CNN-based Facial Affect Analysis on Mobile Devices

Charlie Hewitt
University of Cambridge
cth40@cam.ac.uk

Hatice Gunes
University of Cambridge
hatice.gunes@cl.cam.ac.uk

ABSTRACT

This paper focuses on the design, deployment and evaluation of Convolutional Neural Network (CNN) architectures for facial affect analysis on mobile devices. Unlike traditional CNN approaches, models deployed to mobile devices must minimise storage requirements while retaining high performance. We therefore propose three variants of established CNN architectures and comparatively evaluate them on a large, in-the-wild benchmark dataset of facial images. Our results show that the proposed architectures retain similar performance to the dataset baseline while minimising storage requirements: achieving 58% accuracy for eight-class emotion classification and average RMSE of 0.39 for valence/arousal prediction. To demonstrate the feasibility of deploying these models for real-world applications, we implement a music recommendation interface based on predicted user affect. Although the CNN models were not trained in the context of music recommendation, our case study shows that: (i) the trained models achieve similar prediction performance to the benchmark dataset, and (ii) users tend to positively rate the song recommendations provided by the interface. Average runtime of the deployed models on an iPhone 6S equates to ~45 fps, suggesting that the proposed architectures are also well suited for real-time deployment on video streams.

CCS CONCEPTS

• **Human-centered computing → Mobile devices**; **Interactive systems and tools**;

KEYWORDS

Affective computing, mobile computing, intelligent user interfaces, facial affect analysis, emotions, arousal, valence, music recommendation.

1 INTRODUCTION

1.1 Motivation and Contributions

Affective computing has historically remained confined to laboratory settings, typically only involving small studies and with little in the way of large-scale practical application. Recent advances in deep machine learning techniques and increasing availability of large, in-the-wild datasets have led to improved performance in affect recognition tasks such as prediction of emotion, valence and arousal from facial images in real world scenarios, not just in constrained environments. The ubiquitousness of mobile devices with advanced sensors, including high quality cameras, means that the application of affective computing technologies to end-user applications is now a real possibility [31].

This paper aims to explore the feasibility of obtaining state-of-the-art facial affect analysis from a captured facial image using machine learning approaches within the constrained environment of a mobile device, as well as how readily the output of these models can be used within a mobile application. To this aim, we developed the Emosic application which prompts the user to take an image of their face and predicts the displayed prominent facial affect in terms of an emotion category—neutral, happy, sad, surprised, afraid, disgusted, angry or contemptuous—as well as levels of valence (i.e., how positive/negative the displayed affect is) and arousal (i.e., how active/inactive the displayed affect is) using convolutional neural network (CNN) models. Based on the predicted user affect, the Emosic application presents a number of recommended songs to the user. Essentially our work has two primary contributions: (i) design and comparative evaluation of three CNN architectures for mobile affect analysis using the newly available AffectNet dataset [26]; and (ii) demonstration of deployability of the trained models for music recommendation.

The Emosic application is intended as a case study and a proof-of-concept that emotionally intelligent user interfaces (EIUI) on mobile devices are now feasible using modern machine learning approaches and large, in-the-wild datasets. Although the CNN models were not trained in the context of music recommendation, our case study shows that: (i) the trained models overall achieve similar prediction performance to the benchmark dataset, and (ii) users positively rate the song recommendations provided by the interface.

The rest of the paper is organised as follows. A summary of related work is presented in the remainder of Sec. 1. Preprocessing of facial image data and the design, training and evaluation of the proposed convolutional neural network (CNN) models for facial affect analysis and recognition are described in Sec. 2. A detailed description of the application implementation, along with illustrations, is provided in Sec. 3. Conclusions and discussion together with suggestions for future work are presented in Sec. 4.

The source code for the application and the machine learning setup is available on GitHub [14].

1.2 Related Work

Commercially available tools for real world affect analysis and recognition are fairly limited. Affectiva [23] is the most established company, offering a number of successful applications, for example in adaptive children's computer games, automatic tagging on the Imgur image hosting site and assessing viewer reception of television adverts. Microsoft are also trialling Emotion API [25] which offers similar functionality, though has so far seen little in terms of real world applications.

Small-scale deployments of automatic affect recognition have generally focussed on video games [22], medical applications [24, 35] and analysis of driver emotions [1]. There has so far been very little development of EIUIs. This may in part be due to user reluctance based on privacy concerns [32], as well as technological limitations.

To date, mobile affective computing has mostly remained limited to activity monitoring based on accelerometer data and calls, SMS and application usage [31, 42] with only two examples involving input from the device camera [31]. Collecting and analysing visual data is generally considered a high-computational task with the need for wide deployment of cameras. However, this trend is bound to change with the availability of new hardware, datasets and machine learning approaches.

The recently released AffectNet dataset [26] is a very large (450,000 images), in-the-wild annotated dataset for training automatic affect recognition models. The dataset provides annotations for 8 emotion categories, valence and arousal on a continuous scale from -1 to 1, facial bounding boxes and 66 facial landmarks. Previous in-the-wild datasets were generally smaller and did not include annotations of valence and arousal. For instance, FER-2013 [10] included 35,000 images with 7 emotion categories, FER-Wild [27] included 25,000 similarly annotated images and EmotioNet [6] contained 100,000 images with 23 emotion categories. The increased availability of these large annotated datasets of facial images enables further developments in the field of affective computing.

Existing machine learning approaches typically do not consider model size as an important attribute in architecture design. General image classification architectures such as InceptionV3 [38], ResNet50 [13] and VGG16 [34] result in data files of significant size; 90MB, 97MB and 528MB respectively. The CNNEmotions architecture [20] designed for the task of emotion classification is certainly too large for any realistic mobile application (475MB), as is VGGFace [30]. The only architecture specifically designed with mobile deployment in mind is MobileNet [15] which, at 16.4MB is certainly reasonable for mobile deployment[1].

Implementations of light-weight CNN architectures for affect analysis and recognition have focussed primarily on real-time classification, and therefore often produce smaller models as a side effect of this. [5], [33] and [9] specify models that achieve quite high classification accuracy (~60% on FER2013) for frames in video feeds in real time, with file sizes generally smaller than 30MB.

2 FACIAL AFFECT RECOGNITION

2.1 Considerations and Design

Given the goal of mobile deployment, the final model size must remain reasonable for inclusion in a mobile application. Google and Apple both impose limits on app size in the Play Store and AppStore respectively. For installation over cellular network, Apple limits apps to 150MB (100MB before Sep 2017) and for all apps Google imposes a limit of 100MB (50MB before Sep 2015). Most affective computing models included in a user application are likely to augment interaction (e.g., EIUI) rather than providing the primary functionality of the app. The storage space used by these models should therefore remain well under this 100MB limit.

Cloud offload might be seen as the obvious solution to the issue of constrained resources, but for the proposed work users' facial imagery is unavoidably involved, so privacy becomes an immediate concern. Due to both privacy concerns and concerns regarding latency, local execution is the preferred course of action.

In order to emulate a more complex application with multiple models we include two separate models—one for emotion classification and one for valence/arousal prediction—despite the clear possibility to exploit the similarity of these tasks and use a single network with multiple outputs for this application. Consequently, we impose a maximum model size of 15MB for the proposed application. The total contributed file-size should therefore remain less than 30MB for the two models, making the app approximately 50MB in size overall.

The time and computational resource taken to obtain predictions from images are also a factor to consider on mobile devices. However, this is expected to be of little issue for the models designed and implemented in this work, given the simplicity inherent in architectures of this size.

Inspired by previously established networks, three CNN architectures are designed and evaluated: (1) a design similar to AlexNet [19] using a series of convolution layers with incrementally smaller kernels interspersed with max-pooling layers, (2) an architecture based on VGG16 [34] with stacked 3×3 convolution layers interspersed with max-pooling layers, and (3) a network based on MobileNet [15] utilising depth-wise separable convolutions to maximise spatial efficiency.

All CNN models are implemented using Keras [7] and trained on an NVIDIA GeFore GTX 1080 Ti GPU using TensorFlow [11].

2.2 Preprocessing and Training

The AffectNet dataset [26] contains images of a highly heterogeneous nature. The dataset is divided by its creators into training and validation sets, and the test set labels are not yet available for research purposes. To be able to compare our results to that of the baseline as reported in [26], we follow the predefined dataset partitions.

In order to produce suitable images for input to a CNN the faces are cropped and resized to 128×128 pixels. The facial bounding box annotations provided by AffectNet are used for this purpose. Only manually annotated images are used[2]. For emotion classification all images annotated with invalid emotions (8: none, 9: uncertain and 10: no-face) are discarded leaving a total training set of 287,651 images and a validation set of 4000 images. For valence/arousal regression all images with invalid annotations, indicated using a value of -2, are discarded leaving a training set of 320,739 images and a validation set of 4500 images.

Weighted-loss is used for emotion classification to account for the imbalance in the training set as this achieved the best results in the baseline paper [26] (compared with up- and down-sampling). For valence/arousal regression data imbalance is again a problem resulting in over-fitting and potentially reduced performance. The mean annotations of the training set are 0.19 and 0.09 for valence and arousal respectively, while for the validation set are -0.16 and 0.30. Attempting to rectify this by down-sampling did little to improve performance so the full training set is used.

Randomised data augmentation is used for the training set with potential for images to be rotated by up to 20 degrees, translated by up to 10% (in both x- and y-directions) and flipped in the x-direction. All image data is normalised from $[0, 255]$ to $[0, 1]$ to increase the speed of training.

[1] All sizes relate to pre-trained CoreML [2] models available from CoreML Store [39].

[2] AffectNet also includes a large number of images automatically annotated by models trained on the manually annotated images.

The Adam optimiser [17] is used throughout with suggested parameters $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$, this is due to its design focus for machine learning tasks on large datasets. Batch size is maximised in order to best encapsulate the varied nature of the data and therefore improve training; 400 for architectures 1 and 2, and 250 for architecture 3, limited by available memory on the training hardware.

All classification models are trained over 24 epochs. As there is a strong correlation between valence/arousal and emotion, transfer learning can be exploited to produce the required valence/arousal models more easily. As such, the output layers of the trained emotion classifiers can be removed and replaced with appropriate output layers for the regression task (described below). The resulting models are then fine-tuned over 16 epochs. Both training times were chosen based on the details provided in the AffectNet baseline paper [26] and resulted in a plateau in validation loss towards the end of training.

2.3 Architecture 1: AlexNet Variant

This architecture is inspired by AlexNet [19], including a series of incrementally smaller convolution kernels starting at 9 × 9 and reducing to 3 × 3 with 2 × 2 max-pooling layers in between each convolution block and two fully connected (dense) layers prior to the output layer. There is a 0.2 Gaussian dropout after each pooling layer and a 0.5 dropout after each dense layer. Unlike the AlexNet architecture, each convolution block is constructed from a conventional 2D convolution layer followed by a batch normalisation layer [16] and a ReLU activation layer [28]. This helps to provide regularisation and faster training. The architecture is also shallower and narrower than the original AlexNet design in order to minimise model size. The full architecture specification is given in Table 1; the output layer contains 8 nodes with soft-max activation for emotion classification and 2 nodes with linear activation for valence/arousal regression.

2.4 Architecture 2: VGGNet Variant

This architecture is fairly similar to the AlexNet inspired design above, though it uses the principle behind VGG16 [34] of stacked 3 × 3 convolution kernels to capture larger image structure. The convolution blocks described for Arch. 1 above are again used, interspersed with max-pooling layers and followed by two fully connected layers before the output layer. As above, each pooling layer is followed by a 0.2 Gaussian dropout and there is a 0.5 dropout after each dense layer. The full architecture is given in Table 2, it is also narrower and shallower than typical VGGNet implementations in order to conserve space.

2.5 Architecture 3: MobileNet Variant

This architecture is inspired by MobileNet [15], which leverages 3 × 3 depth-wise separable convolution layers followed by 1 × 1 conventional convolution layers to retain high performance while minimising architectural complexity. This results in far smaller, tunable, network architectures perfect for deployment to mobile devices. Depth-wise separable convolution (DConv) blocks as described in [15] are used, with the full architecture given in Table 3. The reduced layer-wise complexity allows for a much deeper model which also retains good width. The output layer remains as above, but no pooling layers are present (stride in convolution layers is instead used for down-sampling) other than the final global average pooling layer which replaces the conventional fully connected layers. This pooling layer is followed by a dropout at rate 0.3.

Table 1: CNN architecture 1: AlexNet variant.

Type	Shape	Output
Conv	9 × 9 × 16	128 × 128 × 16
MaxPool	2 × 2	64 × 64 × 16
Conv	7 × 7 × 32	64 × 64 × 32
MaxPool	2 × 2	32 × 32 × 32
Conv	5 × 5 × 64	32 × 32 × 64
MaxPool	2 × 2	16 × 16 × 64
Conv	3 × 3 × 128	16 × 16 × 128
MaxPool	2 × 2	8 × 8 × 128
Conv	3 × 3 × 128	8 × 8 × 128
MaxPool	2 × 2	4 × 4 × 128
Flatten	2048	–
2×Dense	1024	–
Dense	8 or 2	1 label or 2 floats

Table 2: CNN architecture 2: VGGNet variant.

Type	Shape	Output
2×Conv	3 × 3 × 16	128 × 128 × 16
MaxPool	2 × 2	64 × 64 × 16
2×Conv	3 × 3 × 32	64 × 64 × 32
MaxPool	2 × 2	32 × 32 × 32
2×Conv	3 × 3 × 64	32 × 32 × 64
MaxPool	2 × 2	16 × 16 × 64
2×Conv	3 × 3 × 128	16 × 16 × 128
MaxPool	2 × 2	8 × 8 × 128
2×Conv	3 × 3 × 128	8 × 8 × 128
MaxPool	2 × 2	4 × 4 × 128
Flatten	2048	–
2×Dense	1024	–
Dense	8 or 2	1 label or 2 floats

Table 3: CNN architecture 3: MobileNet variant.

Type	Shape	Stride	Output
Conv	3 × 3 × 32	2	64 × 64 × 32
DConv	3 × 3 × 64	1	64 × 64 × 64
DConv	3 × 3 × 128	2	32 × 32 × 128
DConv	3 × 3 × 128	1	32 × 32 × 128
DConv	3 × 3 × 256	2	16 × 16 × 256
DConv	3 × 3 × 256	1	16 × 16 × 256
DConv	3 × 3 × 512	2	8 × 8 × 512
5×DConv	3 × 3 × 512	1	8 × 8 × 512
DConv	3 × 3 × 1024	2	4 × 4 × 1024
DConv	3 × 3 × 1024	1	4 × 4 × 1024
GlobalAvePool	1024	–	–
Dense	8 or 2	–	1 label or 2 floats

2.6 Evaluation Results

All architectures are evaluated on the AffectNet validation set (the test set is not publicly available) using the metrics provided for the baselines in [26]. Human annotator agreement for emotion classification on AffectNet is just over 60%.

For emotion classification accuracy (ACC), F1-score (F1), Cohen's kappa [8] (KAPPA), Krippendorff's alpha [18] (ALPHA), area under precision-recall curve (AUCPR) and area under ROC curve (AUC) are used. For valence/arousal prediction RMSE, Pearson's correlation coefficient (CORR), sign agreement metric [29] (SAGR) and concordance correlation coefficient [21] (CCC) are used.

Emotion classification results are presented in Table 4. The table shows that the VGGNet variant outperforms the AlexNet variant and the MobileNet variant in all metrics. It also outperforms the baseline in all but accuracy and F1 which are equalled at 58%.

Valence/arousal regression results are shown in Table 5. As with emotion classification, the VGGNet variant provides the best results of the three proposed architectures, for both valence and arousal prediction, though only marginally. All proposed architectures perform better for arousal than for valence, also outperforming the baseline. In contrast, the baseline performs significantly better for valence than arousal, also outperforming all proposed architectures.

2.7 Analyses and Discussion

The increased spatial efficiency of the MobileNet variant, and consequently its greater depth and width, do surprisingly little to improve the performance of the model over the AlexNet and VGGNet variants. There are many potential reasons for these results. One possible explanation is that facial affect might rely on edge related features which are typically captured by max pooling, but MobileNets only use average pooling. Another potential reason is the slight variation in model size (the VGGNet variant is slightly bigger), or the increased use of dropout in VGGNet.

All models have a file size close to the goal of 15MB, with the VGGNet variant being the largest at 15MB and the MobileNet variant the smallest at 13.2MB. All of these remain viable for mobile deployment as described in the considerations, and have performance close to the baseline for the AffectNet dataset.

Table 4: Emotion classification performance metrics for each architecture against weighted-loss baseline.

	Baseline	Arch. 1	Arch. 2	Arch. 3
ACC	**0.58**	0.56	**0.58**	0.56
F1	**0.58**	0.56	**0.58**	0.56
KAPPPA	0.51	0.50	**0.52**	0.50
ALPHA	0.51	0.50	**0.52**	0.50
AUCPR	0.56	0.61	**0.62**	0.60
AUC	0.82	**0.90**	**0.90**	0.89

Table 5: Valence (V) and arousal (A) regression performance metrics for each architecture against weighted-loss baseline.

	Baseline		Arch. 1		Arch. 2		Arch. 3	
	V	A	V	A	V	A	V	A
RMSE	**0.37**	0.41	0.41	0.39	0.41	**0.37**	0.42	0.38
CORR	**0.66**	0.54	0.59	0.53	0.62	**0.56**	0.59	0.53
SAGR	0.74	0.65	0.73	0.74	**0.75**	**0.75**	0.73	0.74
CCC	**0.60**	0.34	0.54	0.43	0.57	**0.48**	0.55	0.47

Table 6: Emotion classification confusion matrix of the VGGNet variant for the AffectNet validation set.

	N	H	Sa	Su	Af	D	An	C
N	247	7	52	60	11	22	34	67
H	20	358	6	26	4	15	4	67
Sa	63	9	279	22	38	41	37	11
Su	33	22	15	298	97	20	7	8
Af	21	6	32	72	320	32	12	5
D	29	9	36	24	31	316	42	13
An	71	4	39	22	29	98	216	21
C	71	56	12	21	3	33	26	278

Table 6 shows the confusion matrix of the VGGNet variant, the best performing architecture, providing a classification breakdown for **N**eutral, **H**appy, **Sa**d, **Su**prised, **Af**raid, **D**isgusted, **An**gry and **C**ontemptuous for the validation set containing 500 examples of each emotion. We observe that *happiness* has the highest rate of correct classifications (72%), while *anger* has the lowest with just 43% correct, often being confused with *disgust*. In the literature, anger and disgust are known to be confused because of the facial action units they share [40].

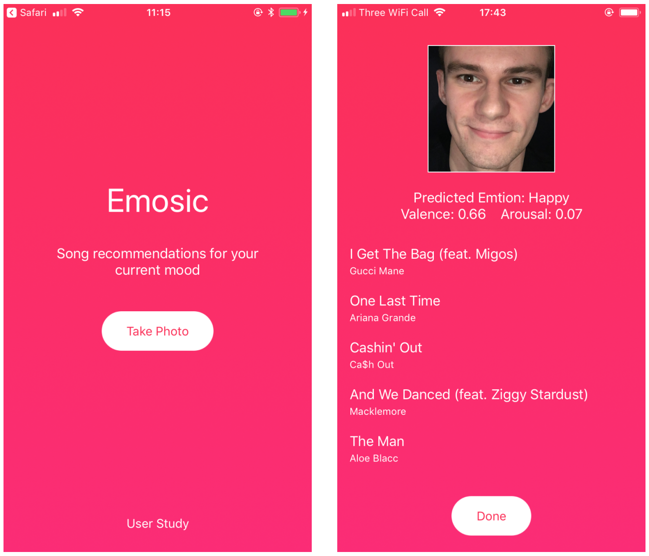
3 MOBILE MUSIC RECOMMENDATION

3.1 User Interface

The Emosic mobile application is implemented in Swift for the iOS platform and has a very simple interface as shown in Fig. 1. The user opts to take a photo, which prompts the native camera interface to be presented using the front-facing camera. Once the user has taken a photo of their face, the emotion, valence and arousal are predicted and the results are presented on the screen in Fig. 1b. In normal operation, the predicted emotion, valence and arousal are shown to the user along with the top five recommended songs. Clicking on each song opens it in the Spotify app [36] for the user to listen to.

3.2 Facial Affect Analysis

In order to be used for prediction within the iOS app, the highest performing Keras models described in Sec. 2 (i.e., Arch. 2) are converted for use with Apple's CoreML framework [2]. Apple provides an open source Python tool-kit, Coremltools [4], for this purpose. Almost no additional modification is required for the models to function within the iOS app, only minor preprocessing of the input image data. To match the input format described in Sec. 2.2, Apple's Vision framework [3] is used to determine the bounding box for the user's face, this is then cropped and resized to the required 128 × 128 pixels. The image data can then be fed directly to the applicable model after conversion to pixel buffer format.



Emosic

Song recommendations for your current mood

Predicted Emtion: Happy
Valence: 0.66 Arousal: 0.07

I Get The Bag (feat. Migos)
Gucci Mane

One Last Time
Ariana Grande

Cashin' Out
Ca$h Out

And We Danced (feat. Ziggy Stardust)
Macklemore

The Man
Aloe Blacc

(a) Main screen. **(b) Song recommendation.**

Figure 1: Emosic app user interface.

3.3 Song Recommendation

Song recommendations are obtained using the Spotify Web API recommendations service [37]. This service provides a REST endpoint which can be given up to five seed genres, a modality (major or minor) and numeric values for valence and energy (taken to be analogous to arousal) between 0 and 1. A list of songs which best match the inputs are then returned in JSON format. Seed genres are determined using a predefined mapping from the emotion predicted by the CoreML classifier (one of the basic eight) to a list of five seed genres. Predicted valence and arousal from the CoreML regression model are used directly after translation from the output range of $[-1, 1]$ to the required $[0, 1]$. The desired modality is taken from the sign of the valence prediction, positive being major and negative minor.

3.4 User Study

The user study is built directly into the app and can be accessed from the bottom of the screen as shown in Fig. 1a. The user is first given instructions regarding the study and information about data retention, they are then presented with the screens shown in Fig. 2 in order from left to right.

Firstly, the user is presented an emotion which they are to emulate (Fig 2a). Then they take a photo using the native camera interface. The recommended songs are presented with a 5-star rating input (Fig. 2b). Subsequently, a self-annotation screen for valence and arousal (Fig. 2c) is displayed. This sequence of screens is displayed for ten distinct emotions: neutral, delighted, happy, miserable, sad, surprised, angry, afraid, disgusted and contemptuous. These emotion categories are chosen (i) to correspond closely with the eight emotions of AffectNet that the models are trained with, and (ii) to provide some notable variance along valence and arousal dimensions.

Table 7: User study results.

(a) Emotion classification.

ACC	0.49
F1	0.45
KAPPPA	0.41
ALPHA	0.41
AUCPR	0.60
AUC	0.88

(b) Valence/Arousal prediction.

	Valence	Arousal
RMSE	0.40	0.40
CORR	0.74	0.44
SAGR	0.74	0.65
CCC	0.68	0.39

Table 8: User study emotion classification confusion matrix.

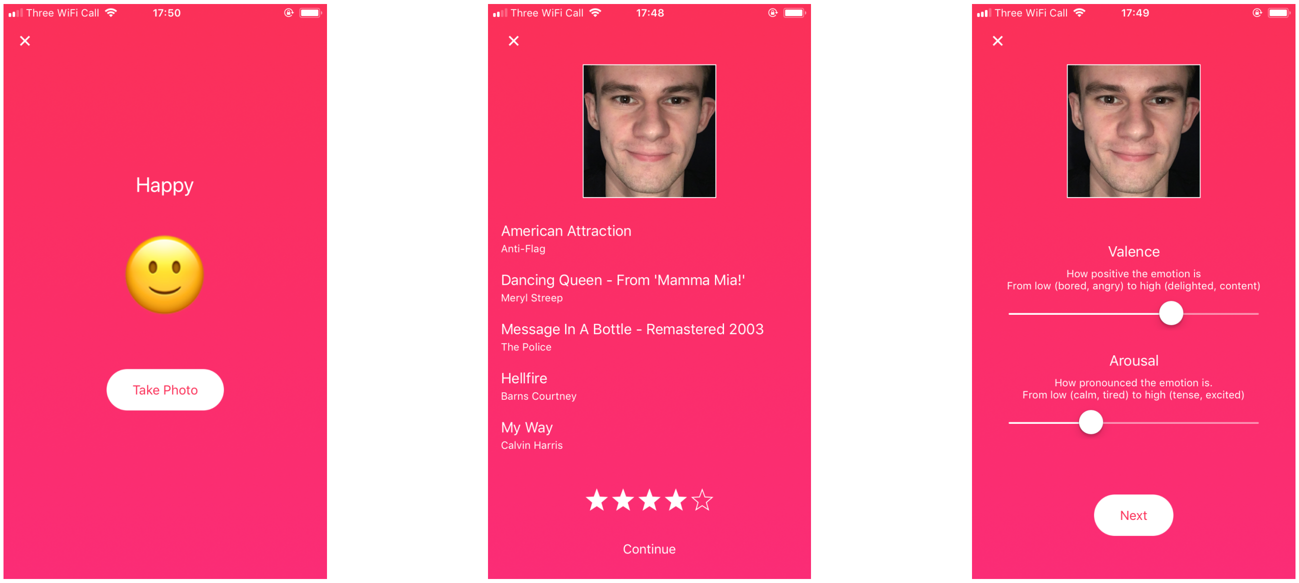
	N	H	Sa	Su	Af	D	An	C
N	3	1	0	3	0	0	1	2
H	0	9	0	0	1	0	0	0
Sa	2	0	7	0	0	0	0	1
Su	0	0	0	4	6	0	0	0
Af	1	0	1	1	7	0	0	0
D	0	1	1	0	1	7	0	0
An	1	0	2	0	2	3	1	1
C	2	1	2	1	1	2	0	1

The study was completed by 10 participants (4 female, 6 male) aged between 19 and 54 (mean 27, std 14). The results are broken down as in Sec. 2.6, with recognised emotion classes compared against instructed emotions, and valence/arousal predictions compared against self-annotated values. Many participants reported that the valence/arousal annotation scheme was not intuitive, this has likely led to discrepancies in what is taken to be ground-truth for the study, particularly for arousal.

The predictions made by the deployed models in the user study are evaluated using the same metrics as for the AffectNet dataset described in Sec. 2.6. The recognition and classification results are shown in Tables 7a and 7b. The confusion matrix of emotion classification for the user study is given in Table 8. When measuring classification performance, in order to obtain a balanced set which is directly comparable to the results in Sec. 2.6, the results for *happy* and *sad* are represented using the emotions *delighted* and *miserable*.

3.5 Analyses and Discussion

Emotion classification results are slightly worse than for the AffectNet validation set and vary greatly between emotions. *Happy* has an accuracy of 90% while *contempt* is at just 10%. *Surprise* is often misclassified as *fear*, and *anger* as *disgust*, similar to the results for AffectNet in Sec. 2.6. Participants reported that some emotions were difficult to emulate in this context (e.g. *contempt*) which may have caused variation in what is deemed to be ground-truth. Valence/arousal prediction is more successful, broadly matching the AffectNet results. Valence has a notably improved CORR and CCC scores of 0.74 and 0.68, while arousal has generally worse performance, more closely matching the AffectNet baseline.



American Attraction
Anti-Flag

Dancing Queen - From 'Mamma Mia!'
Meryl Streep

Message In A Bottle - Remastered 2003
The Police

Hellfire
Barns Courtney

My Way
Calvin Harris

Valence
How positive the emotion is
From low (bored, angry) to high (delighted, content)

Arousal
How pronounced the emotion is.
From low (calm, tired) to high (tense, excited)

(a) Instructing the user which emotion to emulate.

(b) Song recommendation with rating input.

(c) Self annotation of displayed facial affect with valence and arousal sliders.

Figure 2: User study interface.

Average runtime of the deployed CoreML models on an iPhone 6S was 22.4ms across ten runs. This equates to approximately 45 fps, suggesting that the models are well suited for real-time deployment on video streams.

The average rating for song recommendations was 3.17, indicating that users had a generally positive view of the application's functionality. Emotion specific ratings are shown in Table 9. Participants reported that emotions with clearer connotations (e.g. *happy*, *sad*) were easier to interpret musically and therefore easier to rate, while emotions with less clear connotations (e.g. *contempt*, *disgust*) were reported as being difficult to interpret. The lower ratings for the latter may be due to users finding them harder to relate to any music, or these emotions being classified correctly less often and therefore ending up with inappropriate music recommendation. The fact that the user rating was slightly higher (3.31) when emotion was classified correctly than when the predicted emotion was incorrect (3.02) supports the second of these suggestions.

A possible solution to the above mentioned issue is using a music-specific emotion model for labelling and assessment. One such model is derived from the Geneva Emotion Music Scale (GEMS) and has been developed for musically induced emotions [41]. It consists of nine emotional scales: wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness. Zentner et al. [41] compared the discrete approach, the dimensional approach and the GEMS approach and reported that participants preferred to report their emotions using the GEMS approach. Therefore, future studies focusing on affect-based music recommendation should take this into consideration. However, this would require a large dataset acquired in this specific context (which is not yet available).

Table 9: User study song recommendation ratings.

Neutral	3.6
Delighted	3.5
Happy	3.9
Miserable	3.2
Sad	3.7
Surprised	3.2
Angry	2.7
Afraid	2.4
Disgusted	2.6
Contemptuous	2.9
Average	3.2

During the user study, participants were also asked about their privacy concerns with the application. Most responded that they would not like the photos to be saved, and would prefer that data remained local to the device. Some also mentioned that they would like the uses for the data to be clearly stated and agreed to, though overall level of concern seemed to be significantly less than might be expected based on previous literature [32].

4 CONCLUSIONS AND FUTURE WORK

In this paper three CNN architectures for facial affect analysis have been designed and evaluated with the aim of minimising storage requirements for mobile deployment. These models gave comparable results to the current baseline when evaluated on the AffectNet dataset [26].

The best-performing models (i.e., VGGNet variants for emotion classification and for valence/arousal prediction) were deployed in a music recommendation application with an average execution time of 22.4ms (~45fps), also suitable for real-time applications. A user study was conducted to assess their real-world performance; the results showed that the deployed models provide results that are similar to the evaluation results obtained on the AffectNet dataset. Additionally, the users reported to be generally happy with the application's functionality. These results support the proposition that EIUIs are an area of great potential within affective computing and are now becoming increasingly feasible in a real-world setting.

The functionality of the Emosic application could easily be integrated into a fully-featured music application such as Spotify [36] and expanded to great effect. For example, determining a user's affect each time they manually choose a song would allow a model to be built up over time to provide tailored predictions for that specific user.

However, it is important to note that a user is unlikely to be as expressive as they are prompted to be in this study, which may reduce the application's effectiveness. The recent rise of wearables might provide a solution to this [12], as a number of modalities useful for affective computing, such as heartbeat, are now more readily available within mobile applications. These could easily be incorporated into multi-modal models along with accelerometer or usage activity data to improve the accuracy and reduce invasiveness of emotion recognition in a mobile setting.

It is not difficult to see how emotionally intelligent behaviour analysis could be expanded to many other application domains, though privacy issues need to be given care and consideration. Emphasis will need to be placed on clearly explaining what such applications will be doing, and keeping computation local with as little long-term data retention as possible.

To facilitate wide-scale adoption of EIUI, continued research into very efficient (both in terms of file-size and computation) deep neural network architectures will be required. Google's MobileNets [15] are a very promising start in this respect. Alternative structures such as InceptionV3 [38] may provide a more efficient basis than the options presented in this paper and lower floating point precision could be an easy way to cut model size, though the impact on performance might be significant. It is likely that major developments will need to be driven by popular smart-phone manufacturers at an OS level, as is already beginning to happen with Apple's CoreML [2] and Vision frameworks [3].

REFERENCES

[1] I. Abdić et al. 2016. Driver Frustration Detection from Audio and Video in the Wild. In *Proc. of IJCAI*. 1354–1360.

[2] Apple Inc. 2017. CoreML Framework. https://developer.apple.com/documentation/coreml

[3] Apple Inc. 2017. Vision Framework. https://developer.apple.com/documentation/vision

[4] Apple Inc. and contributors. 2017. CoreML Community Tools. https://github.com/apple/coremltools

[5] O. Arriaga, M. Valdenegro-Toro, and P. Plöger. 2017. Real-time Convolutional Neural Networks for Emotion and Gender Classification. *CoRR* abs/1710.07557 (2017).

[6] F. Benitez-Quiroz, R. Srinivasan, and A. Martinez. 2016. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. In *Proc. of CVPR*.

[7] F. Chollet et al. 2015. Keras. https://github.com/fchollet/keras

[8] J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[9] D. Duncan, G. Shine, and C. English. 2017. Facial Emotion Recognition in Real Time. (2017). http://cs231n.stanford.edu/reports/2016/pdfs/022_Report.pdf

[10] I. Goodfellow et al. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* 64, Supplement C (2015), 59 – 63.

[11] Google Inc. and contributors. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

[12] H. Gunes and H. Hung. 2016. Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block. *Image Vision Comput.* 55, P1 (2016), 6–8.

[13] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015).

[14] C. Hewitt. 2018. Emosic. https://github.com/friggog/Emosic

[15] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* abs/1704.04861 (2017).

[16] S. Ioffe and C. Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015).

[17] D. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[18] K. Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement* 30, 1 (1970), 61–70.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[20] G. Levi and T. Hassner. 2015. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. In *Proc. of ICMI*. 503–510.

[21] L. I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268.

[22] G. Mark, D. Alan, and A. Jen. 2005. Affective Videogames and Modes of Affective Gaming: Assist Me, Challenge Me, Emote Me. In *Proc. of the 2005 DiGRA Int'l Conf.*

[23] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard. 2013. Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected "In-the-Wild". In *Proc. of CVPR*. 881–888.

[24] D. Messinger, L. Duvivier, Z. Warren, M. Mahoor, J. Baker, A. Warlaumont, and P. Ruvolo. 2014. Affective computing, emotional development, and autism. In *The Oxford Handbook of Affective Computing*.

[25] Microsoft Corporation. 2017. Microsoft Emotion API. https://azure.microsoft.com/en-gb/services/cognitive-services/emotion/

[26] A. Mollahosseini, B. Hasani, and M. Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* PP, 99 (2017), 1–1.

[27] A. Mollahosseini, B. Hassani, M. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. 2016. Facial Expression Recognition from World Wild Web. *CoRR* abs/1605.03639 (2016).

[28] V. Nair and G. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proc. of the ICML*. 807–814.

[29] M. Nicolaou, H. Gunes, and M. 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.

[30] O. Parkhi, A. Vedaldi, A. Zisserman, et al. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.

[31] E. Politou, E. Alepis, and C. Patsakis. 2017. A survey on mobile affective computing. *Computer Science Review* 25, Supplement C (August 2017), 79 – 100.

[32] C. Reynolds and R. Picard. 2005. Evaluation of affective computing systems from a dimensional metaethical position. In *Proc. of Augmented Cognition Conf.* 22–27.

[33] J. Schwan, E. Ghaleb, E. Hortal, and S. Asteriadis. 2017. High-performance and lightweight real-time deep face emotion recognition. *SMAP* (07 2017), 76–79.

[34] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[35] A. Singh, N. Bianchi-Berthouze, and A. Williams. 2017. Supporting Everyday Function in Chronic Pain Using Wearable Technology. In *Proc. of CHI*. 3903–3915.

[36] Spotify AB. 2008. Spotify iOS App. https://itunes.com/app/spotify

[37] Spotify AB. 2017. Spotify Web API. https://developer.spotify.com/web-api/

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015).

[39] Thwis Inc. 2017. CoreML Store. https://coreml.store

[40] M. Wiggers. 1982. Judgments of facial expressions of emotion predicted from facial behavior. *Journal of Nonverbal Behavior* 7, 2 (1982), 101–116.

[41] M. Zentner, D. Grandjean, and K. R. Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494–521.

[42] S. Zhang and P. Hui. 2014. A survey on mobile affective computing. *CoRR* abs/1410.1648 (2014).