International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 2, April, 2021, pp. 359-368
Journal homepage: http://iieta.org/journals/ts

Machine Learning-Based Emotional Recognition in Surveillance Video Images in the Context of Smart City Safety

Pan Li[1,2], Zhaojun Zhou[3*], Qingjie Liu[1], Xiaoye Sun[1,2], Fuming Chen[1], Wei Xue[1]

[1] School of Information Engineering, Institute of Disaster Prevention, Sanhe 065201, China
[2] Hebei Key Laboratory of Earthquake Disaster Prevention and Risk Assessment, Sanhe 065201, China
[3] School of Continuing Education, Institute of Disaster Prevention, Sanhe 065201, China

Corresponding Author Email: lipan@cidp.edu.cn

https://doi.org/10.18280/ts.380213

Received: 19 November 2020
Accepted: 28 January 2021

Keywords:
machine learning (ML), convolutional neural network (CNN), face expression identification, emotional identification, smart city safety

ABSTRACT

The effective extraction of deep information from surveillance video lays the basis for smart city safety. However, the surveillance video images contain complex targets, whose expression changes are difficult to capture. The traditional face expression recognition methods or sentiment analysis algorithms have a poor application effect. Based on machine learning (ML), this paper explores the emotional recognition in surveillance video images in the context of smart city safety. Firstly, the potential textures of surveillance video images were extracted under multi-order double cross (MODC) mode, and the optical flow features of facial expressions were detected in these images. Next, a facial expression recognition model was constructed based on the DeepID convolutional neural network (CNN), and an emotional semantic space was established for the face images in surveillance video. The proposed method was proved effective through experiments. The research results provide a reference for emotional recognition in images of other fields.

1. INTRODUCTION

The most effective way to ensure smart city safety is to effectively extract deep information from surveillance video, and nip the behaviors threatening urban safety in the bud [1-4]. With dawn of the age of visual big data, sentimental analysis on image data has become a new hotspot in the field of image processing and interpretation [5-8]. The demand for semantic analysis on image emotions creates an application space for machine learning (ML) techniques like unsupervised learning and deep learning. Meanwhile, the development of ML techniques opens a new path for interpreting and recognizing the emotional semantics of people in surveillance video images.

Traditionally, the potential behavioral intentions and objective psychological activities are studied based on the data collected through questionnaire surveys and the physiological signals obtained through interviews. But the traditional approaches have a low accuracy and a high subjectivity [9-12]. Sakurai et al. [13] carried out integral projection and Fisher analysis of boundaries, and identified the presence/absence of negative emotions (e.g., anxiety, depression, and anger) on human faces; optical flow features were combined with windmill pattern features to analyze human face behaviors, and an evaluation system of human face emotions and behaviors was established based on computer vision libraries like OpenCV and OpenFace.

The wide application of artificial intelligence (AI) and computer vision has gradually pushed up the emotional recognition rate of facial expression recognition and speech emotional recognition [14-18]. Lee et al. [19] recognized facial expressions on static face images and face image series, and constructed a two-channel weighted CNN coupled with an attention mechanism and a bidirectional long short-term memory (BiLSTM) network, which effectively improves the recognition accuracy of expressions.

Emotional identification networks tend to have a poor generalization ability, if the expression dataset is very small. Joseph and Geetha [20] expanded the generative adversarial network (GAN) into an emotional recognition network, and experimentally verified that the expanded network does better in generalization and emotional recognition rate, after being trained with the expanded dataset. Considering the close correlation between bus accidents and driver's negative emotions, Kalsi and Rai [21] designed a vehicle intelligent alarm terminal for the recognition of driver's negative emotions, provided a satisfactory design scheme, and demonstrated the effectiveness of the visualized terminal through ten-fold cross-validation tests.

Currently, emotional recognition is mostly implemented by lab simulation, scale method, and physiological signal detection. The common defect of these methods is that all of them are grounded on the real-time monitoring data on the subjects' emotions [22-27]. To reduce the complex illumination interference on face images, Boubenna and Lee [28] put forward an illumination enhancement algorithm with adaptive attenuation quantification. Kang and Yoon [29] constructed a multi-structure, variable-parameter expression identification model, which retains the sequential features of facial expressions, and solves the vanishing gradient problem caused by too many network layers. Fan et al. [30] extracted the sequential features of facial expression images after block-based preprocessing, quantified the associations between various facial expressions and emotions with the emotional index, and realized the continuous description of discrete facial expressions and effective recognition of the corresponding emotions. Balouchian and Foroosh [31] identified the micro-facial expressions by an end-to-end deep neural network, finetuned the parameters of the pretrained model through transfer learning to make up for the small size of sample set, and reduced the imbalance between different classes of micro-expression data with the focal loss function.

Overall, the interpretation of emotional semantics based on facial expressions is often constrained by the associations between the themes, expressions, and emotions of the subjects. In surveillance video images, the subjects are even more complex, and their expression changes are difficult to capture. The traditional face expression recognition methods or sentiment analysis algorithms have a poor application effect. To solve the problem, this paper discusses the ML-based emotional recognition in surveillance video images in the context of smart city safety. Section 2 extracts the potential textures of surveillance video images under multi-order double cross (MODC) mode, gives a detection method for the optical flow features of facial expression changes, and combines to two types of features to identify the facial expression features in surveillance video images. Section 3 constructs a facial expression recognition model based on the DeepID convolutional neural network (CNN), and sets up an emotional semantic space for the face images in surveillance video. Our method was proved to be effective and accurate through experiments.

2. DETECTION OF FACIAL EXPRESSION FEATURES IN SURVEILLANCE VIDEO IMAGES

2.1 Extraction of potential image textures under the MODC mode

To acquire more facial expression information in surveillance video, this paper resorts to the facial texture descriptor under the MODC mode to extract the potential textures from surveillance video images. The texture extraction mainly encompasses three steps: sampling, filtering, and encoding. Under the MODC mode, the sampling direction can be determined by the unique code calculated by:

$$COD_j = \psi\left(D_j, E_j\right) + \psi\left(E_j, F_j\right) \tag{1}$$

Let P be the central pixel in a local sampling area; D_j, E_j, and F_j be the points on three inner circles centering at P with different radii, respectively; J_P, J_{Dj}, J_{Ej}, and J_{Fj} be the gray values of points P, D_j, E_j, and F_j, respectively. Then, we have:

$$\psi(O,W) = R\left(J_O, J_P\right) \times 2 + R\left(J_W - J_O\right) \tag{2}$$

COD_j can be decomposed into the codes of two orthogonal planes, i.e., COD_X and COD_Y:

$$COD_X = \sum_{j=0}^{3} COD_{2j} \times 4^j \tag{3}$$

$$COD_Y = \sum_{j=0}^{3} COD_{2j+1} \times 4^j \tag{4}$$

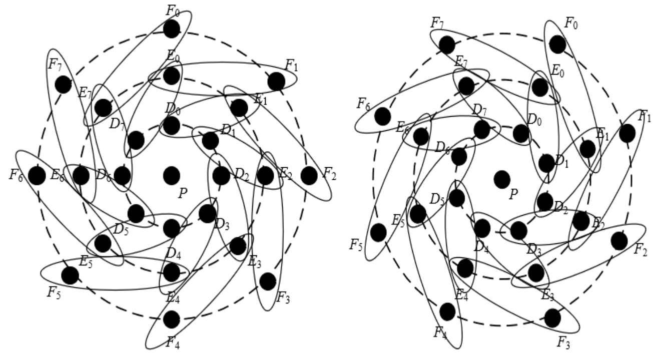
COD_X and COD_Y are connected in series to form COD_j. To extract more potential textures from face video images, every point on each inner circle must be multidirectional in feature description. Here, the three inner circles centering at P with different radii are rotated. Depending on the rotation direction, the feature description methods were divided into left SL and right SR. SL_j can be expressed as:

$$SL_j = \begin{cases} R\left(J_{D_j} - J_P\right) \times 2 + R\left(J_{E_j} - J_{D_{j+1}}\right) + R\left(J_{E_j} - J_P\right) \times 2 + R\left(J_{F_i} - J_{E_{j+1}}\right), 0 \le j \le 6 \\ R\left(J_{D_j} - J_P\right) \times 2 + R\left(J_{E_j} - J_{D_j}\right) + R\left(J_{E_j} - J_P\right) \times 2 + R\left(J_{F_j} - J_{E_j}\right), j = 7 \end{cases} \tag{5}$$

SL can also be decomposed into the codes of two orthogonal planes, i.e., SL_X and SL_Y:

$$SL_X = \sum_{j=0}^{3} SL_{2j} \times 4^j \tag{6}$$

$$SL_Y = \sum_{j=0}^{3} SL_{2j+1} \times 4^j \tag{7}$$



(a) *SL* (b) *SR*

Figure 1. Sketch diagram of rotation transform for *SL* and *SR*

The rotation of the left SL feature description is illustrated in Figure 1(a). Obviously, SL_X and SL_Y are connected in series to form SL. SL_X means the circle on the layer is rotated by 45° in the clockwise direction; SL_Y means the circle on the layer is rotated by 45° in the counterclockwise direction. Similarly, SR_j can be expressed as:

$$SR_j = \begin{cases} R\left(J_{D_j} - J_P\right) \times 2 + R\left(J_{E_{j+1}} - J_{D_j}\right) + R\left(J_{E_j} - J_P\right) \times 2 + R\left(J_{F_{j+1}} - J_{E_j}\right), 0 \le j \le 7 \\ R\left(J_{D_j} - J_P\right) \times 2 + R\left(J_{E_j} - J_{D_j}\right) + R\left(J_{E_j} - J_P\right) \times 2 + R\left(J_{F_j} - J_{E_j}\right), j = 0 \end{cases} \tag{8}$$

SR can also be decomposed into the codes of two orthogonal planes, i.e., SR_X and SR_Y:

$$SR_X = \sum_{j=0}^{3} SR_{2j} \times 4^j \tag{9}$$

$$SR_Y = \sum_{j=0}^{3} SR_{2j+1} \times 4^j \tag{10}$$

Similarly, SR_X and SR_Y are connected in series to form SR. The rotation of the right SR feature description is illustrated in Figure 1(b), where SL_X means the circle on the layer is rotated by 45° in the clockwise direction; SL_Y means the circle on the layer is rotated by 45° in the counterclockwise direction.

After the rotation transform, the textures in the regions of interest (ROIs), i.e., human faces, of surveillance video images can be described in greater details, making the detectors more sensitive to the facial emotions in these images.

2.2 Detection of optical flow features for facial expression changes

The subtle changes of facial expressions can be characterized by the change law of the pixel intensity between video frames in the time domain. Here, the first frame of the video frame series is defined as the reference frame to be compared with every other frame in the series. Let (*a*, *b*, *h*) and (*a*+*ds*, *b*+*de*, *h*+*dh*) be the coordinates and time of the target pixel in the reference frame and the contrastive frame, respectively. After a period of *dh*, the target pixel has a displacement of *ds* and *de* in the horizontal and vertical directions of the two-dimensional (2D) plane, respectively. The facial expression images of adjacent frames need to satisfy the conservation of gray value:

$$J_h(a,b) = J_{h+dh}(a+ds, b+de) \tag{11}$$

Under the premise of constant brightness of the optical flow, moderate moving amplitude, and spatial consistency, the components of the optical flow vector in directions *a* and *b* are denoted as $\frac{da}{dh} = s_a$ and $\frac{db}{dh} = s_b$, respectively. Through Taylor expansion of formula (11), the basic equation of the optical flow can be obtained as:

$$\frac{\partial J}{\partial a}\frac{\partial a}{dh} + \frac{\partial J}{\partial b}\frac{\partial b}{dh} + \frac{\partial J}{\partial h}\frac{\partial h}{dh} = 0 \tag{12}$$

Formula (12) can be rewritten as a matrix:

$$\left[J_a J_b\right]\begin{bmatrix} s_a \\ s_b \end{bmatrix} = -J_h \tag{13}$$

Formula (13) shows that the components of the optical flow vector in directions *a* and *b* need to be solved. Under the constraint of smooth movement, the minimum value *C* meeting the condition can be derived from formulas (12) and (13):

$$C = \min\left\{\iint \left(J_a s_a + J_b s_b + J_h\right)^2 dadh\right\} \tag{14}$$

Under the constraint of smooth movement, only the normal vector of the optical flow can be obtained for the target pixel. The optical flow field becomes too smooth, as the local information of images is ignored. To capture the refined features of facial expression changes, the nonuniform smoothness constraint can be adopted:

$$min\left\{\iint\left[s_a^2 + s_b^2 + e_a^2 + e_b^2 + \mu\left(J_a s_a + J_b s_b + J_h\right)^2\right]dadh\right\} \tag{15}$$

Formula (15) can be simplified by the Euler equation:

$$min\left\{\iint G\left(s, e, s_a, s_b, e_a, e_b\right)dadh\right\} \tag{16}$$

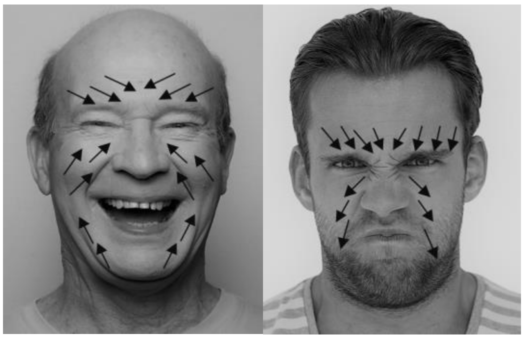
The Euler equation corresponding to formula (16) can be expressed as:

$$\begin{cases} G_s - \frac{\partial G_{s_a}}{\partial a} - \frac{\partial G_{s_b}}{\partial b} = 0 \\ G_e - \frac{\partial G_{e_a}}{\partial a} - \frac{\partial G_{e_b}}{\partial b} = 0 \end{cases} \tag{17}$$

Let *μ* be the degree of the smoothness constraint. Formulas (16) and (17) can be combined into:

$$\begin{cases} \nabla^2 s = \mu J_a\left(J_a s + J_b e + J_h\right) \\ \nabla^2 e = \mu J_b\left(J_a s + J_b e + J_h\right) \end{cases} \tag{18}$$

Formula (18) shows that the *μ* should take a small value if the images are very noisy. In general cases, however, the image data are discretized to solve the corresponding optical flow histogram. Finally, the optical flow feature at time *h* can be described by a 2D vector $[e_a^{h,} e^h_b]^T$.



(a) Pleasure (b) Anger

Figure 2. Optical flow features of different facial expressions

Figure 2 presents the optical flow features of different facial expressions. It can be seen that the optical flow features only converge in the ROIs with expression changes. Therefore, the optical flow vector can be calculated in two parts: module value and angle. The module value OV_k of the optical flow eigenvector of the *k*-th frame can be expressed as:

$$OV_k = \sqrt{a_k^2 + b_k^2} \tag{19}$$

where, a_k and b_k are the components of the optical flow vector of the *k*-th frame in directions *a* and *b*, respectively. Let ξ_j be the angle of the optical flow of the *k*-th frame. Then, the optical flow angle of an expression frame in the facial expression image series can be solved by the anti-trigonometric function:

$$\begin{cases} \xi_{1\text{-}k} = arctan\left|\frac{b_k}{a_k}\right| \\ \xi_{2\text{-}k} = \frac{\pi}{2} + arctan\left|\frac{b_k}{a_k}\right| \\ \xi_{3\text{-}k} = \pi + arctan\left|\frac{b_k}{a_k}\right| \\ \xi_{4\text{-}k} = \frac{3\pi}{2} + arctan\left|\frac{b_k}{a_k}\right| \end{cases} \tag{20}$$

Formula (20) provides the optical flow angles of expression frames in the facial expression image series in four quadrants.

2.3 Feature fusion

For the post-rotation facial expression features under the MODC mode, the peak and mean of the feature difference between surveillance video images after smoothing are denoted as Q_M and Q_A, respectively; the proportional coefficient, which falls in [0, 1] is denoted as δ. Then, any segment with facial expression changes can be pinpointed through threshold and peak detection:

$$VE = Q_A + \delta \times \left(Q_M - Q_A \right) \tag{21}$$

Let OV_j and α_j be the module value and angle of the optical flow after smoothing, respectively. To intuitively explain the detection of the module value and angle of the optical flow at the appearance of facial expressions, the optical flow features can be mapped to the polar coordinate system:

$$\begin{cases} p_j = OV_j cos\alpha_j \\ q_j = OV_j sin\alpha_j \end{cases} \tag{22}$$

Since the OV_j value changes in an inverted U-shaped curve, the position farthest from the origin in the entire micro-expression segment can be defined as the climax frame. Let OV_M and M_T be the module value of the optical flow eigenvector of the climax frame, and the preset module value, respectively. Then, the threshold can be obtained by multiplying M_T with OV_M. Thus, the starting frame and ending frame of a segment with facial expression changes can be determined based on the threshold γ:

$$\begin{cases} OV_j > M_T \cdot OV_M \\ \left| \alpha_j - \alpha_{j-1} \right| < \gamma \end{cases} \tag{23}$$

The micro-expressions in the video segment can be detected by integrating the post-rotation features of the MODC mode and optical flow. Firstly, the presence of optical flow features in the surveillance video samples needs to be verified. If the starting and ending frames of a segment with facial expression changes are both zero, it is necessary to verify the post-rotation features of the MODC mode and optical flow. In the end, the feature detection results in the two steps should be merged by:

$$F = F_{LS} \,||\, F_{COD} \tag{24}$$

3. RECOGNITION AND EMOTIONAL CLASSIFICATION OF FACIAL EXPRESSIONS IN SURVEILLANCE VIDEO IMAGES

3.1 Model construction

The effective extraction of expression features in face images is the key to the recognition of facial expressions and the association between expressions and emotions. After extracting the facial expression features from surveillance video images, it is necessary to establish a model to recognize facial expressions, and classify them into different categories of emotions. Deep learning can automatically label complex features on unsupervised data. This paper designs a facial expression recognition model based on the DeepID CNN (Figure 3).

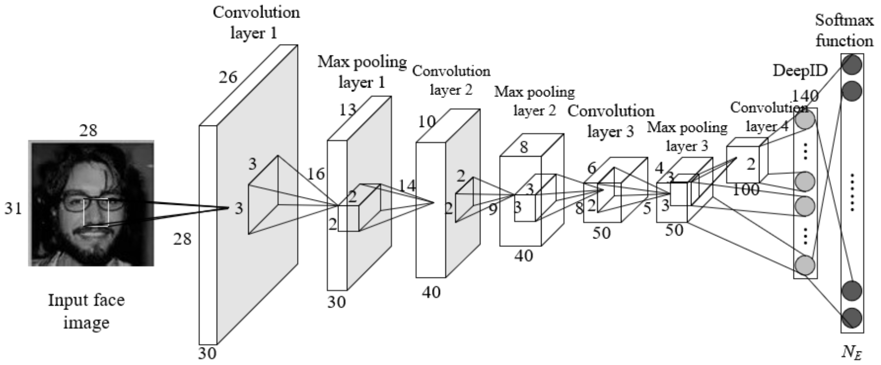


Figure 3. Structure of DeepID CNN-based model

There are four convolution layers in the model. The first three convolution layers adopt convolution-pooling structures with 30, 40, and 50 convolution kernels, respectively. The fourth convolution layer adopts a convolution-fully connected structure with 100 convolution kernels. The four convolution layers are responsible for feature extraction, dimensionality reduction, and fusion of advanced features for facial expression classification. The output layer uses the softmax function. Every layer in the model is activated by the rectified linear unit (ReLU) function. The pooling operations are all max pooling.

To ensure the global completeness of image features while minimizing information loss, the fully-connected layer can be expressed as:

$$O_j = max\left(0, \sum_i I_{ij}^1 * \omega_{ij}^1 + \sum_i I_i^2 * \omega_{ij}^2 + \varepsilon_j \right) \tag{25}$$

Let N_E be the number of emotions for feature classification. The facial expressions can be allocated to multiple classes by the softmax function:

$$O_i' = \frac{e^{O_i^*}}{\sum_{i=1}^{N_E} e^{O_i^*}} \tag{26}$$

3.2 Construction of emotional semantic space

To associate recognized human expressions with emotions, the first task is to set up an emotional semantic space for the face images in surveillance video, that is, to establish the mapping between emotional semantic annotations and low-level features of facial expressions (Figure 4). Let g be the mapping function. Then, we have:

$$\left(B_1, B_2, ..., B_{N_E}\right) = g\left(A_1, A_2, ..., A_{N_E}\right) \tag{27}$$

The mapping function g is the nonlinear classifier of the CNN. To associate recognized human expressions with emotions, this paper builds up the pleasure-arousal-dominance (PAD) model, which is often adopted to describe emotions in the field of psychology. The PAD model has a total of three dimensions: "pleasure-displeasure" P reflecting the positivity/negativity of emotion; "arousal-nonarousal" A reflecting the neurophysiological activation level; "dominance-submissiveness" D reflecting the controlling and dominant nature of the emotion. Together, the three dimensions show the specific emotions corresponding to the recognized facial expressions.

The emotional semantic space was constructed in the following steps: (1) Investigate the data samples of facial expression images, and sort out the adjective pairs that vaguely represent the emotions of the facial expressions in these images; (2) Obtain accurate and clear adjective pairs through cognition tests on the fuzzy adjective pairs; (3) Build a database of emotional semantics corresponding to the facial expressions in the images; (4) Analyze the adjectives in the database by multivariate method, and establish the emotional semantic space for the face images in surveillance video.

Before finalizing the emotional semantic space, the emotional adjectives must be further processed. Here, the 50 pairs of adjectives characterizing the facial expressions are fully expanded in terms of emotion. Table 1 shows 20 of the adjective pairs. The expansion was carried out by extending the emotional concepts horizontally and vertically or expressing them with adverbs of degree. Each emotional adjective was further broken down to five levels, namely, strongly positive, slightly positive, neutral, slightly negative, and slightly negative. The weights of the five levels are 0, 0.25, 0.5, 0.75, and 1, respectively. In this way, the database of adjective pairs was expanded by 4 times.

Step (4) was detailed as follows: Let Φ be the total number of subjects; $EV_{\varphi\text{-}FI\text{-}l}$ be the evaluation by subject φ for the l-th adjective pair of the surveillance video image FI. Then, the mean evaluation score $ES_{FI\text{-}l}$ can be calculated by:

$$ES_{FI-l} = \frac{1}{\Phi}\sum_{\phi=1}^{\Phi} EV_{\phi-FI-l} \tag{28}$$

Then, the mean evaluation score $ES_{FI\text{-}l}$ can be normalized into:

$$ES'_{FI-l} = \frac{1}{\sqrt{\sum_{FI=1}^{N_{FI}}\left(ES_{FI-l}-\overline{ES_l}\right)^2}}\left(ES_{FI-l}-\overline{ES_l}\right) \tag{29}$$

Finally, the common factor θ and factor loading matrix Ψ were solved through factor analysis based on $ES'_{FI\text{-}l}$. In this way, the dimensionality of the emotional semantic space was reduced. After that, the space has a certain orthogonality:

$$ES = \psi\theta + \tau \tag{30}$$

Table 1. Some adjective pairs in the database of emotional semantics

1. Angry - calm	8. Touched - scornful	15. Decadent - positive
2. Happy - sad	9. Uneasy - relieved	16. Disgusted - attracted
3. Serious - relaxed	10. Surprised - unsurprised	17. Fevered - vacuous
4. Scared - laid-back	11. Ashamed - proud	18. Anxious - light-hearted
5. Depressed - confident	12. Stressed - hysteric	19. Mad - quiet
6. Disappointed - gratified	13. Impassioned - indifferent	20. Panic - leisured
7. Helpless - excited	14. Disdainful - worshipful	

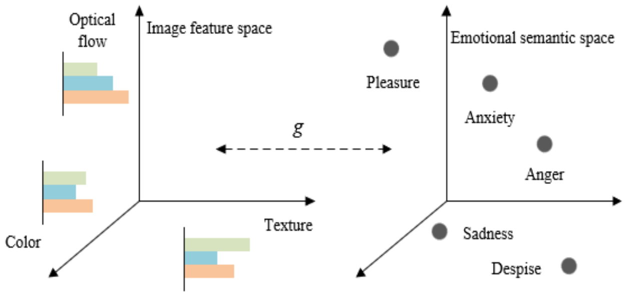


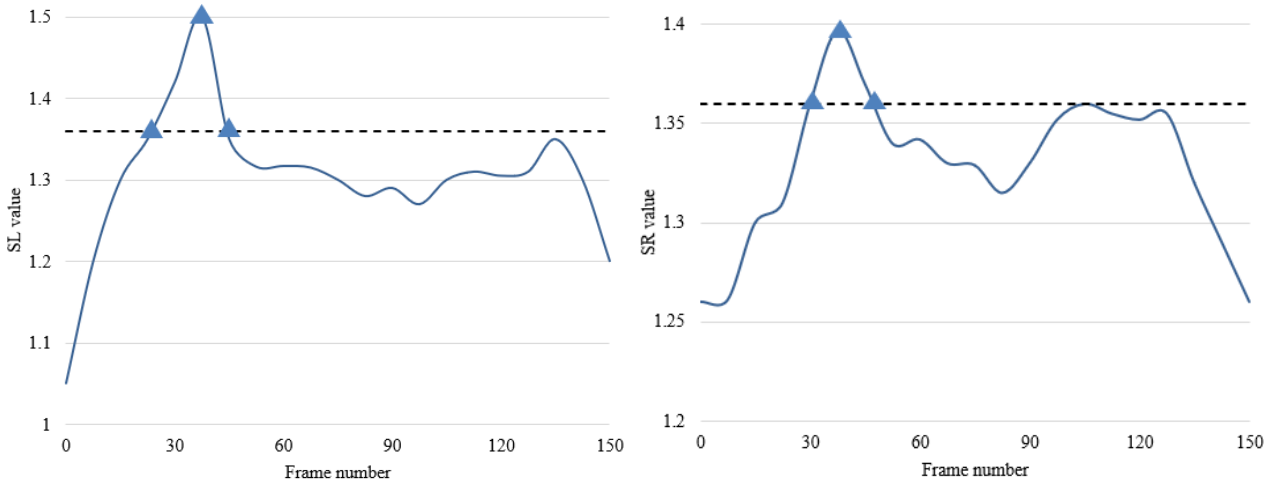
Figure 4. Mapping between emotional semantic annotations and low-level features of facial expressions

4. EXPERIMENTS AND RESULTS ANALYSIS

Figure 5 presents the detection results on facial expression features in surveillance video images by left SL feature descriptor (a) and right SR feature descriptor (b) under the MODC mode. Apparently, the starting, climax, and ending frames of the expression change segment obtained by the left SL feature descriptor were the 27th, 37th, and 45th frames, respectively; e starting, climax, and ending frames of the expression change segment obtained by the right SR feature descriptor were the 30th, 36th, and 43rd frames, respectively. Meanwhile, the manually labeled expression change segment of the same images starts and ends at the 29th and 49th frames. Hence, the results detected by our method fall in the allowable error range. This means the proposed feature detection method is applicable to the surveillance video images with short durations and micro-expression changes.

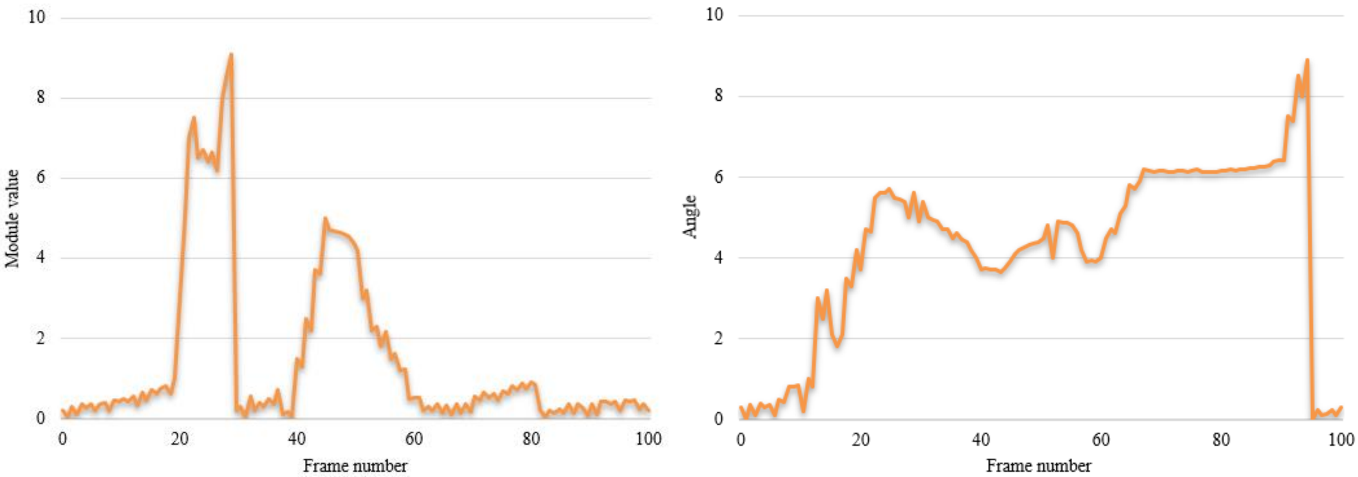
Figure 6 shows how the module value and angle of optical flow change in the ROI (eyebrows) of the video frame series with expression changes. It can be seen that, when the subject had emotional changes and adjusted his/her facial expression, the module value of optical flow would fluctuate, while the optical flow angle would not change greatly. The results are consistent with the information transmitted from the pixels of eyebrows in actual image frames. Thus, the proposed detection method for optical flow features is effective in detecting such expression changes.

The traditional facial expression detection algorithms are defected, because they only consider optical flow features and textures. To accurately recognize the subtle movements of facial muscles in face images, it is important to fully consider numerous texture differences, and remain sensitive to the dynamic features of adjacent frames in the video frame series. Otherwise, it would be impossible to detect any change of expressions or emotions.



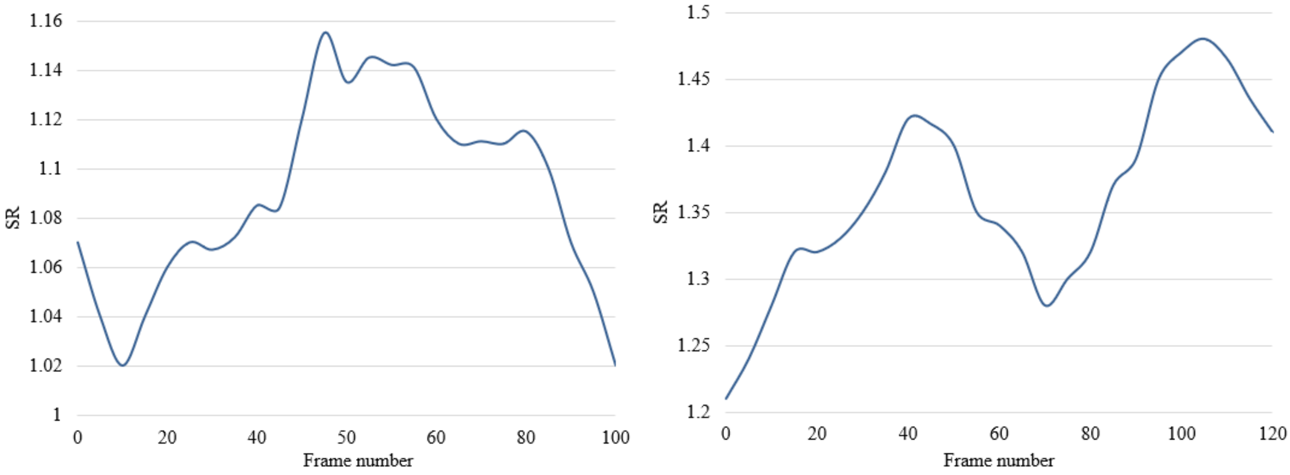
(a) *SL*　　(b) *SR*

Figure 5. Detection results of left *SL* feature descriptor and right *SR* feature descriptor under the MODC mode



(a) Module value　　(b) Angle

Figure 6. Curves of module value and angle of optical flow



(a) *SL*　　(b) *SR*

Figure 7. Optical flow features detected by left *SL* feature descriptor and right *SR* feature descriptor

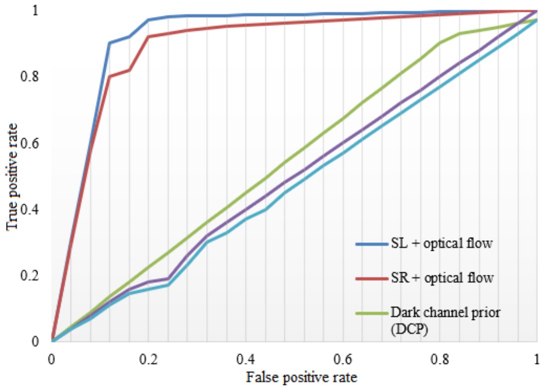
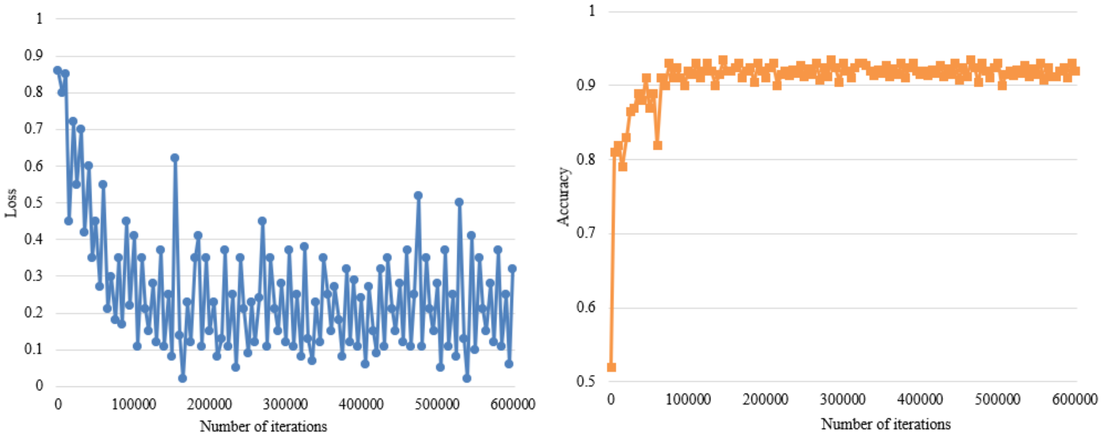


Figure 8. Receiver operating characteristic (ROC) curves of different expression feature extraction methods

Figure 7 shows the optical flow features detected by left *SL* feature descriptor and right *SR* feature descriptor. In Figure 7(a), the facial expression in the sample changed between the 49th and the 58th frames; In Figure 7(b), the facial expression changed more complicatedly, yet the descriptor detected obvious peaks around the 42nd frame and the 105th frame. This proves that the facial expression changes extracted based on feature fusion are very effective, and contain rich texture information.

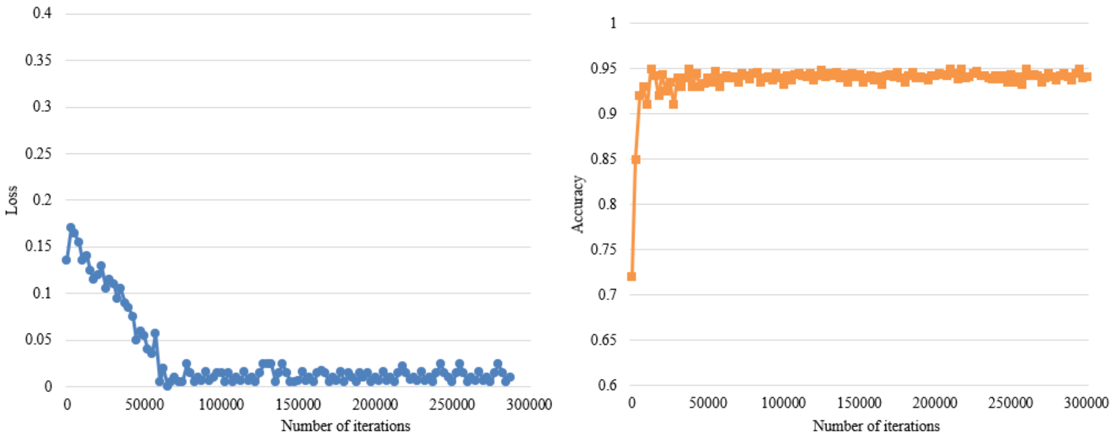
The original surveillance video segments were used to verify the effectiveness of our expression feature extraction method. An image library was created based on a surveillance video, including 300 segments of expression changes of 32 subjects. To demonstrate its superiority, our method was compared with DCP, HOG, and LBP. The ROC curve was selected to evaluate the performance of each method (Figure 8). The area under the curve (AUC) of each method is recorded in Table 2. It can be seen that the proposed *SL* + optical flow and *SR* + optical flow methods achieved the highest AUCs, reflecting the efficiency and accuracy of our method.

The proposed model for recognition and emotional classification of facial expressions was tested under the same environment as the expression feature detection model. The comparative tests were carried out on the graphic processing unit (GPU) server of our lab computer. The facial expressions were classified into four kinds of emotions: happy, angry, calmness, and scared. The proposed model was trained into three different versions: the original DeepID CNN, the network coupled with our expression feature detection method, and the network further coupled with our emotional semantic space. Figures 9, 10, and 11 present the loss and accuracy of the facial expression recognition and classification by the three versions, respectively. Table 3 reports the test results of the three versions. Despite its effectiveness in facial expression recognition and classification, the original DeepID CNN saw large fluctuations in the convergence curve of its loss, leaving a large room for improvement. After introducing our expression feature detection method and emotional semantic space, the facial expression recognition and classification became more accurate, and the loss function converged more stably.



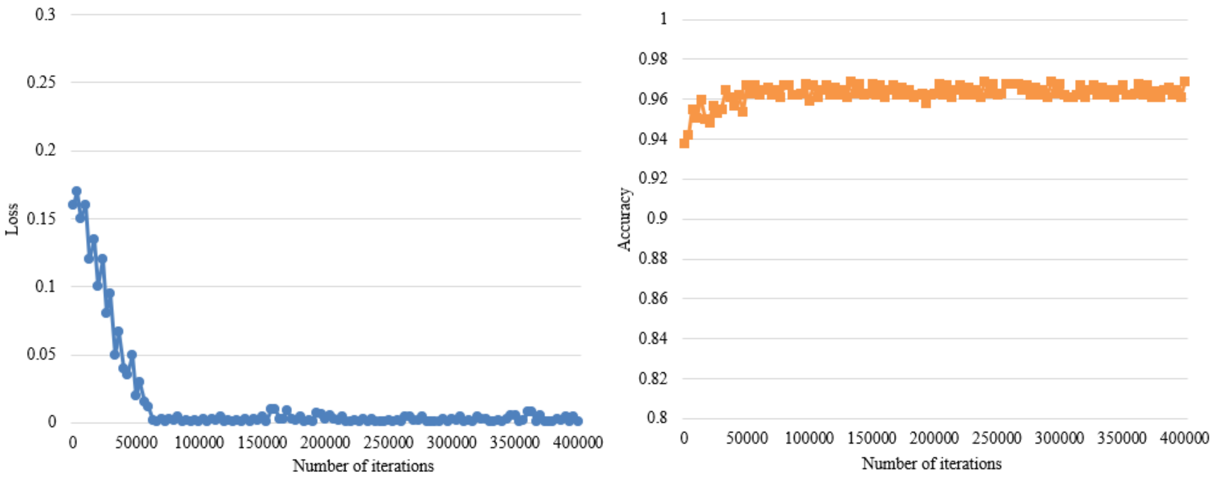
(a) Loss (b) Accuracy

Figure 9. Recognition and classification performance of the original DeepID CNN



(a) Loss (b) Accuracy

Figure 10. Recognition and classification performance coupling our expression feature detection method



(a) Loss (b) Accuracy

Figure 11. Recognition and classification performance further coupling emotional semantic space

Table 2. AUCs of different expression feature extraction methods

***SL* + optical flow**	***SR* + optical flow**	**DCP**	**HOG**	**LBP**
0.931	0.915	0.568	0.537	0.592

Table 3. Test results of the three versions

Version	**Loss**	**Recognition accuracy**	**Classification accuracy**
Original DeepID CNN	0.315	91.11%	90.08%
Network coupled with our expression feature detection method	0.307	92.53%	91.43%
Network further coupled with our emotional semantic space	0.322	95.37%	96.32%

5. CONCLUSIONS

In the context of smart city safety, this paper investigates the emotional recognition in surveillance video images based on the ML. The authors put forward a method for extracting the potential textures of surveillance video images under the MODC mode, and fused the extracted textures with the optical flow features, which reflect the facial expression changes. In this way, the facial expressions were detected in surveillance video images, and their features were extracted thoroughly. To recognize facial expressions and associate them with emotions, the authors developed a facial expression recognition model based on the DeepID CNN, and introduced the emotional semantic space to the original network. Through experiments, the detection results of left *SL* + optical flow and right *SR* + optical flow were obtained. The results prove that the facial expression changes extracted based on feature fusion are very effective, and contain rich texture information. Moreover, our method was found to be efficient and accurate through comparison against DCP, HOG, and LBP. Finally, different versions of our network were designed for facial expression recognition and emotional classification tests. It was learned that the model became more accurate and stable in classification, after introducing our expression feature detection method and emotional semantic space.

ACKNOWLEDGMENT

This paper was supported by the Fundamental Research Funds for the Central Universities (Grant No.: ZY20210302).

REFERENCES

[1] Zhang, W., He, X.Y., Lu, W.Z. (2020). Exploring discriminative representations for image emotion recognition with CNNs. IEEE Transactions on Multimedia, 22(2): 515-523. https://doi.org/10.1109/TMM.2019.2928998

[2] Rao, T.R., Xu, M., Liu, H.Y., Wang, J.Q., Burnett, I. (2016). Multi-scale blocks based image emotion classification using multiple instance learning. Proceedings - International Conference on Image Processing, ICIP, 2016 IEEE International Conference on Image Processing, ICIP 2016 – Proceedings, pp. 634-638. https://doi.org/10.1109/ICIP.2016.7532434

[3] Bhattacharya, A., Choudhury, D., Debangshu, D. (2016). Emotion recognition from facial image analysis using composite similarity measure aided bidimensional empirical mode decomposition. 2016 IEEE 1st International Conference on Control, Measurement and Instrumentation, CMI 2016, 2016 IEEE 1st International Conference on Control, Measurement and Instrumentation, CMI 2016, pp. 336-340. https://doi.org/10.1109/CMI.2016.7413766

[4] Park, M. W., Ko, D., Hwang, H., Moon, J., Lee, E.C. (2017). Image classification using color and spatial frequency in terms of human emotion. In Advanced multimedia and ubiquitous engineering. Springer, Singapore, 91-96. https://doi.org/10.1007/978-981-10-5041-1_16

[5] Stolar, M. N., Lech, M., Bolia, R. S., Skinner, M. (2017). Real time speech emotion recognition using RGB image classification and transfer learning. In 2017 11th

International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, pp. 1-8. https://doi.org/10.1109/ICSPCS.2017.8270472

[6] Yang, J., She, D., Sun, M. (2017). Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network. In IJCAI, 3266-3272.

[7] He, G.W., Liu, X.F., Fan, F.F., You, J. (2020). Image2Audio: Facilitating semi-supervised audio emotion recognition with facial expression image. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June 2020, v2020-June, p3978-3983, Proceedings - 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020. https://doi.org/10.1109/CVPRW50498.2020.00464

[8] Kim, H.R., Kang, H., Lee, I.K. (2016). Image recoloring with valence-arousal emotion model. In Computer Graphics Forum, 35(7): 209-216. https://doi.org/10.1111/cgf.13018

[9] Huang, Y., Lu, H. (2016). Deep learning driven hypergraph representation for image-based emotion recognition. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 243-247. https://doi.org/10.1145/2993148.2993185

[10] Tarasov, A.V., Savchenko, A.V. (2018). Emotion recognition of a group of people in video analytics using deep off-the-shelf image embeddings. In International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, pp. 191-198. https://doi.org/10.1007/978-3-030-11027-7_19

[11] Kundu, T., Saravanan, C. (2017). Advancements and recent trends in emotion recognition using facial image analysis and machine learning models. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT). IEEE, pp. 1-6. https://doi.org10.1109/ICEECCOT.2017.8284512

[12] Basu, A., Routray, A., Shit, S., Deb, A. K. (2015). Human emotion recognition from facial thermal image based on fused statistical feature and multi-class SVM. In 2015 Annual IEEE India Conference (INDICON), pp. 1-5. https://doi.org 10.1109/INDICON.2015.7443712

[13] Sakurai, S., Narumi, T., Katsumura, T., Tanikawa, T., Hirose, M. (2015). Basic study of evoking emotion through extending one's body image by integration of internal sense and external sense. In International Conference on Human Interface and the Management of Information, Springer, Cham, 433-444. https://doi.org/10.1007/978-3-319-20612-7_42

[14] Kyriakou, K., Kleanthous, S., Otterbacher, J., Papadopoulos, G.A. (2020). Emotion-based stereotypes in image analysis services. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 252-259. https://doi.org/10.1145/3386392.3399567

[15] Padhy, N., Singh, S.K., Kumari, A., Kumar, A. (2020). A literature review on image and emotion recognition: proposed model. Smart Intelligent Computing and Applications, 341-354. https://doi.org/10.1007/978-981-32-9690-9_34

[16] Yang, Y., Jia, J., Wu, B., Tang, J. (2016). Social role-aware emotion contagion in image social networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).

[17] Lim, L., Khor, H.Q., Chaemchoy, P., See, J., Wong, L.K. (2020). Where is the Emotion? Dissecting A Multi-Gap Network for Image Emotion Classification. In 2020 IEEE International Conference on Image Processing (ICIP), pp. 1886-1890. https://doi.org/10.1109/ICIP40778.2020.9191258

[18] Jin, K., Wei, Z. (2020). Research on image emotion classification based on gene expression. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) IEEE, 1: 1203-1207. https://doi.org/10.1109/ITNEC48623.2020.9085053

[19] Lee, J.H., Kim, H.J., Cheong, Y.G. (2020). A multi-modal approach for emotion recognition of tv drama characters using image and text. In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp) IEEE, pp. 420-424. https://doi.org/10.1109/BigComp48618.2020.00-37

[20] Joseph, A., Geetha, P. (2020). Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. The Visual Computer, 36(3): 529-539. https://doi.org/10.1007/s00371-019-01628-3

[21] Kalsi, K.S., Rai, P. (2018). A classification of emotion and gender using local biorthogonal binary pattern from detailed wavelet coefficient face image. In Optical and Wireless Technologies, Springer, Singapore, 83-93. https://doi.org/10.1007/978-981-10-7395-3_9

[22] Peng, S., Zhang, L., Winkler, S., Winslett, M. (2018). Give me one portrait image, I will tell you your emotion and personality. In Proceedings of the 26th ACM international conference on Multimedia, pp. 1226-1227. https://doi.org/10.1145/3240508.3241384

[23] Narula, V., Feng, K., Chaspari, T. (2020). Preserving privacy in image-based emotion recognition through user anonymization. In Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 452-460. https://doi.org/10.1145/3382507.3418833

[24] NASRI, M., Hmani, M.A., Mtibaa, A., Petrovska-Delacretaz, D., Slima, M.B., Hamida, A.B. (2020). Face emotion recognition from static image based on convolution neural networks. In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1-6. https://doi.org/10.1109/ATSIP49331.2020.9231537

[25] Nasir, M., Dutta, P., Nandi, A. (2020). Recognition of changes in human emotion from face image sequence using triangulation induced Barycentre-Orthocentre paired distance signature. In 2020 International Conference on Computational Performance Evaluation (ComPE), pp. 101-106. https://doi.org/10.1109/ComPE49325.2020.9200068

[26] Kang, D., Shim, H., Yoon, K. (2018). A method for extracting emotion using colors comprise the painting image. Multimedia Tools and Applications, 77(4): 4985-5002. https://doi.org/10.1007/s11042-017-4667-0

[27] Uchida, M., Akaho, R., Ogawa, K., Tsumura, N. (2018). Image-based non-contact monitoring of skin texture changed by piloerection for emotion estimation. In Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics. International Society for Optics and Photonics, 10501: 1050110G. https://doi.org/10.1117/12.2284297

[28] Boubenna, H., Lee, D. (2018). Image-based emotion

recognition using evolutionary algorithms. Biologically Inspired Cognitive Architectures, 24: 70-76. https://doi.org/10.1016/j.bica.2018.04.008

[29] Kang, D., Yoon, K. (2017). An emotion-based image color modification system. Modelling, Identification and Control.

[30] Fan, Y., Yang, H., Li, Z., Liu, S. (2018). Predicting image emotion distribution by emotional region. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1-9. https://doi.org/10.1109/CISP-BMEI.2018.8633190

[31] Balouchian, P., Foroosh, H. (2018). Context-sensitive single-modality image emotion analysis: A unified architecture from dataset construction to CNN classification. In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1932-1936. https://doi.org/10.1109/ICIP.2018.8451048