Predicting and Improving the Recognition of Emotions

Ayesha Hakim
Waikato Institute of Technology (WINTEC), New Zealand
ayesha.hakim@wintec.ac.nz

Abstract

The technological world is moving towards more effective and friendly human computer interaction. A key factor of these emerging requirements is the ability of future systems to recognise human emotions, since emotional information is an important part of human-human communication and is therefore expected to be essential in natural and intelligent human-computer interaction. Extensive research has been done on emotion recognition using facial expressions, but all of these methods rely mainly on the results of some classifier based on the apparent expressions. However, the results of classifier may be badly affected by the noise including occlusions, inappropriate lighting conditions, sudden movement of head and body, talking, and other possible problems. In this paper, we propose a system using exponential moving averages and Markov chain to improve the classifier results and somewhat predict the future emotions by taking into account the current as well as previous emotions.

Keywords

principal component analysis, short time mood learner, markov chain, exponential moving average

1. Introduction

Humans tailor their interpersonal relationships by recognising emotions. This helps them cope with specific situations, such as realising when somebody else is annoyed. Research findings specify the importance of emotions in learning, decision making, and rational thinking (Picard, R. W. 1995). Given the importance of this for human communication, in order to communicate intelligently with us, computers will need the ability to recognise, express, and respond to our emotions (Jeon, M. 2017), as well as to synthesise their own.

Automatic emotion recognition using facial expressions is a cutting edge research area and day by day new better methods are being introduced, but what we noticed is that all of these methods rely mainly on the apparent facial expressions. This recognition mostly ignores the history of emotions, in other words:

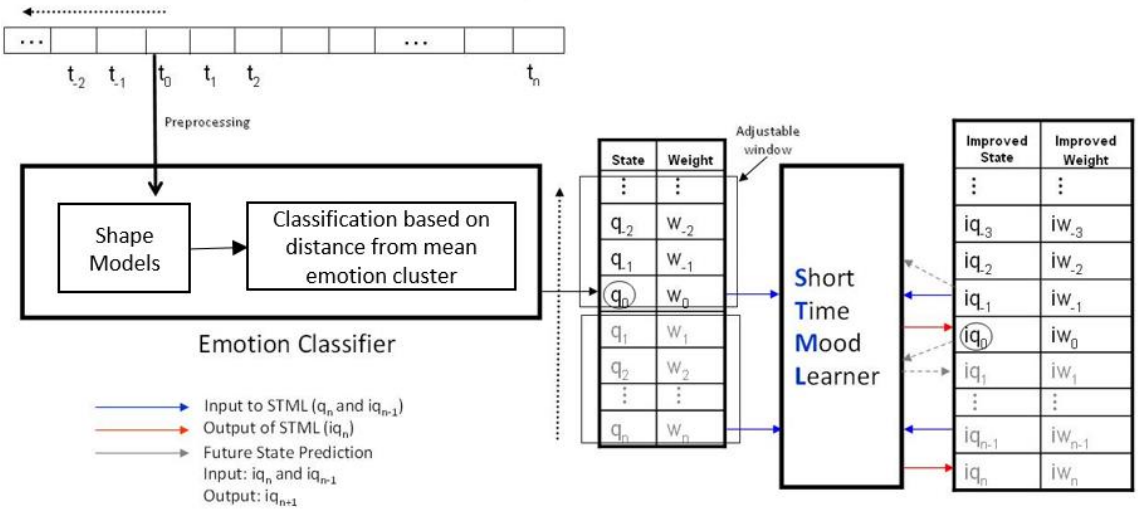


Figure 1: The proposed system. For explanation, see text.

the ‘mood’ of a person. As a result, these methods are completely dependent on the classifier results that are based on the current facial expression. However, the results of classifier may be badly affected by the noise including occlusions, inappropriate lighting conditions, sudden movement of head and body, talking, and other possible problems.

In this paper, we propose a system (Figure 1), which improves the classifier results by taking into account the current as well as previous emotions before giving a final outcome. The proposed system attempts to predict the possible future emotion on the basis of the current emotional state.

2. Methodology

2.1 Dataset

In order to train and evaluate our system, we use the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) (Busso, C., et al., 2008), collected by the Speech Analysis and Interpretation Laboratory (SAIL) at University of South California (USC). In this dataset, ten actors were recorded using a high speed cameras capturing 120 frames per second. One of the actors had a set of reflective markers on their face (Figure 2) and the 3D position of these markers was tracked with very high accuracy. We based our analysis on the locations of these marker points.

The videos of the actors were watched by three independent human evaluators who labelled the data into categorical labels (neutral, happiness, anger, sadness, surprise, disgust, fear, frustration, and excitement) as well as psychological data about emotion intensity (valence, activation, and dominance). However, the evaluators did not always agree. Each dialog consists of almost 25 utterances/turns of each actor with an average of 4.5 seconds each. An utterance is a sentence or similar period during which one actor talks continuously. It is these utterances that were annotated by the three human evaluators into categorical labels (neutral, happy, angry, sad, surprise, disgust, fear, frustration, and excitement) as well as psychological data about emotion intensity (valence, activation, and dominance). Although nine emotional labels were used by the humans, we chose to consider only six of them, as for the missing emotions (disgust, surprise, and fear) there was insufficient data. For further details on any part of the data capture and labelling, see (Busso, C., et al., 2008).

It was assumed that within an utterance, there is no emotion transition (e.g., from happy to excited), however, the evaluators were allowed to select more than one emotion category to describe mixtures/blends of emotions, which are more common in natural communication. The estimated confusion matrix between the assigned categorical labels shows that there is an overlap between happiness and excitement as well as anger and frustration. Neutral, disgust, and anger are often get confused with frustration. Also, sadness is often confused with frustration and neutral.

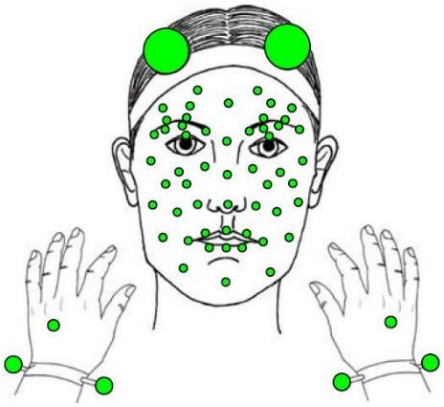


Figure 2: Layout in IEMOCAP data set. 53 markers attached to actor’s face (Busso, C., et al., 2008).

The evaluation-level limitation of the selected dataset is associated with its utterance-based annotation technique. Assuming the emotional content did not change much within an utterance, the same label (or mixture of labels) was assigned to each frame in that utterance. The 'silent' frames and those containing sounds of active listening like 'mmhh' were not annotated at all. Another problem of the chosen dataset is that the human evaluators were sometimes inconsistent in labelling the data, since each one of them perceived and evaluated the emotions associated to an utterance in his own way. Due to the subjective nature of emotions, the inter-evaluator inconsistency is a common problem of all emotion datasets.

2.2 Data Preprocessing

We used the locations of the marker points in 3D as the basis of our analysis, and chose 5,000 frames of each of five emotions (happiness, excitement, anger, frustration, and sadness) and neutral state for each of the ten actors to form a training set of 300,000 frames. It should be noted that the marker points were already aligned to make the nose marker at the centre of each frame that removed any translation effects. The rotational effects were compensated by multiplying each frame by a rotational matrix. For details about markers alignment, refer to (Busso, C., et al., 2008).

For the training set, we took frames from the utterances where all three experts agreed. We used six emotions rather than the full nine as for the missing emotions (disgust, surprise, and fear) there was insufficient data, sometimes as little as 2,000 frames in total.

Out of the six selected emotions, two (frustration and excitement) are the candidate basic emotions (Ortony, A., & Turner, T., J., 1990) and (Ekman, P., 1999). For the testing set, there was no such condition of agreement by all three experts while choosing the frames.

Each frame of the dataset contains the motion capture information of 61 markers in 3 dimensions, so the training data was of size 300,000 * 183 dimensions. We reduced the dimensionality of the data for each frame in three ways:

- Markers not on the face (such as the head and hands) were excluded.
- Markers that did not move significantly (such as eyelids and nose) were removed.
- Sets of markers that moved together (such as, points on the chin and forehead) were replaced by a single point at the centre of the set.

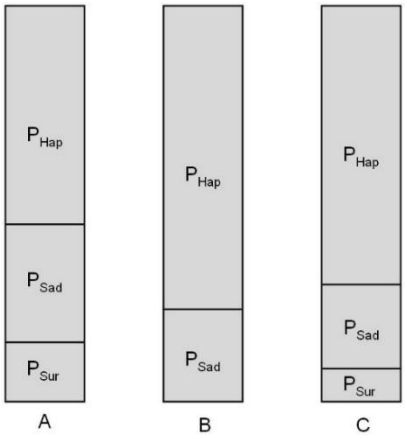


Figure 3.1: Comparison of two ways (B and C) to change the probabilities in (A).

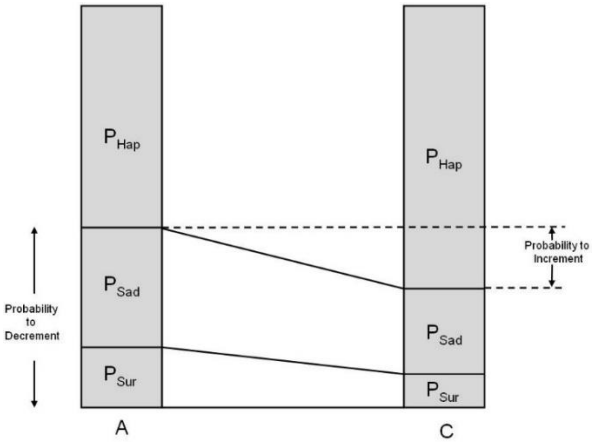
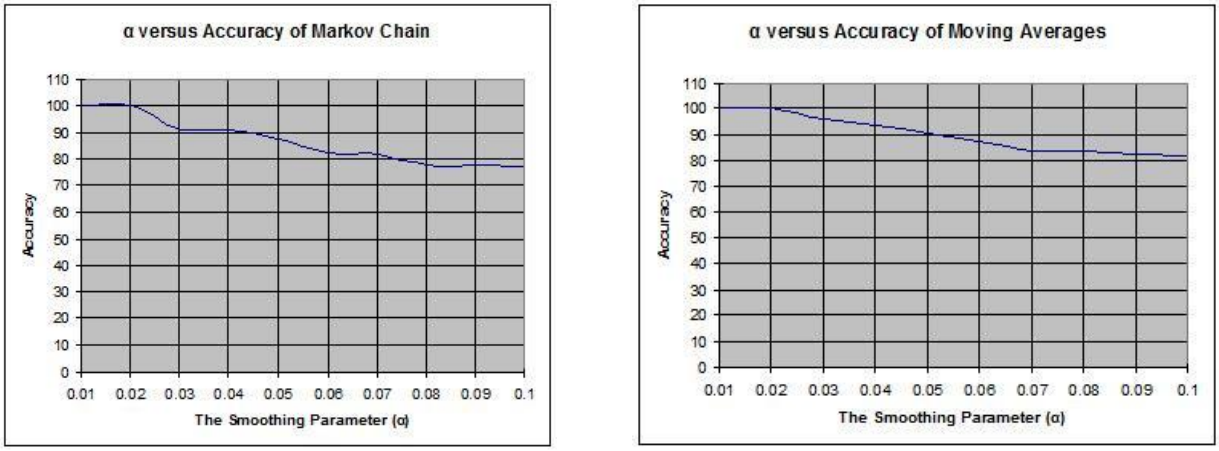


Figure 3.2: The ratio of change in probabilities.



(a) (b)

Figure 4: Accuracy decreases by increasing the smoothing parameter (α).

The proposed system consists of two components: an Emotion classification based on shape models and the Short Time Mood Learner (STML) as shown in Figure 1. Each of these is discussed below.

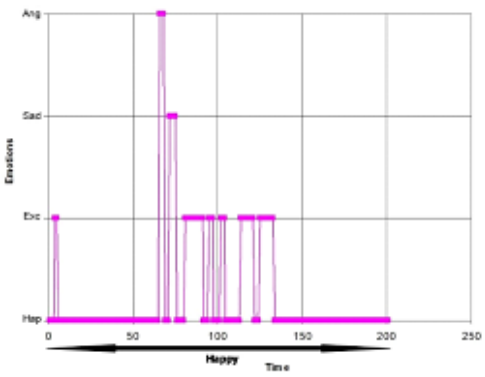
2.3 Emotion Classifier

Based on the data from IEMOCAP we had sets of facial points with an associated emotion label. As described in Section 2.2, we are using the 3D locations of 28 marker points on the face for emotion recognition and analysis. We then used Principal Component Analysis (PCA) to develop a model of only one actor at a time that we called the speaker-dependent model.

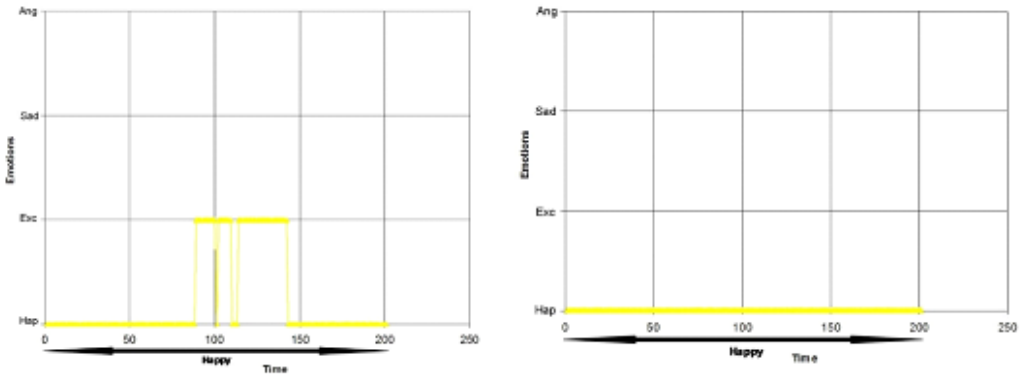
The speaker-dependent model is trained on 5000 frames of each emotion of one actor only to form a training set of 30,000 frames. We then used PCA to develop a face model of the given training set based on the method described in (Hakim, A., Marsland, S. R., & Guesgen, H. W., 2013). We noticed that the first 7 PCs covered 93% of the total variation of the training data out of which the first PC, which covers almost 50% of the total variation, was describing the upward and downward movement of the mouth points. This movement of lips was experimentally shown to be highly correlated with talking, which is not directly connected with emotion recognition, and not much else, and so we discarded the first PC. Consequently, we chose to use six PCs (2-7) of the face model for our analysis. For details about the effects of each principal component on the mean face, refer to (Hakim, A., Marsland, S. R., & Guesgen, H. W., 2013). We transformed the training data into the 6-D space of the selected six principal components. Each data point was then labelled with the majority vote of the three human experts, so that the training set consists of 30,000 points, each labelled with one of six emotions in the 6-D space.

To evaluate the speaker-dependent face model, we chose the continuous testing frames of the same actor on which the model was trained. That was the ideal situation as there was only one face involved leading to no speaker variability and the first PC correlated to talking was removed leading to no lexical variability. For classification of a test frame, it was transformed into the 6-D space of the dependent face model. We then computed the Mahalanobis distance between the test frame and the six emotion clusters. The test frame is classified as the emotion with the minimum distance among the six emotion clusters.

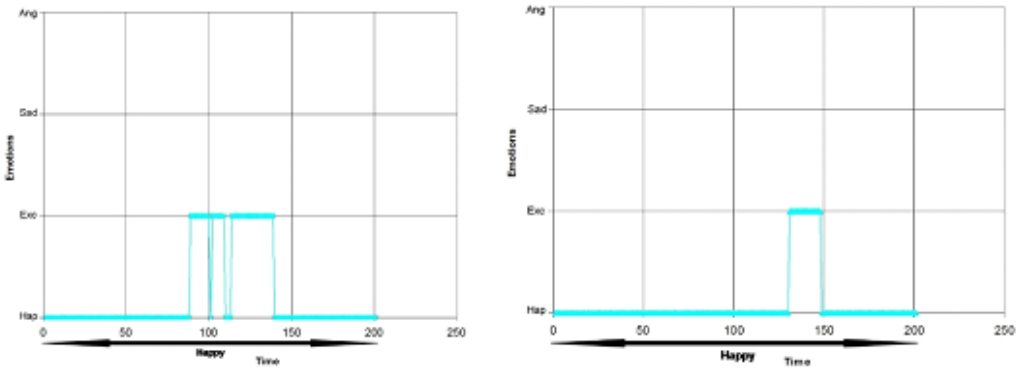
Since, the data is recorded at the speed of 120 frames per second so the classifier is taking into account very minute changes which are not detectable by human eyes (since human eye can process ≈10-12 frames per second). Also, since in the naturalistic communication, the speakers are talking and moving their faces,



a) The Classified Emotions (Hap: Happy, Exc: Excited, Ang: Angry)



b) Smoothing by Moving Averages



c) Smoothing by Markov chain

Figure 6: The affect of different values of α on smoothing by moving averages and Markov chain on 200 frames of “Happy” emotion.

their expression is not always clearly recorded. Due to the same reason, the human evaluators often misclassified an utterance into a different expression than the context of the dialog. This is what we noticed in the classifier results that frames that are annotated as “happy” frames are not always happy; these are sometimes misclassified as excited and frustrated.

This problem is tackled by the proposed system by deciding the emotion category on the basis of current as well as previous emotional states. Based on the experiments, we found that the classification done by

the proposed system is more stable than that of the classifier alone. The results are described in Section 3.

2.4 Short Time Mood Learner (STML)

Short Time Mood Learner is based on a continuous temporal process which is taking into account the current as well as previous emotion based by the history of previous emotions. Based on the reasons discussed in Section 2.3, we cannot always trust the classifier results alone. The proposed method improves the classified results by removing the noise caused by sudden movement of head, face and that caused by misinterpretation of facial expression due to recording data at a very high speed.

We applied two methods to test our proposed method: the exponential moving average (EMA) (Everett, J. E., 2011) and discrete-time Markov chain (Gardiner, C. W., 1986). This paper aims to describe and compare these methods on the time series of human emotions.

2.4.1 Exponential Moving Average (EMA)

Exponential moving average, also called exponentially weighted moving average (EWMA), was first introduced in (Roberts, S. W., 1959), as the geometric moving average which gives the greatest weight to the recent observation, and decreases the weight of all previous observations in geometric progression from the most recent back to the start. Since then, it has been widely used to measure drifts from the target, quality control and forecasting by exponential smoothing (Everett, J. E., 2011). For further explanation of EMA for time-series analysis, refer to (Diebold, F. X., 1998), (Box, G. E. et al., 2015), and (Ramjee, R. et al., 2002).

We propose to use EMA as a tool to minimise the noise of classified emotions by assigning higher weights to the recent observations. Firstly, SMA often lags behind the original data since it is based on past records and secondly our current emotion mostly depends on our nearest past emotion (Levine, L. J. et al., 2001) so it is necessary to assign higher weight to the most recent emotions (short-time mood). Mathematically, EMA can be defined as,

$$\mathrm{EMA} = S_t = \alpha * P_t + (1 - \alpha) * S_{t-1}$$

where α is the weight of the current probability P_t is the current probability, S_{t-1} is the previous average value (Everett, J. E., 2011).

In EMA, the probabilities can only get the value of 0 or 1, while the weights are continuously varying on the basis of classified emotions. For instance, if a frame is classified as “happy”, its probability is set to 1 and all other probabilities are set to 0. On the basis of these probabilities, the average value (s_t) is calculated, which is then shifted to the next calculation. EMA also enables the system to predict the future emotion by slightly changing the formula such as,

$$S_{t+1} = \alpha * P_t + (1 - \alpha) * S_t$$

Once we get the current emotion, EMA can forecast the future one by using the current probability and average.

2.4.2 The first-order Markov Chain

The Markov chain is a discrete-time stochastic process assuming the Markov property. By discrete-time process, we mean the process indexed by countable number of time points. Markov property is such that for each t_n, $P(q(t_n)|q(t_{n-1})|q(t_{n-2})|\dots|q(t_0))$ is equal to $P(q(t_n)|q(t_{n-1}))$, where t represents time, $q(t_n)$ is the state of the chain at time t_n. The Markov chain based on this Markov property is called *first*-order Markov chain. It may be extended to n-order chain depending on the specific problem (Gardiner, C. W., 1986). We are using the *first*-order Markov chain, such that the current emotion $q(t_0)$ is based only on the previous emotion $q(t_{-1})$ and not on the long sequence of previous emotions. One of the reasons for this is

discussed in the previous section. Another reason is that the system is continuously learning and even if we look at only one previous emotion, it would still make the decision based on the history of emotions (short-time mood) after proper training. Moreover, it simplifies the process to get a computationally efficient method.

Basically, it is a game of rewards and punishments. The system is initialized with an arbitrary data. For instance, if the current test frame is classified as “happy”, there is a high probability of it being classified as “happy” than “sad” in the next frame. Similarly, there is more probability of a test frame being classified as “surprise” after “happy” than classified as “sad”. This arbitrary data is required to start the chain process and the system would refine itself by updating based on the classifier result.

The Markov emotion is obtained by assigning the current classified emotion and the previous Markov emotion to STML (as presented in Figure 1). STML checks the probabilities such that if the probability of classified emotion given the previous Markov emotion is greater than or equal to a threshold, the classified emotion is the considered to be the right emotion and will be assigned as the Markov emotion. Otherwise the new Markov emotion will be the same as the previous one. Unlike EMA where the probabilities can only get the fixed value of 0 or 1, in Markov chain the probabilities of emotion’s transitions are continuously varying.

Since the data is recorded at a rate of 120 frames per second, we may improve the efficiency of the system by skipping k number of frames, where the value of k can be controlled depending on the application of the system.

In our experiments, we chose to select $k = 3$ to calculate the Markov emotion. The system calculates the rewards on the basis of current Markov emotion and the current probabilities. In order to avoid the normalization problems, the probabilities should increase asymptotically based on the following formula:

$$P_t = P_{t-1} + \alpha(1 - P_{t-1}) \qquad (1)$$

where P_t is the current probability, P_{t-1} is the previous probability and α is the smoothing parameter equal to the product of number of occurrences of the specific emotion and the amount of reward (which is currently a constant parameter).

Based on this formula, the greater the number of occurrences, the bigger would be the reward. However, the system is not so generous; it will punish the probabilities of other relevant emotions pro rata. Refer to Figure 4 (a), consider a series of three emotions $\{Hap, Sad, Sur\}$, if A represents the initial probabilities of emotions and P_{Hap} is increased (by eq. 1), the other two probabilities P_{Sad} and P_{Sur} should be decreased to produce the distribution shown in C, rather than B. The ratio of change in probabilities is presented graphically in Figure 4 (b).

Like EMA, the Markov chain also enables the system to predict the future emotion. As mentioned above, the system is currently using first-order Markov chain, so it just needs the current emotion in order to predict the future one. The emotion with the highest probability would be considered as the expected future emotion. Depending on the type of application for which the system is being used, it may be updated to second or higher order Markov chain.

2.4.3 How to choose the value of the smoothing parameter ($\boldsymbol{\alpha}$)?

The literature reveals that the value of $\boldsymbol{\alpha}$ should be chosen based on the purpose for which exponential moving average is being used (Everett, J. E., 2011). In both methods, $\boldsymbol{\alpha}$ is the controlling factor such that by increasing the value of $\boldsymbol{\alpha}$ the system would accept the change easily, while by decreasing it would make the model resistant to accept any change. In Markov chain, $\boldsymbol{\alpha}$ refers to the amount of reward as discussed before. Depending on the purpose of using the system and the level of acceptable noise, the value of $\boldsymbol{\alpha}$ may be changed accordingly. In our case, the purpose is to smooth the classifier results which appear as a result of noise, we should not accept the sudden changes which appear for very short duration. We made several experiments to observe the effect of different values of $\boldsymbol{\alpha}$ on the test frames that have been classified as one emotion by all three evaluators.

Figure 5 shows the effect of different values of α on the accuracy of two methods (EMA and Markov chain). Figure 6 shows more clearly the result of choosing two different values of α on the test frames

classified as “happy” emotion. This effect remains the same for the frames containing emotion’s transitions. It is evident that the smaller value of α smoothes the classifier results better in our dataset.

3. Results

By comparing the results of EMA and Markov chain, we found that EMA is better than Markov chain in accuracy as well as the execution time, if we choose proper controlling parameters. EMA smoothes the data containing 2515 frames, with the accuracy of 96.5% in 25 milliseconds per frame while the accuracy of Markov chain process is 91% in the total time of 41 milliseconds per frame. The performance of both methods is tested on a 64-bit Operating system with Intel Core 2 Duo CPU @ 2 GHz and RAM of 2 GB.

Figure 7 shows the comparison graph of classified and improved emotion by Markov process and moving average technique on emotion’s transitions. Figure 8 shows the results of emotion prediction using EMA and Markov chain. Both of these figures are representing the results of test frames that have been classified as a series of Happy (195 frames), Angry (531 frames), Sad (340 frames), Angry (264 frames), Sad (984 frames), and happy (200 frames). Both of them behave quite similarly, however EMA is a bit stable.

Up till now, the proposed system is tested on human emotions recognized by facial changes, but it is not restricted to it. As the Figure 1 shows that the proposed system is “classifier friendly”. By using the proposed system, the results of any classifier which is able to recognise emotions or action units reliably may be improved based on its application.

4. Conclusions

The last few decades in human computer interaction have seen enormous advances in the field of affective computing. Extensive research has been done on automatic human emotion’s recognition using facial expressions but the literature reveals that the results of all of these methods rely mainly on the classifier’s response. However, the classifier’s results may be badly affected by noise due to occlusions, inappropriate lighting conditions, sudden movement of head and body, talking and several related problems. There is a high need to improve the classifier results before further processing.

We proposed a model which minimise this noise by taking into account the current emotion as well as the short time *mood* of the person. We have tested this model on the IEMOCAP dataset using two methods: exponential moving averages and the first-order Markov chain. These methods have successfully improved the results of classifier by choosing the appropriate controllable parameters which include the window size and the smoothing parameter (α). We have shown the effect of different values of α on the accuracy of smoothing our dataset. The proposed model is quite flexible in the choice of the classifier and the parameters may be adjusted on the basis of specific application. This flexibility has increased the scope of the proposed model beyond this particular problem of human emotion's recognition.

Furthermore, as suggested by Moltchanova, E., & Bartneck, C., (2017), individual differences effect the recognition and perception of emotions. Including the individual differences in the proposed model might improve the recognition and prediction of future emotions. This might be used to generate personalised emotion responses based on the individual differences.

Acknowledgements

This work was supported by funds from the Higher Education Commission (HEC) Pakistan. The authors would like to thank Massey University for providing support throughout the research.

References

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335.
Dalgleish, T., & Power, M. (Eds.). (2000). Handbook of cognition and emotion. John Wiley & Sons.
Diebold, F. X. (1998). Elements of forecasting. South-Western College Pub.
Ekman, P. (1999). Basic emotions. Handbook of cognition and emotion, 45-60.
Everett, J. E. (2011). The exponentially weighted moving average applied to the control and monitoring of varying sample sizes. WIT Transactions on Modelling and Simulation, 51, 3-13.
Gardiner, C. W. (1986). Handbook of stochastic methods for physics, chemistry and the natural sciences. Applied Optics, 25, 3145.
Hakim, A., Marsland, S., & Guesgen, H. W. (2013, September). Statistical modelling of complex emotions using mixture of von mises distributions. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 517-522). IEEE.
Jeon, M. (2017). Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods. In Emotions and Affect in Human Factors and Human-Computer Interaction (pp. 3-26).
Levine, L. J., Prohaska, V., Burgess, S. L., Rice, J. A., & Laulhere, T. M. (2001). Remembering past emotions: The role of current appraisals. Cognition & Emotion, 15(4), 393-417.
Moltchanova, E., & Bartneck, C. (2017). Individual differences are more important than the emotional category for the perception of emotional expressions. Interaction Studies, 18(2), 161-173.
Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? Psychological review, 97(3), 315.
Picard, R. W. (1995). Affective computing.
Ramjee, R., Crato, N., & Ray, B. K. (2002). A note on moving average forecasts of long memory processes with an application to quality control. International Journal of Forecasting, 18(2), 291-297.
Roberts, S. W. (1959). Control chart tests based on geometric moving averages. Technometrics, 1(3), 239-250.

Copyright