Smile Intensity Detection in Multiparty Interaction using Deep Learning

Philine Witzig
Disney Research Los Angeles
Glendale, California, USA
philine.witzig@gmail.com

James Kennedy
Disney Research Los Angeles
Glendale, California, USA
james.r.kennedy@disney.com

Cristina Segalin
Disney Research Los Angeles
Glendale, California, USA
cristina.segalin@disney.com

***Abstract*—Emotion expression recognition is an important aspect for enabling decision making in autonomous agents and systems designed to interact with humans. In this paper, we present our experience in developing a software component for smile intensity detection for multiparty interaction. First, the deep learning architecture and training process is described in detail. This is followed by analysis of the results obtained from testing the trained network. Finally, we outline the steps taken to implement and visualize this network in a real-time software component.**

***Index Terms*—smile intensity, deep learning, groups, multiparty interaction**

I. Introduction

As social animals, humans are primed to assess and respond to social signals in others [22]. Emotions are key to establishing and maintaining relations between humans and their environment, including others [2]. Consequently, emotions are often used as a basis for decision-making in human-human interactions; "the emotional signals of another guide action" [2]. This carries over into interactions between humans and agents [28], [36], [37]. The quality of emotion regulation in an interaction directly relates to the quality of social interactions [24]. For artificial agents to effectively regulate emotion and use emotion as the basis for decisions, fast and accurate emotion detection is an essential precursor.

A great deal of research has previously been conducted into emotion recognition using many different techniques [3], [32], [35]. More recently, attention has turned toward deep learning for emotion recognition [16], [17]. While much progress is being made, the majority of these existing works focus on developing new algorithms for emotion detection, and show results benchmarking against datasets. It is less common to discuss the deployment of these algorithms in online systems, particularly in cases where hardware resources might be limited.

In this paper, we present a smile intensity machine learning model, but also describe the deployment in a responsive software component. Our work is specifically targeted toward an agent interacting with groups of people. Dealing with groups of people increases the complexity of the deployment due to the varied, and increased numbers of people in the scene. This requires additional computational resources, which are not always available at the point of deployment, placing constraints on the software design. By discussing the deployment process and associated challenges, we contribute our lessons learned and potential solutions.

II. Related Work

Several deep learning models have been proposed for arousal/valence classification (e.g., [21]), however, smile detection is a less explored problem. Smile detection can be used to motivate agent behavior, or to further inform other affective detection components and reasoning. Detecting different smile intensities allows agents to respond appropriately based on the strength of the input expression from one or more users. This will enable the design and creation of more natural interactions.

Whitehill et al. developed a smile detector using Support Vector Machines (SVM) with Gabor features [38], [39]. The results showed the promise of smile intensity detection, but uses source data with coarse annotation ('happy' vs. 'not happy'), as well as mostly images, rather than videos. Consequently, information carried in the temporal nature of the smile is lost. Several other approaches using SVMs and Viola-Jones classifiers have also been developed [4], [33]. However, using coarse annotation, then relying on the decision values as a proxy for intensity has been questioned [10].

Other related work has involved the UvA-NEMO database [7]. This work has mostly focussed on categorizing genuine vs. posed smiles, rather than the intensity of the expression [8], [25]. However, this could be a useful dataset to explore smile intensity in isolation.

The expression intensity detector developed by Dhall and Goecke [5], [6] includes detection of smiles and laughs, but also has a coarse annotation of smiles: neutral, small smile, large smile. When teeth are showing, the images are labelled as laughs instead of smiles, possibly because the source data is image rather than video, so the audio signal is not available. For our use case, we would like more fine-grained categories of smiles to enable more nuanced detection, and subsequently, more nuanced agent reactions to the user input. We also make a distinction between a smile, which has no auditory component, and a laugh, which does. The work in [5] includes an interesting approach to extend the detector to estimate group emotion by factoring in social context of user orientations in a scene. This model place higher weights on

978-1-7281-3891-6/19/$31.00

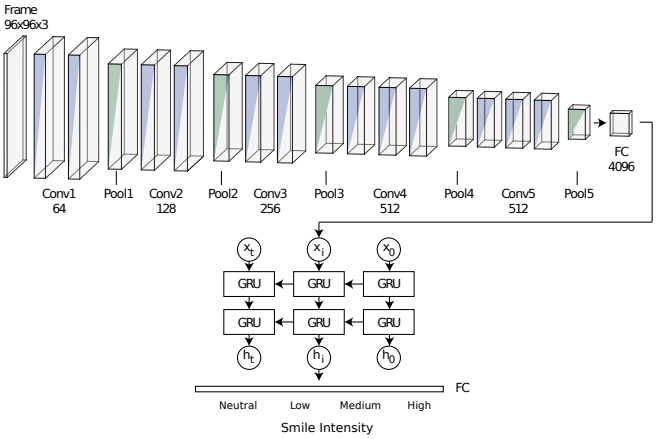


Fig. 1: CNN-RNN architecture based on [21] and [29] for visual features to predict smile intensity categories.

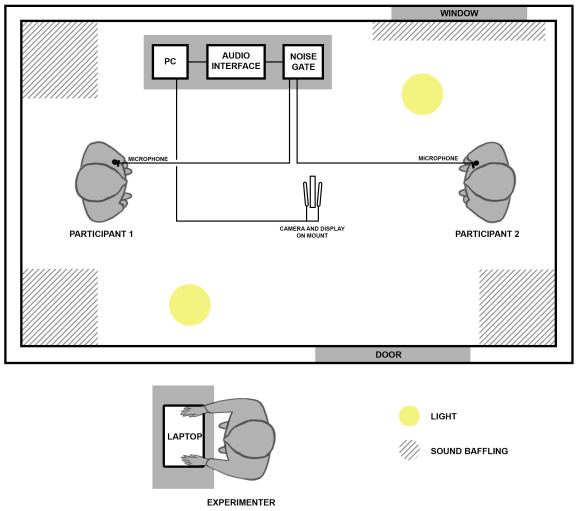


Fig. 2: Data collection setup.

larger faces (i.e., those closer to the camera). [6] takes this approach further, to also consider where people are in the group and other social factors.

Especially in working and entertainment environments, multi-party human-computer interaction systems play an important role [9], [20], [34]. One of the main difficulties in this research area is the demand for real-time applications, which becomes even more challenging in multiparty scenarios as the computation time potentially increases with every person added to a scene. There have been some previous approaches in designing real-time affect prediction systems for EEG data or eye gaze and speech [23], [27]. However, to the best of our knowledge, none of these works provide real-time solutions for multi-party situations. We therefore propose a DNN architecture for smile intensity detection and present its capability of producing predictions in multi-party scenarios in real-time, with limited hardware resources.

III. Network Architecture

The architecture uses a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). This architecture seeks to exploit the temporal aspects of the input data. The network is based on the work of Kollias and Zafeiriou [21] which had success in predicting arousal and valence from videos. The final version used in our paper is shown in Fig. 1. The network takes an RGB image of a face which is passed through a deep CNN and then an RNN. The CNN is based on [29] and is pre-trained on the VGGFACE database [29]. The RNN input uses the output of the first fully connected layer from the CNN, and employs two layers of Gated Recurrent Units (GRU). Each hidden layer of the RNN had 128 hidden units. Unlike [21], we do not employ a final attention layer after the GRU layers as this did not provide improvements in our case; instead they feed directly into a fully connected layer.

An additional key difference between the visual network we trained and that of [21] is motivated by the difference in data annotation and formulation of the problem. Our data is annotated categorically, so can now be treated as a classification problem. Accordingly, we change the loss function to use Cross Entropy Loss and softmax.

IV. Dataset

A. Data Collection

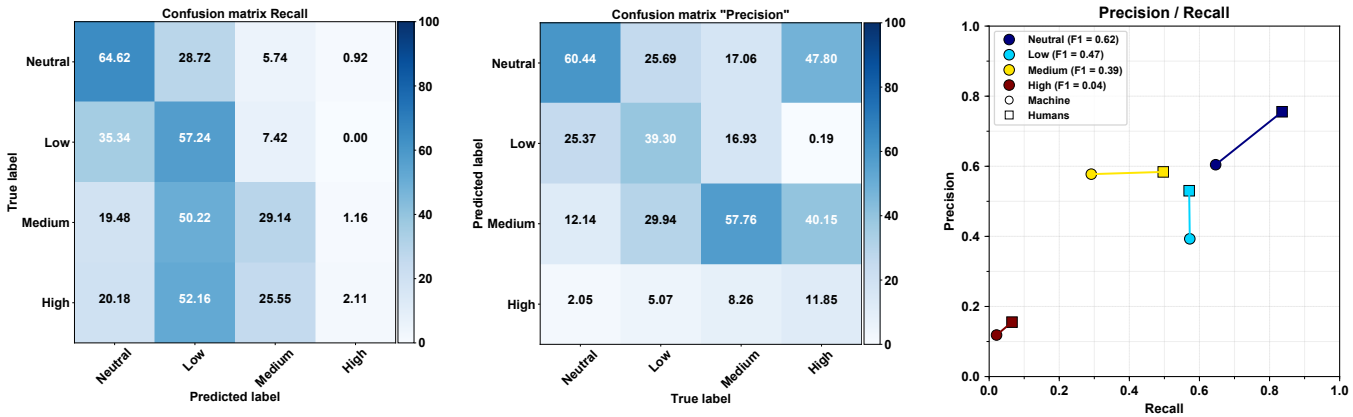
Several datasets for affect detection already exist, such as MELD [31], Aff-Wild [42], and AMI [26]. However, for the context of this work, interaction between a human and a virtual agent, we specifically wanted a dataset that contained spontaneous, non-acted expressions, in a conversational scenario. As such, we collected our own data through dyadic interactions.

Participants were given prompts designed to elicit emotion and were asked to then discuss for around 2 minutes before another prompt was shown. Prompts include questions such as 'What is the most embarrassing situation you have been in?', and 'What is the funniest story you have experienced?', inspired by [14]. This would continue until around 10 minutes of interaction had been collected. Including a short period before the beginning of the prompts where the participants would introduce themselves to one another, the average interaction length was M=14 minutes, SD=2.07 minutes. For this paper, a total of 15 interaction pairs were annotated, resulting in a total time of 6.80 hours of annotated data. Video frames were captured at a resolution of 1280x720, at 30 frames per second. Correspondingly, 734462 frames form the dataset.

The room was set up such that the participants would face each other, as shown in Fig. 2. A stand in the center of the room was used to mount an Intel RealSense camera facing toward each participant, so only one participant is visible in the camera frame. The upper body and face of the participants were included in the camera frame.

B. Data Annotation

The data was annotated by a team of 4 trained annotators using ELAN [41]. The annotators were asked to focus visually



(a) (b) (c)

Fig. 3: Confusion matrices for recall (*a*) and precision (*b*) of the trained smile intensity classifier. Performance is compared to human performance (*c*), with circles representing the machine classifier, and squares representing humans, for each class.

on facial expressions using the activation of the zygomaticus, pulling the lip corners, lifting the cheeks, and activation of the orbicularis oculi (around the eyes). An intensity level is given to each smile segment and is used to delineate the segments. A smile segment starts at the frame at which the expression starts, and stops either at the beginning of a smile of a different level, or at the end of the transition to a non-smile expression. The levels in increasing intensity order are: *Subtle*, *Low*, *Medium* and *High*. The *Subtle* level is included to capture instances where a low activation of the muscles co-occurs with another expression. It is intended to make sure all expressions are captured, but limit the time needed to decide whether this is a smile co-occurring with another expression or not, which can be a challenging task.

The annotators were trained by an experimenter experienced in annotating smiles from interaction data. They were provided with a training manual, and multiple sessions of feedback and comparison between their annotations. The training was performed on a short video from the same capture setup as the collected dataset (a video created when piloting the interactions).

In order to estimate the inter-rater agreement, six files were annotated by a second coder (i.e., another annotation team member who had not seen that file previously). Cohen's kappa was calculated with, and without taking the intensity levels into account. Calculating segment overlap without considering intensity levels, we obtain good results: κ=0.581. Omitting the "subtle" smile segments (which had less strict annotation criteria), κ increases to 0.672. The *Subtle* class caused problems in classification, possibly due to the lack of agreement between annotators, so is omitted from the following results. This will be returned to in the discussion.

C. Data pre-processing

We pre-process each RGB frame by detecting faces using the dlib CNN face detector [18]. Each detected face is resized to 96x96 pixels and normalized to $[-1, 1]$ range as in [21]. In 2.9% of the frames, no faces were detected, so these frames were removed from the dataset. In analyzing the data, it became clear that there was an imbalance between the annotated classes. The data balance was: *Neutral* - 45.8%, *Low* - 27.6%, *Medium* - 20.3%, *High* - 6.3%.

V. Experiments and Results

A. Training Details

The network was trained using PyTorch [30]. Weights are initialized using the PyTorch implementation of the uniform Glorot initialization [11]. For optimization we use Adam [19], dropping the learning rate every epoch by a factor of 0.97. We trained the network for 10 epochs with learning rate 1e-5, a batch size of 320 and a sequence length of 20 for the RNN. As stopping criteria we used early stopping: during training, the model is evaluated on a holdout validation dataset after each epoch. If the performance of the model on the validation dataset starts to degrade (e.g. loss begins to increase or F1 measure begins to decrease), then the training process is stopped. We split the dataset into 27 videos for training (with 3 used for validation), and 3 for testing (i.e., an 80:10:10 split).

B. Evaluation

In this section we present the classification results of our approach. To measure the quality of our classifier, we use confusion matrices, precision-recall curves, and level of agreement between annotations and predictions at the frame level.

A confusion matrix, used in multi-class classification, is a square matrix whose rows contain the normalized distribution of a predicted class for all ground truth instances of a single class, i.e., each entry (i, j) represents the fraction of ground truth instances of class i that are predicted as j, and is commonly summarized by its diagonal mean. The confusion matrix works well when classes are well balanced within a dataset, but fails when they are imbalanced, as is often the case with detection problems. In this case, the diagonal effectively measures the recall of each class but fails to emphasize

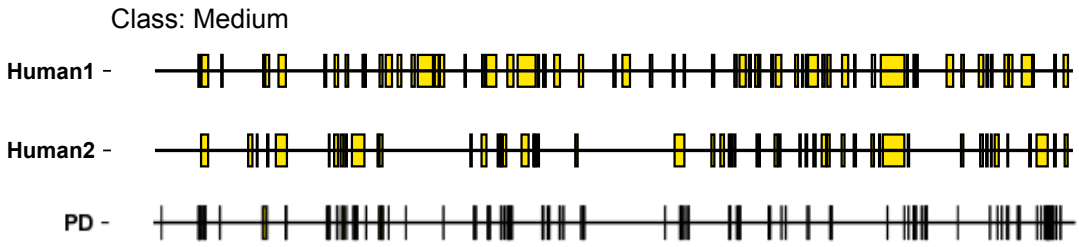


Fig. 4: Comparison between two human annotators and the classifier predictions (*PD*) for the *Medium* smile intensity class on one of the test videos. The timeline of the video is represented in a frame-by-frame manner, from left to right.

false positive instances which get absorbed into the grab-class *Neutral*, which contains the large majority of data points. To account for this, one must also look at the 'dual' confusion matrix, where entry (i, j) represents the fraction of predicted instances of class i that belong to class j according to the ground truth, in which case the diagonal effectively measures the precision of the class. The confusion matrices in Fig. 3, show that the classifier has difficulty in distinguishing the *Low* class from the *Medium* and *High* classes. *High* is the most confused class, probably because it is underrepresented in the dataset.

Precision-recall curves are used for measuring detection performance for a single class. We plot $precision = TP/PP = 1 - FP/PP$, which compares the false positives to the total number of predicted positives, against $recall = TP/(TP + FN)$, which is the same as the true positive rate but on the other axis. Recall favors the minimum number of false positives and false negatives with respect to the number of true positives. Fig. 3c shows the precision/recall performance of our system (circles) and human annotators (squares), connected by a line to show how far they are from each other. Our method reaches same precision for the *Medium* class and the same Recall for the Low class when compared to human annotators. The performance between our system and the annotators is also similar for the *High* class, probably meaning that it is hard to distinguish it from the medium class. However, overall, there is a discrepancy between the performance of our classifier and human-level performance.

Human performance is a good indicator for what to expect from automatic detection algorithms. Detections will not be perfect, due to smile ambiguity and imperfections in the ground truth annotations, but ideally they should achieve at least as good performance as humans. In Fig. 4 we show the comparison between our system predictions (PD) and two human annotators for the *Medium* class on a test video. We can observe that the two humans annotate this class in a similar way, although there are some segments where they strongly disagree. This is probably due one of the annotators segmenting the class more heavily, while the other has softer criteria, and therefore fewer class segmentations. Our system does a good job in predicting the class when using Human1 as ground truth, also detecting it for those frames where Human2 did not. This shows that the intensity can be subjective, leading to a system that learns to predict the classes according to the way the coder chose as a reference.

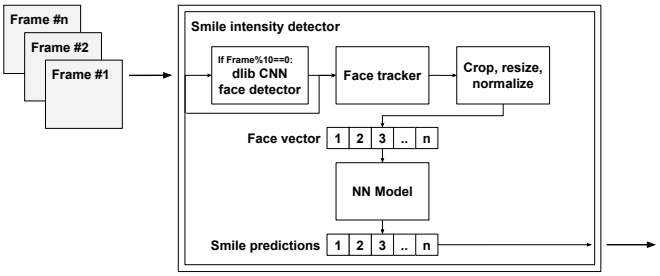


Fig. 5: Deployed software component architecture. RGB images are received by the component either via a TCP/IP connection or a locally connected camera. The component outputs predictions for each face in the frame, with a consistent identifying label.

VI. Online Software Component

In order to use the detection model in a multiparty interaction, the software requires consideration in order to maximize processing speed and to handle multiple faces. The resulting component can be seen in Fig. 5, where the video frames are first passed through face detection and a tracker.

The face detection uses a CNN from dlib [18] and is run every 10 frames. This is because it is relatively slow, so running every frame becomes an expensive operation. Once the detector has returned the faces, the tracker will add the faces from subsequent frames to a vector of the faces. Because the face detector runs once every 10 frames, this is the maximum length of time for a new face in the scene to be detected; a reasonable length for our use case, as we assume smile intensity changes don't occur faster.

The initial approach we adopted was to feed faces to the model for inference in a sequential manner. That is, we take a time sequence from a single face and run inference, before sending the next face. We process all faces in a frame, and then do the same on the next frame, and so on. With a single Graphical Processing Unit (GPU), when more than one face is present in the frame, the processing time is slower than the camera frame rate. This leads to a time lag in the results, or the need to drop frames. Having a single GPU is a reasonable constraint for a deployed system. To mitigate

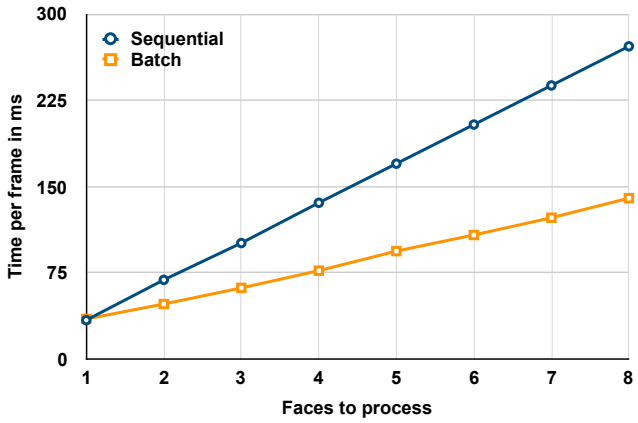


Fig. 6: Total processing time when our network processes each face in an input frame sequentially, or collates them to process as a batch. As can be seen, when the number of faces increases, the batch processing offers a substantial reduction in processing time.

this issue, we collect the vectors of these faces into a batch and feed a batch to the model, as would be done at training time. This technique has a downside in that we must predefine this batch size (and therefore the maximum number of faces we can process), but the upside is that the speed of processing becomes substantially faster when there are more faces (Fig. 6). With a single face to process, both techniques can run at 30fps (i.e., the speed the frames arrive), however, with 8 faces, the sequential approach achieves 4fps, compared to 7fps with the batch approach.

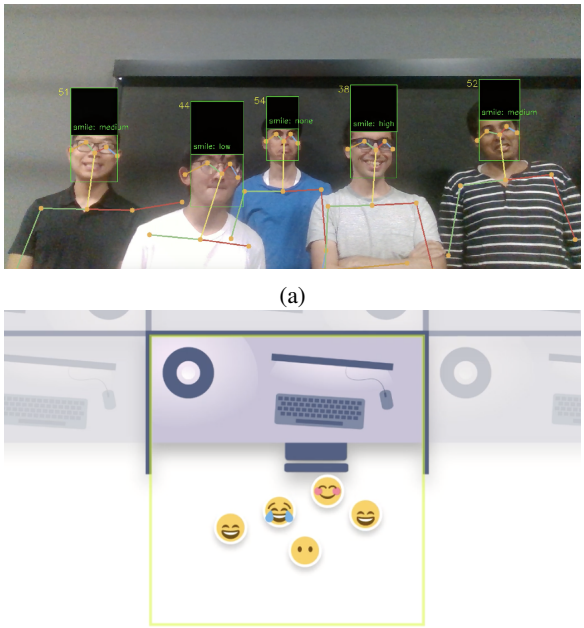
Other techniques for improving the speed of inference include weight pruning and quantization [13], and use of teacher-student networks [15]. However, these techniques either involve spending time to train and evaluate further models, or a reduction in precision. Using the technique of batching is a simple way to improve the component prediction speed.

To visualize the output of the predictions for groups of people, we created a web-based interface. Fig. 7 shows this interface and the underlying image data with the prediction label added. The interface maps from the real-world space to a virtual representation. Each individual in the camera view is replaced with an emoji showing their current smile intensity. This dynamically updates as new predictions are produced.

VII. Discussion and Future Work

In this paper we presented details of the smile intensity classifier that we trained using a deep neural network. We then presented details of using this classifier in a software implementation designed to classify the smiles of multiple users interacting with an agent, in real-time.

The results show that the classifier performance is reasonable, but is not yet comparable to human agreement levels. This could in part be due to noise in the training data from the human raters, with some subjectivity in the annotation process. Whilst we obtain good agreement between annotators



(a)

(b)

Fig. 7: (a) Raw input image with face tracking and prediction overlays. (b) Corresponding web visualization of network predictions. Note that the mapping from (a) to (b) factors in the camera perspective, so the view is flipped horizontally.

(Cohen's κ=0.672), there is clearly room for improvement. Part of this disagreement likely stems from the annotation of smiles when they co-occur with other expressions or facial movements, such as when the interlocutors are talking. These expressions are harder to classify due to the movement of the mouth. This is the probable reason for the reduced agreement when including the *Subtle* class, which was designed to capture many of these cases. In experiments where the network was trained and included the subtle class, both the recall and precision of this class were particularly poor.

One way to improve the overall output could be to experiment with multimodal fusion. Using audio, particularly when individuals are talking, may be a better indicator of smiles co-occurring with speech, than just using the visual modality alone. This also raises questions about early or late fusion of modalities. Prior work suggests that early fusion may lead to an improvement in results for emotion recognition tasks [12], [40], however, these approaches do not use deep learning. This would present an interesting avenue for future work.

Additional future work could include further exploration of the visual modality, with comparison to an existing SVM approach as a baseline for this data, or re-training with the model used here on an existing dataset. This would provide greater context for the results and could highlight where additional improvements might be made. We would postulate that the imbalance of the data here, and the relatively low numbers of individuals in the dataset prevented the network from greater generalization. In particular, more examples of

the *High* smile class would be beneficial.

The annotation procedure is highly time-consuming, so we did not annotate the entire dataset collected. While the entire dataset consists of around 100 different individuals, we only annotated 30 of those. Instead of annotating entire interactions for a smaller number of individuals, we may have found better results by annotating part of the interactions for a larger number of individuals. Alternatively, multiple annotators could label the same videos, and the model could be trained only on the segments that the annotators agree on. This may provide a more robust classification performance, as the subjectivity of the training data would be reduced.

Our current software component can be used to predict the smile intensity of multiple people in a group in real-time. This can be used to drive artificial agent decision making and behavior. However, this could be further improved by considering the emotion displayed in the entire scene, as in [6]. An agent should have some understanding of the interacting users not only as individuals, but as a social group, for producing the optimal behavior. Our current input data of two individuals interacting does not easily lend itself to solving this kind of problem. However, an interesting next step could be to consider the dynamics between the two individuals in the interaction to improve the smile predictions. Multiparty datasets such as the SALSA dataset would potentially provide the kind of rich social scene for further exploration of group emotion, also considering temporal factors [1]. However, accurate annotation of facial expression in the data may be difficult from the camera angles. Collecting and annotating videos of groups remains a challenge in furthering this line of research.

REFERENCES

[1] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2015.

[2] Joseph J Campos, Rosemary G Campos, and Karen C Barrett. Emergent themes in the study of emotional development and emotion regulation. *Developmental Psychology*, 25(3):394, 1989.

[3] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[4] Oscar Déniz, M Castrillon, J Lorenzo, L Anton, and Gloria Bueno. Smile detection for user interfaces. In *International Symposium on Visual Computing*, pages 602–611. Springer, 2008.

[5] Abhinav Dhall and Roland Goecke. Group expression intensity estimation in videos via gaussian processes. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3525–3528. IEEE, 2012.

[6] Abhinav Dhall, Roland Goecke, and Tom Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[7] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *Proceedings of the 12th European Conference on Computer Vision*, pages 525–538. Springer Berlin Heidelberg, 2012.

[8] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279–294, 2015.

[9] Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald Petrick. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 3–10. ACM, 2012.

[10] Jeffrey M Girard, Jeffrey F Cohn, and Fernando De la Torre. Estimating smile intensity: A better way. *Pattern Recognition Letters*, 66:13–21, 2015.

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[12] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443. IEEE, 2005.

[13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[14] Louise Heron, Jaebok Kim, Minha Lee, Kevin El Haddad, Stephane Dupont, Thierry Dutoit, and Khiet Truong. A dyadic conversation dataset on moral emotions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 687–691. IEEE, 2018.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[17] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691. IEEE, 2013.

[18] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Katrin Kirchhoff and Mari Ostendorf. Directions for multi-party human-computer interaction research. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*, 2003.

[21] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[22] Megan N Kozak, Abigail A Marsh, and Daniel M Wegner. What do I think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, 90(4):543, 2006.

[23] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from eeg signals. *IEEE Transactions on Affective Computing*, 9(4):550–562, 2018.

[24] Paulo N Lopes, Peter Salovey, Stéphane Côté, Michael Beers, and Richard E Petty. Emotion regulation abilities and the quality of social interaction. *Emotion*, 5(1):113, 2005.

[25] Bappaditya Mandal, David Lee, and Nizar Ouarti. Distinguishing posed and spontaneous smiles by facial dynamics. *CoRR*, abs/1701.01573, 2017.

[26] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[27] Jonny O'Dwyer, Ronan Flynn, and Niall Murray. Continuous affect prediction using eye gaze and speech. In *2017 IEEE International Conference on Bioinformatics and Biomedicine*, pages 2001–2007. IEEE, 2017.

[28] Maike Paetzel, James Kennedy, Ginevra Castellano, and Jill Fain Lehman. Incremental acquisition and reuse of multimodal affective behaviors in a conversational agent. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI '18, pages 92–100, New York, NY, USA, 2018. ACM.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[32] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1136–1139. IEEE, 2006.

[33] Keiji Shimada, Yoshihiro Noguchi, and Takio Kuria. Fast and robust smile intensity estimation by cascaded support vector machines. *International Journal of Computer Theory and Engineering*, 5(1):24, 2013.

[34] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2*, pages 766–773. ACM, 2002.

[35] Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE, 2006.

[36] Alessandro Vinciarelli and Anna Esposito. The handbook of multimodal-multisensor interfaces. chapter Multimodal Analysis of Social Signals, pages 203–226. ACM and Morgan & Claypool, 2019.

[37] Alessandro Vinciarelli and Alex Sandy Pentland. New social signals in a new interaction world: the next frontier for social signal processing. *IEEE Systems, Man, and Cybernetics Magazine*, 1(2):10–17, 2015.

[38] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Developing a practical smile detector. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, volume 2, 2008.

[39] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Toward practical smile detection. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2106–2111, 2009.

[40] Matthias Wimmer, Björn Schuller, Dejan Arsic, Bernd Radig, and Gerhard Rigoll. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proceedings of the 3rd International Conference on Computer Vision Theory and Applications*, pages 145–151, 2008.

[41] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1556–1559, 2006.

[42] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild' challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–41, 2017.