AUTOMATIC EMOTION CLASSIFICATION: AN APPROACH USING EULERIAN VIDEO MAGNIFICATION AND ARTIFICIAL NEURAL NETWORKS

Vitor Quaresma S. de Hollanda Ramos*, Ana Paula G. S. de Almeida*, Carla M. C. C. e Koike*, Dianne M. Viana†, Flavio de B. Vidal*

*Department of Computer Science, University of Brasilia, Brasilia - Distrito Federal, Brazil

†Department of Mechanical Engineering, University of Brasilia, Brasilia - Distrito Federal, Brazil

Email: vitor.q.s.h.r@gmail.com, anapaula.gsa@gmail.com, ckoike@unb.br, diannemv@unb.br, fbvidal@unb.br

Abstract— Researches using facial emotion detection, which involves making a computer being able to perceive the emotional state of a human user. In this work, we propose an initial approach to develop an automatic emotion classification using Eulerian Video Magnification and Artificial Neural Networks (ANN). The proposed approach creates a descriptor vector from processed Eulerian Magnification videos and uses an ANN classification to achieve a suitable accuracy rate for isolated and combined emotions.

Keywords— Emotion detection, Face analysis, Eulerian Video Magnification.

1 Introduction

Facial emotion recognition deals with computers automatically perceiving the emotional state of a human user. This research area has been gaining a lot of attention recently due to its potential application in many fields, such as medical engineering (Leon et al., 2005), psychology (Yick et al., 2015), robotics (Tsai et al., 2009) and forensics applications (Fairhurst et al., 2014). According to (Sanchez-Mendoza et al., 2015), a dynamic facial behavior can be very rich source of information for revealing emotions and much of the information of the emitter's message can be inferred by the facial expression. For these reasons, facial emotion recognition tasks are considered an interesting challenge for computer vision industrial and research areas. Many methods have been proposed by different research groups trying to achieve an efficient approach to detect human facial emotions using computer vision.

In this work, we propose a robust system able to perform automatic emotion classification using Eulerian Video Magnification and Artificial Neural Networks. Section 2 describes the main related works on automatic emotion detection. The proposed methodology and initial results are presented in Section 3 and 4, respectively. Conclusions and further work are discussed in Section 5.

2 Related Works

Emotion classification is currently receiving a lot of attention from researchers due to its potential application in many fields as psychology, human behavior studies and human-computer interface. The process of emotion classification is still a challenging task in computer vision. Currently, the main focus is on systems that perform the classification based on the combination of information provided by human voice (Alva et al., 2015)(Reney and Tripathi, 2015) and features extracted from human face captured by a camera, as described in (Su et al., 2014).

The Facial Action Coding System (FACS) (Ekman et al., 2002) is the best known and the most commonly used system developed for human observers to describe facial activity in terms of visually observable facial muscle actions (called Action Units, AUs). With FACS, human observers uniquely decompose a facial expression into one or more of 44 AUs that produced the expression in question. According to (Pantic and Rothkrantz, 2003), many works have applied several techniques, as Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Bayesian Networks, when trying to achieve an efficient emotion classification system.

(Bouhabba et al., 2011) presents a real time approach to emotion recognition using an automatic facial feature tracker to perform face localization and feature extraction. The facial feature displacements in the video stream are used as input to a Support Vector Machine classifier. The proposed method is evaluated in terms of recognition accuracy for a variety of interaction and classification scenarios. In (Jagadiswary et al., 2011), a segmentation method is proposed that can handle even degraded images acquired in less constrained conditions. First of all, the information from eye elements (sclera) is used to generate a new type of feature that measures the proportion of these eye elements in each direction. After processing eye elements, the procedure is suitable for real-time applications.

In the approach described in (Fan et al., 2014), the authors focused on emotion detection in angry-neutral speech, which is common in recent studies of Automatic Emotion Variation Detection (AEVD). A novel framework for AEVD using Multi-scaled Sliding Window (MSW- AEVD) is proposed to assign an emotion class to each window-shift by fusion of decisions of all the sliding windows containing the shift. Using the parallel fusion strategy, they achieve a *F-score* about 92%. Unfortunately, not much additional information is described, which is crucial to explain the high *F-score* rates achieved.

In the study described in (Oh and Hong, 2014), a face region is identified through a combination of the ***YCbCr*** color space model and the Maximum Morphological Gradient Combination (MMGC) image. The search region for the facial component detection is limited to the recognized face region. The facial components are detected using the histogram method, the blob labeling method, and the MMGC image. A specialized emotion dataset (eNTERFACE'05) is used to evaluate the proposed methodology and to achieve accuracy around 81%.

In this paper, we propose an early approach to the emotion recognition problem. A particular feature of this proposal lies in that no additional information, besides the face video, is necessary to proceed the recognition.

3 Proposed Methodology

The approach proposed in this work is based on Eulerian Video Magnification, as described in (Wu et al., 2012). Using this technique, it is possible to extract small micro motion variations in a video sequence. This amplification is necessary to improve the magnified features about the face in the video, defined as facial micro-expressions, which can then be used to generate descriptors and allow the training of an Artificial Neural Network (ANN) to classify and recognize the emotions. The proposed methodology is composed of steps described in the Figure 1 flowchart.

The following Sections describe detailed descriptions of each stage shown in the flowchart.

3.1 Face Detection

In this approach, the individual face is the main region of interest in the image. To capture this area, a face detection algorithm is used to detect and crop it. This process results in a new video that only has the region of interest in it, removing unwanted information for the future steps. We use the Viola-Jones's with Haar-Cascade face detection algorithm, as described in (Viola and Jones, 2004).

3.2 Eulerian Video Magnification

The Eulerian Video Magnification (Wu et al., 2012) amplifies motion and color changes in the input video, allowing the use of information that was not available without the magnification. Different types of magnification could be used, as it

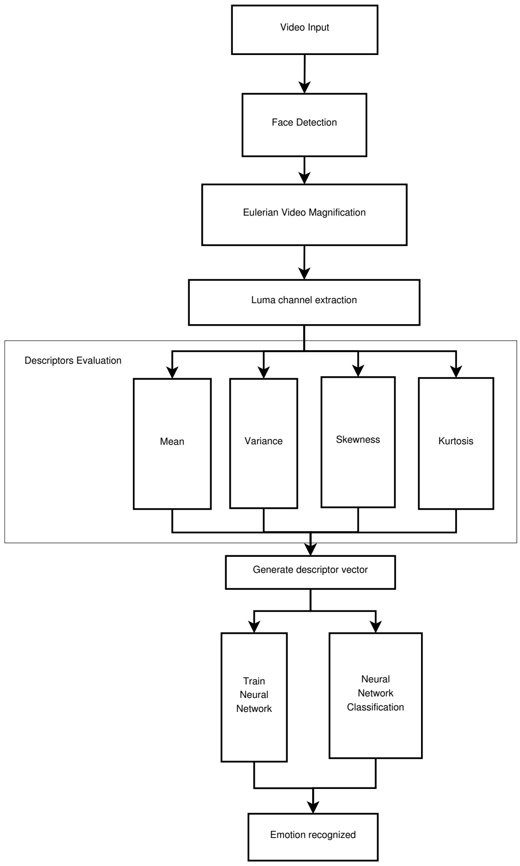
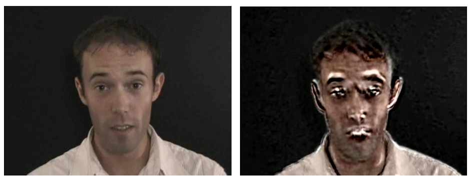


Figure 1: Flowchart describing the proposed methodology.

is possible to change the type of spatial process and the type of temporal processing.

The spatial types of processing employed here are the Laplacian and Gaussian methods. The Gaussian spatial processing only works with an ideal time processing, due to an impossibility of performing the magnification frame by frame along with the filters creation.

The Spatial Laplacian processing, on the other hand, allows different types of temporal processing, as *Butter* and *Infinite Impulse Response* (IIR), that creates Laplacian pyramids and use the filters created by these pyramids while processing the video, being faster than the ideal temporal processing. For the proposed methodology, the spatial processing used is the Spatial Laplacian. The Figure 2 illustrates the Eulerian Video Magnification on an example.



(a) Frame from the input video. (b) Frame from the magnified video.

Figure 2: Sample frames from an input video.

3.3 Luma Component

After the Eulerian Video Magnification stage, it is necessary to obtain the Luma component (Y). The magnified video is then used as an input and each frame in *RGB* color space is converted to the color system *YCbCr*, which results in one value of the component for each pixel of each frame. Preliminary tests have shown that the Luma component channels many quantities of the magnified facial micro-expression, and for that reason it is used here to generate the descriptors.

Since the components *Cb* and *Cr* are poor in these magnified micro-expressions,(they carry similar information among distinct emotions), they are discarded as descriptors candidates to emotion classification stage.

3.4 Descriptors Evaluation

With Luma values of all pixels of each frame, four elements are used in the descriptor vector: *mean*, *variance*, *kurtosis* and *skewness*. These values are employed to analyze the magnified video, and it results in four prototype variables for the classification stage.

The prototype variable *mean* (μ) is the mean (Equation 1) of the Luma component from all pixels of all frames in the analyzed video. The *variance* variable (Equation 2) measures how far the Luma component spreads out, and when the *variance* value is zero it indicates that all values are identical. Since the values of the Luma components can be modeled as a discrete function, the *variance* can be computed as shown in Equation 2.

$$\mu = \frac{\sum_{i=1}^{n} l_i}{n}, \tag{1}$$

where l_i is the Luma component of the pixel i and n is the number of pixels in a frame.

$$Var = \frac{\sum_{i=1}^{n} (l_i - \mu)^2}{n} \tag{2}$$

The *skewness* is obtained from the measure of symmetry of the probability function. The value of *skewness* can be positive, negative, or even undefined. When the *skewness* is equal to zero, it means that the probability function is symmetric. If a distribution has a positive *skewness*, the tail on the right side is longer or bigger than the tail on the left side, while a negative *skewness* value indicates that the tail is longer or bigger on the left side. The *skewness* (S) is evaluated using the following equation:

$$S = E[(\frac{L-\mu}{\sigma})^3], \tag{3}$$

where μ is the *mean*, σ is the standard deviation, L is a random variable from Luma component and E is the expectation operator. For the emotion classification, the *skewness* feature of the magnified video is used to increase the discrimination between similar emotions in the prototype vector. For instance, happiness and surprise have huge similarity in the *mean* and *variance* values.

The *kurtosis* measures how much a distribution is similar to the Gaussian distribution based on the peaks of that distribution. If the *kurtosis* has a positive result, it has more peaks than the Gaussian distribution, while a negative one has less peaks. The computation of the *kurtosis* (K) is shown in Equation 4.

$$K[L] = \frac{\mu_4}{\sigma^4}, \tag{4}$$

where μ_4 is the fourth moment about the *mean* and σ is the standard deviation.

3.4.1 Descriptor Vector

In order to use the prototype variables and achieve a descriptor vector for training an ANN, an input vector is created containing the four values of *mean*, *variance*, *skewness* and *kurtosis*. All these prototype variables are normalized, with values between $[0,1]$.

In the example shown in Table 1, it is possible to verify that each emotion for this person is distinct when employing the descriptor vector described above in support to classify the emotions.

3.5 Artificial Neural Network

In order to use the proposed descriptor vector (Subsection 3.4) for emotion classification, an ANN was trained with a reduced set of videos where the expressed emotions are known.

Two types of ANNs were tested: Multilayer Perceptron (MLP) and Self-Organizing Maps (SOM), as described in (Haykin, 1998). The MLP with a feed-forward network showed the best results.

The Multilayer Perceptron was trained with the *Levenberg-Marquardt* algorithm (Yu and Wilamowski, 2011) and it has four layers. The first layer carries the inputs descriptor vector to the classifier, and no computation is realized here. The first and the second hidden layers have at least 100 neurons, while the fourth layer is the output layer with only one neuron. With the exception of the input layer, the majority of the neurons have hyperbolic tangent transfer functions. It was already shown that feed-forward neural networks with only one hidden layer can approximate arbitrarily well any function with a finite numbers of discontinuities (Demuth et al., 1993). However, exhaustive tests have shown that, in our experiments, two hidden layers perform better.

In the training stage, the network outputs for emotions classes were evaluated using an output

Values	Mean	Variance	Skewness	Kurtosis
Anger	0.6671	0.1285	1	0.8591
Disgust	0	0.1344	0.0188	0.0346
Fear	0.9709	0.0394	0.2032	0.6035
Happiness	0.2937	0.4583	0.1962	0.6066
Sadness	0.9797	0.6001	0.2928	0.8056
Surprise	0.8718	0.1309	0.2362	0.7925

Table 1: Values of mean, variance, skewness and kurtosis from a person normalized.

range equally distributed between $[-1,1]$, in according to number of tested emotions. For example, for six emotions classes, the output range is as described in Equation 5, where $\alpha(T)$ is the ANN output and T is the threshold of activation function in the feed-forward algorithm.

$$\alpha(T) = \begin{cases} \text{Anger, if } -1 \leq T \leq -0.65 \\ \text{Disgust, if } -0.65 \leq T \leq -0.33 \\ \text{Fear, if } -0.33 \leq T \leq 0 \\ \text{Happiness, if } 0 \leq T \leq 0.33 \\ \text{Sadness, if } 0.33 \leq T \leq 0.65 \\ \text{Surprise, if } 0.65 \leq T \leq 1 \end{cases} \tag{5}$$

3.6 *Classification using an Artificial Neural Network*

To validate the emotion classification, random videos with the unlabeled emotion were used. The descriptor vector of these videos was created and used as input in the ANN, which recognize the emotion.

The validation is made as a comparison between the real emotion in video and the recognized emotion, generating a confusion matrix to determine the effectiveness of the proposed methodology.

4 PRELIMINARY RESULTS

The proposed implementation has been evaluated for a public emotion video dataset called *eNTERFACE'05* (Martin et al., 2006), composed of video files from 44 subjects expressing 6 classes of emotions defined by 5 different phrases for each emotion: Anger, Happiness, Sadness, Fear, Surprise and Disgust. The complete dataset is composed of 1320 input videos.

The video content consists of a person speaking a phrase related to an assigned emotion and the execution time is around 10 seconds. In the tests, the audio was discarded, but the background is unchanged. Samples from the used video dataset are shown in Figure 3.

In the preliminary tests, four test scenarios to emotion classification are proposed and evaluated following the flowchart described in Figure 1. For all scenarios, we are using the training/evaluation strategy 70/30 (70% of whole dataset is used in ANN training and the remains 30% in evaluation stage).

4.1 *Scenario 1*

The Scenario 1 is composed of two emotion classes, picked randomly in a dataset. In order to evaluate the efficiency of the algorithm, the confusion matrix to Scenario 1 is described in Table 2.

Emotion Detected	Surprise	Fear
Surprise	1	0
Fear	0	1

Table 2: Confusion matrix of the randomly picked emotions Surprise and Fear in Scenario 1.

In despite of the simplicity of Scenario 1, the confusion matrix in Table 2 showed that the ANN was able to perform classification with no difficulty.

4.2 *Scenario 2*

This scenario consists in three emotion classes, also chosen randomly. The confusion matrix to Scenario 2 is shown in Table 3.

Emotion Detected	Happiness	Anger	Fear
Happiness	1	0	0
Anger	0.33	0.66	0
Fear	0	0	1

Table 3: Confusion matrix of the emotions Happiness, Anger and Fear to Scenario 2.

For this scenario, the slight increase in complexity with the addition of one emotion class resulted in a grown difficulty of ANN when classifying the emotions happiness and anger.

4.3 *Scenario 3*

Scenario 3 contains five emotions classes and the resulting confusion matrix (Table 4) showed that the discriminative capabilities of the ANN have increased. The ANN was able to classify correctly most of input videos, but the results are worse in the case of emotions fear and anger. Even with an increase in the number of layers and neurons, the ANN configuration was not able to perform this classification.

(a) Sentence 1 of Anger. (b) Sentence 1 of Disgust. (c) Sentence 1 of Fear. (d) Sentence 1 of Happiness. (e) Sentence 1 of Sadness. (f) Sentence 1 of Surprise.

Figure 3: Sample frames from videos of different emotions from the person 1.

After analyzing the video with these two emotions classes, it was found that the main reason for of the proposed approach classification was that, after performing the video magnification process, the resulting video generated a much higher amount of noise, probably due to the high motion and the low frame rate. This noise does not allow the proposed descriptor vector to perform an appropriate classification.

4.4 Scenario 4

In Scenario 4, all emotions classes from the dataset are used and the results are presented in the confusion matrix in Table 5.

Evaluating the results shown in Table 5, from obtained confusion matrix the classification capacity of the artificial neural network has been compromised. In various tests, it was observed that in according to the emotion class used as input, the output would be a possible one up to six emotions (see happiness classification results in Table 5). In these situations, it was observed that the capability of generalization and accuracy, of the input vector using only four variables in proposed methodology would not be sufficient to achieve good results.

5 Conclusions

In this paper, we presented an initial strategy to recognize different emotions in a video using Eulerian Video Magnification and ANNs. Based on the achieved results, it is possible to confirm that the proposed methodology can perform the classification in a restrict group of emotions (only six classes).

It is possible to observe that some emotions have a similar behavior, as Anger and Disgust, hindering the classification. As well as there are emotions that are completely different, as Anger and Surprise, thus being easier to be classified.

Comparing the confusing matrix from all analyzed people, it is notable that the algorithm works better with videos from the same person than with the videos from all the people together. In our initial tests, that behavior was expected, since different people express their emotions in distinct ways, even though they are saying the same phrase.

Future works towards an improvement of the proposal include using a larger group of emotions and more people to be tested, as well as to verify other approaches as convolutional neural networks (CNNs) in order to avoid issued related to an usual ANNs.

References

Alva, M., Nachamai, M. and Paulose, J. (2015). A comprehensive survey on features and methods for speech emotion detection, *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, pp. 1–6.

Bouhabba, E., Shafie, A. and Akmeliawati, R. (2011). Support vector machine for face emotion detection on real time basis, *Mechatronics (ICOM), 2011 4th International Conference On*, pp. 1–6.

Demuth, H., Beale, M., Demuth, H. and Beale, M. (1993). Neural network toolbox for use with matlab.

Ekman, P., Friesen, W. and Hager, J. (2002). The facial action coding system, *Research Nexus eBook*.

Fairhurst, M., Erbilek, M. and Li, C. (2014). Enhancing the forensic value of handwriting using emotion prediction, *Biometrics and Forensics (IWBF), 2014 International Workshop on*, pp. 1–6.

Fan, Y., Xu, M., Wu, Z. and Cai, L. (2014). Automatic emotion variation detection using multi-scaled sliding window, *Orange Technologies (ICOT), 2014 IEEE International Conference on*, pp. 232–236.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*, 2nd edn, Prentice Hall PTR, Upper Saddle River, NJ, USA.

Jagadiswary, D., Appasami, G. and Rajesh, S. (2011). Eye features normalization and face emotion detection for human face recognition, *Electronics, Communication and Computing Technologies (ICECCT), 2011 International Conference on*, pp. 64–68.

Emotion Detected	Disgust	Happiness	Anger	Fear	Sadness
Disgust	1	0	0	0	0
Happiness	0	0.5	0.5	0	0
Anger	0.2	0.2	0.2	0.2	0.2
Fear	0	0	1	0	0
Sadness	0	0	0	0	1

Table 4: Confusion matrix of the emotions Disgust, Happiness, Anger, Fear and Sadness to Scenario 3.

Emotion Detected	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	0.219	0.041	0.333	0.0417	0	0.366
Disgust	0.05	0.013	0.057	0	0.864	0.014
Fear	0.251	0.029	0.128	0.126	0.295	0.168
Happiness	0.029	0.359	0.196	0.194	0.025	0.194
Sadness	0	0.55	0.008	0.357	0.075	0.002
Surprise	0	0	0	0	0.5	0.5

Table 5: Confusion matrix of the emotions Anger, Disgust, Fear, Happiness, Sadness and Surprise to Scenario 4.

Leon, E., Clarke, G., Sepulveda, F. and Callaghan, V. (2005). Real-time physiological emotion detection mechanisms: Effects of exercise and affect intensity, *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pp. 4719–4722.

Martin, O., Kotsia, I., Macq, B. and Pitas, I. (2006). The enterface' 05 audio-visual emotion database, *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pp. 8–8.

Oh, B.-H. and Hong, K.-S. (2014). A study on facial components detection method for face-based emotion recognition, *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*, pp. 256–259.

Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction, *Proceedings of the IEEE*, pp. 1370–1390.

Reney, D. and Tripathi, N. (2015). An efficient method to face and emotion detection, *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*, pp. 493–497.

Sanchez-Mendoza, D., Masip, D. and Lapedriza, A. (2015). Emotion recognition from mid-level features, *Pattern Recognition Letters* **67, Part 1**: 66 – 74. Cognitive Systems for Knowledge Discovery.

Su, B.-H., Fu, P.-W., Lin, P.-C., Shih, P.-Y., Lin, Y.-C., Wang, J.-F. and Tsai, A.-C. (2014). A spoken dialogue system with situation and emotion detection based on anthropomorphic learning for warming healthcare d, *Orange Technologies (ICOT), 2014 IEEE International Conference on*, pp. 133–136.

Tsai, C.-C., Chen, Y.-Z. and Liao, C.-W. (2009). Interactive emotion recognition using support vector machine for human-robot interaction, *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pp. 407–412.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection, *Int. J. Comput. Vision* **57**(2): 137–154.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F. and Freeman, W. T. (2012). Eulerian video magnification for revealing subtle changes in the world, *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* **31**(4).

Yick, Y. Y., Buratto, L. G. and Schaefer, A. (2015). The effects of negative emotion on encoding-related neural activity predicting item and source recognition, *Neuropsychologia* **73**: 48 – 59.

Yu, H. and Wilamowski, B. M. (2011). Levenberg–marquardt training, *Industrial Electronics Handbook, 2nd Edition*, Vol. 5, CRC Press, pp. 12–1.