A Quality Adaptive Multimodal Affect Recognition System for User-Centric Multimedia Indexing

Rishabh Gupta * [1,2], Mojtaba Khomami Abadi * [1,3],
Jesús Alejandro Cárdenes Cabré [1], Fabio Morreale [3], Tiago H. Falk [2], Nicu Sebe [3]
[1] Sensaura Inc., Montréal, Canada
[2] INRS-EMT, University of Quebec, Montréal, Canada
[3] Department of Information Engineering and Computer Science, University of Trento, Italy
rishabh.gupta@emt.inrs.ca, khomamiabadi@disi.unitn.it,
jesus.cardenes@sensauratech.com, morreale@disi.unitn.it, falk@emt.inrs.ca,
sebe@disi.unitn.it

ABSTRACT

The recent increase in interest for online multimedia streaming platforms has availed massive amounts of multimedia information that need to be indexed to be searchable and retrievable. User-centric implicit affective indexing employing emotion detection based on psycho-physiological signals, such as electrocardiography (ECG), galvanic skin response (GSR), electroencephalography (EEG) and face tracking, has recently gained attention. However, real world psycho-physiological signals obtained from wearable devices and facial trackers are contaminated by various noise sources that can result in spurious emotion detection. Therefore, in this paper we propose the development of psycho-physiological signal quality estimators for unimodal affect recognition systems. The presented systems perform adequately in classifying users affect however, they resulted in high failure rates due to rejection of bad quality samples. Thus, to reduce the affect recognition failure rate, a quality adaptive multimodal fusion scheme is proposed. The proposed scheme yields no failure, while at the same time classify the users' arousal/valence and liking with significantly above chance weighted F1-scores in a cross-user experiment. Another finding of this study is that head movements encode liking perception of users in response to music snippets. This work also includes the release of the employed dataset including psycho-physiological signals, their quality annotations, and users' affective self-assessments.

Keywords

Affective Computing; Multimedia Indexing; Cross-User; User-Centric; Implicit Affective Tagging; Psycho-Physiological Signals; Signal Quality; Multimodal Interaction; Decision Fusion

*these authors contributed equally to the implementation of this work.

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912059

1. INTRODUCTION

The burgeoning number of platforms for online multimedia streaming/storage (e.g. YouTube, Netflix) has resulted in generation of a huge database of multimedia content online. In 2015, approximately 400 hours of videos were uploaded every minute on YouTube[1]. This enormous amount of data is not limited to the online realm, the availability of portable devices storing thousands of music tracks and pictures has brought massive amounts of multimedia information in our pockets. However, the huge amount of generated multimedia information needs to be indexed to be searchable and retrievable by the users.

Multimedia retrieval systems that rely on user-generated labels for indexing are potentially biased by subjective judgements and/or intentions[11]. Moreover, manual tagging of multimedia content interrupts the user experience process. Therefore, it is necessary to automate the process of multimedia indexing through implicit tagging. The classical multimedia indexing relies on cognitive indexing procedures which are based on concepts to characterize the multimedia content, such as locations, objects, and events. Whereas, a recent approach, the so-called affective indexing, depends on the emotions generated by the multimedia content [15]. The implicit affective indexing technique is expected to provide more detailed and meaningful information regarding users experience with multimedia [11, 5]. Previously, affective tags have been used for indexing multimedia content for improving information retrieval and recommendation systems [14, 5].

This work presents a multimodal approach on a hard implicit affective-indexing problem. We tackled cross-user and user-centric implicit affective-tagging of (weakly-affective) and short music snippets when the information sources are damaged by various noise artifacts. We achieved significantly above chance results for classification of user perceptions on affective music snippets and we also found that head movements encode *liking* perception in response to music snippets. We made the dataset publicly available for research community so that other researcher can improve our methods.

2. RELATED WORK

Two general approaches of generating affective tags for multimedia content are (i) using the information content of multimedia [3] and (ii) using the human affective perception (via detecting users' emotions) to tag the perceived content (implicit user-centric approach [4]). Certainly a hybrid approach [6, 8] could be success-

[1] www.reelseo.com/hours-minute-uploaded-youtube/

Table 1: Features used for quality estimation and affect recognition for each modality.

Modality	Quality Estimation	Affect Recognition
NeuroSky EEG	Statistical measures (such as, mean, median, skewness, kurtosis) for EEG data, power spectral features in ranges (0-1, 0-4, 4-8, 8-12, 12-30, 30-50, 58-62 Hz)	Band powers for δ, θ, α, β and γ bands, statistical measures for cognitive measures provided by NeuroSky
ECG	Statistical measures for heart rate and heart rate variability, power spectral features from ECG for ranges (5-15, 5-40, 0-40, 1-40 Hz) and their ratios	Wavelet based power spectral features over ECG and HRV, statistical measures for the spectral features and Poincare features
GSR	Statistical measures for raw GSR signal, and GSR signal band-passed between ranges (0-0.08, 0.08-0.2, 1-2, 1-2, 10-20, 20-30 Hz)	Power spectral features, rise time, fall time and zero crossing rate for very low frequency (≤ 0.08 Hz) and low frequency ($0.08 - 0.2$ Hz) components of the signal
Face/Head-pose	Features encoding information regarding lips thickness, ratios, such as upper lip to lower lip thickness, eye brows width to lips width	Statistical measures, rise time, drop time and zero crossing rate for head's pitch, yaw and roll

fully utilized. This work follows the second approach in a multimodal scheme.

Affective computing techniques have been successfully utilized for detection of users' emotions in response to multimedia content [17, 6, 8, 15] by leveraging the information of a plethora of modalities including facial expressions, gestures, body postures, voice, heart activities, electrodermal signals, and brain responses.

Since many of the underlying affective patterns in the above mentioned modalities are highly subjective (i.e. they significantly vary from one user to another), most of the state of the art user-centric emotion recognition studies focus on subject-based emotion recognition when the training and the test data for evaluation of a scheme comes from the same user [6, 8, 15]. However, a scalable implicit affective indexing system should be able to classify human emotions for any *unseen* user. Such framework where the test data is from unseen users is called *cross-user*.

A multimodal system for affect recognition is expected to perform better than a unimodal system, as reported in previous studies [6, 2, 8, 15], which can be attributed to the fusion of complimentary information provided by each modality.

However, the signals from the above mentioned modalities are often contaminated with various sources of noise [13], that significantly hinders the task of affect recognition. Particularly, in real-world biomedical signals, for e.g. signals obtained from wearable devices, the problem of noise contamination is more exaggerated.

This paper presents the first steps towards the validation and development of a user-centric multi-modal quality adaptive affect recognition system for cross-user implicit affective indexing of multimedia content.

3. MATERIALS AND METHODS

3.1 Experimental Setup

We recruited a total of thirty-three participants (with age distribution of 29.7 ± 5.4 years, 21 males) for the study. The participants were asked to listen to music excerpts, originally generated by *Robin* [10], an algorithmic composer that generates western classical-like music with affective connotation in real time. The stimuli are weakly affective being evident from the fact that facial expressions in the dataset are negligible if not absent.

The participants experienced music excerpts using a AKG K512 headphone inside a silent room at the University of Trento, Italy. At the beginning of the experiment, four training excerpts were played to the participants to make the subjects familiar with the task. During the experiment, participants were presented with twenty thirty-two seconds long music excerpts in random order. While experiencing the music excerpts, participants' physiological signals, including electrocardiography (ECG), galvanic skin response (GSR), frontal electroencephalography (EEG), and facial videos were recorded. The ECG and GSR signals were recorded using the Shimmer sensors, whereas for recording EEG, we used a NeuroSky Mindwave headset. Moreover, participants' facial videos were recorded using an A4Tech webcam at a resolution of 640 X 480 pixels and the SDM face alignment method[16] is employed to detect users' head poses and facial landmark tracks. All the recordings are available online via the dataset website[2].

In order to measure valence, arousal and liking, participants were asked to rate them on three seven-point semantic differential scales, from 1 (negative, relaxing or unlike) to 7 (positive, exciting or like). To record the levels of arousal, valence and liking, participants were asked to type in numbers between 1-7 on a keyboard after listening to each excerpt. Moreover, to reduce emotional bias, a sequence of randomly generated notes were played, from a set of five 15 second long pre-recorded snippets, between each music excerpt. The obtained subjective scores showed substantial inter-rater agreement as measured using intra-class correlation (ICC), where ICC for arousal, valence and liking was 0.79, 0.63 and 0.81, respectively.

3.2 Signal Quality Estimator Development

Towards developing a quality adaptive affect recognition system signal quality estimators (SQEs) were developed for each modality. In order to realize the SQEs, first, the quality of the signals was assessed by two expert annotators, forming the ground truth. We employed the Cohen's kappa to measure inter-rater agreement over the annotations of the experts and measures of 0.96, 0.98, 0.94, 0.73 were observed over the quality annotations for EEG, Facial detections, ECG, GSR, respectively. The first three are indicating *almost perfect agreement (≥ 0.9)* and the agreement on GSR quality is *substantial (≥ 0.7)*. The marginal disagreement on the GSR quality could be due to two issues: (i) GSR has a slow response so that the structures of GSR signals sometimes are not recognizable over short recordings, and (ii) noise artifacts that are present on the signal sometimes induce patterns that are similar to the structure of GSR responses. In cases of disagreement between the two annotators, the annotation of the second expert is employed. Features listed in Table 1 were extracted, as they encode signal quality information for each modality. Finally, a bagging classifier with decision trees as a base estimator was implemented to differentiate between good/bad quality signals for each modality. The classification performance was validated using a leave-one-subject-out cross-validation i.e., the classifier was trained on samples from all the subjects except the one used for testing. Moreover, the performance of the SQEs was assessed using a weighted F1-score [12], as it accurately measures the classifier performance for a highly imbalanced classification problem.

3.3 Affect Classifier Development

As a next step, features (listed in Table 1, column labeled 'Affect recognition') that encode affect [6], were extracted. The affect en-

[2]mhug.disi.unitn.it/wp-content/QAMAF/QAMAF.html

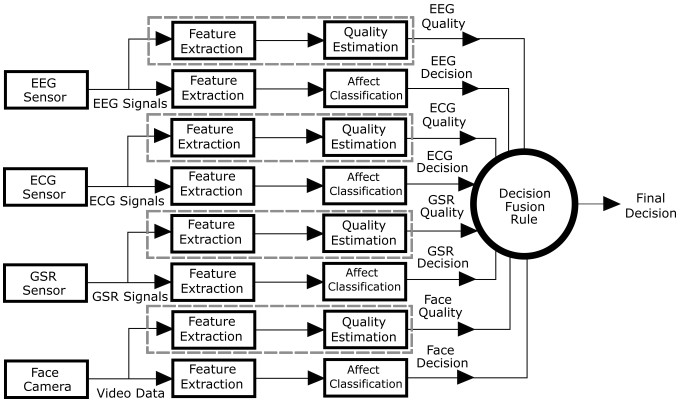

Figure 1: Quality adaptive multimodal decision level fusion schema.

coding features were then employed with linear support vector machine (SVM, C = 1.0) and Naive Bayes (NB) classifiers for three binary classification problems of differentiation between high/low valence, arousal and liking, respectively. For validation of the developed classifiers, a cross-subject classification approach was adopted using a leave-one-subject-out cross-validation [1]. Due to high intra-variability among human signal patterns, cross-subject affect recognition is a *harder*, but more *generalizable*, problem than subject specific affect recognition (as used in [6]). In this study we validated our results using cross-subject cross-validation.

In our proposed quality adaptive affect recognition scheme, during the development of affect binary classifiers, good quality samples form the train data and SQEs are applied to assess the quality of the test samples. The SQEs provided the reliability of the predictions thus, assisting in rejecting the bad quality samples. For accurate measurement of classifier performance for an imbalanced classification problem, we computed weighted F1-scores [12]. The performance of each classifier developed above was tested for significance against random voting using a paired *t-test*.

3.4 Quality Adaptive Multimodal Fusion

The quality adaptive multimodal decision fusion affect recognition system in this study is developed by modifying the decision fusion scheme presented in [9]. In our study, we used a decision fusion classifier for multimodal fusion as depicted in Fig. 1. In decision level fusion, a linear combination of the individual unimodal classifiers' outputs is calculated as the output. The decision fusion can be implemented (i) using an equal weight scheme i.e., all modalities used for affect recognition being given equal weights or (ii) using an optimal weight scheme i.e., the weight for each modality can be optimised given a set of training data [9]. The weights indeed encode the relative importance of the unimodal classifiers in calculation of the final output. The formulation and implementation of the fusion scheme that is employed in this study are described below.

3.4.1 Formulation

For the decision level fusion scheme, the fusion classification probability $p_0^x \in [0,1]$ for each class $x \in \{1,2\}$ can be denoted by

$$p_0^x = \sum_{i=1}^{N} \alpha_i p_i^x q_i t_i \tag{1}$$

where, i is the index of a particular modality used for affect recognition, N is the number of modalities used, α_i are the weights corresponding to each modality ($\sum_{i=1}^{N} \alpha_i = 1$), q_i [3] corresponds to the quality of the respective modality and t_i is the normalized training set performance for a particular modality, such that the fusion probabilities for all classes sum up to 1, and is given by

$$t_i = \frac{F_i}{\sum_{i=1}^{N} \alpha_i q_i F_i} \tag{2}$$

where, F_i is the F1-score obtained on the training set using a particular modality and $F_i \in [0,1]$. Then, it can be shown that,

$$p_0^1 + p_0^2 = \sum_{i=1}^{N} \alpha_i q_i t_i = \sum_{i=1}^{N} \left(\frac{\alpha_i q_i F_i}{\sum_{i=1}^{N} \alpha_i q_i F_i} \right) = 1 \tag{3}$$

3.4.2 Implementation

The quality adaptive fusion scheme described above was implemented using equal weights for EEG, ECG, GSR and head pose. Therefore, the weights used for fusion were $\alpha_i = 0.25$ and the class probabilities from each single modality is given by,

$$p_Q^x = 0.25 \times (p_{ee}^x q_{ee} t_{ee} + p_{ec}^x q_{ec} t_{ec} + p_{gs}^x q_{gs} t_{gs} + p_{hp}^x q_{hp} t_{hp}) \tag{4}$$

where subscripted 'Q' denotes quality adaptive system, abbreviations ee, ec, gs and hp denote EEG, ECG, GSR and head pose, respectively. Whereas, for non-adaptive multimodal fusion was developed using similar decision rule while excluding the quality term 'q_i' from the equation resulting in,

$$p_{nQ}^x = 0.25 \times (p_{ee}^x t_{ee} + p_{ec}^x t_{ec} + p_{gs}^x t_{gs} + p_{hp}^x t_{hp}) \tag{5}$$

where subscripted 'nQ' denotes non-adaptive system for multimodal fusion.

4. RESULTS

The SQEs for each modality performed adequately for each modality, as the weighted F1-scores were as follows: EEG - 0.93, ECG - 0.95, Face/headpose - 0.86 and GSR - 0.78, while the weighted F1-score for random voting for each modality was 0.62. Moreover, the performance of the affect classifiers (both, quality adaptive and non-adaptive) are reported in Table 2, which can be compared to the baseline weighted F1-score of 0.50, obtained from random voting for valence, arousal and liking. It was observed that, using the quality adaptive affect recognition system, for arousal, all the modalities performed significantly better than chance. Moreover, for valence EEG, ECG and face/headpose produced significant results whereas, for liking only face/headpose using a SVM classifier performed significantly better than chance. The non-adaptive affect recognition systems resulted in very few significant results. However, it should be noted that *uni-modal* quality adaptive systems had higher failure rate due to sample rejections. The non-adaptive multimodal decision fusion produced significantly better than chance classification performance for the three subjective dimensions of valence, arousal and liking, using the NB classifier whereas, no significant results were obtained using SVM classifier. Furthermore, the quality adaptive multimodal decision fusion also produced significantly better than chance classification performance for valence, arousal and liking, using the NB classifier whereas, using SVM classifier significant results were observed only for valence classification. Moreover, the sample rejection rate was brought down significantly, to 0%, using the quality adaptive multimodal fusion.

5. DISCUSSION AND CONCLUSION

The developed SQEs performed well for each modality (weighted F1-score of about 0.90). However, for GSR the obtained weighted F1-score was relatively lower than the other modalities

[3] $q_i \in \{1,0\}$: i.e. good or bad according to the output of the SQE binary classification for the i^{th}modality

Table 2: Classification results for the self-assessment of valence (V), arousal (A) and liking (L) using a leave-one-subject-out cross-validation schema. The weighted F1-scores significantly higher than chance level (0.50) are highlighted (superscripted $* : p < 0.05$). The table also lists the percentage of samples rejected in quality adaptive schema for each modality.

Modality	Classifier	Quality Adaptive				Non-Adaptive		
		A	V	L	Rejections	A	V	L
NeuroSky EEG	SVM	**0.57***	0.53	0.54	22.58%	0.52	0.50	0.51
	NB	**0.56***	**0.57***	0.53		0.52	0.50	0.51
ECG	SVM	**0.59***	**0.57***	0.53	18.33%	0.57*	0.52	0.50
	NB	0.54	**0.57***	0.54		0.55	0.55*	0.51
GSR	SVM	0.52	0.52	0.46	13.79%	0.51	0.52	0.52
	NB	**0.55***	0.48	0.54		0.54	0.52	0.53
Headpose	SVM	0.55*	**0.58***	**0.58***	5.76%	**0.56***	0.54	**0.58***
	NB	0.50	0.55*	0.53		0.53	0.52	0.55
Decision Fusion	SVM	**0.60***	**0.59***	**0.58***	**0%**	0.57*	0.54	**0.58***
	NB	**0.57***	**0.58***	**0.56***		0.56*	0.56*	0.55*

suggesting that better features could aid in improving the GSR signal quality estimation. Moreover, the advantage of using a quality adaptive affect recognizer is evident from Table 2, where the quality adaptive affect recognizer produced more number of significant results compared to a non-adaptive affect recognizer. However, a higher percentage of sample rejections for *uni-modal* quality adaptive systems resulted in their higher failure rate. The efficacy of multi-modal fusion techniques, both quality adaptive and non-adaptive, is evident as both techniques produced significant results for all three affective dimensions. The quality adaptive multimodal fusion has an added advantage of decreasing the failure rate resulting from quality adaptive uni-modal systems and achieving slightly higher performances. Moreover, the results reported in Table 2 validate the performance of our approach for *cross-user* affect recognition in noisy recordings (e.g. due to noise in environment) for multimedia implicit affective indexing.

Furthermore, it is worth noting that quality adaptive arousal and valence classifiers performed significantly above chance on all the modalities except on valence recognition using GSR that is in corroboration with the finding in [7]. It is worthy to mention that low unimodal performances on GSR could be due to the fact that GSR responses are slow. Therefore, GSR is an unsuitable modality for an experiment with short recordings like ours.

According to the observed results, a link between participants' head-pose (e.g. following the rhythm of music) and the likeability of a music excerpt was observed as the liking classifiers developed using head-pose features performed significantly above chance. Head-pose also significantly encodes valence and arousal perceptions.

In the study presented here, we developed a quality adaptive multimodal affect recognition system for cross-user and user-centric implicit mutlimedia indexing. The multimodal system was developed using data from four different modalities of EEG, ECG, GSR and face/headpose videos while users experienced affective music generated by an algorithmic composer, *Robin*. The signal quality for each modality was estimated first using a bagging classifier, which was followed by affect recognition. The quality adaptive uni-modal affect recognition performed better than chance however, these systems resulted in high failure rate due to bad quality sample rejection. Towards decreasing the failure rate of the uni-modal affect recognition systems, we proposed a quality adaptive multimodal decision fusion rule, giving equal weights to each modality, which performed adequately for affect recognition while lowering the failure rate. However, to improve the classification performance, quality adaptive modality weight optimisation should be explored in future studies. We release the dataset for the community, so that researchers in the community would improve our baseline results employing more innovative signal-quality adaptive methods.

6. ACKNOWLEDGEMENTS

This research was funded and supported by Sensaura Inc. The dataset was collected at the University of Trento, Italy and we acknowledge the great support we received during the data collection. We do also acknowledge Mr. Jean-Philip R. Poulin for his support at Sensaura Inc.

7. REFERENCES

[1] M. Esterman, B. J. Tamber-Rosenau, Y.-C. Chiu, and S. Yantis. Avoiding non-independence in fmri data analysis: leave one subject out. *Neuroimage*, 50(2):572–576, 2010.

[2] R. Gupta, K. ur Rehman Laghari, and T. H. Falk. Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization. *Neurocomputing*, 174(PB):875–884, Jan. 2016.

[3] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[4] M. Khomami Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe. User-centric affective video tagging from meg and peripheral physiological responses. In *ACII*, 2013.

[5] M. Khomami Abadi, J. Staiano, A. Cappelletti, M. Zancanaro, and N. Sebe. Multimodal engagement classification for affective cinema. In *ACII*, 2013.

[6] M. Khomami Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, July 2015.

[7] J. Kim and E. AndrŐ. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, Dec 2008.

[8] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, Jan 2012.

[9] S. Koelstra and I. Patras. Fusion of facial expressions and {EEG} for implicit affective tagging. *Image and Vision Computing*, 31(2):164 – 174, 2013.

[10] F. Morreale, R. Masu, and A. De Angeli. Robin: an algorithmic composer for interactive scenarios. *Proceedings of 10th Sound and Music Comp.*, 2013.

[11] M. Pantic and A. Vinciarelli. Implicit human-centered tagging [social sciences]. *IEEE Signal Processing Magazine*, 26(6):173–180, November 2009.

[12] 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. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, Nov. 2011.

[13] T. Penzel, B. Kemp, G. Klosch, A. Schlogl, J. Hasan, A. Varri, and I. Korhonen. Acquisition of biomedical signals databases. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):25–32, May 2001.

[14] M.-K. Shan, F.-F. Kuo, M.-F. Chiang, and S.-Y. Lee. Emotion-based music recommendation by affinity discovery from film music. *Expert Systems with Applications*, 2009.

[15] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, Jan 2012.

[16] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[17] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan 2009.