Real-time Mobile Facial Expression Recognition System – A Case Study

Myunghoon Suk and Balakrishnan Prabhakaran
Department of Computer Engineering
The University of Texas at Dallas, Richardson, TX 75080
{mhsuk, praba}@utdallas.edu

Abstract

This paper presents a mobile application for real time facial expression recognition running on a smart phone with a camera. The proposed system uses a set of Support Vector Machines (SVMs) for classifying 6 basic emotions and neutral expression along with checking mouth status. The facial expression features for emotion recognition are extracted by Active Shape Model (ASM) fitting landmarks on a face and then dynamic features are generated by the displacement between neutral and expression features. We show experimental results with 86% of accuracy with 10 folds cross validation in 309 video samples of the extended Cohn-Kanade (CK+) dataset. Using the same SVM models, the mobile app is running on Samsung Galaxy S3 with 2.4 fps. The accuracy of real-time mobile emotion recognition is about 72% for 6 posed basic emotions and neutral expression by 7 subjects who are not professional actors.

1. Introduction

Increased and extensive use of the mobile camera embedded in smart phones has spawned a wide variety of personal and business applications. The potential role of computer vision technologies to emerging application with smart phone cameras has undergone substantial changes in the user interface beyond the basic function of the camera for taking pictures. For instance, "Face Unlock" feature is already in use in many of smart phones. These kinds of applications for smart phones have become feasible due to increasing computation power in mobile devices and effective solutions associated with computer vision. Automatic facial expression recognition and emotion recognition have been actively studied in a variety of area such as human-computer interaction, robotics, games, education and entertainment. Widespread smart phones have aroused considerable interest in human interaction through users' emotion on smart phones. However, even a couple of years ago, researchers preferred to handling all heavy computer vision jobs on a remote high performance server (by transmitting images or video frames from mobile phones to remote servers) instead of processing them on a mobile device itself. But with the swift advances in processing power and memory, even the real time video processing related to computer vision is within the bounds of possibility in smart phones. Nevertheless, relatively low computation power of a mobile phone compared to ordinary PC still makes it difficult to directly adopt solutions related to facial expression recognition or emotion recognition simply from PC to mobile platform.

We propose an efficient approach for real time video based facial expression recognition running on mobile devices. The proposed system recognizing a user's emotion through a mobile camera in real time without any communication delay to remote servers can be a good starting point for the emergence of a diversity of mobile services and applications. We have tested our system using the *Samsung Galaxy S3* running Android 4.3 *Jelly Bean* with a frame rate of average 2.4 *fps* and the mobile app is available for download (See the link in the Section 5.5).

Our contributions in the paper are as follows. In order to handle the lower processing power in mobile devices (compared to their desktop counterparts), we have proposed very simple but effective approaches for identifying neutral and peak expression video frames. These approaches are based on a set of SVM classifiers that use ASM features for classification of neutral expression frame, as well as for the facial emotion recognition (happy, sad, angry, etc). We show that though the system is able to handle only around 3 frames per second, it is still able to achieve the goal (of facial emotion recognition) with a high degree of accuracy.

The rest of this paper is organized as follows. In Section 2, we describe related works. In Section 3, we explain the proposed system overview. In Section 4, we describe our methods in details. In Section 5, we show the performance of our system, and finally we conclude this paper and give future works in Section 6.

2. Related Work

Many studies on facial expression recognition and analysis have been carried out for a long time because facial ex-

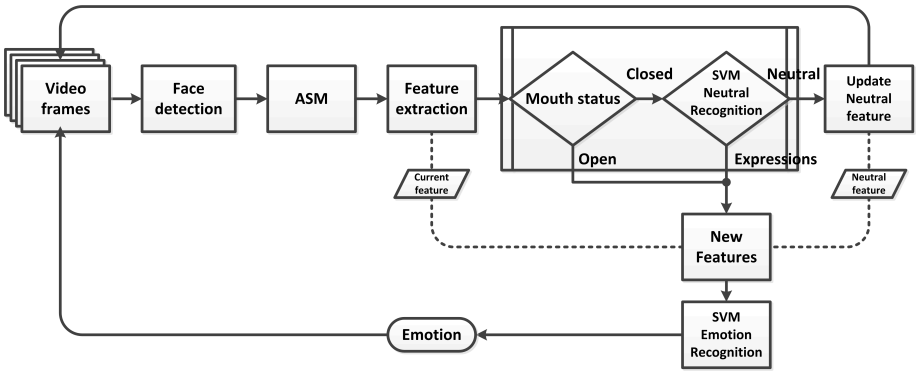

Figure 1: The proposed system architecture.

pressions play an important role in natural human-computer interaction as one of many different types of nonverbal communication cue. Paul Ekman *et al*. postulated six universal emotions (anger, disgust, fear, happiness, sadness, and surprise), and developed Facial Action Coding System (FACS) for taxonomy of facial expressions [7]. Their significant works have formed the basis of the existing automatic facial expression recognition systems.

For automatic facial expression recognition system, a variety of research approaches have been proposed. According to types of facial features that the proposed systems take more interest, they extract geometric features, appearance features, or a hybrid of geometric and appearance features on targeting face. For example, Active Shape Models (ASMs) are one of the most popular statistical models fitted to a new face image which can be successfully used for good geometric features such as measurements among coordinates of landmarks on the face [5]. On the other hand, Gabor wavelets representation and local binary patterns (LBP) are successful appearance features with changes of the facial appearance, e.g. wrinkles and furrows. For hybrid features of shape and appearance, the Active Appearance Model (AAM) is a well-known method of good performance [6].

In addition, there are different approaches for classifying facial expressions in video sequences depending on spatial and spatio-temporal information. For the frame-by-frame approaches relying on static image or only a frame of video sequences without temporal information, diverse classifiers such as Neural Network (NN) [11, 17], Bayesian Network (BN) [4], rule-based classifiers [18], Support Vector Machine (SVM) [2, 12] achieve good results for facial expression recognition. On the other hand, spatio-temporal approaches result in better performance in video sequences, compared to spatial approaches without temporal information. Above all, Hidden Markov Models (HMM) is one of the most popular classifiers among spatio-temporal approaches and works well for facial expression recognition in [4, 2, 10]. Although most systems are obviously interested in achieving high performance in terms of accuracy rate of recognition, they have no great concern about mobile platform.

To the best of our knowledge, real time video based facial expression recognition system that we propose is a rare work for mobile platform. In [1], authors present a use-case scenario, an eBook Reader application where a user can control it by using facial expressions as a natural interaction. But they used a few facial gestures with limited Facial Action Units that are more relevant to eyes, not various facial expressions. In [9], authors present a facial expression system for a smart phone. They employed AAM with a Difference Of Gaussian (DOG) for fixing illumination variation problems. But they showed experiment results with only 4 classes such as sadness, surprise, neutral, and happiness. In [15], authors proposed a system recommending multimedia content to user by understanding a user's current emotional state.

3. Proposed System Overview

We have proposed a real-time mobile application for facial expression recognition system (See Figure 1). Our proposed system follows main steps similar to ordinary facial expression recognition systems such as face detection, feature extraction, and facial expression classification. However, there are other steps to generate features from displacement between neutral and expression frames.

Although dynamic features lead to better results than spatial and static features for facial expression recognition in video sequences, extracting the dynamic features requires additional process such as temporal segmentation. However, the application on mobile device with low computation power needs to be optimized by improving the performance of speed in order to run the application smoothly. Considering this issue, our proposed system adopts a simple and reliable approach such as SVM using dynamic features without complicated temporal segmentation.

The proposed system does not identify the peak of ex-

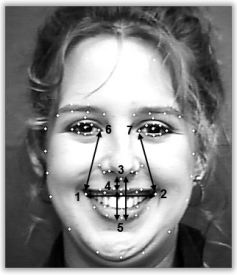

Figure 2: Example of landmarks by STASM and some points and distances as features.

pressions accurately in video sequences though it detects neutral expressions. In other words, the system distinguishes between neutral expression and non-neutral expression frames in video sequences of a facial expression instead of complex temporal segmentation. Therefore, the expression frames can include all non-neutral expressions in the stages of onset, offset and peak of a facial expression.

For creating dynamic features for facial expressions in video sequences, we employ the changes between neutral features and current features. Comparing with displacement of frame-to-frame basis, these dynamic features between neutral and current frames (not neutral frame) make the system less sensitive to frame-to-frame variations. Moreover, even if the application is running with the low frame rate of around 2.4 *fps*, the system usually does not look missing peak frames in practice (most natural expressions reach a peak within a second.) because it can catch at least some non-neutral frames between previous neutral expression frame and the next neutral frames. Therefore, in the case of the system with a low frame rate such as mobile devices, this kind of simple approach will be efficient in the performance of speed without dropping accuracy while the complex approach accurately finding peak frames may be useless in the system with a low frame rate.

As Figure 1 shown, we employ both the SVM model and mouth status to detect neutral expressions in a frame. If the current frame has a neutral expression, the current features are saved as neutral features for creating dynamic features later. If a non-neutral expression is found in the current frame, the system generates new dynamic features by displacement between the saved neutral features and current features. Then the dynamic features are fed into SVM models for facial expression recognition. The next Section 4 explains in details of each process.

4. Methods of Facial Expression Recognition

4.1. Face Detection and Feature Extraction

First, one frame in video stream is grabbed on the mobile device. In a frame, the face is found by Haar cascade facial detection module. Through ASM module implemented with STASM [16], 77 facial landmarks are located on a face, and then based on x, y coordinates of landmarks, 13 high-level facial shape features are generated and normalized. Some of features that we show in Figure 2 are the same points and distances as those used in [8].

Figure 3: Examples of CK+ dataset: (left to right) anger, disgust, fear, happiness, sadness, and surprise expression.

If this face has a neutral expression, the system keeps current features as neutral features which will be used as base features of calculating displacement between features later. Otherwise, if it is not a neutral expression, the system generates new features by calculating the displacement between current features and neutral features.

4.2. Neutral and Expressions Classification

The basic classifier used in the proposed system is a set of SVMs that have been already well known for pattern recognition tasks including both face recognition and facial expression recognition. For developing the mobile application, we use the open source library called libsvm [3].

First, a SVM model as Neutral classifier with Linear kernel function is built by training with neutral faces and others from CK+ dataset which we details in Section 5.1.

In the neutral expression recognition module, the status of the mouth as one of the 13D features is checked. If the mouth is open on detected face, intuitively we determine the face shows no neutral expression (which includes some expressions such as happiness, surprise, and partial anger, disgust, and fear). But if the mouth is closed, the system needs to double-check if it is a neutral expression or not by using the SVM classifier for distinguishing between "neutral" class and "non-neutral" class.

During expression frames, dynamic features by the displacement are used as an input for SVM classifiers of 6 emotions such as anger, disgust, fear, happiness, sadness, and surprise. The SVM classifier is built with radial basis function kernel from CK+ dataset with 6 emotions.

5. Experimental Results

5.1. Facial Expression Dataset

For training SVM classifiers of our proposed system, we use the extended Cohn-Kanade (CK+) database [14]. The CK+ database has 593 video sequences from 123 subjects. Each video sequence in CK+ dataset has been already segmented temporally from neutral frame and to the peak frame of facial expressions for about 10-60 frames. Figure 3 shows examples of 6 emotions in CK+ dataset.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Expressions	0.660	0.116	0.843	0.660	0.740	0.772	0.772
Neutral	0.884	0.340	0.734	0.884	0.802	0.772	0.708
weighted avg.	**0.775**	**0.231**	**0.787**	**0.775**	**0.772**	**0.772**	**0.715**

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Expressions	0.809	0.055	0.933	0.809	0.867	0.877	0.847
Neutral	0.945	0.191	0.840	0.945	0.889	0.877	0.822
weighted avg.	**0.879**	**0.125**	**0.885**	**0.879**	**0.878**	**0.877**	**0.834**

Table 1: The accuracy results for Neutral recognition: (top) Linear kernel function, (bottom) RBF kernel function.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
anger	0.778	0.045	0.745	0.778	0.761	0.866	0.612
disgust	0.831	0.044	0.817	0.831	0.824	0.893	0.711
fear	0.72	0.011	0.857	0.72	0.783	0.855	0.64
happiness	0.928	0.021	0.928	0.928	0.928	0.953	0.877
sadness	0.679	0.025	0.731	0.679	0.704	0.827	0.525
surprise	0.964	0.027	0.93	0.964	0.947	0.969	0.906
weighted avg.	**0.858**	**0.03**	**0.857**	**0.858**	**0.857**	**0.914**	**0.763**

Table 2: The accuracy results for facial expression recognition on CK+ dataset.

	Frame Resolution	# of Frames	Optical Flow	ASM	SVM Neutral	SVM 6 Emotions	Total Frame
Nexus 4	480x320	119	69.4±14.3	416.3±74.5	1.1±0.38	0.9±0.67	526.5±79.24
Galaxy S3	640x480	146	59.1±7.29	318.3±9.79	1.0±0.15	0.6±0.24	420.2±14.97

Table 3: Average computation time and standard deviation (milliseconds) taken in the process of modules of the proposed system. Two different smart phones (*Nexus 4* and *Galaxy S3*) are tested.

For the classification of neutral expression, we used 309 neutral frames and 327 peak frames of 7 emotions for training. For the classification of facial expressions, we took 309 neutral frames and 309 peak frames except contempt emotion of 7 emotions. 309 samples consist of 45 from anger, 59 from disgust, 25 from fear, 69 from happiness, 28 from sadness and 83 from surprise.

5.2. Neutral Expression Recognition

The SVM model for neutral expression recognition is trained with CK+. We show Table 1 for comparison of the accuracy of different kernel functions (top: Linear, and bottom: RBF kernel function). Both of results are obtained by 10-folds cross validation. The SVM classifier with Linear kernel has 77.5% of the accuracy while the RBF has 87.9% of accuracy. Although the SVM classifier with RBF Kernel results in better accuracy, we use the Linear kernel function for our proposed system. Though the SVM classifier with RBF is more optimized to CK+ dataset, the SVM classifier with Linear kernel is more flexible that it is more suitable for applications that deal with data from outside the training data set (as in our case of using the mobile facial expression application with users not present in CK+ dataset).

	An	Di	Fe	Ha	Sa	Su	Ne
An	**41.4**	32.9	10.0	2.9	2.9	5.7	4.3
Di	4.3	**87.1**	2.9	2.9	1.4	1.4	0.0
Fe	7.1	4.3	**57.1**	2.9	20.0	8.6	0.0
Ha	0.0	0.0	0.0	**98.6**	1.4	0.0	0.0
Sa	15.7	17.1	5.7	0.0	**30.0**	2.9	28.6
Su	2.9	1.4	1.4	1.4	1.4	**91.4**	0.0
Ne	0.0	0.0	0.0	0.0	0.0	0.0	**100.0**

Table 4: Emotion classification confusion matrix as result of facial expression recognition (%) on *Samsung Galaxy S3*.

Figure 4: Screenshot of Mobile app for real time facial expression recognition.

5.3. Facial Expression Recognition

We show the accuracy results for 6 facial expressions recognition in Table 2. We obtained average 85.8% of accuracy from 10-folds cross validation with features created between neutral frame (the first frame of video sample) and apex frame (the last frame of video sample) in CK+ dataset. As Table 2 shown, sadness has lowest accuracy among emotions because the expression is usually done with small movements and very ambiguous with some other expression such as anger, disgust, and fear.

5.4. Evaluation of Real-time Emotion Recognition Accuracy

Our proposed system is developed on an Android smartphone. The experiment is carried out on *Samsung Galaxy S3* (CPU: Quad-core 1.4 GHz Cortex-A9, GPU: Mali-400MP, Android 4.3 *Jelly Bean*). We tested with *HTC One* (CPU: Quad-core 1.7 GHz Krait 300, GPU: Adreno 320, Android 4.4 *KitKat*) and *Samsung Galaxy S4* (CPU: Quad-core 1.9 GHz Krait 300, GPU: Adreno 320, Android 4.4 *KitKat*) as more recent smartphones. However, they failed to run the app with an issue that since a *VideoCapture* function of *OpenCV* Library 2.4.8 is implemented using non-public Android API, it may not support the cameras of particular smartphones on which the vendor modified this part of Android OS. That's why *Google Nexus 4* have no issue though it uses the same versions of Android 4.4 and *OpenCV* library 2.4.8. Figure 4 shows a screenshot of the mobile app running on *Galaxy S3*.

For the evaluation of real-time emotion recognition accuracy on our proposed system, we tested with 7 subjects who were asked to perform 7 expressions (See Figure 5) on the front-facing camera of the smartphone. They are not professional actors and made posed facial expressions, not

(a) Anger (b) Disgust (c) Fear (d) Happiness (e) Sadness (f) Surprise (g) Neutral

Figure 5: Screenshot of facial expressions captured in real time Mobile app.

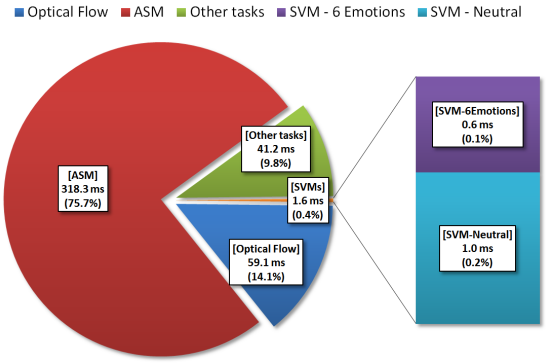

Figure 6: Computation time in modules (Optical flow, ASM, SVM-neutral and SVM-emotions) in a grabbed frame when running on *Samsung Galaxy S3*.

spontaneous expressions. We just considered the expressions of peak frames except transitional expressions. 70 expressions (10 times per an expression) are taken from each subject. The confusion matrix for emotion classification is shown in Table 4. The average accuracy of 7 expressions is about 72% (average 92% of accuracy for disgust, happiness and surprise, whereas average 43% of accuracy for anger, fear and sadness). One apparent reason why the expressions such anger, fear and sadness are recognized with low accuracy is because the subjects are not professional actors and they had to perform unclear expressions, uncertain about what to do for the expressions such as fear or anger. But they were able to perform clearly expressions such as disgust, happiness and surprise. The "sadness" is the most difficult expression to be correctly classified because it is very confused with other expressions such as anger, disgust, and fear. Also since the "sadness" usually has closed mouth, it is easily classified to neutral class. However, the expressions with both open mouth such as happiness and surprise, and closed mouth such as neutral are easier to be classified correctly.

5.5. Evaluation of Processing Time

For evaluating computation time of modules on the proposed system, we tested with two different smartphones such as *Google Nexus 4* (CPU: Quad-core 1.5 GHz Krait, GPU: Adreno 320, Android 4.4 *KitKat*) and *Samsung Galaxy S3*. Computation times are collected during performing facial expressions for about 1 minute. Because different mobile devices have different camera capabilities such as picture preview size and ratio, we selected different preview size and ratio of video images. In Table 3, we show average computation time and standard deviation (milliseconds) taken in the process of the sub-modules.

Also we show a pie chart for comparing relative computation times of sub-modules in a grabbed frame of video sequences on *Galaxy S3* in Figure 6. During a frame, the ASM module including face detection and landmarks fitting is the most time-consuming task with average $318.3ms$ (75.7%). The SVM modules take small time between 0.1% and 0.2%. The rest of total time with tasks such as handling an image and graphics and is average 9.8%. In addition, we compare the results with different parameters for face detection embedded in ASM module in Figure 7. When searching a face in an image, we can set the minimum size of the face (e.g., 25% or 50% width of a frame) to be found. Searching smaller faces in an image spends more time while larger faces take less time relatively. Therefore, compared to the result with face width 25% on *Galaxy S3*, the results with face width 50% on both *Galaxy S3* and *Nexus 4* show less processing time for ASM module.

Evaluating the Possible Use of Optical Flow in Mobile Phones: In this evaluation, we added and compared a Lucas-Kanade's optical flow module [13] because we want to compare how much the optical flow affects in the performance of speed to the system.

As opposed to the general system with a stationary camera, the hand-held mobile devices have a problem such as camera shake causing relative head movements. Head movements occurring in spontaneous facial expressions typically lead to degradation of facial emotion recognition accuracy, as the recognition is based on displacements of

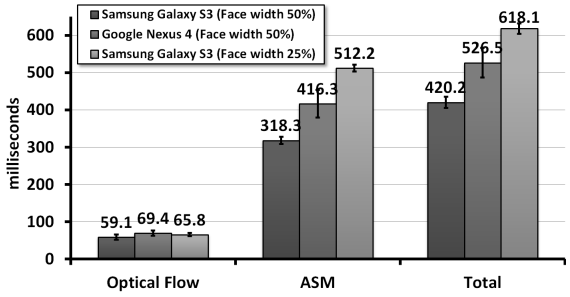

Figure 7: Computation time comparison in different parameters such as minimum width of face - 25% and 50% on *Galaxy S3*, and 50% on *Nexus 4*.

different facial features as detected by ASM. To handle this negative effect from camera shake, the optical flow vector for head movement detection may possibly be helpful. In our experiments, the Lucas-Kanade optical flow module takes average $59ms$ in a frame. This time is nearly 14% of the total time spent for emotion recognition (see Figure 6). Because this additional task of optical flow slows down the system, the integration with the optical flow vector will be taken up as a future work.

Demo of the Proposed Mobile App: The proposed mobile application for facial expression recognition can be downloaded from: `https://www.dropbox.com/s/716dlnwka4irc8y/EmotionRecognitionRT.apk`

6. Conclusions and Future Work

We present an efficient approach for real-time facial expression recognition running on smart phones. The proposed system employs fast and accurate methods such as SVM models with dynamic features created and extracted from ASM. After the landmarks from ASM are fitted on detected face in each frame, the system recognizes whether or not the face has an expression with the SVM classifier for neutral face detection along with mouth status. While the system keeps updating neutral features, it creates new dynamic features with displacement between latest neutral feature and current feature if the face is recognized as non-neutral expressions, and the system returns the recognized expression by the SVM classifiers for facial expression recognition with the dynamic features. The proposed system is so simple and efficient that it runs smoothly on the mobile devices with good performances in terms of speed and recognition accuracy as we evaluated the computation time and accuracy on different smartphones in Section 5.

In the future work, the proposed system may be integrated with a variety of mobile applications or systems without any remote servers (e.g., the system recommending multimedia content in [15])

References

[1] B. Anand, B. Navathe, S. Velusamy, H. Kannan, A. Sharma, and V. Gopalakrishnan. Beyond touch: Natural interactions using facial expressions. In *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, pages 255–259, Jan 2012.

[2] M. S. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. J. Sejnowski, and J. R. Movellan. Automatic analysis of spontaneous facial behavior. In *IN THE EYE OF THE*. Oxford University Press, 2001.

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[4] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models–their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, Jan. 1995.

[6] G. Edwards, C. Taylor, and T. Cootes. Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 300–305, 1998.

[7] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[8] O. Houstis and S. Kiliaridis. Gender and age differences in facial expressions. *Eur J Orthod*, 31(5):459–66, 2009.

[9] G.-S. Jo, I.-H. Choi, and Y.-G. Kim. Robust facial expression recognition against illumination variation appeared in mobile environment. In *Proceedings of the 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering*, CNSI '11, pages 10–13, Washington, DC, USA, 2011. IEEE Computer Society.

[10] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[11] C. L. Lisetti and D. E. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the Eleventh International Florida Artificial Intelligence Research Society Conference*, pages 328–332. AAAI Press, 1998.

[12] G. Littlewort, I. Fasel, M. S. Bartlett, and J. R. Movellan. Fully automatic coding of basic expressions from video. Technical report, Tech. rep.(2002) U of Calif., S.Diego, INC MPLab, 2002.

[13] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[14] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101, June 2010.

[15] M. B. Mariappan, M. Suk, and B. Prabhakaran. Facefetch: A user emotion driven multimedia content recommendation system based on facial expression recognition. *2013 IEEE International Symposium on Multimedia*, 0:84–87, 2012.

[16] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, ECCV '08, pages 504–513, Berlin, Heidelberg, 2008. Springer-Verlag.

[17] C. Padgett and G. W. Cottrell. Representing face images for emotion classification. In M. Mozer, M. I. Jordan, and T. Petsche, editors, *NIPS*, pages 894–900. MIT Press, 1996.

[18] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881 – 905, 2000.