Towards Cheaper Tourists' Emotion and Satisfaction Estimation with PCA and Subgroup Analysis

Lucas Maris*, Yuki Matsuda*‡, Ramin Sadre†, Keiichi Yasumoto*‡
* Nara Institute of Science and Technology, Nara, Japan
† Université catholique de Louvain, Louvain-la-Neuve, Belgium
‡ RIKEN Center for Advanced Intelligence Project AIP, Tokyo, Japan
Email: {lucas.maris.lo3, yukimat, yasumoto}@is.naist.jp, ramin.sadre@uclouvain.be

***Abstract*—Smart tourism leverages ubiquitous sensors to recognise the state of tourists and provide them with a better-tailored sightseeing experience. We previously reported on our EmoTour system [1], which uses behavioural cues and audiovisual data collected during sightseeing to estimate tourists' emotional status and satisfaction levels. Some of this data is however not exceedingly convenient to collect, as eye-gaze trackers for instance are not widely available nor usually worn by regular tourists. In this paper, we explore different possibilities to both improve our previous results and lessen the cost of data collection, to work towards a system that is better suited for real-world applications. Using Principal Component Analysis dimensionality reduction, we show how leaving out either or both of eye-gaze tracker and physiological wristband sensor data can have little to no impact on the quality of predictions, and improve on our previously reported classification and regression scores. We also apply this new method to explore differences in emotional responses according to participants' nationality, age, and gender.**

***Index Terms*—smart tourism, emotion recognition, satisfaction estimation, principal component analysis, subgroup analysis**

I. Introduction

To achieve smart tourism, highly context-aware systems are necessary, as the emotional feedback of tourists provides useful insights to offer them an experience that is tailored to their taste. Currently, tourists' emotions and satisfaction levels are most commonly collected through user reviews and surveys. These usually aim to use previous tourists' experiences to improve the experiences of future tourists. This not only entails challenges regarding user incentivisation, but also is not suitable for any sort of dynamic recommendation system [2].

Our previous study [1] aimed to address this, by multimodally sensing tourists during sightseeing tours using various wearable devices to measure their eye, head and body movement, as well as their facial and vocal expressions and physiological data, and asking them to report their emotional status and satisfaction level at different times of the tour. This data was then used to train a neural network in a supervised manner to estimate subjects' emotional status and satisfaction level.

While multimodal approaches appear intuitively more desirable, often yielding better results, they don't come without drawbacks. First, they increase the number of sensors needed for data collection, in turn increasing the overall cost of the system. Second, multimodal systems also worsen the burden on the user, who is expected to carry and manage more devices. As the purpose of our application is specifically to lessen the burden on users by passively collecting data, it appears important to assess the relevance of used modalities.

In order to do so, this study aims to see how leaving out the most expensive sensors used in the original system affects prediction performances, i.e., how relevant eye movement and physiological data are within our system. We expect this to give us insights on which of its components has the largest impact on prediction results, ultimately allowing us to decide whether the cost of a particular sensor is justified by the improvements in results it leads to.

We apply Principal Component Analysis (PCA) to our data, both to mitigate the possible drop in performance from leaving out modalities and to get an overview of the features that offer the highest variability. We use it as a dimensionality reduction method and compare the quality of predictions when varying the data's dimensionality. We also conduct subgroup analysis in order to see whether differences between participants such as gender, nationality, age, and experiment location have any meaningful impact on our system's key modalities.

The subject of this paper is thus two-fold: to assess if and how much prediction metrics worsen when leaving out data that is collected through more expensive sen-

sors, while considering PCA dimensionality reduction to improve those results, and to compare how separating the dataset in different subsets affects results. By considering these questions, we want to improve the usability of our system for real-world applications.

Our results[1] indicate that (1) PCA dimensionality reduction does indeed improve our system's prediction accuracy, (2) both eye movement and physiological data can be dropped with little to no impact on prediction accuracy at suitable PCA levels, (3) satisfaction estimation benefits most from eye-tracking data, while emotion recognition profits most from physiological data, (4) the limited size of the EmoTour database does not really allow any conclusions on the relationship between specific participant subsets and relevant modalities.

II. Related Work

The following sections describe related work surrounding emotional status estimation in systems similar to ours, through context-aware unimodal or multimodal sensing. We also give an overview of the methods we used in this study.

A. Emotion recognition and context-aware sensing

Traditionally, collecting the emotional status and satisfaction level of individuals has been done through questionnaires or online user reviews [3], [4]. While de facto standards, these methods are quite burdensome for users, requiring them to plough through long lists of questions or to write out a few paragraphs about their experience. Moreover, to avoid a biased distribution of reviews, incentives must be devised for all sorts of participants to respond, to avoid any selection bias [5].

To address these issues, context-aware sensing has been investigated in many different studies, in order to get real-time information about users' emotional status and satisfaction level. Previous unimodal systems have used audio sensors [6], [7], cameras [8] or accelerometer data [9] for this purpose. Multimodal systems such as [10], [11], combining different feature types, have also been proposed, and these usually achieve higher accuracies.

B. Principal Component Analysis

Principal Component Analysis [12], [13] is one of the oldest and most widely used data analysis techniques, capable of drastically reducing the dimensionality of large datasets while preserving as much information as possible contained in the data it is applied to. Its application to a given dataset creates new, uncorrelated variables, called Principal Components (PCs), which are simple linear combinations of the existing variables in such a way that they successively maximise the data's variability. PCA can help increase the interpretability of large datasets and decrease model complexity, which can reduce overfitting, improve the quality of predictions and reduce their computational cost.

[1]These findings were submitted as a part of the author's master's thesis, available here: http://hdl.handle.net/2078.1/thesis:35586

TABLE I
Overview of the EmoTour dataset

Place	#Participants (#Sessions)	Experiment dates
Ulm, Germany	18 (149)	Dec. 2017 - Aug. 2018
Nara, Japan	5 (40)	Jan. 2018
Kyoto, Japan	24 (263)	Mar. 2019
Total	47 (452)	Dec. 2017 - Mar. 2019

C. Subgroup Analysis

Subgroup analysis [14] is a widely used technique in the medical field aiming to explore how a shared characteristic between some participants in a study can affect their reaction to a given treatment. Beyond the medical field, it allows to explore how considering different subgroups within the participant pool might lead to different results when applying given data analysis methods to them, independently of the other subgroups.

Our previous study [15] explored differences between participants depending on whether they were of Japanese or Russian nationality, as studies have proven that different culture groups will differ in their ways of expressing emotion [16], [17]. Differences in gender identity [18] and age [19] also affect the way people express themselves, hence we here want to extend our subgroup analysis to include not only nationality, but also gender and age of participants.

III. Method

This section goes over the system developed in our previous study [1], onto which this paper builds, and the modifications this paper brings to our previous model.

A. Overview of the EmoTour system

The data for the EmoTour dataset was collected by voluntary participants equipped with various wearable devices along predefined touristic routes. Each of these routes was divided into sessions, including at least one sight each, during which the data from the participants' wearable devices was continuously recorded, and after each of which participants were asked to record a selfie video to briefly comment on their enjoyment of this session. Table I gives an overview of the previously collected dataset.

1) Devices and Modalities: In order to build a multimodal dataset, different wearable devices were used in order to cover a wide array of modalities. Participants were equipped with:

- a Pupil Core eye tracker [20], which was used to record their eye gaze data, more specifically the intensity of their eye movements and various statistics on the eyes' positions (*Pupil* features).
- a SenStick sensor board [21] mounted on one side of the eye-tracking device, which features both an accelerometer and a gyroscope, amongst other sensors, and from which data about head and body movements were extracted (*SenStick* features).
- a smartphone, which participants used to record brief selfie videos after each session to express their current mood. Two types of features were extracted from these videos. The ones we refer to as *avg(Selfie)* features are vocal expressions (low-level audio descriptors) and facial expression (Action Units) averaged over recordings and extracted using the openSMILE [22] and the OpenFace [23] softwares, respectively. New, high-level features were derived from the avg(Selfie) features by feeding them into pre-trained models tasked with predicting arousal and valence scores for each recording, and we refer to these derived features as *high(Selfie)* features.
- an Empatica E4 wristband [24], used to measure participants' physiological signals, most notably their heart rate, electrodermal activity, and body skin temperature (*Empatica* features).

Further information about the feature extraction process can be found in previous works [1], [25]. In total, 284 features are used: 80 Pupil features, 28 SenStick features, 80 avg(Selfie) features, 72 high(Selfie) features and 24 Empatica features.

2) Labels and ground truths: The collected data is labelled manually by participants, which were asked to input their emotional status and satisfaction level after each session through a simple smartphone application.

Emotional states are defined using Russell's circumplex model of affect [26], which arranges emotional states along its two Valence and Arousal axes in a two-dimensional space. Participants were asked to pick one of the nine emotions defined through this map, which were streamlined into three more general emotion levels: Positive, Neutral and Negative.

Satisfaction levels are defined using a 7-point Likert scale, ranging from 0 (fully unsatisfied) to 6 (fully satisfied), with 3 corresponding to a neutral satisfaction level.

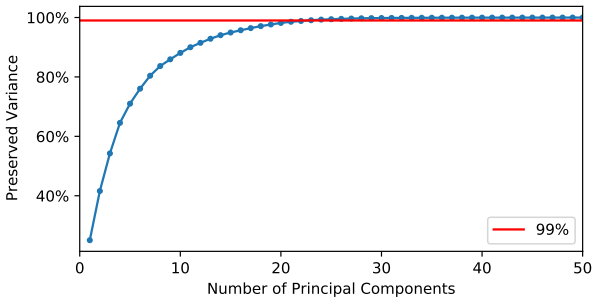

Fig. 1. Data variance captured by the x first principal components.

B. Modelling

1) Feature sets: In this paper, we use our previously collected dataset to investigate the effect of leaving out features obtained through costly sensors, by comparing the prediction results of a simple neural network model on different feature sets. As both the Pupil Core eye tracker and the Empatica E4 wristband are expensive research devices, we consider different feature sets leaving out their data. In total, 4 feature sets are considered: one using all available features, one using all but the Pupil features, one using all but the Empatica features, and one using only the SenStick & Selfie features, i.e., the cheapest sensors. Each of these sets is declined into 2 variants, depending on whether they use the avg(Selfie) features or the high(Selfie) features.

2) Prediction model: As in previous work [1], we build two machine learning models, one to estimate tourists' emotions from data through a 3-class classification task, the other one to predict tourists' satisfaction levels as a regression task along a 7-point Likert scale. These models consist of a single layer neural network, with an input layer of dimensionality equal to the number of features that are being used, a hidden layer of 30 neurons, and an output layer of 3 neurons for the classification task or 1 neuron for the regression task. We apply feature-level fusion to the features obtained through the different considered modalities.

The performance of these models is evaluated using Unweighted Average Recall (UAR) for emotion estimation and using Mean Absolute Error (MAE) for satisfaction estimation. The aim is to maximise UAR and minimise MAE. Models are trained using 10-fold cross-validation, while a Leave-One-Out methodology is used for testing models, by building a model that is trained on all data but the data relative to a specific participant, and tested on this left-out data, for every participant. Reported results correspond to the average evaluation metric value of all models evaluated on their respective test sets.

TABLE II
BEST RESULTS FOR EMOTIONAL STATUS PREDICTIONS ON EACH FEATURE SET, WITH RESPECTIVE AMOUNT OF KEPT PRINCIPAL COMPONENTS.

	avg(Selfie)		high(Selfie)	
	#PC	*UAR*	*#PC*	*UAR*
SenStick + Selfie + Pupil + Empatica	40	0.589	20	0.576
SenStick + Selfie + Empatica	30	**0.602**	15	0.583
SenStick + Selfie + Pupil	60	0.580	30	0.572
SenStick + Selfie	60	0.600	10	**0.585**
Previous works*	-	0.451	-	0.428

* Scores reported in Fedotov2020 and EmoTour2018, respectively.

3) Dimensionality reduction: To both improve the accuracy of our predictions and streamline our high-dimensional dataset, we standardise and apply PCA to every feature set before feeding it through our models. To decide an appropriate number of principal components to be considered in the following section, and thus of how much dimensionality reduction to apply, we look at the amount of variance principal components capture, as illustrated in Fig. 1.

From this, it appears that very few principal components suffice to capture most of the variance in our data: 16, 23 and 34 PCs suffice to capture 95%, 99% and 99.9% of the variance in the data, respectively. In the following, we hence consider results when keeping a number of principal components $\in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70\}$.

IV. RESULTS

In this section, we first report how results vary when applying PCA to various subsets of the input data, reducing their dimensionality to only include the x best principal components. We then use our best-performing transformations to the input data to investigate differences observed between subgroups in the dataset.

A. Performances after PCA dimensionality reduction

1) Performance baseline: As a performance baseline, results from our previously published works are reported with our results, and referred to as EmoTour2018 [1] and Fedotov2020 [25]. These results were not obtained on the exact same data/feature sets but were chosen for being the available reported results with datasets that most closely match those considered here.

2) Emotional status predictions: Table II reports the best UAR scores achieved for each of the feature sets when training models that use 5 to 70 PCs. It appears quite clearly that the use of PCA improves previously achieved emotion prediction performances for all of the considered feature sets. As reported in [25], avg(Selfie) features tend to perform better than their high-level counterparts, making it safe to focus on these more easily obtainable features. It is however interesting to note that feature sets that include high(Selfie) instead of avg(Selfie) appear to benefit most from a drastic dimensionality reduction.

As for reducing the cost of the sensors used in our system, results seem to suggest that leaving out the Pupil features actually has a positive impact on predictions. Indeed, the highest prediction accuracy is obtained when using SenStick, avg(Selfie) and Empatica features and keeping 30 PCs, with an UAR score of 0.602. The cheapest feature set, consisting just of SenStick and avg(Selfie) features, allows for predictions nearly as good when using 60 PCs specifically, with an UAR score of 0.600.

Capital modalities for emotional status predictions thus comprise SenStick and avg(Selfie) features, with Empatica features enabling a slight performance bonus. Keeping around 60 PCs tends to lead to maximal performances overall.

3) Satisfaction level predictions: Fig. 2 reports all the MAE losses for models trained on each of the feature sets with a number of principal components varying from 5 to 70. From these graphs, it is less immediately apparent that the use of PCA improves previously achieved satisfaction prediction performances, as all obtained results are systematically worse than those published in [25], but better than those from [1]. Nonetheless, a clear upwards trend is visible in both graphs, suggesting that reducing the number of kept PCs helps to improve performances. Again, avg(Selfie) features consistently perform better than their high-level counterparts, so the rest of our analysis refers to Fig. 2(a).

Best prediction accuracies are obtained when using all feature types (black line) and keeping 15 to 20 PCs, the latter achieving a MAE of 1.100. Keeping only the cheapest features, i.e., SenStick and avg(Selfie) features (red line), barely worsens results when keeping 15 PCs or less, with 15 PCs scoring a MAE of 1.116. Unlike emotion estimation, it appears that Pupil features (green line) are consistently more important than Empatica features (orange line), with especially good performances at 10 kept PCs, for a MAE of 1.105. When keeping a number of PCs around the threshold of 50, using only SenStick and avg(Selfie) features (red line) consistently and significantly outperforms other feature combinations.

While satisfaction level predictions thus benefit from all modalities, using only the SenStick and Selfie features allows for nearly the same performances, with the addition of Pupil features being slightly helpful but not unmissable. A low number of PCs generally suffices for maximal performances.

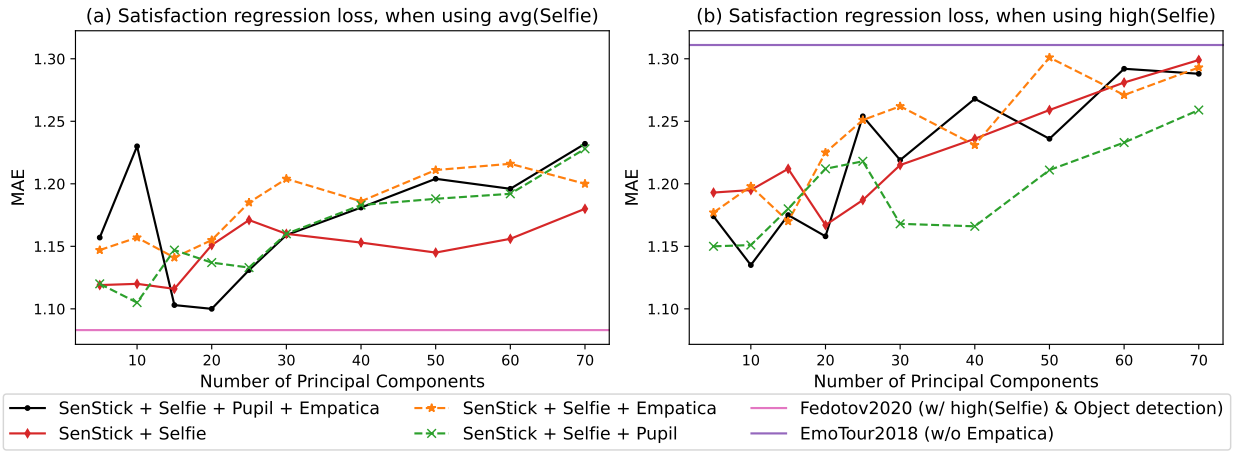

Fig. 2. Results for satisfaction level predictions, in function of the amount of principal components considered.

B. Dataset subset performances

Our last study [1] noted that emotion and satisfaction estimation accuracy can vary according to the cultural backgrounds of participants. In order to build onto that observation, we extended this subgroup analysis to all plausible subgroups that can be delimited within our dataset. In this section, we report the estimation scores when applying some of the best performing transformations described in Section IV-A, in order to get a feel for how different subgroups may express emotions through different means. Each subgroup was used to train 4 different models, half of them tasked with estimating emotional status, the other half tasked with estimating satisfaction level. For each task, one of the models uses only the cheapest available features (SenStick and avg(Selfie)), while the other uses a third modality (Empatica for emotions, Pupil for satisfaction). All models use some level of PCA dimensionality reduction. All UAR and MAE evaluation scores are reported in Table III.

1) Subsets by nationality: Data was split in three subgroups: Japanese (187 samples), Russian (106 samples), Other (159 samples). The last subgroup includes data samples from 10 different Asian nationalities. Estimation accuracies when evaluating each model on the subsets lie within the same range, with no subset clearly outperforming the others. When considering models stripped of just one modality (lines 1 & 3), it appears performances benefit from having nationality-specific models.

2) Subsets by gender: Data was split in two subgroups: Male (360 samples) and Female (92 samples). One could expect the smallest of these subgroups to lead to worse model performance, but it turns out all but one of the models perform better when trained exclusively on Female samples. This observation is tempered by the fact that differences are not very significant for satisfaction estimation.

3) Subsets by age: Data was split in three subgroups: 22-23 years old (139 samples), 24-25 years old (142 samples), 26+ years old (171 samples). These subgroups were chosen purely for their approximately even distribution of data samples. It appears the 26+ subgroup outperforms the others in 3 out of 4 models, although the difference is not very significant for satisfaction satisfaction.

4) Subsets by location: Data was split in three subgroups: Ulm (149 samples), Nara (40 samples) and Kyoto (263 samples). As expected, the Nara samples alone are too few to accurately train models, and systematically perform worse than the other subgroups. Kyoto samples always outperform the others, which can be explained by the fact that they are the only samples that (hardly) miss any sensor data [25] and were collected on three consecutive days, as opposed to the other subgroups.

TABLE III
RESULTS OF SUBGROUP ANALYSIS

			Nationality			Gender		Age			Location		
		All	*Japanese*	*Russian*	*Other*	*Male*	*Female*	*22-23*	*24-25*	*26+*	*Ulm*	*Nara*	*Kyoto*
Emotions	*30 PCs w/ Empatica*	0.602	0.619	**0.642**	0.606	0.582	**0.621**	**0.621**	0.605	0.597	**0.612**	0.607	**0.612**
(UAR)	*60 PCs*	0.600	**0.579**	0.541	0.556	**0.557**	0.513	0.555	0.567	**0.596**	0.549	0.545	**0.585**
Satisfaction	*10 PCs w/ Pupil*	1.105	**1.096**	1.106	1.097	1.119	**1.113**	1.117	1.123	**1.116**	1.109	1.118	**1.100**
(MAE)	*15 PCs*	1.116	1.129	1.128	**1.121**	1.124	**1.119**	1.124	1.123	**1.116**	1.129	1.130	**1.102**

All 4 considered models use SenStick and avg(Selfie) features on top of the features mentioned in the table.

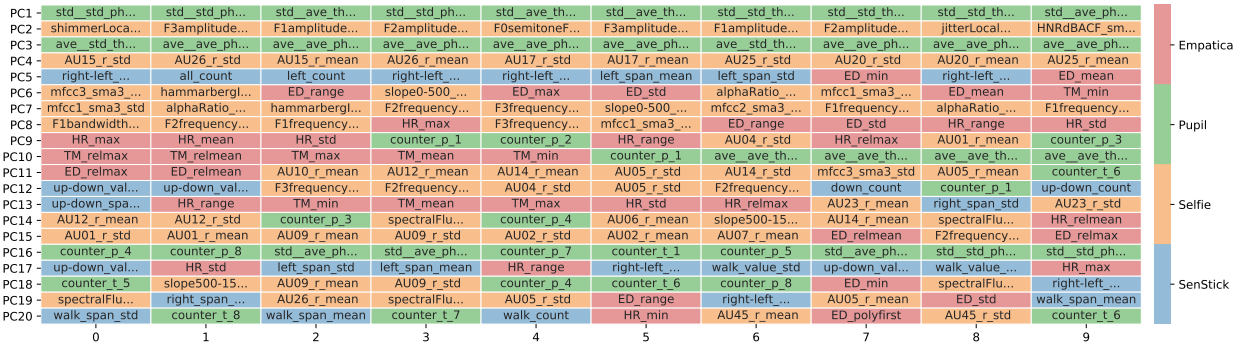

Fig. 3. The 10 most contributing features to the 20 most important PCs, ranked by contribution importance.

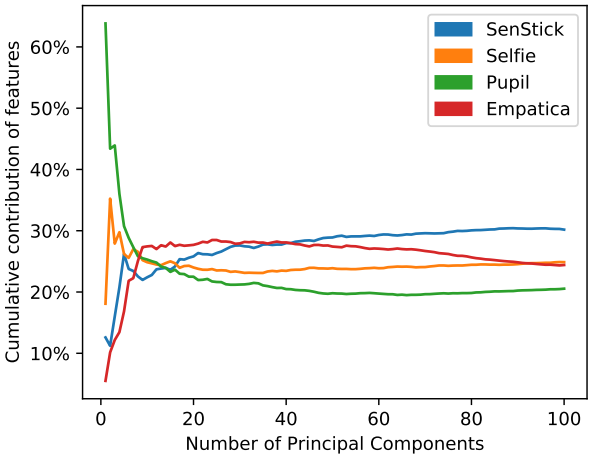

Fig. 4. Cumulative contribution of each feature type when the number of kept PCs increases.

V. Discussion

A. Composition of Principal Components

To get a better understanding of the features that are considered most prominent by PCA, i.e., of the features with highest variability, we illustrate the composition of the 20 first principal components in Fig. 3. The table shows the 10 most contributing features to each of these PCs, ranked by contribution importance. It appears both Pupil and Empatica features are prominently used by the PCA transformation, with PC1, PC3 and PC16 being majoritarily defined in terms of Pupil features, and Empatica features contributing heavily to many PCs as well. Taken alone, this view would suggest these features provide valuable, highly varied data, whose inclusion is likely to positively affect prediction results on any model using this data.

However, as previously noted in Section IV-A, their omission does not necessarily worsen prediction accuracies, and can sometimes even improve them. This can be attributed in part to the lack of consideration of PCA for the sample's labels, as well as to the biased view offered by Fig. 3. Indeed, as there are far more Pupil features than there are, say, SenStick features (80 versus 28, respectively), it is quite natural for them to appear more in this table.

Fig. 4 aims to balance this observation, by plotting the cumulative importance of feature types, i.e., the cumulative sum of the magnitude of their contributions to PCs, normalised by the number of features provided by that feature type, for any number of kept PCs. On this graph, it can be seen Pupil features quickly lose their prominent influence on PCs as their number grows, while Empatica features lose importance at a much steadier rate. Selfie features see their influence vary, while SenStick features consistently augment their influence on PCs.

B. Limitations of present subgroup analysis

The results reported in Section IV-B, while somewhat insightful, were obtained on a previously collected dataset, which was never specifically designed to include varied samples for every subgroup considered here, nor for other possible subgroups that could be deemed sensible (e.g., occupation, education) but are not considered here for lack of labelled data.

This is reflected by an imbalance in subgroups, and is likely to not only hamper prediction results but also the insights that can be taken away from such an analysis. Amalgamating many nationalities into a single subgroup or considering such small differences in age between participants is unlikely to have a meaningful impact on prediction results.

Further experiments would need to take place in order to deepen this aspect of our analysis. To expand on how nationality and culture affects participants' expressions, the inclusion of a wide group of, e.g., western participants could prove interesting. Collecting more female samples and samples that do not fit this simplistic binary opposition, or conducting new experiments on a set of participants with larger age differences would likely improve the quality of such an analysis.

VI. Conclusion

With the aim of improving the EmoTour system, which multimodally senses tourists in order to estimate their emotional status and satisfaction level for context-aware smart tourism applications, we analysed the effect of partitioning our previously collected dataset on estimation performances. We have shown that PCA dimensionality reduction can vastly improve the results of our machine learning models, and that the use of expensive sensors such as the Pupil Core eye tracker and the Empatica E4 wristband can be avoided for certain tasks: emotion classification does not necessarily benefit from eye-tracking data, while satisfaction estimation can prove just as good without physiological data. Using subgroup analysis, we explored how nationality, age and gender identity can affect how tourists express themselves, and how that affects prediction results. Our results suggest differences in expression both between Japanese and Russian, and male and female participants, but we lack the data to thoroughly confirm these differences. Further work in this direction should drastically expand the participant pool to include much more varied profiles, and might benefit from extra labels pertaining to, e.g., occupation or education.

Acknowledgement

This study was supported in part by JST PRESTO under Grant No. JPMJPR2039 and JSPS KAKENHI Grant Number JP21H03431, JP22H03648.

References

[1] Y. Matsuda, D. Fedotov, Y. Takahashi, Y. Arakawa, K. Yasumoto, and W. Minker, "EmoTour: Estimating emotion and satisfaction of users based on behavioral cues and audiovisual data," *Sensors*, vol. 18, no. 11, 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/11/3978

[2] M. Hidaka, Y. Kanaya, S. Kawanaka, Y. Matsuda, Y. Nakamura, H. Suwa, M. Fujimoto, Y. Arakawa, and K. Yasumoto, "On-site trip planning support system based on dynamic information on tourism spots," *Smart Cities*, vol. 3, no. 2, pp. 212–231, 2020.

[3] J. Alegre and J. Garau, "Tourist satisfaction and dissatisfaction," *Annals of Tourism Research*, vol. 37, no. 1, pp. 52–73, 2010.

[4] C. F. Chen and F. S. Chen, "Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists," *Tourism Management*, vol. 31, no. 1, pp. 29–35, 2010.

[5] I. Marinescu, N. Klein, A. Chamberlain, and M. Smart, "Incentives can reduce bias in online reviews," *Economics of Networks eJournal*, 2018.

[6] H. Kaya, A. A. Karpov, and A. A. Salah, "Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines," in *Advances in Neural Networks - ISNN 2016*, 2016, pp. 115–123.

[7] W. Y. Quck, D.-Y. Huang, W. Lin, H. Li, and M. Dong, "Mobile acoustic emotion recognition," in *Region 10 Conference (TENCON), 2016 IEEE*. IEEE, 2016, pp. 170–174.

[8] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.

[9] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu, "Emotion recognition based on customized smart bracelet with built-in accelerometer," *PeerJ*, vol. 4, p. e2258, 2016.

[10] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[11] B. Resch, A. Summa, G. Sagl, P. Zeile, and J.-P. Exner, "Urban emotions – geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data," in *Progress in Location-Based Services 2014*, 11 2014, pp. 199–212.

[12] K. P. F.R.S., "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[13] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, vol. 24, pp. 498–520, 1933.

[14] D. I. Cook, V. J. Gebski, and A. C. Keech, "Subgroup analysis in clinical trials," *Medical Journal of Australia*, vol. 180, no. 6, p. 289–291, Mar 2004.

[15] Y. Matsuda, D. Fedotov, Y. Arakawa, H. Suwa, W. Minker, and K. Yasumoto, "Analysis of tourists' nationality effects on behavior-based emotion and satisfaction estimation," in *4th International Conference on Imaging, Vision & Pattern Recognition (IVPR '20)*, 2020, pp. 1–7.

[16] J. T. Stanley, X. Zhang, H. H. Fung, and D. M. Isaacowitz, "Cultural differences in gaze and emotion recognition: Americans contrast more than chinese," *Emotion*, vol. 13, no. 1, pp. 36–46, February 2013.

[17] J. Miehle, K. Yoshino, L. Pragst, S. Ultes, S. Nakamura, and W. Minker, "Cultural communication idiosyncrasies in human-computer interaction," 09 2016, pp. 74–79.

[18] J. Miehle, W. Minker, and S. Ultes, "What causes the differences in communication styles? a multicultural study on directness and elaborateness," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), May 2018.

[19] R. Khawar, F. Malik, S. Maqsood, T. Yasmin, and S. Habib, "Age and gender differences in emotion recognition ability and intellectual functioning," *Journal of Behavioral Sciences*, 01 2014.

[20] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct, 2014, pp. 1151–1160.

[21] Y. Nakamura, Y. Arakawa, T. Kanehira, M. Fujiwara, and K. Yasumoto, "Senstick: Comprehensive sensing platform with an ultra tiny all-in-one sensor board for iot research," *Journal of Sensors*, vol. 2017, 2017.

[22] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. ACM, 2010, pp. 1459–1462.

[23] T. Baltrušaitis, P. Robinson, and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[24] Empatica Inc., "Empatica E4," https://www.empatica.com/research/e4/.

[25] D. Fedotov, "Contextual time-continuous emotion recognition based on multimodal data," Ph.D. dissertation, Ulm University, 2020.

[26] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.