Automatic Estimation of Enjoyment Levels during Cardiac Rehabilitation Exercise

Haolin Wei
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
haolin.wei2@mail.dcu.ie

Kieran Moran
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
kieran.moran@dcu.ie

Noel E O'Connor
Insight Centre for Data Analytics,
Dublin City University
Dublin, Ireland
Noel.OConnor@dcu.ie

ABSTRACT

Cardiovascular disease (CVD) is the leading cause of premature death and disability in Europe and worldwide. Effective Cardiac Rehabilitation (CR) can significantly improve mortality and morbidity rates, leading to longer independent living and a reduced use of health care resources. However, adherence to such an exercise programme is generally low for a variety of reasons such as lack of time and how enjoyable the CR programme is. In this work, we proposed a method for automatic enjoyment estimation during an exercise which could be used by a clinician to identify when a patient is not enjoying the exercise and therefore at risk of early dropout. In order to evaluate the proposed method, a database was captured where participants perform various of CR exercises. Three set of facial features were extracted and were evaluated using seven different classifiers. The proposed method achieved 49% average accuracy in predicting five different enjoyment level on the newly collected database.

CCS CONCEPTS

• **Applied computing → Health care information systems**;

KEYWORDS

Enjoyment Recognition; Cardiac Rehabilitation; Affective Computing

ACM Reference Format:
Haolin Wei, Kieran Moran, and Noel E O'Connor. 2018. Automatic Estimation of Enjoyment Levels during Cardiac Rehabilitation Exercise. In *3rd International Workshop on Multimedia for Personal Health and Health Care (HealthMedia'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3264996.3265003

1 INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of premature death (30% of all deaths) and disability in Europe and worldwide (WHO), costing the EU economy almost €196 million [16]. Effective Cardiac Rehabilitation (CR) programmes can significantly improve mortality and morbidity rates, leading to longer independent living and a reduced use of health care resources. However, it is well known that the adherence rate for such programmes is generally low for a variety of reasons including lack of time and financial constrains [24]. The study carried out by [8] also indicated that patient enjoyment of a CR programme is also a key factor affecting adherence rates.

Enjoyment usually refers to the affective or mental state of having delight or pleasure in certain activities or experiences. In recent years, many methods and systems have been proposed for automatic affective state recognition of users. These methods and systems differ in what features are being used, which machine learning technique is chosen and what labels are being predicted. However, to the best of our knowledge, there has been no research carried out to explore the use of automatic enjoyment recognition in an CR exercise scenario. Although different features from audio and biomedical modalities have been widely used for affective state recognition, due to the nature of an exercise scenario, there is not much verbal and non-verbal communication involved and as suggested by [6] obtaining the accurate measurement from biomedical sensors is still affected by human physical activities. As a result, in this paper, we focus on investigating an approach to automatic enjoyment estimation using visual features. The paper is divided as follows: In Section 2, we present a brief review of the related works on automatic affect and enjoyment recognition; Section 3 presents the details of our novel database along with the proposed feature extraction process and the selected machine learning techniques. Section 4 shows the experimental results on the new database. The paper concludes in Section 5 with a summary of our work and possibilities for further extensions.

2 RELATED WORK

In recent years, research in the field of automatic recognition of affective state has received a lot of attention. [23] showed an early attempt to analyse facial expressions automatically by tracking the motion of twenty identified points on the face. With advances in machine learning, [10] proposed to recognise the six basic emotions (happiness, sadness, fear, disgust, anger and surprise) from face images using a neural network. The recent work developed by [9] have achieved 78.61% average accuracy over six basic emotions on the MMI dataset [18] using a deep learning approach. The work carried out by [17] presented one of the first attempts to continuously recognise spontaneous affect in the valance-arousal parameter space using both visual and vocal cues. Very little research has focused on automatic enjoyment recognition. [12] developed a real-time system using single-channel EEG signal to quantify user enjoyment level elicited by media content. [14] proposed an

HealthMedia'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5982-5/18/10…$15.00
https://doi.org/10.1145/3264996.3265003

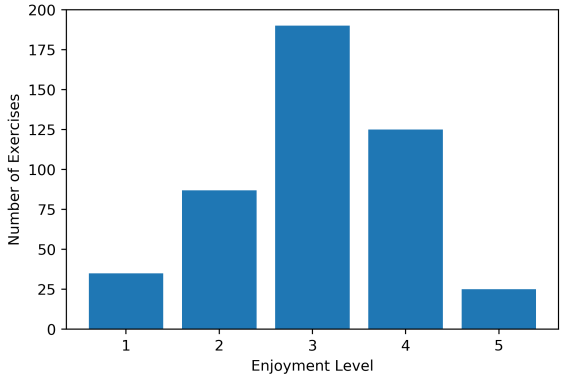

Figure 1: Enjoyment distribution of the collected database.

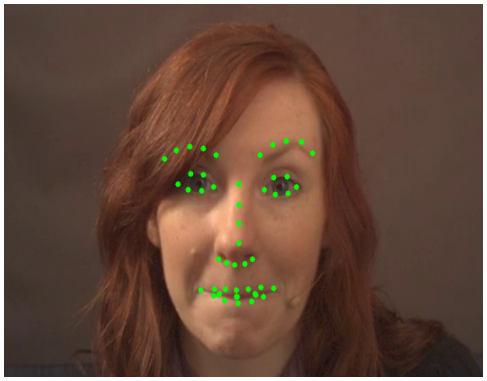

Figure 2: Sample Landmarks.

event fusion based approach for classifying the enjoyment-episodes within the audiovisual Belfast Story Telling Corpus.

3 METHODOLOGY

This section describes the data and methods used for our study. The new enjoyment database employed in this study is first introduced. Methods used to extract features from video recordings are then described, followed by a brief discussion of the different classification methods that are used to perform automatic enjoyment recognition in this work.

3.1 Data Acquisition

The Microsoft Kinect V2 sensor was used to capture the video data for the participants while they were doing the exercises. The Kinect V2 sensor is able to capture the video data at 1920×1080, 30 frame per seconds. The Kinect sensor was connected to a workstation PC via USB 3 and was placed on a table 2 meters away from the participants.

In total 46 participants (26 females, 20 males) from the MedEx Wellness programme and the School of Health and Human Performance of Dublin City University, Ireland, were recruited. Participants were told that the purpose of this data capture is to develop an automatic action detection system using video data. Not knowing the real objectives of the experiment avoided having participants exaggerate or mask their true affective state [15]. The participants were divided into 4 groups. 15 Cardiac Rehabilitation (CR) exercises were selected for each group and the participants were asked to perform each exercise for 1 minute mimicking an avatar performing the exercise displayed on the computer monitor. At the end of each exercise, the participant was asked to rate the enjoyment level using a 5-Likert scale with 1 being not enjoyable at all and 5 being very enjoyable. The enjoyment level distribution for the database is shown in Figure 1.

3.2 Feature Extraction

Facial expressions could indicate a person's affective state, intentions and ultimately, elicit other people's response. For instance, [22] suggested that it is possible to infer other people's affective state just by looking at the individual's face, without any complementary information such as voice or gesture, indicating that the face could be the most effective communication tool for understanding enjoyment levels. In this work, we choose to use facial landmark features for automatic enjoyment level estimation. The facial landmarks were detected using the IntraFace library based on the Supervised Descent Method [26]. In total, 49 facial landmarks were detected for each frame in each video as shown in Figure 2. Each landmark is represented by a 2D image coordinates and piecewise interpolation was used to cope with missed detections, due to large head movement and motion blur. In total, three sets of features were extracted based on the facial landmarks obtained from the previous step. Each feature set was first extracted at frame level and the mean and variance were calculated as the video level features. The first set of features include the mean and variance of the raw landmark positions. This generates 196 features for each video. The second set of features include the mean and variance of the aligned landmark positions using Procrustes Analysis [5]. This generates 196 features for each video. The third set of features use the geometric features proposed in [21]. It includes: i) the difference between the aligned landmarks and the mean shape, and also between previous and the current frame; ii) the Euclidean distances (L2-norm) and the angles (in radians) between the points in left eye and left eyebrow region, right eye and right eyebrow region and the mouth region; iii) the Euclidean distance between the median of the stable landmarks and each aligned landmark in each frame. The mean and variance of the geometric features are then calculated for each video. This generates 732 features for each video.

3.3 Classification

In total, seven commonly used machine learning techniques are selected to evaluate the performance of enjoyment level prediction. This includes Nearest Neighbours, Support Vector Machine (SVM), Decision Tree, Random Forest, Neural Net, AdaBoost and Quadratic Discriminant Analysis. The efficiency of the proposed classifiers have already been proved by other researches for affect recognition using facial features [1–3, 11, 13, 20, 25]. All training was performed using the Scikit-Learn machine learning library [19]. For the Nearest Neighbours classifier, the number of neighbours was set to 5 with uniform weight. For the SVM, a linear kernel was selected and the

complexity parameter was optimised with values in the $[10^{-4} - 10^{0}]$. For the decision tree classifier the max depth was set to 20. For the Random Forest classifier the number of trees in the forest was set to 10 with the max depth of the tree was set to 20. The Neural Net classifier was set to consist of 1 hidden layer with 100 neurons. For the Adaboost classifier, the decision tree base classifier is used with maximum number of estimators set to 20. Finally, for Quadratic Discriminant Analysis, the threshold used for rank estimation was set to 0.001.

4 RESULTS AND DISCUSSION

Due to the low number of samples in certain classes, 5-fold cross validation was used to evaluate the performance of each classifier. The performance is measured as the averaged classification accuracy across 5 folds. Before the experiments, the data was first shuffled. Then the training data and the testing data was selected randomly to ensure the training and testing data have similar data distribution. The average classification accuracy is shown in Table 1. Among all combinations of different features sets and classifiers, the random forest classifier with geometric features achieved the best result at 49% followed by the aligned landmarks features at 47%. Comparing the performance of different classifiers, the random forest achieved the best results followed by the SVM classifier. The normalised confusion matrix for random forest classifier using geometric feature is shown in Figure 3 and it can be seen that enjoyment level 3 and 4 perform well compared to the other enjoyment levels. This could be caused by the fact that the captured database consists of an imbalanced number of samples for different enjoyment levels as shown in Figure 1. To further explore this issue, the over-sampling technique was used. In particular, three oversampling techniques were evaluated including random sampling, the Synthetic Minority Oversampling Technique (SMOTE) [4] and Adaptive Synthetic (ADASYN) sampling [7]. The results for each oversampling method are shown in Figure 4, 5 and 6. As can be seen, both SMOTE and ADASYN oversampling methods improved the classification accuracy on enjoyment level 1 and 2, but decreased the performance on enjoyment 3 and 4. An explanation could be that the captured data between different enjoyment levels are not discriminative since only one label is given to the entire exercise. For instance the same participant who has a high enjoyment level at the beginning of the exercise but low enjoyment level at the end of the exercise might reports different enjoyment levels when he/she has a low enjoyment level at the beginning but high enjoyment level at the end, however, the features extracted for these two exercises might be similar.

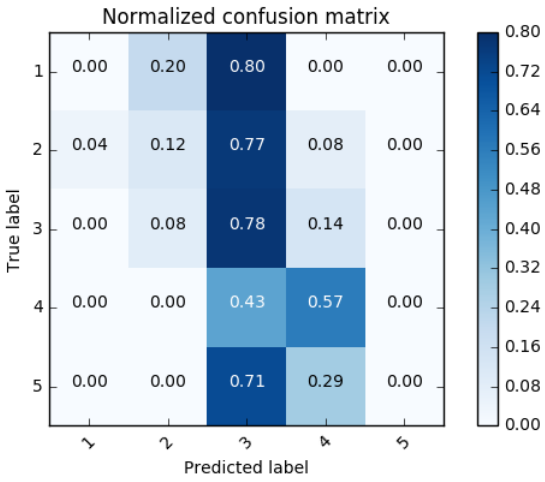

Figure 3: Normalised Confusion Matrix for Random Forest Classifier using geometric feature (CA=0.52).

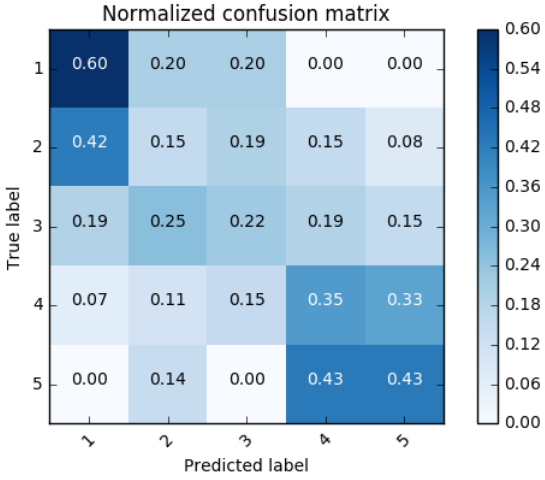

Figure 4: Normalised Confusion Matrix for Random Forest Classifier using geometric feature and Random upsampling (CA=0.27).

5 CONCLUSION

In this paper, a method for automatic estimation of enjoyment level in a CR exercise scenario was proposed. A novel enjoyment database for CR exercise was captured and was used to evaluate the performance of different visual feature sets and classifiers combinations. Due to the imbalanced database, oversampling methods were further investigated. In future work, feature fusion and new features using temporal information could be investigated to see if this can further improve the recognition accuracy. In addition, representative sample selection could also be explored to overcome the imbalanced problem.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant NO.: SFI/12/RC/2289. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation Action under Grant Agreement No.: 643491.

REFERENCES

[1] Ralph Adolphs. 2002. Neural systems for recognizing emotion. *Current opinion in neurobiology* 12, 2 (2002), 169–177.

Table 1: Classifier performances. Best accuracy is achieved using Random Forest classifier with geometric features

	Raw Landmarks	Aligned Landmarks	Geometric Features
Nearest Neighbours	35%	45%	36%
Support Vector Machine	43%	43%	43%
Decision Tree	31%	37%	35%
Random Forest	41%	47%	**49%**
Neural Net	28%	36%	33%
AdaBoost	37%	39%	41%
Quadratic Discriminant Analysis	43%	43%	27%

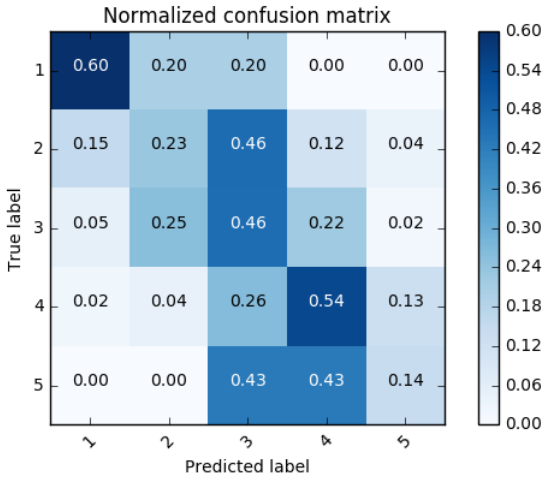

Figure 5: Normalised Confusion Matrix for Random Forest Classifier using geometric feature and SMOTE up-sampling (CA=0.43).

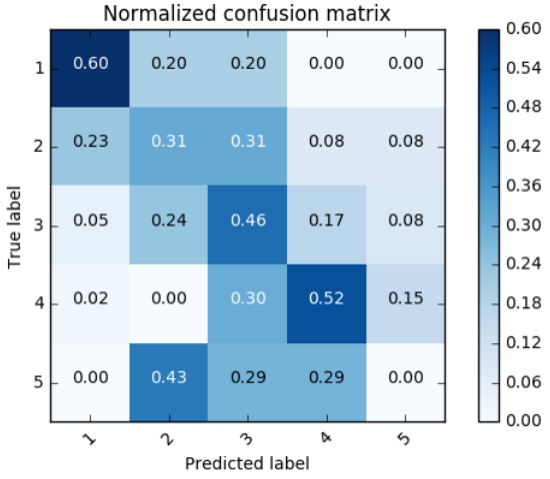

Figure 6: Normalised Confusion Matrix for Random Forest Classifier using geometric feature and ADASYN up-sampling (CA=0.43).

[2] Jeremy N Bailenson, Emmanuel D Pontikakis, Iris B Mauss, James J Gross, Maria E Jabon, Cendri AC Hutcherson, Clifford Nass, and Oliver John. 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International journal of human-computer studies* 66, 5 (2008), 303–317.

[3] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 205–211.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] John C Gower. 1975. Generalized procrustes analysis. *Psychometrika* 40, 1 (1975), 33–51.

[6] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[7] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 1322–1328.

[8] Natalie A Johnson and Richard F Heller. 1998. Prediction of patient nonadherence with home-based exercise for cardiac rehabilitation: the role of perceived barriers and perceived benefits. *Preventive Medicine* 27, 1 (1998), 56–64.

[9] Dae Hoe Kim, Wissam Baddar, Jinhyeok Jang, and Yong Man Ro. 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing* (2017).

[10] Hiroshi Kobayashi and Fumio Hara. 1991. The recognition of basic facial expressions by neural network. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*. IEEE, 460–466.

[11] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication* 53, 9-10 (2011), 1162–1171.

[12] Zhen Liang, Hongtao Liu, and Joseph N Mak. 2016. Detection of media enjoyment using single-channel EEG. In *Biomedical Circuits and Systems Conference (BioCAS), 2016 IEEE*. IEEE, 516–519.

[13] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. 2000. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems* 31, 3 (2000), 131–146.

[14] Florian Lingenfelser, Johannes Wagner, Elisabeth André, Gary McKeown, and Will Curran. 2014. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 377–386.

[15] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek. 2011. 3D corpus of spontaneous complex mental states. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 205–214.

[16] Peter Scarborough Mike Rayner Jose Leal Ramon Luengo-Fernandez Alastair Gray Melanie Nichols, Nick Townsend. 2012. *European Cardiovascular Disease Statics 2012 edition*. https://www.escardio.org/static_file/Escardio/Press-media/press-releases/2013/EU-cardiovascular-disease-statistics-2012.pdf.

[17] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[18] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. 2011. Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 1449–1456.

[21] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1335–1336.

[22] J. Segal. 2008. *The Language of Emotional Intelligence: The Five Essential Tools for Building Powerful and Effective Relationships*. McGraw-Hill Education. https://books.google.ie/books?id=wOefbfeiIfcC

[23] Motoi Suwa, Noboru Sugie, and Keisuke Fujimora. 1978. A preliminary note on pattern recognition of human emotional expression. In *International joint conference on pattern recognition*. 408–410.

[24] Rod S Taylor, Hayes Dalal, Kate Jolly, Tiffany Moxham, and Anna Zawada. 2010. Home-based versus centre-based cardiac rehabilitation. *The Cochrane database of systematic reviews* 1 (2010), CD007130.

[25] Yubo Wang, Haizhou Ai, Bo Wu, and Chang Huang. 2004. Real time facial expression recognition with adaboost. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, Vol. 3. IEEE, 926–929.

[26] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 532–539.