Emotion Recognition using Spatiotemporal Features from Facial Expression Landmarks

Hamid Golzadeh
School of Engineering & Applied Science
Aston University
Birmingham, UK
golzadeh@aston.ac.uk

Diego R. Faria
School of Engineering & Applied Science
Aston University
Birmingham, UK
d.faria@aston.ac.uk

Luis J. Manso
School of Engineering & Applied Science
Aston University
Birmingham, UK
l.manso@aston.ac.uk

Anikó Ekárt
School of Engineering & Applied Science
Aston University
Birmingham, UK
a.ekart@aston.ac.uk

Christopher D. Buckingham
School of Engineering & Applied Science
Aston University
Birmingham, UK
c.d.buckingham@aston.ac.uk

***Abstract*—Emotion expression is a type of nonverbal communication (i.e. wordless cues) between people, where affect plays the role of interpersonal communication with information conveyed by facial and/or body expressions. Much can be understood about how people are feeling through their expressions, which are crucial for everyday communication and interaction. This paper presents a study on spatiotemporal feature extraction based on tracked facial landmarks. The features are tested with multiple classification methods to verify whether they are discriminative enough for an automatic emotion recognition system. The Karolinska Directed Emotional Faces (KDEF) [1] were used to determine features representing the human facial expressions of angry, disgusted, happy, sad, afraid, surprised and neutral. The resulting set of features were tested using K-fold cross-validation. Experimental results show that facial expressions can be recognised correctly with an accuracy of up to 87% when using the newly-developed features and a multiclass Support Vector Machine classifier.**

Keywords—Facial Expressions, Emotion Recognition, Spatiotemporal Features, Classification, SVM, RFC, SAG.

I. INTRODUCTION

Interactions among human beings are facilitated by interpretation of emotions, which can be expressed in many ways, including body language, voice intonation and facial expressions. Machine interpretation can use technologies such as electroencephalography to pick up emotions in voices [2] but there are easier practical methods for examining facial expressions.

It is said that there are seven forms of human emotions that are recognisable in faces across different cultures [3]: disgust, contempt, happiness, anger, surprise, fear and sadness. Facial expression recognition (FER) therefore plays a very important role in improving the quality of human communications and can be usefully exploited by machines. For example, at airports, FER can be used as a method of security check to investigate unexpected emotional states of travellers or when investigating suspected criminals. FER can also have medical applications such as assessing the reactions of patients before or after surgery with respect to pain, stress, or anxiety. In mental health, computers can play an important role in gaining information from people who are reluctant to talk to a human due to stigma surrounding mental-health problems [4]. Avatars are becoming important components of e-mental health interventions [5] and can help improve engagement. One way is asking the right questions in the right way, which was a central motivation for developing the myGRaCE self-assessment version [6] of GRiST (www.egrist.org) for early detection of risks such as suicide and violence. However, virtual avatars also need to show appropriate emotional responses during interactions if they are to maximise therapeutic benefit and the research reported in this paper is an important step towards that goal.

Automatic recognition of human emotions is a difficult task for at least the two following reasons: (i) a large database of training (labelled) images with realistic emotions (not acting) does not exist; (ii) static images that are a single point in time are not easy to classify with any confidence because facial expressions quickly change and the transitions between image frames are important pieces of information. In this paper, we present a study on facial expression recognition by which emotions are recognised automatically when using a dataset and that can be applied dynamically within live video. Multiple machine learning techniques such as Random Forest Classification (RFC), Support Vector Machines (SVM) and linear regression with Stochastic Average Gradient (SAG) have been tested. A set of spatiotemporal features based on tracked facial landmarks is presented and tested with multiple classifiers to verify its discriminability. Thus, the main contributions of the paper are highlighted as follows:

- A new and effective set of spatiotemporal features based on 1D and 2D distances among facial landmarks, log-covariance, angles, derivatives, log-energy and angular velocities.

- Experimental tests and comparison of multiple state-of-the-art machine learning algorithms for emotion classification to validate the effectiveness of the feature set selected from an affective facial expression dataset. A system has been developed by which human emotions can be detected in different lighting conditions, scenes and angles in real-time.

The remainder of this paper is structured as follows.

Section II presents related work followed by the dataset description in Section III. Section IV introduces the methods adopted in this study. Section V presents the results attained on the KDEF dataset, and the conclusions and future work are addressed in Section VI.

978-1-5386-7097-2/18/$31.00 @2018 IEEE

II. Related Work

One of the main challenges in image processing and facial expression recognition is face detection and tracking. Various approaches can be found in the literature to solve this problem. For instance, the work presented in [12] introduced a bioinspired algorithm for face recognition. An important task in their face detection algorithm is to identify the presence of a face in a particular area of the image. The genetic algorithm approach includes a pre-processing stage to decrease the scale of the image, reduce noise, and enhance edges. Their algorithm had an acceptable performance but only with images without a complex background. However, it can recognise faces even with beards, glasses, long hair, etc. Another classical technique is the well-known Haar-like features [14], which have succeeded well in this type of automatic detection. The problem with FER is how to reveal emotions. Two different people can indicate their emotions in completely different ways, which means their facial expressions are not equivalent. Shih-Chung Hsu *et al*. [7] successfully recognised facial expressions according to four phases: neutral, onset, apex, offset. In each phase, the facial expression is recognised via a specific method such as a hybrid approach in the apex phase for recognising facial expressions, Gabor filter for obtaining facial features in the neutral stage, and SVM for recognising Action Units in the onset phase.

Barbara Gonsior *et al*. [8] explored facial expression influence on human-robot interaction by implementing an experimental setup in which EDDIE, a robot head, has a conversation with participants. The structure of the robot's dialog comes from *Akinator*, a web-based gaming application. The robot tries to guess the thought of a person by asking various questions. The robot replies to the facial expression of the participant in different ways such as mirroring, ignoring or showing a facial expression according to a physiological model. After the game, the participant is required to answer a questionnaire which consists of two parts. The first part measures the social robot's acceptance by the user and the second part evaluates five human-robot interaction concepts: animacy, perceived intelligence, anthropomorphism, likeability and perceived safety. The results support the hypothesis that empathy between human and robot and the subjective perception of task-performance are heavily influenced by the robot's behaviour during the interaction.

The authors concern in [9] is about spoofing attacks that weaken the process of face recognition using faces that are not real. Currently, there are three approaches in performing liveness detection: face texture, challenge and response, and joining some biometrics for liveness detection. However, these methods are not successful in an unconstrained environment. Therefore, they proposed a method based on Image Quality Assessment (IQA) parameters, which has been successfully validated within an unconstrained environment. By utilizing visual information obtained by a robot, Cid et al [10] introduce a system for imitating and recognising facial expressions. A Bayesian approach was proposed to estimate human emotion via facial expression recognition. The information acquired updates the robot's knowledge about the individuals and can therefore be used for future interactions. At the same time, a twelve degrees of freedom robotics face restricts the facial expressions of humans. The results indicated good quality of imitation and detection when utilizing an avatar within various scenarios.

Jake Bruce *et al*. [11] had a novel interface for controlling a drone using facial expressions and no other instrumentation. Users were able to control and direct the drone on a favourite 3D path at wide latitude and outside the users' line of vision, with minimum practice in a few minutes. The drone used facial expression recognition to guide its flight after the user has defined a list of different pre-trained facial expressions for the directions. There are three different suggested paths available for this application: direct line, ellipse, and orbit. In the prototype, facial expression recognition was also able to detect unauthorized or unregistered users and thus could refuse to obey their instructions.

The work reported in this paper differs from previous research by finding appropriate discriminative spatiotemporal features and classification methods for automatic emotion recognition based on facial expressions. Extensive tests on the KDEF dataset were conducted to find an appropriate method and feature set.

III. KDEF Dataset

The Karolinska Directed Emotional Faces (KDEF) [1] is publicly available and contains 4900 images of human facial expressions from 70 people: 35 males and 35 females with ages between 20 and 30 years old. The images have no moustache, eyeglasses or earrings and no visible make-up, with each person displaying seven emotional expressions. Each expression is captured from five different angles (straight, full left profile, full right profile, half right profile, and half left profile). These images have 562 x 762 pixels, resolution of 72 x 72 dpi, 32-bit colours, and JPEG file format with compression quality of 94%. Our research selected 1470 images for training and testing.

IV. Methods

A. Spatiotemporal Features from Facial Expressions

Before the features extraction, we have detected the face based on Haar-like features followed by landmarks detection and tracking using the python DLib library, which is based on face alignment with an ensemble of regression trees [15]. A total of 68 points of facial landmarks were detected on locations of the eyebrows, eyes, nose, lips and the contour of the face. Our features are based on the movements of facial muscles, i.e. the locations where these points change over time, which is an extension of previous works [20], [21]. The idea herein is to cover facial landmarks and form a connected graph, so as the density of graphs is different in each facial expression due to the facial motion. In order to achieve this goal, the well-known Delaunay triangulation method was used to join a set of points (landmarks) aiming at making a triangular mesh. Thus, after the graph is connected, we have obtained 109 triangles formed on the human face as shown in Fig. 1.

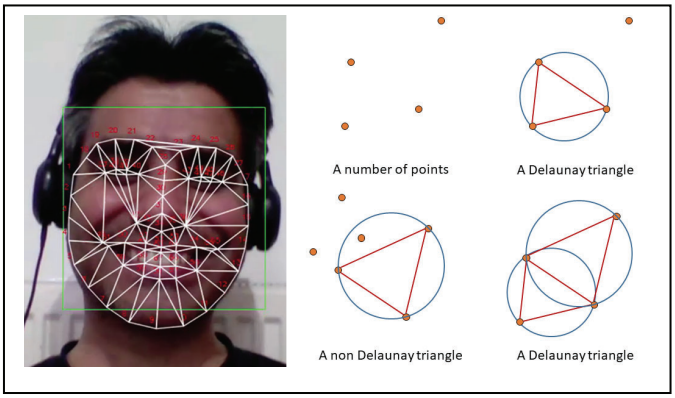

Fig. 1: Example of face and landmarks detection and the Delaunay triangulation to form a mesh.

Then, a set of geometrical features have been computed. It is based on Euclidean distances $\delta_{pq} = \sqrt{(p_2 - p_1)^2 + (q_2 - q_1)^2}$ among all landmarks, thus, obtaining a square matrix of 68 x 68 related to all landmark distances. As the distance between each point with itself is zero, the matrix diagonal is null.

By removing the null diagonal, the final matrix is M = 67 x 68. Then, we have applied a normalization step for all matrix elements, $M = \frac{M_{ij} - \min(M)}{\max(M) - min(M)}$. The normalization step is important because the subject's distance from the camera affects the distance between the facial landmarks. Then, the log-covariance of M has been applied on M as follows:

$$lcM = \mathrm{U}(\mathrm{logm}(\mathrm{cov}(M))) \quad (1)$$

where *lcM* is a resulting vector containing the upper triangular elements (2278) of the matrix after computing the matrix logarithm over the covariance matrix M; U(.) is a function to return the upper triangular elements; logm(.) is the matrix logarithm function; and the covariance matrix is given by $cov(M) = cov_{ij} = 1/N \sum_k^N (x_{ik} - \mu_i)(x_{kj} - \mu_j)$. The rationale behind log-covariance is the mapping of the convex cone of a covariance matrix to the vector space by using the matrix logarithm so that it does not lie in Euclidean space, i.e., the covariance matrix space is not closed under multiplication with negative scalars.

The types of feature are computed after detecting the triangles among all landmarks. Given all three angles of each triangle in the mesh, a total of 3 angles x 109 triangles = 327 angles were obtained. The angles were calculated with the law of cosines (a.k.a. cosine rule), which is given by:

$$c^2 = a^2 + b^2 - 2ab\cos(\gamma) \quad (2)$$

where γ denotes the angle contained between sides of lengths a and b and opposite the side of length c.

Then, we computed the 1D Euclidean distances $d = |p_2 - p_1|$ among all x and y landmarks coordinates, independently. The other two subsets of features are acquired similarly to (1), but the inputs are the 1D Euclidean distances computed given the facial landmarks. Thus, we keep the upper triangular elements of the matrix after the log-covariance computation for the x and y coordinates.

In order to add the temporal factor during features extraction, we have used two consecutives frames, i.e. images on time instant *t-1* and *t*. Then we computed the following derivative for all types of aforementioned features:

$$v = \frac{f^t - f^{t-1}}{\Delta t} \quad (3)$$

where v is the resulting value, f represents a specific type of feature at time t and t-1 and $\Delta t = 1/30$ (camera frame rate). Applying this technique, we noticed the best features were attained when computing the current frame to the initial one, which is usually the neutral face. This happens due to the diversity of the motions of each of the seven emotions when compared to the neutral face. For the tests on the KDEF dataset, we have considered the neutral face as the precedent (t-1) image.

The log-energy given the derivative of the resulting elements $\{i=1, \ldots, n\}$ of the log-covariance matrix was computed as a new feature as follows:

$$ln_e = \sum_{i=1}^{n} \log\left(\left(\frac{lcM_i^t - cM_i^{t-1}}{\Delta t}\right)^2\right) \quad (4)$$

With these types of features presented, we acquired over 10K feature values due to the number of elements given the resulting matrices (log-covariance, distances, etc.). This high number of features can be reduced around 1 to 4% by applying features selection algorithms.

B. Classification Methods

In this work, classical state-of-the-art machine learning methods for classification such as Random Forest Classifier (RFC), Support Vector Machines (SVM), linear regression with Stochastic Average Gradient (SAG), Random Tree Classifier (RTC) and Naïve Bayes Classifier (NBC) were used to classify our set of features. The primary goal of testing multiple classification methods is to check whether the presented features are discriminative enough to classify facial emotional expressions even with more simplistic machine learning techniques.

The SVM finds a line (hyperplane) which classifies the training data set into different classes. Due to existence of lots of linear hyperplanes, SVM uses the method of margin maximization. In other words, SVM tries to keep a maximum distance between the different classes that are involved. SVM parameters that are used in this work are: *kernel* = linear, *probability* = '*true*' and *class_weight* = '*balanced*'. We also used a different training method for the multiclass SVM. In particular, an improved version of *Platt's Sequential Minimal Optimization* (SMO) was used to train the SVM [17], [18].

The linear regression tries to find the relationship between two variables and how the changes of one of them affect another one. In fact, there are two types of variables, dependent and independent, where any change in an independent variable shows the impact on the dependent one. The independent and dependent variables are referred as explanatory and the factor of interest or predicator, respectively. The algorithm of Stochastic Average Gradient (SAG) was proposed by Mark Schmidt *et al.* [16] to optimise smooth convex problems on

finite data sets at high speed based on a randomized variant of the incremental aggregated gradient. The authors implemented SAG for L2-regularized logistic regression. Stochastic gradient methods are often used to solve the problem of optimising a finite sample average: $\min_{x\in\mathbb{R}^p} g(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. Thus, the SAG iterations take the form:

$$x^{k+1} = x^k - \frac{\alpha^k}{n}\sum_{i=1}^{n} y_i^k, \quad (5)$$

where at each iteration a random index i_k is selected, and we set:

$$y_i^k = \begin{cases} f_i'(x^k), & i = k. \\ y_i^{k-1}, & \text{otherwise} \end{cases}. \quad (6)$$

The step incorporates a gradient with respect to each function.

Random forests are a group of decision trees. Many decision trees are combined in random forest with the aim of decreasing the risk of overfitting. Generally, when there are lots of trees in the forest, it becomes more robust. The same idea applies to the random forest classifier (RFC), where there is higher number of trees in the forest, and the accuracy improves as well. For both, regression and classification tasks, the same algorithm of random forest can be applied. In this work, the RFC parameters are considered as *n-estimators* = 7 and *criterion* = 'entropy'.

NBC was considered in our study as well. It is within a family of more simplistic probabilistic classifiers based on Bayes' theorem: $P(H \mid E) = \frac{P(E \mid H)\,P(H)}{\sum_j P(E|H)P(H)}$, i.e. a formula of conditional probability based on hypothesis H and evidence E. However, the NBC makes a strong (naive) independence assumptions between the features. The NBC combines a decision rule in its model, usually the maximum a posteriori (MAP) as follows:

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} P(C_k)\prod_{i=1}^{n} P(x_i|C_k). \quad (7)$$

where $P(C_k)$ is the prior model over the classes C and $P(x_i \mid C_k)$ is the distribution (likelihood) given the features model.

V. Results

We have adopted the leave-one-out cross validation and k-fold cross validation tests over the KDEF dataset for 70 individuals who performed 7 different emotions. There are 210 images for each emotion group, with exception for surprise, which has 206 images, because four images were not processed since the face detection algorithm could not find the faces for features extraction and classification. Figure 2 indicates the number of images for which the face and facial landmarks have been correctly detected in each emotion group. In our system, the recognition confidence must be over 50% to consider a correct class given that we have 7 different classes of emotion. Note that a confidence level can influence the accuracy performance because some classifiers only check the highest score/probability even with confidence below 50%.

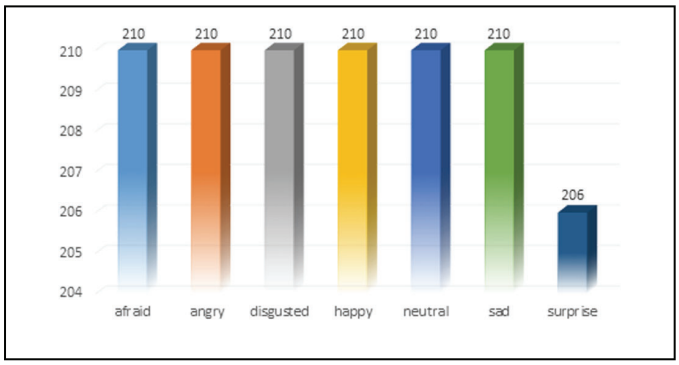

Fig. 2: Number of detected faces in each group of emotion. The vertical axis represents the number of images from the KDEF dataset. The horizontal axis shows the groups of emotion. On top of each bar we have the number of detected faces using all aforementioned libraries for detection and tracking.

TABLE I. Classification Results in terms of Accuracy

Classifier	Correct	Wrong	Accuracy
SVM	1280	186	**87.31%**
RFC	1076	390	73.4%
SAG	1152	314	78.58%
SMO	1138	328	77.76%
RT	701	943	47.81%
NBC	991	475	67.59%

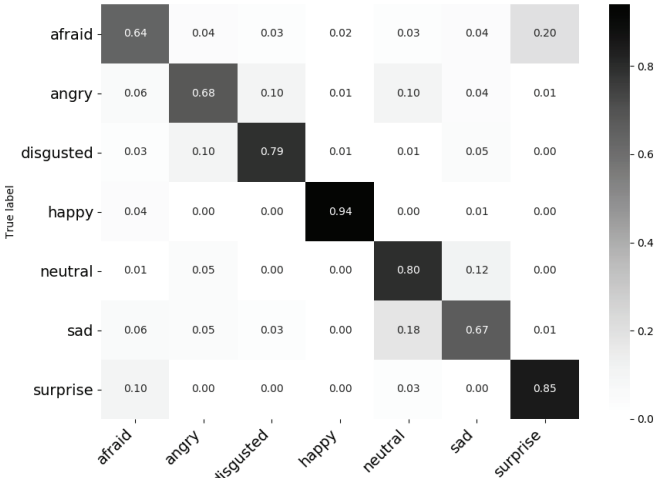

Fig. 3: Overall Confusion matrix for SVM tests.

The results attained using all aforementioned classifiers for emotion recognition are shown in Table 1. In order to better visualise the performance, a confusion matrix for each classifier was generated. Figures 3 to 8 show the results attained for the following classifiers: SVM, RFC, SAG, SMO, RT, and NBC, respectively. The multiclass SVM using the linear kernel implemented using the LibSVM library [19] for python attained the best performance using our set of features on the KDEF dataset as shown in Table 1 and Fig. 3. From the six classifiers tested, most of them achieved an accuracy above 75%, which

shows that the presented features are discriminative enough for emotion recognition.

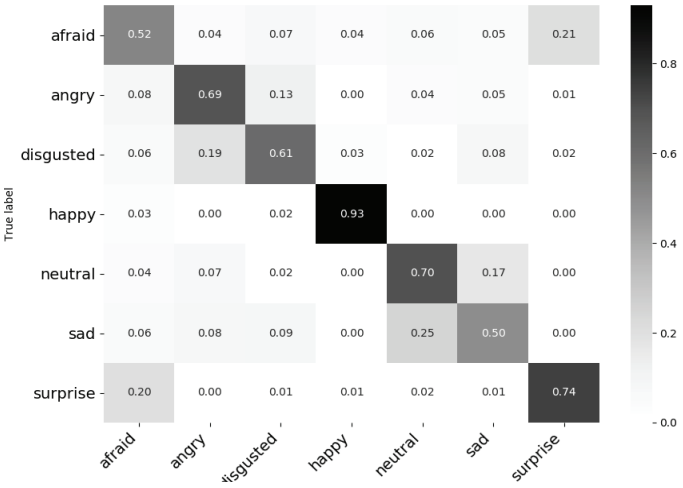

Fig. 4: Overall Confusion matrix for RFC tests.

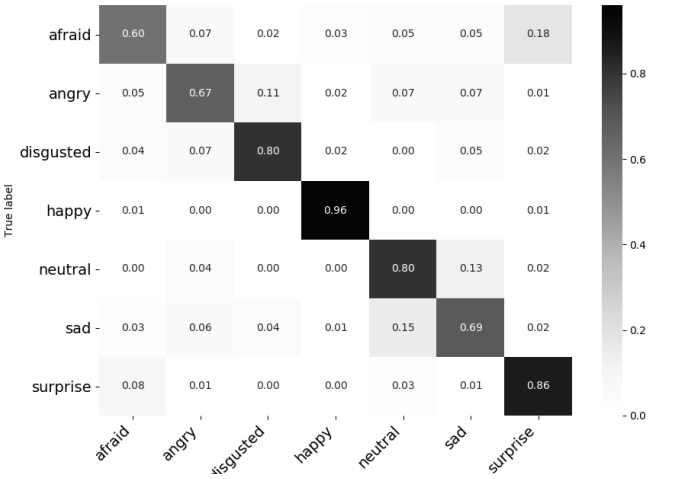

Fig. 5: Overall Confusion matrix for SAG tests.

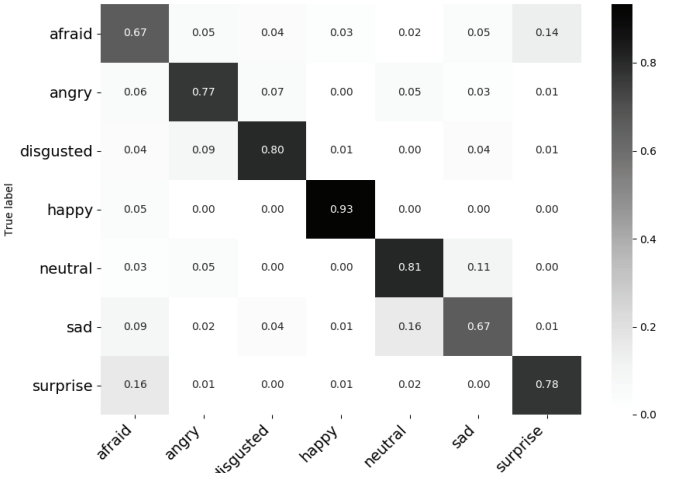

Fig. 6: Overall Confusion matrix for SMO tests.

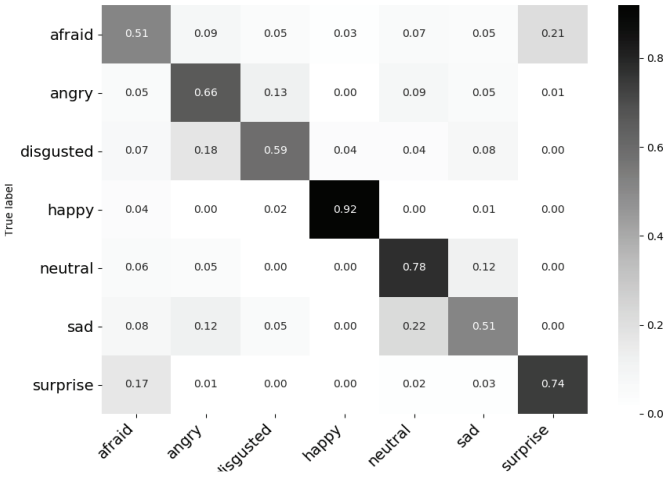

Fig. 7: Overall Confusion matrix for RT tests.

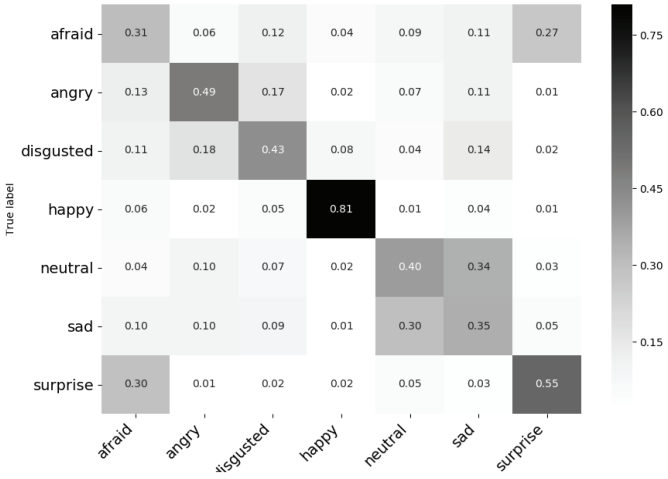

Fig. 8: Overall Confusion matrix for NBC tests.

VI. Conclusion

Emotional expression recognition is an important and challenging topic that has widely perceived benefits in various domains such as mental health, security, medicine, and general social communication software. This paper presented a study on recognising emotions using facial expressions. It defined a set of spatiotemporal features in order to discriminate seven classes of emotions. Experiments using multiple classification methods were run to verify whether the features are discriminative enough for emotion recognition. Results on the challenging KDEF dataset show that from the six classifiers tested, the multiclass SVM with a linear kernel attained the best performance in terms of accuracy (87.31%). Future work will be focused on (i) reducing the feature set required for accurate classification and (ii) development of the methods for real-time application to avatars in health care and particularly mental health where appropriate emotional responses are critical to the success of engagement and interventions.

AKNOWLEDGMENT

This research was part supported by the EIT Health GRaCE-AGE grant number 18429 awarded to C. D. Buckingham.

REFERENCES

[1] D. Lundqvist, A. Flykt, & A. Öhman (1998). "The Karolinska Directed Emotional Faces – KDEF". Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.

[2] P. Abhang, S. Rao, B. W. Gawali, and P. Rokade. (2011). "Emotion recognition using speech and eeg signal: a review," International Journal of Computer Applications, vol. 15, pp. 37–40.

[3] P. Ekman (1971). "Universals and cultural differences in facial expressions of emotion". Nebraska, USA: Lincoln University of Nebraska Press.

[4] M. L. Gale, A. Rizzo, J. Gratch, S. Scherer, G. Stratou, J. Boberg, L-P. Morency (2017). "Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers". Frontiers in Robotics and AI, Vol. 4.

[5] C. R. Imogen, E. Foenander, K. Wallace, M. J. Abbott, M. Kyrios, N. Thomas (2016). "What Role Can Avatars Play in e-Mental Health Interventions? Exploring New Models of Client–Therapist Interaction". Frontiers in Psychiatry, Vol. 7.

[6] C. D. Buckingham, A. Adams, L. Vail, A. Kumar, A. Ahmed, A. Whelan, E. Karasouli (2015). "Integrating service user and practitioner expertise within a web-based system for collab- orative mental-health risk and safety management". Patient Education and Counseling, pp. 1189-1196, .

[7] S.-C. Hsu, H.-H. Huang (2017). "Facial Expression Recognition for Human-Robot Interaction" IEEE International Conference on Robotic Computing (IRC).

[8] B. Gonsior, S. Sosnowski, C. Mayer, J. Blume, B. Radig, D. Wollherr and K. Kuhnlenz (2011). "Improving Aspects of Empathy and Subjective Performance for HRI through Mirroring Facial Expressions". IEEE International Symposium on Robot and Human Interactive Communication, pp. 350-356.

[9] S. L. Fernandesa, G. J. Bala (2016). "Developing a Novel Technique for Face Liveness Detection", Procedia Computer Science, v 78, pp. 241-247.

[10] F. Cid, J. A. Prado, P. Bustos and P. Nunez (2012) "Development of a facial expression recognition and imitation method for affective HRI" Workshop of Physical Agents.

[11] J. Bruce, J. Perron, and R. Vaughan (2017) "Ready-Aim-Fly! Hands-Free Face-Based HRI for 3D Trajectory Control of UAVs", Conference on Computer and Robot Vision (CRV).

[12] V. N. Dhage, P. L. Ramteke (2014) "Human Face Detection Using Genetic Algorithm: A Review", International Journal of Science, Engineering and Technology Research (IJSETR), 3(ISSN: 2278 – 7798), pp. 752-755.

[13] Dezyre (2018) Top 10 Machine Learning Algorithms, Available at: https://www.dezyre.com/article/top-10-machine-learning-algorithms/202 (Accessed: May 11, 2018).

[14] Viola and Jones, "Rapid object detection using a boosted cascade of simple features", Computer Vision and Pattern Recognition, 2001.

[15] V. Kazemi, J. Sullivan (2014). "One millisecond face alignment with an ensemble of regression trees". 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[16] M. Schmidt, N. Le Roux, F. Bach (2017). "Minimizing Finite Sums with the Stochastic Average Gradient". Mathematical Programming, v.162, issue 1–2, pp 83–112.

[17] J. C. Platt, (1999). "12 fast training of support vector machines using sequential minimal optimization". Adv. in kernel methods, pp.185-208.

[18] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. Murthy, (2001). "Improvements to Platt's SMO algorithm for SVM classifier design". Neural computation, 13(3), pp.637-649.

[19] C.-C. Chang and C.-J. Lin, (2011)."LIBSVM: A library for support vector machines," ACM TIST, http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[20] D. R. Faria, M. Vieira, F. C. C. Faria (2017). "Towards the Development of Affective Facial Expression Recognition for Human-Robot Interaction". ACM PETRA'17: 10th International Conference on PErvasive Technologies Related to Assistive Environments (NOTION: Human Behaviour Monitoring, Interpretation and Understanding), pp. 300-304.

[21] D. R. Faria, M. Vieira, F. C. C. Faria, C. Premebida (2017). "Affective Facial Expressions Recognition for Human-Robot Interaction", IEEE RO-MAN'17: IEEE International Symposium on Robot and Human Interactive Communication, pp. 805-810.