Recognition and Intensity Estimation of Facial Expression Using Ensemble Classifiers

Hiroki Nomiya [1] , Shota Sakaue [2] , Teruhisa Hochin [3]

[1] *Information and Human Sciences, Kyoto Institute of Technology,*
Goshokaido-cho, Matsugasaki, Sakyo-ku, Kyoto, 606-8585, Japan
E-mail: nomiya@kit.ac.jp

[2] *Graduate School of Information Science, Kyoto Institute of Technology,*
Goshokaido-cho, Matsugasaki, Sakyo-ku, Kyoto, 606-8585, Japan

[3] *Information and Human Sciences, Kyoto Institute of Technology,*
Goshokaido-cho, Matsugasaki, Sakyo-ku, Kyoto, 606-8585, Japan
E-mail: hochin@kit.ac.jp

Abstract

Facial expression recognition (FER) has been widely studied since it can be used for various applications. However, most of FER techniques focus on discriminating typical facial expressions such as six basic facial expressions. Spontaneous facial expressions are not limited to such typical ones because the intensity of a facial expression varies depending on the intensity of an emotion. In order to utilize FER for real-world applications, therefore, it is necessary to discriminate slight difference of facial expressions. In this paper, we propose an effective FER method to recognize spontaneous facial expressions using ensemble learning which combines a number of naive Bayes classifiers. In addition, a method to estimate the intensity of facial expression is also proposed by using the classification results of the classifiers. The effectiveness of these methods are evaluated through an FER experiment and an experiment to estimate the intensity of facial expressions using a data set including spontaneous facial expressions.

Keywords: Facial expression recognition, Facial expression intensity estimation, Feature selection, Ensemble learning.

1. Introduction

Facial expression recognition (FER) can be utilized for various applications such as retrieval of images and videos, robotics, and healthcare. It thus attracts attention and has been widely studied to improve the accuracy and efficiency of FER. FER is required, for example, to efficiently retrieve interesting scenes from large-scale video databases and to develop a robot which can communicate with human beings in real time. In FER, therefore, it is important to precisely and quickly recognize complex and subtle facial expressions and to handle personal difference of facial expressions.

Most of FER methods construct FER models by utilizing some feature values computed from facial images.[1–10] Therefore, the feature values have great influence on the performance of FER. It is, however, difficult to determine effective facial features due to a tradeoff between recognition accuracy and computational cost. Moreover, an enormous number of feature values can be defined using a wide variety of

facial features (e.g., positions of facial parts or textures of a face). These are great obstacles to manually determine feature values.

In this paper, we propose a method to automatically determine facial features effective for FER. The proposed method selects useful facial features on the basis of positional relation between facial feature points. The facial feature points are salient points on a face such as end points of eyebrows, eyes, a mouth, and nasolabial folds. A facial feature point is represented as a point on two-dimensional space. A feature value of our facial feature is defined based on the angle between two line segments consisting of three facial feature points. It can be obtained by a simple computation and thus can be obtained very fast.

Because a single facial feature is too simple to recognize complex facial expressions, we introduce ensemble learning[11] and combine several facial features to enhance the recognition accuracy. We first prepare a large number of facial features by simply enumerating all possible combinations of facial feature points. Feature selection is then performed to select a small number of useful facial features. For each facial feature, an FER model is constructed on the basis of naive Bayes classifier. The FER models of all the selected facial features are finally integrated into a single FER model by means of weighted voting.

Most of existing FER methods can only discriminate several types of facial expressions, typically the six basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise). Furthermore, evaluation experiment is conducted by using a data set containing only deliberate facial expressions.[1−8] Considering that recognizing spontaneous facial expression is required for real-world application and that spontaneous facial expression can vary depending on the intensity of the emotion, it is important not only to discriminate facial expressions but to estimate the intensity of facial expressions.

In order to estimate the intensity of facial expressions, we utilize the weight of each classifier used in the weighted voting. This means that no additional computation cost is required to estimate the intensity of facial expressions.

We conduct several experiments to evaluate the proposed method from the viewpoints of the accuracy and efficiency of the FER and estimation of facial expression intensity. A data set consists of several video clips including spontaneous facial expressions is used for the experiment.

The remainder of this paper is organized as follows. Section 2 describes related works. Section 3 shows the facial features and explains the feature selection method. Section 4 elaborates the FER method on the basis of ensemble learning. Section 5 introduces the method to estimate the intensity of facial expressions. Section 6 evaluates the proposed method through an experiment. Finally, Section 7 concludes this paper.

2. Related Works

Typical FER methods construct recognition models on the basis of facial features obtained from facial images. The facial features are roughly divided into appearance-based features and geometric-based features.[1]

The appearance-based features are based on the textures on a face. For example, Gabor wavelet[2], local binary pattern (LBP)[3] are used as the appearance-based feature. There is an FER method combining Gabor filters and LBPs.[4] The appearance-based features are generally computed from the intensity values of image pixels on a face. They are thus sensitive to the change of lighting condition and the difference of image quality.

The geometric-based features are represented by the shapes and/or positional relations of salient facial parts (e.g., eyebrows, eyes, and a mouth). For example, they can be obtained by forming a 3-dimensional model of a face.[5,6] The 3D model leads to accurate recognition since it can precisely describe the shapes of the facial parts. However, it generally requires high computational cost. The problem of the computational cost can be resolved by introducing a more concise 2D facial model.[7,8] In the 2D facial model, for example, the positional relations between salient points on some facial parts (e.g., end points of eyes and eyebrows) are used as facial features. These facial features are fully con-

cise to make the computational cost reasonable. It is, however, difficult to determine appropriate facial features because they should be manually defined. In this paper, we propose an FER method on the basis of geometric-based feature. The proposed method provides the automatic selection method of facial features with low computational cost.

Most of existing FER methods including the aforementioned ones are only able to judge whether a certain facial expression appears or not. It is, however, insufficient because the facial expressions can vary depending on the intensity of emotions. We thus propose a method not only to discriminate facial expressions but also to estimate the intensity of facial expressions. There are few methods to estimate the intensity of facial expressions[10] and it seems to be difficult to apply them to large-scale databases because of the problem of computational cost. The proposed method solves this issue by providing the method to estimate the intensity of facial expressions with fairly low computational cost.

3. Facial Feature

The proposed method uses a number of facial features computed on the basis of positional relation of several salient points on a face (we call these points *facial feature points*).

3.1. Facial feature points

We use 59 facial feature points as shown in Fig. 1. They consist of salient points on left and right eyebrows (10 points), left and right eyes (22 points), a nose (9 points), a mouth (14 points), and left and right nasolabial folds (4 points). They are obtained by using a publicly available software called Luxand FaceSDK (version 4.0).[13]
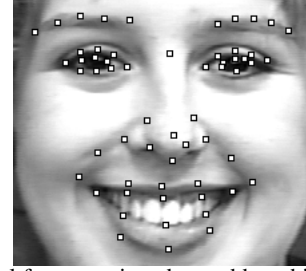

Fig. 1. Facial feature points denoted by white squares (this facial image is from Cohn-Kanade AU-Coded Facial Expression Database[12]).

3.2. Feature values

The feature value used in the proposed method is computed as the cosine of the angle ($\cos\theta$) between two line segments formed by three facial feature points p_i, p_j, and p_k as shown in Fig. 2.
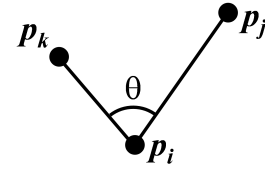

Fig. 2. A facial feature.

Fig. 3 shows an example of the facial feature when p_j and p_k are the end points of the mouth and p_i is the center point of the left eye.
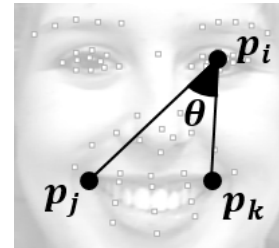

Fig. 3. An example of a facial feature.

The facial feature value $f_{i,j,k}$ computed from p_i, p_j, and p_k is defined as Eq. (1).

$$f_{i,j,k} = \frac{X_{ji}X_{ki} + Y_{ji}Y_{ki}}{\sqrt{X_{ji}^2 + Y_{ji}^2}\sqrt{X_{ki}^2 + Y_{ki}^2}}. \tag{1}$$

Here, X_{ji}, X_{ki}, Y_{ji}, and Y_{ki} are defined as Eq. (2).

$$X_{ji} = x_j - x_i, \; X_{ki} = x_k - x_i, \qquad (2)$$
$$Y_{ji} = y_j - y_i, \; Y_{ki} = y_k - y_i.$$

where x_a and y_a are the x- and y-coordinates of the facial feature point p_a ($a \in \{i, j, k\}$), respectively.

3.3. Feature selection

A total of 97527 possible facial features can be defined from 59 facial feature points.[*] Using all the possible facial features leads to high computational cost. In addition, there are a lot of useless or redundant facial features. For the purpose of efficient FER, we introduce feature selection to select a small number of useful facial features. The proposed feature selection algorithm is shown in Algorithm 1.

Algorithm 1 Feature selection

1: Divide a training set E into two subsets E_{learn} and $E_{validate}$. E_{learn} is used to construct an FER model. $E_{validate}$ is used to evaluate the constructed model. Note that $E = E_{learn} \cup E_{validate}$ and $E_{learn} \cap E_{validate} = \emptyset$.

2: Do steps 3 to 5 for each possible facial feature $f_{i,j,k}$.

3: Compute feature values for all the training examples in E_{learn}.

4: Construct an FER model using the feature values (how to construct the model is described in Section 4).

5: Compute the recognition accuracy $A_{i,j,k}$ by applying the FER model to all the examples in $E_{validate}$.

6: Select N facial features which have N highest values of $A_{i,j,k}$.

Note that the number of facial features N should be experimentally decided. How to divide E into E_{learn} and $E_{validate}$ should also be experimentally determined.

---

[*] One point which is a common end point of two line segments (i.e., p_i shown in Fig. 2) can be selected from 59 facial feature points. Then, two points can be selected from remaining 58 facial feature points. The number of possible facial features is thus $59 \times {}_{58}C_2 = 97527$.

4. Facial Expression Recognition Method

4.1. Construction of weak classifiers

The proposed FER model is constructed by integrating several classifiers (called *weak classifiers*) into a single more accurate classifier (called *a strong classifier*) on the basis of ensemble learning. By using N facial features selected by the feature selection, N weak classifiers are constructed. That is, a single weak classifier is constructed for a facial feature. In the proposed method, we adopt naive Bayes classifiers as the weak classifiers since they have relatively low computational cost and high accuracy.

However, the weak classifier constructed based on a single facial feature will be too weak to construct a fully accurate strong classifier. To solve this problem, we enhance each weak classifier by incrementally adding the other facial features (including ones which are not selected) to the weak classifier. In the enhancement process, facial features are added to the weak classifier until there is no facial feature which can improve the accuracy of the weak classifier, or the number of facial features added to the weak classifier reaches a predefined number M. This process is shown in Algorithm 2.

4.2. Construction of a strong classifier

A strong classifier is constructed by integrating N weak classifiers using weighted voting. Here, $T_1, \ldots, T_N$ denote N weak classifiers and $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_C\}$ denotes a set of facial expressions to be discriminated. There are, therefore, C types of facial expressions. The prediction of T_i is denoted by $e_i \in \varepsilon$. Then, we define the score S_i of the facial expression ε_i by Eq. (3).

$$S_i = \sum_{j=1}^{N} \alpha_j B(e_j = \varepsilon_i), \qquad (3)$$

where, α_j is the recognition accuracy of T_j for $E_{validate}$. $B(p)$ is a function which returns 1 if p is true and returns 0 otherwise. Hence, it returns 1 if e_j is equal to ε_i. It returns 0 if e_j is different from ε_i.

Algorithm 2 Enhancement of weak classifiers

Input:

F : A set of all the facial features ($|F| = 97527$).

F_{sel} : A set of facial feature selected ($|F_{sel}| = N$).

Output: W : A set of enhanced weak classifiers.

1:    $W \leftarrow \emptyset$.
2:    Select an element f from F_{sel}.
3:    $F_{sel} \leftarrow F_{sel} - f$.
4:    $\Phi \leftarrow \{f\}$.
5:    Construct a weak classifier W_Φ using Φ and compute the recognition accuracy A_Φ of W_Φ for $E_{validate}$.
6:    For each φ ($\varphi \in F - \Phi$), construct a weak classifier using a set of facial features $\Phi \cup \varphi$ and compute its recognition accuracy A_φ for $E_{validate}$.
7:    Obtain φ^* which has the highest accuracy A_{φ^*} (i.e., $\varphi^* = \text{argmax}_\varphi A_\varphi$).
8:    If $A_\Phi \leqslant A_{\varphi^*}$, then $\Phi \leftarrow \Phi \cup \varphi^*$ and go to step 9.
       Otherwise, go to step 10.
9:    If $|\Phi| < M$, then return to step 5.
       Otherwise, go to step 10.
10:   $W \leftarrow W \cup W_\Phi$.
11:   If $F_{sel} = \emptyset$ then output W and finish the algorithm. Otherwise, return to step 2.

The score of ε_i is the sum of the recognition accuracies of the weak classifiers which predict that the facial expression is ε_i. This is a kind of weighted voting using the accuracy as a weight.

The prediction e of the strong classifier is given by Eq. (4).

$$e = \underset{\varepsilon_i}{\text{argmax}}\, S_i. \qquad (4)$$

The prediction of the strong classifier (i.e., the prediction of the FER model) is the facial expression which has the highest score.

5.   Estimation of Facial Expression Intensity

The intensity of the facial expression is estimated on the basis of the score S_i. This is based on the idea that the number of weak classifiers which predict that the facial expression is ε_i is proportional to

the intensity of ε_i. For the convenience of the evaluation of the intensity, the intensity is normalized so that it varies from 0 to 1. We define the intensity I_i of the facial expression ε_i as Eq. (5).

$$I_i = \frac{S_i}{\sum_{j=1}^{C} S_j}. \qquad (5)$$

Since the intensity can be directly computed from the scores, additional computation cost to estimate the intensity is considerably small.

6.   Experiment

The proposed method is evaluated through the experiments of FER and facial expression intensity estimation from the viewpoints of accuracy and efficiency.

6.1.   Data set

A total of 12 video clips were used for this experiment. The facial expressions of 8 subjects (denoted as A, B, C, D, E, F, G, and H) were recorded in the video clips. All of them were male university students. Each of the subjects A, B, C, and D has one video clip. These video clips are referred to as A1, B1, C1, and D1, respectively. Each of the other four subjects has two video clips. These video clips are referred to as E1, E2, F1, F2, G1, G2, H1, and H2, respectively. In the video clips, the facial expressions of the subjects while playing cards were recorded so that spontaneous facial expression should be included. We chose appropriate card games such as Hearts which triggers frequent changes of facial expressions.

The size of each video clip is 640×480 pixels and the frame rate is 30 frames per second. The frame images extracted from the video clips were used for the FER and the estimation of facial expressions. Considering the high frame rate, the frames from each video clip after every 10 frames were used to reduce the computational cost. The video clips recorded are not shown in this paper because of privacy reasons.

Most of the facial expressions observed in these video clips were smiles. In this experiment, therefore, we evaluate the performance of the proposed

method in distinguishing smiles from the other facial expressions (mainly neutral faces) and estimating the intensity of smiles. Table 1 shows the number of frames and the ratio of frames with smiling faces in each video clip. Prior to the experiment, a male university student was judged whether the facial expression was smile or not.

Table 1. Number of frames and ratio of smile in each video clip.

Video	#frames	Ratio of smile
A1	2427	10.5%
B1	1585	29.6%
C1	2004	26.8%
D1	1616	29.4%
E1	1361	16.6%
E2	723	30.8%
F1	2203	14.2%
F2	1730	18.6%
G1	2333	12.9%
G2	1703	10.3%
H1	2302	19.6%
H2	1436	26.5%

6.2. Experimental settings

Each video clip was divided into two halves. The one was used for the training set and the other was used for the test set. That is, we performed two-fold cross validation. The number of facial features N was set from 1 to 10. We randomly chose two thirds of examples in E (i.e., a training set) for E_{learn}. The remaining examples belonged to $E_{validate}$. The parameter M (shown in Algorithm 2) was experimentally set to 5.

6.3. Accuracy of facial expression recognition

The recognition accuracy of the proposed method is evaluated by using F-measure. The F-measure Fm is computed by Eq. (6) on the basis of recall R and precision P.

$$Fm = \frac{2 \cdot R \cdot P}{R + P}. \tag{6}$$

Recall R and precision P are given by Eq. (7) and (8), respectively.

$$R = \frac{|\Omega \cap \widehat{\Omega}|}{|\Omega|}, \tag{7}$$

$$P = \frac{|\Omega \cap \widehat{\Omega}|}{|\widehat{\Omega}|}, \tag{8}$$

where, Ω is a set of examples of smiles and $\widehat{\Omega}$ is a set of examples predicted as the example of smiles.

The recognition accuracy of the proposed method is shown in Fig. 4. The horizontal axis shows the number of facial features N and the vertical axis shows the average F-measure of all the video clips.
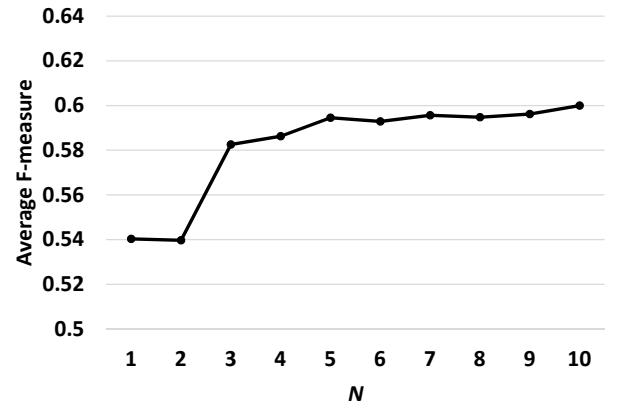

Fig. 4. Facial expression recognition accuracy for each value of N.

The accuracy is improved by adding facial features when $N < 5$. However, the improvement becomes quite small when $N > 5$. From this result, selecting five facial features seems to be adequate.

Next, we compare the proposed method with two existing methods. The one is support vector machine (SVM) which is widely used for various classification problems. The other is random forest which is a kind of ensemble learning method using an ensemble of decision tree classifiers. We used Weka implementations for both methods.[14] We used RBF (radial basis function) kernel for SVM with Gamma = 1 since linear SVM was quite inaccurate. Random forest was performed with the default setting of Weka. The number of facial features N was fixed

to 5. The facial features used in SVM and random forest were the same as those used in the proposed method.

The FER accuracy is shown in Fig. 5 for each video clip. Additionally, the average accuracy of all the video clips is also shown in the figure.

In 9 video clips out of 12 ones, the proposed method performs the best. On average, the proposed method is more accurate than the other methods. In addition, the proposed method seems to be effective for the videos such as A1, E1, and G1. They are relatively difficult examples because the ratios of smile are relatively low (this means the insufficiency of the positive training samples) and weak smiles are often observed in these videos. Therefore, this experimental result shows the effectiveness of the proposed method.

6.4. Efficiency of facial expression recognition

The efficiency of the proposed method is evaluated on the basis of the execution time for FER. Fig. 6 shows the execution time to recognize the facial expressions of all the frames. The execution time for each video clip and the average execution time of all the video clips are shown in the figure. A computer with an Intel Xeon E5 (3.7GHz) processor and 32GB memory was used for this experiment. Note that the time to obtain facial feature points is not included because this process is done by an existing software.
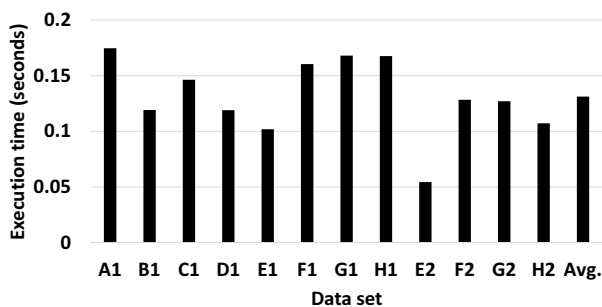

Fig. 6. Execution time for FER.

The execution time varies depending on the video clip due to the difference of the frames in-

cluded in each video clip. On average, it takes approximately 0.132 seconds to process a 10-minute video clip. This result indicates the efficiency of the proposed method. Note that the additional processing time to estimate the intensity of facial expression is ignorable since the additional computation cost is very small as described in Section 5.

6.5. Estimation accuracy of facial expression intensity

Finally, we evaluate the estimation accuracy of facial expression intensity of the proposed method. The accuracy is assessed by comparing the intensity computed by the proposed method with the one manually determined. Manually determined intensity is divided into four levels; "smile does not appear (level 0)," "smile weakly appears (level 1)," "smile moderately appears (level 2)," and "smile strongly appears (level 3)." Prior to the experiment, the intensity is determined by the same university student described in Section 6.1. Since the intensity computed by proposed method varies from 0 to 1, the levels 0, 1, 2, and 3 of the manually-determined intensity are converted into the intensity values 0, $\frac{1}{3}$, $\frac{2}{3}$, and 1, respectively.

Due to space limitation, we only show the estimation result of F2 which has moderate recognition accuracy in Fig. 7. In this figure, horizontal axis shows the frame number (the frame number of the n-th frame from the beginning of the video clip is n). The vertical axis shows the intensity of smile. "Estimated" and "Manual" correspond to the intensity values estimated by the proposed method and manually determined, respectively. For the readability of the intensity values estimated by the proposed method, we show the moving average with a five-frame length window.

In the frames which have high values of manually-determined intensity, the estimated intensity values also tend to be high. The estimated intensity values of most of the other frames are below $\frac{1}{3}$. The proposed method might occasionally fail to accurately estimate the intensity due to the occlusions such as hands or cards in front of a face (in such a case, the positions of the facial feature points
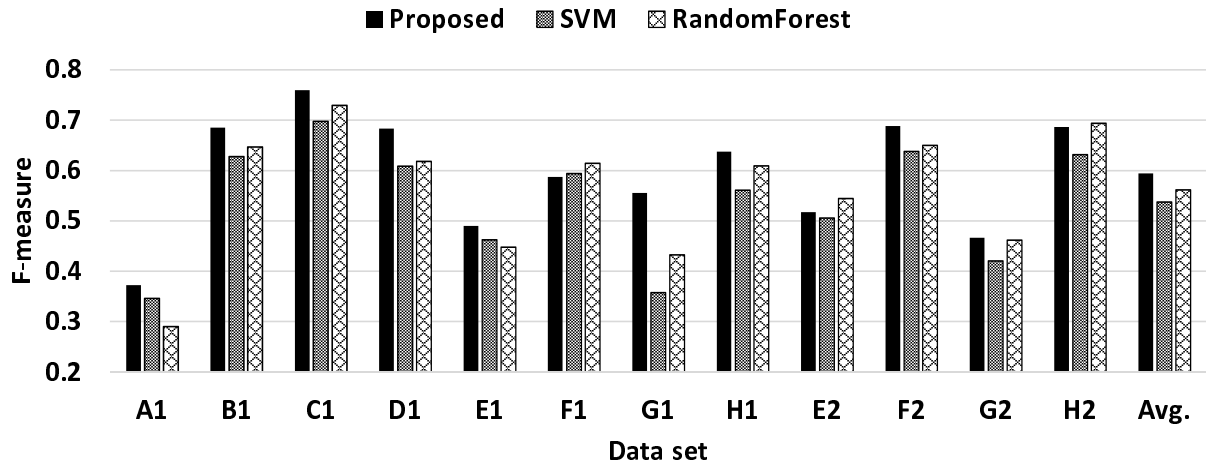
Figure 5: Facial expression accuracy of each method.

become inaccurate). However, considering that the correlation coefficient between the intensity values of "Estimated" and "Manual" is 0.683, the proposed method could appropriately estimate the intensity of smile in many scenes.

7. Conclusion

In this paper, we proposed an FER method and an estimation method of the intensity of facial expressions on the basis of ensemble learning. The proposed method automatically selects useful facial features by means of feature selection. The facial features can be computed fast from the positional relation between the facial feature points. This enables the proposed method to efficiently perform FER and the intensity estimation.

We evaluated the proposed method through the experiment to discriminate and estimate the intensity of spontaneous smiles. The experimental result showed that the average FER accuracy of the proposed method was higher than widely-used classification methods such as SVM. The time to recognize facial expression indicated that the proposed method could quickly perform FER. We confirmed

that the intensity of smile was appropriately estimated to some extent.

In the experiment, we only dealt with the FER of smiles. Evaluating the proposed method for a wider variety of facial expressions is included in the future work. Since all of the subjects were male university student, conducting further experiment with various people (in gender and age) is also the future work.

References

1. Y. Tian, T. Kanade, and J. F. Cohn, Facial Expression Recognition, in *Handbook of Face Recognition*, eds. S. Z. Li and A. K. Jain (London, Springer, 2011), pp. 487–519.
2. M. Lyons and S. Akamatsu, Coding Facial Expressions with Gabor Wavelets, in *Proc. 3rd Int. Conf. Automatic Face and Gesture Recognition*, (Nara, Japan, 1998), pp. 200–205.
3. X. Feng and M. Pietikäinen, Facial Expression Recognition with Local Binary Patterns and Linear Program-
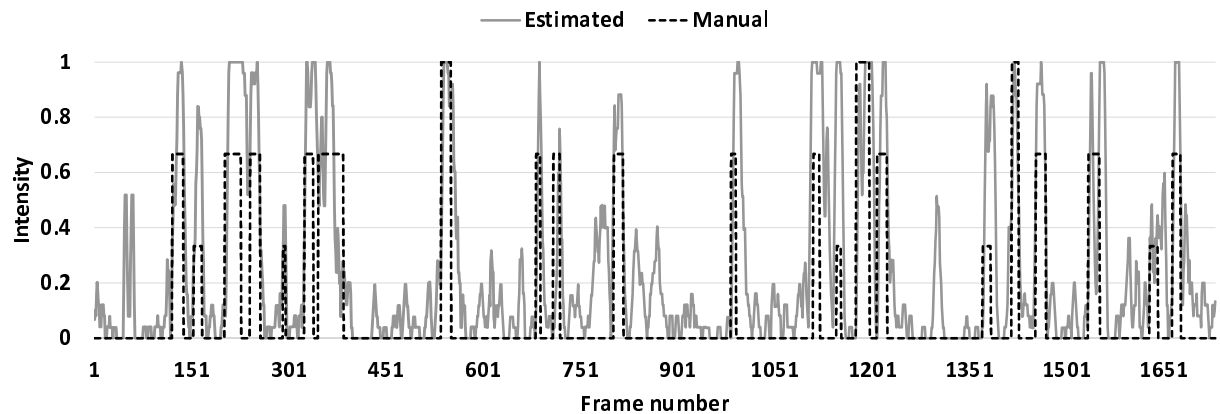
Figure 7: The facial expression intensity estimated for video clip F2.

ming, *Pattern Recognition and Image Analysis* **15** (2) (2005), 546–548.

4. T. R. Almaev and M. F. Valstar, Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition, in *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction*, (Geneva, Switzerland, 2013), pp. 356–361.

5. H. Soyel and H. Demirel, Facial Expression Recognition Using 3D Facial Feature Distances, in *Proc. 4th Int. Conf. Image Analysis and Recognition*, (Montreal, Canada, 2007), pp. 831–838.

6. G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, Recognition of 3D Facial Expression Dynamics, *J. Image and Vision Computing* **30** (10) (2012), 762-733.

7. N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann, Real-time Facial Expression Recognition Using a Fuzzy Emotion Model, in *Proc. Int. Conf. Fuzzy Systems*, (London, UK, 2007), pp. 1–6.

8. I. Hupont, E. Cerezo, and S. Baldassarri, Sensing Facial Emotions in a Continuous 2D Affective Space, in *Proc. Int. Conf. Systems, Man, and Cybernetics*, (Is-

tanbul, Turkey, 2010), pp. 2045–2051.

9. H. Nomiya, A. Morikuni, and T. Hochin, An Emotional Scene Retrieval Framework for Lifelog Videos Using Ensemble Clustering, *J. Software Innovation* **3** (3) (2015), 1–13.

10. O. Rudovic, V. Pavlovic, and M. Pantic, Multi-output Laplacian Dynamic Ordinal Regression for Facial Expression Recognition and Intensity Estimation, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, (Providence, USA, 2012), pp. 2634–2641.

11. T. G. Dietterich, Ensemble Methods in Machine Learning, in *Proc. 1st Int. Workshop on Multiple Classifier Systems*, (Sardinia, Italy, 2000), pp. 1–15.

12. T. Kanade, J. F. Cohn, and Y. Tian, Comprehensive Database for Facial Expression Analysis, in *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, (Grenoble, France, 2000), pp. 46–53.

13. Luxand Inc., Luxand FaceSDK, http://www.luxand.com/facesdk [February 19, 2016].

14. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/ [February 19, 2016].