case of using only CK+ smile images, the accuracy of the model is also higher than that of the latest technology. Our model achieved better accuracy by just using CK+ smile images than state-of-the-art techniques. Visualizations of learned features at various layers and their deconvolutions are also presented for understanding the learning process.

Keyword: Smile intensity; Spatial attention mechanism; Deep learning.

1. Introduction

Deep learning is to learn the internal laws and representation levels of sample data through a deep architecture with multiple processing layers. The processing layer consists of linear or non-linear transformations. At the same time, deep learning uses automatic feature learning and hierarchical feature extraction to replace handcrafted features [1][2]. Convolutional neural networks (CNN) is one of the tools for modeling sparse correlation, and has the characteristics of translation invariance, suitable for image recognition research work [2][3]. This study uses deep CNN architecture to learn features to estimate smile intensity.

1.1. Motivation

As one of the most common facial expressions, smile can indicate one’s psychological state, express one’s sentiment and intention. To date, smile has been applied to every whiff of human’s life, including users’ experience perception [4], students’ mentality [5], photos enhance processing [6], smile shutter, etc. Better human-computer interfaces and affective understanding can be developed by exploring smile intensity. In recent years, scholars have conducted many researches related to human’s smile recognition [7]. As for facial expression analysis, some researchers have made great headway towards smile facial recognition, however, the estimation of smile intensity is still hard to elucidate due to subtle difference between smile intensity level [9].

In this study, we used the Extended Cohn – Kanade (CK+) [13] database for smile intensity categories. It is a challenge for our researchers to evaluate the smile intensity by the image content, as the smile intensity is founded on the muscle deformation and texture features. There are techniques that utilize
facial imagery for estimating facial expression intensity using Paul Ekman's Facial Action Coding System (FACS) and no-FACS techniques. In Facial Action Coding System (FACS), there are 44 facial action units, in which AU6, AU12, AU25 come up with complete smile expression. The Action Units (AUs) are in line with the contraction of specific facial muscles from FACS and their intensities vary [9]. Use a 4-level scale range of intensity 1-4 (or neutral to apex), where 1 (neutral) is the minimum intensity and 4 (apex) is the maximum intensity. Figure 1 gives an demonstration on a 4-level smile intensity.

AUs have a clear facial expression differentiation and definition to facilitate the researches which make use of the psychology research results choosing efficient AUs for different expression recognition. However, this technique has flaws and still suffers from subjective bias. Due to inherent limitations of the empirical method used, the method based on AUs for smile detection is deeply depending on AUs location and motion features computing.

The method based on non-AUs is not longer to analysis face action units. Non-AUs is the method which, after face registration, facial features are straightly extracted from image, and combined with machine learning to establish mapping from image to the smile expressions. For example, By selected wavelets and SIFT [14] descriptors in regions around the landmarks of the whole face (about 60) as appearance features, Girard et al. made use of Support Vector Regression (SVR), binary Support Vector Machines (SVMs) classifiers to estimate smile intensity. However, utilizing machine learning to extract features, conduct significant pre-processing, and classifier training requires domain knowledge. It deeply depends on human intervention for analysis, and cannot effectively combine feature learning and classification into a single model.

In this study, we focus on obtaining higher accuracy by combining facial landmark and spatial attention mechanism into a single deep convolution neutral network.

1.2. Our Contribution
According to the task of smile intensity estimation, combined with the characteristics of small class differences and fuzzy sub class boundaries, we introduce the attention mechanism in fine-grained image recognition to suppress the invalid information learned by convolution layer and enhance the learning of important area features by the network, especially for the feature learning of the area where the facial muscles producing smile face expression are located, and combines the feature point information to approach this book The goal of this paper. In this paper, we use focal loss function to replace the common cross entropy loss, reduce the adverse effect of uneven sample distribution, strengthen the learning of difficult samples, and improve the accuracy of smile intensity estimation.

This research is categorized into the following sections. Section 2 explains some information concerning deep CNNs. Section 3 introduces our layers and architecture of the deep CNN. The experimental results, visualization and performance are showed in Section 4.

2. Deep Convolutional Neural Networks with Spatial Attention Mechanism
CNN is a variant of multilayer perceptron. A convolutional network is a different combination of convolutional layers and fully connected layers. The deep neural network is composed of feed forward
network and back propagation. The former is designed for network and the latter is trained. Errors are propagated back in the network. We can optimize the network by updating the weights and deviations to reduce the loss function. At the same time, in order to improve computational performance and solve the problem of over fitting, the free parameters to be learned can be reduced by sharing parameters (weights and biases) of sparse connections.

Similar to the human attention mechanism, deep learning can also improve the attention of specific areas in the process of training. The strength of smile is mainly determined by the exercise intensity of facial muscle unit, all of these actions take place in the special area of the face. AU6 shows the action of narrowing eyes when smiling face, focusing on the area of eyebrows and eyes; AU12 shows the stretching of the corner muscles of the mouth towards the ear socket, focusing on the mouth; AU25 shows the downward movement of the chin, focusing on the jaw. Attention mechanism often depends on color channel information or spatial scale information. In this paper, the AUs intensity and color information are almost irrelevant in the smile intensity evaluation. Therefore, the spatial attention mechanism is introduced to learn the regional weight and highlight the features that are beneficial to classification.

3. Overall Evaluation Model and Process

The process of the smile intensity assessment is shown in Figure 2. Details of the model are tabulated in Table 1.

![Figure 2. The process of the smile intensity assessment.](image)

| Layer          | Output Shape | Layer          | Output Shape |
|----------------|--------------|----------------|--------------|
| Input          | (270,360,3)  | Reshape        | (270,360,1)  |
| VGG19          | (8,11,512)   | VGG16          | (2,2,512)    |
| BN_1           | (8,11,512)   | BN_2           | (2,2,512)    |
| CONV_1         | (8,11,64)    | CONV_2         | (2,2,64)     |
| CONV_3         | (8,11,16)    | CONV_4         | (2,2,16)     |
| ATTENTION_1    | (8,11,12)    | ATTENTION_2    | (2,2,1)      |
| CONV_5         | (8,11,512)   |               |              |
| MULTIPLY_1     | (8,11,512)   | MULTIPLY_2     | (2,2,512)    |
| AVG_POOLING_1  | 512          | AVG_POOLING_2  | 512          |
| DENSE_1        | 512          | DENSE_2        | 512          |
| CONCATENATE    | 1024         |                |              |
| DENSE_LAST     | 4            |                |              |

Table 1. Configuration of our network.

We designed an attention-VGGNet using a scale spatial attention mechanism. First, single-channel input of size 48×48 is transformed a 3-channel map of size 270×360 with opencv tools. The purpose is to prevent the feature maps calculated by convolution and pooling from being too small in size and losing characterization ability, as well as to adapt to the channel of the network. Then, we can clearly
see that both VGGNet16 and VGGNet19 are used in the model and learn characteristics separately. To understand the distinction, the VGGNet16 integrates the pre-training weight of ImageNet 16 million pictures with a rich feature weight compensates for the fewer number of data sets, and the VGGNet19 is focusing on training samples of smile strength without pre-trained weight. In this structure, the weight information of the attention area is obtained by adding the two learned features. Finally, it provides a fully connected layer output to evaluate smile intensity.

\[
\left( S_{x}, S_{y} \right)
\]

\[
\left( S^{TR}_{x}, S^{TR}_{y} \right)
\]

\[
\left( S_{x}, S_{y} \right)
\]

\[
\left( S^{BL}_{x}, S^{BL}_{y} \right)
\]

\[ S_L \]

**Figure 3.** Attention area search box.

### 3.1. Overview of Layers

We first optimized the convolutional layer. Equation 1 explains the output \( x_{ij}^1 \) by pixel lies in position of \((i, j)\), using a filter of size \( K \) multiply \( K \), and Equation 2 explains the position where \( x_{ij}^2 \) output by pixel of different convolution kernels is located at \((i, j)\). The final pixel output \( x_{ij}^f \) is the average of the sum of the \( x_{ij}^1 \) and the \( x_{ij}^2 \). Meanwhile, \( w_{ab} \) and \( w_{pq} \) are the weights of their pixels, as a mapping result \( y_{(i+a), (j+b)}^{l-1} \) and \( y_{(i+p), (j+q)}^{l-1} \) of receptive filed at position \((i + a), ( j + b) \) or \((i + p), (j + q) \) from last nearest layer.

\[
x_{ij}^1 = \sum_{a=0}^{K-1} \sum_{b=0}^{K-1} w_{ab} y_{(i+a), (j+b)}^{l-1} + B^{l_1}
\]  

(1)

\[
x_{ij}^2 = \sum_{a=0}^{K-1} \sum_{b=0}^{K-1} w_{pq} y_{(i+p), (j+q)}^{l-1} + B^{l_2}
\]  

(2)

\[
x_{ij}^f = \frac{x_{ij}^1 + x_{ij}^2}{2}
\]  

(3)

We use the features extracted by the convolution of different structures to fuse and found that while enhancing the generalization ability of the model, it also enhances the characterization ability of the receiving field of different sizes, improves the robustness of the model, and help ease over-fitting. Translation invariance comes from pooling, down-sampling can help decrease the amount of parameters, refraining from overfitting. The deep CNN depicts a situation where max pooling layer comes after the first, second, third and fifth convolutional layers. The local response normalization (LRN) are applied throughout channels, and the normalization of local input regions enables us to attain lateral inhibition. Normalization is done by dividing each input by:

\[
(1 + \frac{a}{n} \sum_{i} x_{i}^2)^{\beta}
\]  

(4)
where \( n \) is the size of local region and \( \alpha \) and \( \beta \) are basic parameters. By doing so, we normalize the units, which are in the same position but from various channels. The last layer in Figure 3 consists of 4 units (equal to the number of distinct classes). Multinomial logistic loss is computed through the usage of focal softmax loss layer. In this way, the weight parameters that minimizes the loss function can be updated and a single class out of \( K \) mutually exclusive classes is determined as follows.

\[
FL = -\frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{\exp f_{y_i}}{\sum_{k=1}^{K} \exp f_k} \right)
\]

(5)

Softmax takes advantage of the output of connected layer which leads to the probability distribution of each class. The paper proposed sum up total weighted loss over the network, in which loss and gradient computed which forward and backward respectively. Through stochastic gradient descent (SGD) technique, can determine the minimum cost iteratively[19].

\[
L(W) \approx \frac{1}{N} \sum_{i=1}^{N} f_{y_i}(X^i) + \lambda r(W)
\]

(6)

where \( L(W) \) is the stochastic approximation of objective, \( f_{y_i}(X^i) \) is the loss on data instance \( X^i \), \( r(W) \) is the regularization term, and \( \lambda \) is the weight decay for the regularization term.

3.2. Attention Mechanism

Similar to the human attention mechanism, deep learning can also improve the "attention" of specific areas in the process of training. The intensity of smile is mainly determined by the exercise intensity of facial muscle unit (AU6, AU12, AU25). All of these actions take place in the special area of the face. AU6 shows the action of narrowing eyes when smiling face, focusing on the area of eyebrows and eyes; AU12 shows the stretching of the corner muscles of the mouth towards the ear socket, focusing on the mouth; AU25 shows the downward movement of the chin, focusing on the jaw. Attention mechanism often depends on color channel information or spatial scale information. In this paper, the Au unit motion intensity and color information are almost irrelevant in the smiley face intensity evaluation. Therefore, the spatial scale attention mechanism is introduced to learn the regional weight, highlighting the characteristics that are beneficial to classification. In this paper, spatial (based on spatial scale) attention mechanism is introduced to enhance the effect of smiley face intensity evaluation. The principle of this mechanism is explained as follows:

1) For the input smiley face picture \( X \), the feature of convolution operation is recorded as \( W^*X \), and its probability distribution is \( p(X) = \tilde{f}(W^*X) \). The function \( f \) represents the function of network output normalized by softmax.

2) In this paper, a candidate box is set up to represent the attention area. The candidate box is represented by \( (S_x, S_y, S_L) \) It represents the horizontal and vertical coordinates of the center of the candidate box and the distance between the center and the frame.

\[
S_x^{TR} = S_x + S_L, S_y^{TR} = S_y + S_L
\]

(9)

\[
S_x^{BL} = S_x - S_L, S_y^{BL} = S_y - S_L
\]

3.3. Optimization

1) The attention mechanism is to add the weight distribution to the common feature distribution, and multiply the original image by the matrix point generated by the hard coding or logistic regression function of 0,1 to highlight the local feature and generate the attention area, in which \( \theta \in (0,1) \).

\[
X^{atten} = X \bullet M(s_x, s_y, s_L)
\]

(10)
\[ M(s_x, s_y, s_L) = \left[ \theta(X + S^{TR}_X) - \theta(X - S^{BL}_X) \right] \bullet \left[ \theta(Y + S^{TR}_Y) - \theta(Y - S^{BL}_Y) \right] \]

(11)

\[ \frac{\partial M(s_x, s_y, s_L)}{s_x} < 0 \text{ if } (s_x, s_y) \rightarrow (s_x^{TR}, s_y^{TR}) \text{ or } (s_x, s_y) \rightarrow (s_x^{BL}, s_y^{BL}) \]

(12)

2) The loss function is composed of the cross entropy function to distinguish the smiley face region and a function of the sorting part. The smaller the loss value of the sorting function is, the greater the confidence degree of the later candidate box is than that of the first candidate box with the change of the candidate box. K represents the category. When the network tends to converge, the candidate frame becomes smaller and smaller, and the extracted local information is more refined.

3) Hyperparameters: As for larger input, larger convolution filter size is adopted. With layers going deeper, the filter size is reduced gradually. Comparing to higher layers, Layers near the input have fewer filters, which strengthen the ability to subtle recognition. To simplify parameters and lessen burden of the network, find a balance between feature representation and parameter amount. We design the structure as table1 shows.

4) Regularizations: Use a model with 90% verification accuracy in the experiment for early stopping. Sometimes, early stopping may cause underfitting. Overfitting can be eased with dropout method and also its performance can be improved.

4. Experiments
In this section, we explain our dataset, training and testing, visualization of features, and performance analysis.

4.1. Dataset
The Carnegie Mellon University's CK+ facial expressions database is a popular database for the study of facial expressions. The CK+ database obtained 593 video sequences by photographing the facial expressions of 123 subjects. The sequences ranged in duration from 10 to 60 frames, and each
sequence recorded a transition from neutral to peak expression; in the videos, a total of 327 sequences from 118 subjects were labeled with seven basic Facial expression labels (anger, contempt, disgust, fear, happiness, sadness and surprise) according to FACS. 394 Action units (AUs) are the basic units of Facial muscles that make up Facial expressions, defined by psychologists based on the Facial Action Coding System (FACS). All of the Smiles had three action units, AU6, AU12 and AU25. The intensity of the muscle action units ranged from low to high on a scale of 1,2,3,4,5, with 0 representing the missing information.

1) Images in Dataset: This paper makes a detailed investigation on the original data set of CK and finds that among the 593 basic expression sequences, 82 expression sequences simultaneously exist in AU6 and AU12 units, 93 expression sequences exist in AU12 and AU25 units, and 71 expression sequences simultaneously exist in the above three action units. Figure 4 is a sample picture containing three units at the same time. The figure 5 shows its AUs intensity labeling information. The expression includes AU6, AU12, AU16, and AU25 muscle action units.

2) Pre-Processing: The unrelated part of the image are cut off. In order to prevent information missing as down sampling, we resize image 270*360 with interpolation of inter cubic.

4.2. Training and Testing
The data set of 1790 images is divided into three small data sets, which are training, test and validation set as shown in Table 2. The transformed image and its original image are stored in the same data set. For example, if the original image belongs to the training set, the corresponding transformed image will also appear in the training set. The rarer situation is due to the segmentation ratio, the transformed image is separated from the original image, but these situations will not have much impact on training and accuracy.

| Intensities | Train | Validation | Test | Total |
|------------|-------|------------|------|-------|
| neutral    | 343   | 100        | 50   | 493   |
| onset      | 311   | 100        | 50   | 461   |
| boost      | 265   | 67         | 33   | 365   |
| apex       | 321   | 100        | 50   | 471   |

Figure 6. Re-weighted area with attention mechanism.

Figure 6 shows after introducing the attention mechanism, our network learns the key area of smiling face in the picture. Lines one through four show the key areas of the Smile that the attention network learned. It can be seen that the attention network trained in this paper can locate the key action area of AU6, AU12, AU25, and the attention area of the image increases as the intensity of the smiling face.
increases, and the proportion of the total area of the image increases, and the attention area moved down, because the AU25 Chin area moved down more strongly, as a result of the larger facial movements, but also added more noise, which affected the network's assessment of the Smile peak frames. Of the 16 CK+ image, 9 attentional areas included AU6 action areas, 11 attentional areas included AU12 action areas, and 7 included AU25 action areas. Based on the results of attention mechanism extraction, adding attention mechanism is beneficial to the task of smile intensity recognition.

![Validation accuracy](image1.png)

![Validation loss](image2.png)

**Figure 7.** Validation accuracy, validation loss for our CNN with attention mechanism.

**Figure 8.** Validation accuracy, validation loss for our CNN with attention mechanism (4 or 3 - class).

Figure 7 illustrates the validation graph trends during the iteration of training process. The model learns better with the advance of epochs. Through gradually improved accuracy and the reduced loss, we can observe this after each epoch. As the number of epochs increases, the slope of the accuracy curve approaches 0, which means that training can be stopped and the model has converged to the best. In this experiment, the gradual learning rate is reduced by 10%.

As above shows, the model we proposed performs 64% of accuracy in testing sets. We have figured out every image's top (top-1) classification in the dataset and used the probabilities from softmax function in classification. The smile class with the highest probability is allocated to the category.

**Figure 8.** Validation accuracy, validation loss for our CNN with attention mechanism (4 or 3 - class).

4.3. **Visualization**

Figure 9 visualizes the learning result expressed by feature map by each layer with same input. Input image as well as the feature maps from the first convolution, normalization and pooling are displayed in sequence. Each filter produces a different map, and last layer output is next layer input. It is less...
difficult to explain activated images from the first convolution, information becomes abstract when layers increase.

![Feature maps generated from first convolutional layer of our network.](image)

**Figure 9.** Feature maps generated from first convolutional layer of our network.

### 4.4. Performance Analysis

The results of this experiment are showed in this section. We also compare our model’s performance with the performance of antecedent researches for intensity estimation shown in table 3. The results of our experiments on CK dataset for attention VGG and work of [15] [16] [17] denoted as HCRF, HCORF, VSL-CRF and JSV.

1) Models: According to the validation dataset, its hyperparameters can be modulated as long as there is a relatively acceptable architecture. In this experiment, the architecture of different hyperparameters does not always get a consistent response.

2) Confusion Matrix: Figure 5 shows the confusion matrix of smile classifications for CK+ datasets. The numbers along the diagonal just indicate how many images for any category are correctly classified. The figure 5 describes that smiles are more likely to be misclassified, that is, smiles are more likely to be classified into categories with smaller intensity differences than those with larger intensity differences.

| ID          | Top-1 Accuracy | F1-Score | Computation time(s) |
|-------------|----------------|----------|---------------------|
| HCRF        | 37.8           | 34.2     | 0.5833              |
| HCORF       | 15.2           | 12.3     | 0.6072              |
| VSL-CRF     | 29.9           | 28.5     | 1.8285              |
| JWV         | 82.3           | 81.8     | 0.0977              |
| Attention-VGG | 78.0     | 78.49    | 0.5332              |
| Attention-VGG – with feature points | **84.4** | **84.39** | 0.5466 |

### Table 3. Compare the results of different network configurations. The best results are highlighted in bold.

### 4.5. Discussion

1) this method can evaluate the intensity of Smile face more precisely. By introducing attention mechanism, experiments were carried out to visualize the process of deepening the information of smiling face's region weight. The results of comparative experiments show that the attention mechanism enhances the network's assessment of the intensity of the smiling face during the transition period, in which the recognition rate of the frame is increased from 55% to 59% , and the recognition
rate of the enhanced frame is increased from 5% to 40%, however, the recognition rate of apex frame is decreased.

2) the relationship between the distance of facial feature points and the intensity of Smile expression is verified by experiments, and the important feature points are selected to guide the network learning. When using face feature points, the accuracy and F1-score of the proposed method are better than those of other methods in the literature, and the overall recognition rate is 84.4%, apex frame recognition rate increased to 83.3%.

3) the proposed method can evaluate the intensity of smiling face with detecting the feature points, and the recognition rate better than JWV, which reaches 84.4%. The premise of JWV algorithm is that it can locate the feature points accurately, in contrast, the deep learning method in this paper is more robust.

4) the recognition speed of this method is faster than most of the algorithms listed, and the evaluation speed is up to two frames per second, but it has no advantage over JWV.

Acknowledgments
This work is supported by National Natural Science Foundation of China (No.61967010, 41661083), the Provincial Key Research and Development Program of Jiangxi under Grant (No.20181ACE50030).

References
[1] Zhang Kaipeng, Tan Lianzhi, Li Zhifeng. Gender and Smile Classification Using Deep Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 26 June-1 July 2016.
[2] Ivanovsky L, Khryaschev V, Lebedev A. Facial expression recognition algorithm based on Deep Convolutional Neural Network. 21st Conference of Open Innovations Association (FRUCT), 2017.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. in Advances in Neural Information Processing Systems, vol. 25, no.2, 2012, pp. 1097–1105.
[4] Chen, J., Ou, Q., Chi, Z. et al. Smile detection in the wild with deep convolutional neural networks. Machine Vision and Applications, vol.28, no.1-2, 2017, pp.173-183.
[5] Senechal T, Turcot J and Kaliouby R. el. Smile or smirk Automatic detection of asymmetric smiles to understand viewer experience [C]. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp.1-8, 2013.
[6] Chen J, Luo N, Liu Y, and et al. A hybrid intelligence-aided approach to affect-sensitive e-learning [J]. Computing, 98(1), pp.215-333, 2016.
[7] Shah R, Kwatra V. All Smiles: Automatic Photo Enhancement by Facial Expression Analysis [C]. The 12th European Conference on Visual media production (CVMP), pp.1-10, December, 2012.
[8] Jennifer H, Tracey P, and Willibald R. Laughter and Smiling in 16 Positive Emotions. IEEE journal of affective computing, vol. 8, no. 1. 4. 10-2017.
[9] Whitehill J, Littlewort G, Fasel I. Toward practical smile detection [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 31(11), pp.2106-2111, 2009.
[10] J. M. Girard, J. F. Cohn, and F. D. la Torre, Estimating smile intensity: A better way, Pattern Recognition Letters, vol. 66, pp. 13 – 21, 2015.
[11] Zhang Kaipeng, Tan Lianzhi, Li Zhifeng. Gender and Smile Classification Using Deep Convolutional Neural Networks. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 26 June-1 July 2016.
[12] Julio Cesar Batista, Olga R. P. Bellon and Luciano Silva. Landmark-free smile intensity estimation. Workshop Conference Graphics, 2016
[13] Lucey P, Cohn J F, Kanade T, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010:94-101
[14] P. Ekman, W.V. Friesen, J. Hager, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Research Nexus, Salt Lake City, UT, 2002

[15] S.B. Wang, A. Quattoni, P. Moren, et al. Conditional random fields for gesture recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006:1521-1527

[16] M. Kim, V. Pavlovic, Hidden conditional ordinal random fields for sequence classification [C]// Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, 2010:6322: 51-65

[17] R. Walecki, O. Rudovic, V. Pavlovic and M. Pantic, Variable-state latent conditional random fields for facial expression recognition and action unit detection [C] // IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015:1-8.

[18] Kamarol S K A, Jaward M. Joint facial expression recognition & intensity estimation based on weighted votes of image sequences [J]. Pattern Recognition, 2017, 92(3): 25-32.

[19] R. Pradhan, R. S. Aygun, M. Maskey, R. Ramachandran and D. J. Cecil, Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network, in IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 692-702, Feb. 2018, doi: 10.1109/TIP.2017.2766358.