Gaze-enhanced Crossmodal Embeddings for Emotion Recognition

AHMED ABDOU, Technical University of Munich, Germany
EKTA SOOD, University of Stuttgart, Germany
PHILIPP MÜLLER, German Research Center for Artificial Intelligence, Germany
ANDREAS BULLING, University of Stuttgart, Germany

Emotional expressions are inherently multimodal – integrating facial behavior, speech, and gaze – but their automatic recognition is often limited to a single modality, e.g. speech during a phone call. While previous work proposed crossmodal emotion embeddings to improve monomodal recognition performance, despite its importance, an explicit representation of gaze was not included. We propose a new approach to emotion recognition that incorporates an explicit representation of gaze in a crossmodal emotion embedding framework. We show that our method outperforms the previous state of the art for both audio-only and video-only emotion classification on the popular One-Minute Gradual Emotion Recognition dataset. Furthermore, we report extensive ablation experiments and provide detailed insights into the performance of different state-of-the-art gaze representations and integration strategies. Our results not only underline the importance of gaze for emotion recognition but also demonstrate a practical and highly effective approach to leveraging gaze information for this task.

CCS Concepts: • **Computing methodologies → Machine learning**; **Computer vision**; • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: gaze, emotion recognition, multi-modality

ACM Reference Format:
Ahmed Abdou, Ekta Sood, Philipp Müller, and Andreas Bulling. 2022. Gaze-enhanced Crossmodal Embeddings for Emotion Recognition. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA, Article 138 (May 2022), 18 pages. https://doi.org/10.1145/3530879

1 INTRODUCTION

Automatic recognition of emotional expressions is an inherently multimodal task [51, 67]. Many state-of-the-art approaches combine information extracted from multiple modalities, e.g. expressions in a persons' face with speech-based features [56, 59, 66]. While these approaches consistently outperform uni-modal alternatives [19, 56] they share one crucial limitation: They require all modalities to be simultaneously present at both training and test time. However, this assumption rarely holds in application scenarios for emotion recognition. For example, emotion recognition systems have to rely on audio only at test time if used for telephone-based screening in the medical domain [38], or when users are outside the field of view of the camera in a video conference. In contrast, if a user is silent or strong background noise overlays a conversation, video analysis might

Authors' addresses: Ahmed Abdou, Technical University of Munich, Germany, ahmed.abdou@tum.de; Ekta Sood, University of Stuttgart, Germany, ekta.sood@vis.uni-stuttgart.de; Philipp Müller, German Research Center for Artificial Intelligence, Germany, philipp.mueller@dfki.de; Andreas Bulling, University of Stuttgart, Germany, andreas.bulling@vis.uni-stuttgart.de.

2573-0142/2022/5-ART138 $15.00
https://doi.org/10.1145/3530879

be the only usable modality. To combine the advantages of multimodal training with the flexibility of only requiring a subset of these modalities at test time, recent work proposed *cross-modal emotion embeddings* [26]. In this approach, a helper modality (e.g. video) is used during training time to improve the latent representation of a second modality (e.g. audio). This approach was shown to improve performance when only using the second modality (audio) at test time.

At the same time, a large body of work has demonstrated the close link between gaze and emotional expressions [1, 2, 22, 34, 40]. For example, gaze aversion was shown to impair the perception of anger and happiness [2, 10], and embarrassment is connected to more downward gaze than amusement [34]. Despite the importance of gaze, relatively few works have studied emotion recognition based on gaze location and pupil size [5] or combined gaze with other channels of affective information [3, 50, 62]. While the performance improvements demonstrated by these approaches underline the importance of integrating gaze into emotion recognition systems, they either rely on video information only [62] or assume both video-based gaze features and speech input to be available both at training and test time [3, 50]. In particular, so far gaze has not been integrated in multi-modal (i.e. video and audio) emotion recognition approaches that only require unimodal (i.e. either video or audio) data at test time.

In this work, we are first to propose to include an explicit representation of gaze in a cross-modal learning framework for emotion recognition. More specifically, our approach builds on the crossmodal emotion embeddings model (EmoBed) by Han et al. [26]. We augment the visual pipeline of EmoBed with a state-of-the-art gaze feature representation for gaze-based emotion recognition [50]. The resulting novel feature representation of the video modality acts as a helper modality during training for speech-only testing, or is supported by speech during training when performing video-only test evaluations. With a model-level fusion strategy of gaze and facial features, we achieve a F1 score of 45.0 for video-only testing on the One-Minute Gradual Emotion Recognition dataset [9], clearly improving over the previous state of the art by [26]. For audio-only testing, we reach 43.4 F1 with an early fusion approach for gaze integration at training time, also outperforming the previous state of the art [26]. [1]

The specific contributions of our work are threefold: First, we present a novel state-of-the-art crossmodal learning approach for emotion recognition that, for the first time, makes use of an explicit representation of gaze. Second, we conduct experiments on the One-Minute Gradual Emotion Recognition dataset [9], improving over the previous state of the art both for video-only as well as audio-only testing. Third, we perform extensive ablation experiments and report results for different gaze integration strategies, including early versus model-level fusion, as well as different state-of-the-art gaze feature representations [50, 62].

2 RELATED WORK

Our work is related to 1) the connection between gaze and displays of emotions, 2) gaze-based automatic emotion recognition, as well as 3) multimodal approaches to emotion recognition.

2.1 Gaze and Emotions

Eye gaze behavior (coupled with other modalities such as facial expressions) has been studied widely in psychology of emotion expression and perception [30, 36, 41, 42, 63]. Gaze can reveal information about the users attention and intentions [30], and specific eye movement behaviors (e.g., gaze direction) coupled with facial movements [1, 10, 64] are relevant when expressing and perceiving specific kinds of emotion classes. With such facial/eye region expressions, the additional

[1] Code and other supporting material is made publicly available at https://perceptualui.org/publications/abdou22_etra/

use of perceived gaze direction (either directed or averted) has been shown to better assist humans to distinguish between emotion classes [1, 22, 40–42].

Gaze direction has been shown to be associated with "the underlying behavioral intent (approach-avoidance) communicated by an emotional expression" [2]. Milders et al. [42] showed that the gaze direction of another person can affect your emotion recognition accuracy and the intensity by which you perceive the emotional stimuli. Moreover, results showed that averted gaze is a useful feature when detecting fear over happiness or anger and the exact opposite with a direct gaze. In addition, [34] indicated that when humans experience embarrassment, they first avert their gaze and then subsequently additional expressions occur, such as shifting eye, abnormal speech sounds, and smiling.

2.2 Gaze-based Emotion Recognition

Given the strong link between gaze and emotions, a large body of research has explored the use of gaze for automatic emotion recognition. For example, Jaques et al. [31] explored different machine learning algorithms for gaze-based recognition of two emotional states, curious and bored, in an e-learning environment. They found that while predicting curiosity was not possible above chance level, predicting boredom showed more promise. Similarly, [5] trained a shallow feed-forward neural network to predict positive, negative, or neutral emotion classes. To improve performance, they suggested to include additional modalities and gaze features in future work. O'Dwyer et al. [50] explored the use of a larger gaze feature set to train an LSTM network for the task of continuous affect prediction from the RECOLA dataset [55]. They found that their model performed better for arousal prediction when trained on gaze features.

A number of works have provided strong evidence for the benefit of using gaze for emotion recognition. Anwar [4] proposed a method to predict seven basic emotion classes from facial action units and gaze with 93% accuracy. Similarly, Alhargan et al. [3] explored the use of speech and gaze features for affect recognition in a gaming environment and reported top performance when combining both modalities. Their results further showed that gaze features were even more helpful than speech. In a study by O'Dwyer et al. [49], the addition of eye gaze features to speech yielded an improvement of 19.5% for valence prediction and 3.5% for arousal prediction. An opposing pattern was found by O'Dwyer et al. [50] who employed pupillometry and gaze features for valence and arousal estimation on the RECOLA dataset [55]. While eye-based features did not perform well for arousal prediction, the best performance was achieved when combining eye-based features and speech features. Van Huynh et al. [62] presented a neural network approach for emotion recognition based on facial expression and gaze and showed clear improvements when using gaze features to representations of facial expressions. Despite all of these works, to the best of our knowledge, no emotion recognition approach incorporating gaze with both facial expressions and speech has been proposed.

2.3 Multimodal Emotion Recognition

Due to their ubiquity, most works on multimodal emotion recognition have focused on combining audio and video [56, 66], but how to combine them remains an open question. In early fusion, inputs or raw feature representations are merged before they are fed into a joint network [17, 57]. In model-level fusion, each modality is processed by a dedicated network before both intermediate feature representations are merged and then passed through a joint network [16, 54]. Finally, in late fusion, modalities are fused on the level of predictions [29, 66]. Caridakis et al. [14] built a multi-modal discrete emotion classification system based on facial expressions, body movement/gestures, and speech. While all fusion methods improved over monomodal classification, feature fusion provided the best performance. Hu et al. [28] won the EmotiW challenge in 2017 by proposing a novel score function that added model-level fusion at multiple levels within a neural network. [43] obtained state of the art results on the IEMOCAP dataset [12] by separately learning spatio-temporal features using a recurrent neural network. Recently, Schoneveld et al. [59] combined a recurrent neural network with model-level fusion and reported state-of-the-art results on the RECOLA dataset [55]. Although integrating audio and video generally increases emotion recognition performance, a common limitation of all previous methods is that all modalities also need to be present at test time. However, this is rarely the case in real-world applications. To address this limitation, Han et al. [26] proposed EmoBed – an approach that only required a single modality at test time, yet could still profit from both modalities during training. To this end, EmoBed aligned video and audio embeddings in a joint space and used a subsequent network to predict emotion labels from embeddings of either modality. Most recently, [53] proposed Stronger Enhancing Weaker (SEW), a method that allowed to exploit a stronger modality during training to improve test performance of a weaker modality. The crucial difference to EmoBed is that this method cannot improve the test performance of a strong modality by training jointly with a complementary but weaker modality. While these "asymmetric" approaches were shown to improve monomodal testing performance, they disregarded gaze for emotion recognition. In stark contrast, we show that integrating an explicit gaze representation into the visual pipeline of such an asymmetric approach improves test performances both on video- as well as on audio alone. Given that our literature review did not indicate a clear preference for early- or model level fusion of feature representations, we evaluate the performance of both alternatives of joining gaze features with the visual processing stream.

3 METHOD

Our method extends the crossmodal emotion embedding (EmoBed) [26] framework by integrating an explicit gaze representation into the visual pathway (see Figure 1). EmoBed consists of a visual and a speech stream, each of which produce embeddings of the same size. These pathways are trained in two ways. First, a downstream network is trained on the embeddings without knowledge of the embeddings' source modalities to predict the ground truth emotion labels. Second, in a shared embedding space, intra- as well as inter-modal triplet losses are applied to improve and align visual and audio embeddings. This training setup only requires a single modality at test time. Due to the inter-modal triplet loss applied to the embedding space, a second modality supplied at training time is able to improve the embeddings of the test modality.

3.1 Gaze-enhanced Visual Stream

We propose a novel gaze-enhanced visual stream network for EmoBed (see Figure 1). This network incorporates a dedicated gaze representation into the visual stream of EmoBed, mapping both input features to an embedding space in $\mathbb{R}^E$.

3.1.1 Feature fusion. As from our literature review it remained unclear what is the best way to fuse feature representation for emotion recognition, we study two alternatives: (1) early fusion, and (2) model-level fusion.

In early fusion, we concatenate the visual and gaze data before passing it to the encoder. In other words, the early fusion process can be expressed as $e_{vg}^{early} = f_{vg}([x_v; x_g])$ where $f_{vg}(.) : \mathbb{R}^{V+G} \rightarrow \mathbb{R}^E$. V, G, and E denote the dimension of visual data input, gaze data input, and the embedding size, respectively.

In model-level fusion, we employ a separate encoder for each of the two data streams. Their corresponding embeddings are subsequently concatenated, forming an embedding of twice the original size (i.e. $2E$). Finally, we pass this new embedding into a projection layer to transform it to the original embedding size E. In other words, the visual encoder can be expressed as $f_v(.) : \mathbb{R}^V \rightarrow \mathbb{R}^E$,

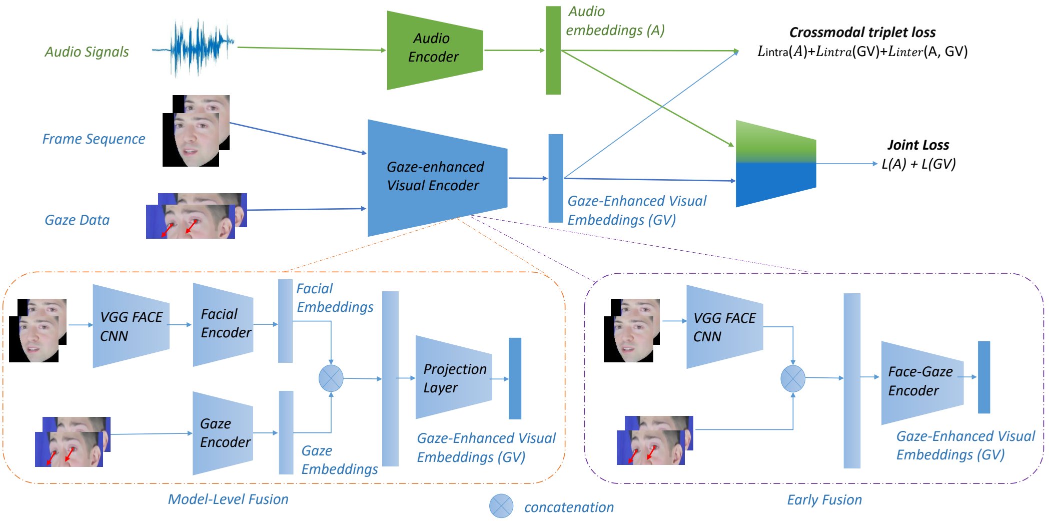

Fig. 1. Overview of our proposed gaze-enhanced crossmodal emotion embeddings method. **Top**: general structure of the crossmodal emotion embeddings framework. Audio and video signals are encoded in separate streams. A crossmodal triplet loss is applied to the embedding space to align embeddings from the two modalities. For classification, the embeddings are fed to a joint classification network that is agnostic of the embeddings' source modality. **Bottom Left**: Our proposed model-level fusion network for integration of gaze information into the visual encoder. Here, VGG FACE features and gaze features are concatenated after application of the encoder networks. **Bottom right:** Our proposed early-fusion network. Here, gaze features are concatenated and subsequently fed through a joint encoder network.

the gaze encoder as $f_g(.) : \mathbb{R}^G \rightarrow \mathbb{R}^E$, and the projection network as $f_p(.) : \mathbb{R}^{2E} \rightarrow \mathbb{R}^E$. After passing the visual input x_v and gaze input x_g into their corresponding encoders we get $e_v = f_v(x_v)$ and $e_g = f_g(x_g)$. Finally, we project them again into the $\mathbb{R}^E$ dimension by $e_{vg}^{model} = f_p([e_v; e_g])$.

3.2 Audio Stream

Similarly, the goal of the audio stream is to map each audio input to an embedding space of the same dimensionality of the gaze-enhanced visual embeddings (i.e. $\mathbb{R}^E$). To this end, we feed the audio input $x_a \in \mathbb{R}^M$ into the audio encoder network which can be expressed as $f_a(.) : \mathbb{R}^M \rightarrow \mathbb{R}^E$. We obtain the resulting audio embeddings $e_a = f_a(x_a)$.

3.3 Triplet Training

Triplet training is applied on the gaze-enhanced visual embeddings and the audio embeddings. The goal of the triplet training is to align the two different types of embeddings in a shared embedding space. The embeddings in the new space, regardless of their input source, should be close to each other if they have the same label, and far from each other otherwise. We quantify the semantic similarity using the euclidean distance. I.e. the similarity between two embeddings e_i and e_j is defined as $d(e_i, e_j) = ||e_i - e_j||_2$.

The triplet loss consists of an intra-modal as well as an inter-modal component. The intra-modality loss ensures that embeddings that carry the same label within the same modality are close to each other and far from each other otherwise. The inter-modality on the other hand ensures that embeddings carrying the same label across modalities are close to each other and far from each other otherwise.

3.3.1 Intra-modality loss. Given a batch of n embeddings A from a single modality, the intra-modality loss is calculated by first computing the nxn pairwise distance matrix. The elements on its diagonal are zero as they represent the distance between each embedding and itself. For each embedding $e \in A$, we followed [26] in the triplet mining technique, i.e. we choose the hardest positive- as well as the hardest negative example. The hardest positive example e^+ is defined as the example with the same label that is farthest from e in the embedding space. Similarly, the hardest negative example e^- is the the example that is closest to e in the embedding space, but has a different label than e. Then, the intra-modalitiy loss of modality A is defined as as:

$$L_{intra}(A) = \sum^{n}(d(e_a, e_a^+) - d(e_a, e_a^-)) \tag{1}$$

3.3.2 Inter-modality loss. Given a batch of n examples with both embeddings e_a of modality A and corresponding embeddings e_b of modality B, we first compute the cross-modality nxn distance matrix. This matrix contains the distances $d(e_a, e_b)$ for all $e_a \in A$ and $e_b \in B$. The elements on its diagonal represent the distances between the e_a and e_b embeddings belonging to the same video instance. Then, with the same strategy for searching for hardest positives and negatives, the inter-modality loss is calculated by

$$L_{inter}(AB) = \sum^{n}(d(e_a, e_b^+) - d(e_a, e_b^-)) \tag{2}$$

3.3.3 Full triplet loss. For two batches of embeddings A and B of size n coming from two different modalities, we calculate the final triplet loss as follows:

$$L_{triplet} = L_{intra}(A) + L_{intra}(B) + L_{inter}(AB) \tag{3}$$

Thus, regardless of their modality, the triplet loss pulls embeddings that have the same label close to each other while it pushes embeddings with different labels apart.

3.4 Joint Training

At the same time when the embedding space is trained with the triplet loss, the embeddings from each modality are passed to a shared classifier. This classifier is agnostic to the source modality of the embeddings. It is trained to classify emotion categories based on the embeddings it receives. That is, given an embedding $e \in \mathbb{R}^E$ the classifier is a function $f(e) = \hat{y}$, where $\hat{y} \in \mathbb{R}$ is the predicted label for e. Thus the total loss of the classifier can be expressed as:

$$L_{joint} = L(A) + L(B) \tag{4}$$

where $L(A)$, $L(B)$ are the cross entropy loss for inputs embeddings A and B.

3.5 Features

3.5.1 Gaze Features. As a basis for gaze feature computation, we make use of OpenFace [8]. OpenFace makes use of a Constrained Local Neural Field (CLNF) landmark detector presented in [7, 65] for eyelids, iris, and the pupil detection. OpenFace is trained using the SynthesEyes dataset [65] which contains photorealistic close-up images of eyes and is widely used for training and evaluating gaze estimation models. The gaze estimation of OpenFace Baltrusaitis et al. [8] was originally evaluated on the MPIIGAZE dataset [68], which is widely used for appearance based gaze estimation and contains realistic in-the-wild images of users taken from webcams. OpenFace reached an angular error of 9.1 in the challenging cross-dataset evaluation scenario, clearly outperforming

Table 1. Base features from OpenFace output and the corresponding statistical features. These 103 features are subset of the feature set in [50]. *LR* refers to a linear regression fitted to the time series of feature values in the window. *Time ratio* is the proportion of time during which a binary feature is detected in the analysis window. *IQR* denotes the interquartile range, i.e. *IQR 2-3* refers to the difference between third and second quartile.

Base feature	Statistical functionals	# Features
gaze angle x, gaze angle y, Δ gaze angle x, Δ gaze angle y, pupil diameter mm	min, max, mean, median, quartile 1, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	60
Δ pupil diameter mm	min, max, mean, quartile 1, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	11
eye blink intensity	max, mean, median, quartile 3, std, IQR 1-2, IQR 2-3, IQR 1-3, LR intercept, LR slope	10
pupil dilation, pupil constriction	time ratio, mean time, max time, total time	8
gaze approach	time ratio, mean time, max time, median time	4
eyes closed, gaze fixation	time ratio, min time, max time, mean time, median time	10

previously proposed methods [8]. Openface was subsequently used for a multitude of tasks, such as mutual gaze prediction for human-robot interaction [58], improving methods for controlling devices with gaze [35], and anticipating averted gaze [47]. As such, Openface is a well established and accurate toolkit for facial landmark detection and gaze estimation, particularly for real world person independent gaze estimation. While a dedicated eye tracking device can reach higher gaze estimation accuracy, this would limit our approach to highly controlled lab-based scenarios.

Based on the gaze estimates obtained by OpenFace, we compute a set of statistical features from the recently proposed gaze feature set by O'Dwyer et al. [50]. We observed that some features of O'Dwyer et al. [50] can have zero variance during a time window, rendering computation of skewness and kurtosis impossible. As a consequence, we restrict our set of features to those from O'Dwyer et al. [50] that are not subject to this issue.

In particular, we extracted a base set of 9 gaze related features, on top of which we computed several statistical functionals, resulting in 103 features in total. The base set consists of 4 numerical and 5 binary features. The numerical features are eye gaze direction in radians in world coordinates averaged for right and left eyes, the pupil diameter estimated from the left eye landmarks, and the eye blink intensity (i.e. Facial Action Unit 45[2]). The binary features are the pupil dilation time, constriction time, eye closed time, gaze approach time, and gaze fixation time. Gaze approach is calculated by taking the derivative of the average depth of the eye landmarks. If the derivative is above zero, it is counted as gaze approach. See Table 1 for an overview over all base features and the applied statistical functionals.

OpenFace outputs pupil diameter in millimeters based on the detected facial landmarks [8]. As we do not assume knowledge of the camera intrinsics, the absolute value of the estimated pupil diameter may be inaccurate, as it is the case in previous work that computed pupil diameter features based on OpenFace output [50]. However, for the task of gaze-based emotion recognition, we are

[2] list of action units and their description https://imotions.com/blog/facial-action-coding-system/

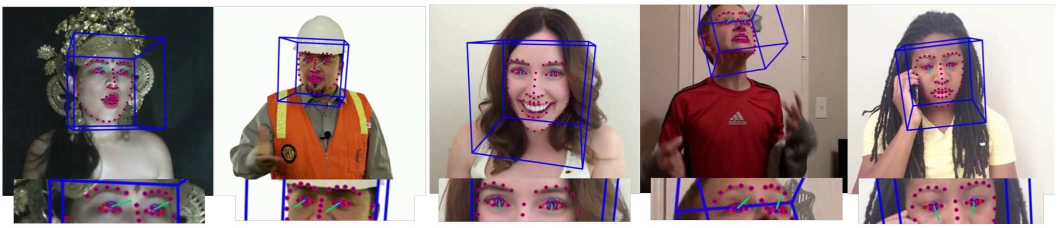

Fig. 2. Facial keypoints and gaze estimates for randomly chosen frames with **high** OpenFace confidence.

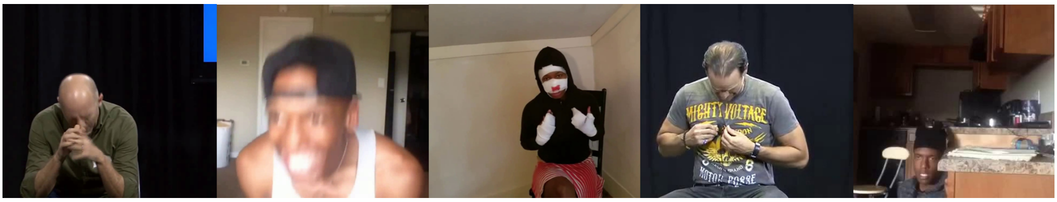

Fig. 3. Facial keypoints and gaze estimates fn randomly chosen frames where landmarks detected with **low** OpenFace confidence

only interested in relative differences in pupil diameter that are preserved even if the absolute value is incorrectly estimated. We still use “millimeter” to be in line with the terminology of [50].

We extract the features on windows of 1 second, resulting in as many feature values per window as there are frames per second in the input video. These window-based features can be used both in the early- as well as the model-level fusion model discussed in Section 3.1.1, as they can be aligned with the frame-wise facial features. In the model-level fusion scenario, we also investigate a version of the gaze features where we average the features that were first computed on the window level across the whole utterance. We suspect this could lead to a more robust gaze feature representation that is less prone to random fluctuations.

3.5.2 Facial Features. We used OpenFace [8] for face detection and face alignment. We pass the aligned face images to the VGG-Face model [52] and extract the the 4096 length feature vector from the "fc-7" layer. If k is the number of frames in the video, this results in (k,4096) feature values for that video.

3.5.3 Audio Features. In line with [26], we extracted the eGemaps feature set [23] using the OpenSMILE toolkit [24]. This results in a 88-dimensional feature vector per video.

4 EXPERIMENTAL EVALUATION

4.1 Training Details

Inspired by Han et al. [26], we used GRU-RNN layers for each modality encoder as well as for the shared classification network. We used a single hidden layer for both the encoders as well as the shared network with 120 units each. We applied L2 regularization of 10^{-4} via weight decay, and chose a batch size of 64. We found that using the Adam optimizer [37] with a 10^{-4} learning rate yielded the best generalization results on the OMG dataset. Moreover, all inputs were standardized to their mean and variance on the training set. Each batch of input data yielded two batches of the same size: One batch containing gaze-enhanced visual embeddings and another batch of audio embeddings. Crossmodal triplet loss is applied on both batches. Furthermore, the batches

Table 2. F1 scores for different approaches and test modalities on the OMG [9] dataset. Scores are averaged across 20 runs with different random intialisations and corresponding standard errors are shown in brackets.

Approach ↓ ‖ Test Modality →	Audio	Video
OMG baseline [9]	33.0	37.0
EmoBed [26]	41.7	43.9
monomodal		
no gaze	41.3 (0.15)	42.8 (0.24)
averaging, model-level fusion	-	43.8 (0.21)
windowing, early fusion	-	43.0 (0.23)
windowing, model-level fusion	-	43.7 (0.20)
crossmodal triplet training		
no gaze	42.9 (0.17)	43.5 (0.25)
averaging, model-level fusion	42.6 (0.19)	**45.0** (0.22)
windowing, early fusion	**43.4** (0.17)	43.7 (0.26)
windowing, model-level fusion	42.8 (0.15)	44.5 (0.20)

were passed in an alternating fashion to the joint classification network. That is, the classification network receives the audio embeddings of batch 1, then the gaze-enhanced video embeddings of batch 1, the audio embeddings of batch 2 and so forth.

4.2 Dataset

We evaluated our gaze-enhanced crossmodal triplet training on the popular One-Minute Gradual Emotional (OMG) Behavior dataset [9]. We opted for this dataset given that it is well suited for the task and has been used in closely related prior work [26]. The OMG dataset [9] consists of 567 monologue videos collected from YouTube with an average video length of one minute and a frame rate of 29 fps. Each monologue video contains various emotional behaviors shown by the person in the video. Each video was subdivided into several utterances with an average length of 8 seconds. The utterances were subsequently annotated by 5 different Amazon Mechanical Turk (AMT) workers each with discrete labels for the six basic emotions according to Ekman and Friesen [21]. Addition of a neutral class resulted in the seven classes: *disgust*, *anger*, *fear*, *happy*, *neutral*, *sad*, and *surprise*. Annotators saw the utterances in sequence, i.e. they could take into account context information from preceding utterances of the same video. As annotations are based on the full array of human behavior channels including speech, gaze, and facial expressions, this dataset is well suited to evaluate our gaze-enhanced crossmodal embedding approach.

Before feature extraction, we removed four utterances from the training set for which neither OpenFace could detect any face nor OpenSMILE could detect any audio signal. This resulted in a training dataset of 2,438 utterances in training and 617 utterances in the validation set. All experiments and results reported in the following have been performed on the validation set given that the test labels are not available for OMG. Using the OMG validation set allows us to directly compare our results with those of [26].

The facial landmark detector of OpenFace 2.0 [8] reported a confidence score of more than 0.9 for 91.2% of all frames in the OMG dataset, indicating that in the vast majority of frames high-quality gaze estimates can be expected. For a qualitative analysis we randomly sampled one frame for each gaze direction (left, right, up, down, straight) from the set of frames with confidence score of at least 0.9 (see Figure 2). All sampled frames show accurate estimates of gaze direction. Figure 3

Table 3. Average of F1 scores for each emotion class for cross modal triplet training. Standard errors across 20 runs are shown in brackets.

Approach ↓ || Emotion →	Neutral	Happy	Sad	Anger	Disgust	Fear	Surpr.
Train Samples	888	713	346	293	112	60	26
Validation Samples	196	207	79	61	48	18	8
test modality: video							
averaging, model-level fusion	49.6(0.5)	55.7(0.4)	32.7(0.8)	29.2(1.3)	4.7(1.3)	0	0
no gaze	48.1(0.6)	52.3(0.4)	29.7(1.5)	32.5(1.5)	11.6(1.7)	0	0
test modality: audio							
windowing, early fusion	49.7(0.4)	49.7(0.5)	34.8(0.8)	25.3(1.5)	0	0	0
no gaze	49.2(0.4)	49.8(0.5)	33.9(0.8)	23.4(1.1)	0	0	0

shows five frames where facial landmark detection failed. Such complete failures occur seldom (2.5% on the training set) and are often due to occlusions, motion, or extreme head poses.

4.3 Results

In line with previous work [9, 26], we use the micro F1 score as our general performance metric. It is calculated by counting the total true positives, false negatives and false positives globally across all classes.

4.3.1 Effect of Gaze Integration. We compare our proposed method with the previous state of the art reported by Han et al. [26] as well as ablated versions in Table 2. For video as test modality, we achieve 45.0 F1 for our crossmodal triplet training method that uses averaged gaze features and model-level fusion. This is a clear improvement over the previous state of the art of 43.9 F1. In the case of audio-only testing, our crossmodal triplet training method with gaze features extracted on windows and early fusion achieves the best result with 43.4 F1. Again, this outperforms the previous state of the art of Han et al. [26] at 41.7 F1.

At the same time, our best gaze integration methods for each test modality improve over their corresponding no gaze ablation at 43.5 F1 for video testing, and 42.9 F1 for audio testing. The "no gaze" ablation represents our implementation of the EmoBed framework. For audio-only testing, it improves over the performance originally reported in Han et al. [26] by 1.2 F1, and for video only testing it is worse by 0.2 F1. While all gaze integration strategies improve over the no gaze triplet training alternative for video testing, the results are more mixed for audio-only testing. Here, only windowed gaze extraction combined with early fusion improves over the no gaze ablation.

In addition to crossmodal training setups, we investigate purely monomodal training. Most importantly, all our crossmodal triplet training models improve over their monomodal counterparts. Furthermore, the pattern of results for different gaze integration strategies in monomodal video models follows the corresponding pattern of results for our crossmodal triplet training models. This indicates that the advantage of averaging and model-level fusion for video-only testing is stable across evaluation conditions.

4.3.2 Analysis of improvements per emotion class. To gain a deeper insight into the performance of our gaze integration method, we conducted a performance analysis per emotion class. In Table 3, we compare the performances of our gaze-enhanced crossmodal triplet training approaches to

Table 4. F1 scores of different gaze feature representations in the crossmodal triplet training formulation on the OMG [9] dataset. Standard errors across 20 runs are shown in brackets.

Gaze Features ↓ || Test Modality →	Audio	Video
O'Dwyer et al. [50]		
averaging, model-level fusion	42.6 (0.19)	**45.0** (0.22)
windowing, early fusion	**43.4** (0.17)	43.7 (0.26)
windowing, model-level fusion	42.8 (0.15)	44.5 (0.20)
Van Huynh et al. [62]		
averaging, model-level fusion	43.1 (0.16)	44.8 (0.24)
windowing, early fusion	43.2 (0.18)	43.6 (0.28)
windowing, model-level fusion	42.8 (0.18)	43.5 (0.23)

the corresponding no-gaze ablations. For video testing, we observe an improvement for our the gaze-enhanced method for the classes neutral (+1.5 F1), happy (+3.4 F1), and sad (+3.0 F1). On the other hand, our gaze-enhanced method obtains a lower F1 score for anger (-3.3 F1) and disgust (-6.9 F1). The improvements are aligned with the number of training samples available for each class. Our gaze-enhanced approach improves for classes with high numbers of training samples and has a disadvantage on classes with low numbers of training samples. A different pattern can be observed in the audio testing case. Here, our gaze-enhanced approach improves for the classes neutral (+0.5 F1), sad (+0.9 F1), as well as anger (+1.9 F1), and is on par with the no gaze baseline for happy (-0.1 F1). These varied behaviors point to the different roles gaze features play when incorporated into the helper modality (audio testing), or when used as a direct basis of test time prediction in video testing (see our Discussion in Section 5.1).

Interestingly, all approaches only reach a very low performance on classes with a low number of training samples (i.e. disgust, fear, and surprise). To check whether this might be due to not yet converged training for these emotion classes, we tried training our model for a large number epochs (250). This did not resolve the issue but lead to overfitting.

4.3.3 Comparing Gaze Feature sets. Our gaze-enhanced visual stream utilizes a set of gaze features proposed for emotion recognition by O'Dwyer et al. [50]. A different set of features was proposed by Van Huynh et al. [62], which to the best of our knowledge, was the only other feature set proposed in the emotion recognition literature that utilizes gaze estimates extracted from standard RGB videos (e.g. via OpenFace [8]). The main difference between these feature sets is that the features of Van Huynh et al. [62] do not contain information about the pupil dilation and constriction, nor eye blinking. On the other hand, they include information about the iris and the average location of the eye landmarks. In detail, 51 statistical features were calculated from a base set of features extracted by OpenFace. These base features include the eye gaze direction vector in 3D and radians averaged for both eyes, height and width of the pupil for both eyes, the eye locations, as well as the vertical distances of the iris landmarks.

Table 4 shows the results of a comparison of our gaze featureset inspired by O'Dwyer et al. [50] with the one by Van Huynh et al. [62] in the crossmodal triplet training approach. In both audio and video testing scenarios, the best performance is achieved using the featureset based on O'Dwyer et al. [50]. While for video-only testing, the features of O'Dwyer et al. [50] are always superior to those of Van Huynh et al. [62], the results for audio-only testing are more mixed.

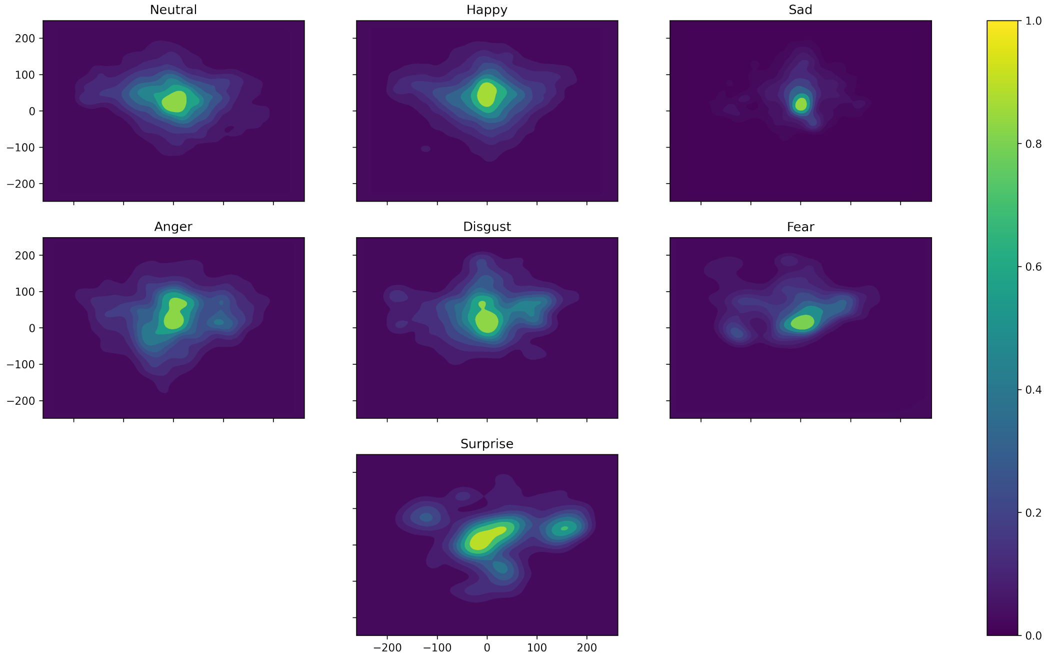

Fig. 4. Visualization of the distribution of OpenFace gaze estimates projected on the camera plane for different emotion classes.

4.3.4 Gaze Feature Analysis. We visualized the distribution of gaze estimates for each emotion class by projecting them on the camera plane. Figure 4 shows the kernel density estimation of these projections, demonstrating the connection between gaze estimates and the emotions expressed in the OMG dataset. For instance, the gaze distribution for Sadness is very concentrated which is in line with passivity that is associated with this emotion. Surprise on the other hand is more spread out.

We additionally performed a feature importance analysis for each emotion class. We used maximum relevance minimum redundancy (mRMR) [18] to rank the features that would most distinguish this class from the others in a one versus all manner. The top 5 features for each class are shown in Table 5. The dominance of the gaze angle based features is in line with previous results on feature importance analysis for arousal and valence regression O'Dwyer et al. [50]. The gaze angle based features, which represent 46.6% of the full feature set, contributed to 71.4% of the top five ranked features in Table 5. Furthermore, pupil diameter, eye blink intensity, and gaze approach account for 10.6%, 9.7%, and 3.8% of the full feature set, while contributing to the top ranked features by 17.1%, 8.5%, and 2.8% respectively.

5 DISCUSSION

5.1 Achieved Performance

In our experiments, gaze integration into the visual stream of the crossmodal emotion embeddings (EmoBed) [26] framework clearly improved test performances for both video-only as well as audio-only testing. The precise gaze integration method affected performances differently for different test modalities, i.e. we did not find a single best gaze integration method that is superior across all

Table 5. Top five features selected by mRMR for each emotion class.

Neutral	Happy	Sad	Anger	Fear	Disgust	Surprise
gaze angle y median	gaze angle y quartile 3	gaze angle x SD	Δ gaze angle x max	gaze angle y LR intercept	Δ gaze angle y quartile 1	gaze angle x IQR 2 3
Δ pupil diameter mm LR slope	gaze angle x median	Δ pupil diameter mm mean	Δ pupil diameter mm mean	Δ gaze angle x LR slope	Δ pupil diameter mm LR intercept	Δ gaze angle x LR slope
Δ pupil diameter mm IQR 2-3	Δ gaze angle y median	Δ gaze angle y LR intercept	Δ gaze angle x LR intercept	Δ gaze angle y quartile 3	eye blink intensity LR slope	Δ pupil diameter mm LR intercept
eye blink intensity median	Δ gaze angle x IQR 2-3	Δ gaze angle y quartile 3	Δ gaze angle y LR intercept	gaze angle x LR intercept	gaze angle y quartile 1	Δ gaze angle x median
Δ gaze angle x LR intercept	gaze approach time secs mean	gaze angle x IQR 1-3	Δ gaze angle x IQR 1-3	eye blink intensity median	gaze angle y IQR 1-2	Δ gaze angle y min

scenarios. When testing on video, we reached the best performance when averaging gaze features across the whole input videos and performing model-level fusion. In contrast, when testing on audio, we achieved best performance for window-based features and early fusion. This difference might be the result of the different roles gaze features play in each of these test scenarios. In the visual testing scenario, prediction is performed directly on gaze inputs. A smaller feature representation that is less subject to noise (i.e. averaged across the whole input video) might help in this scenario. On the other hand, when gaze features are part of the helper modality, a certain degree of randomness might be beneficial as it can act as a regularizer in the embedding space. This interpretation is in line with our analysis of performances achieved in each emotion class (Table 3). Here, in contrast to the overall improvement, gaze integration for video testing decreased performance for emotion classes with a low number of samples (disgust and anger). The combination of the low number of samples with the higher number of features introduced due to gaze integration might be the reason for impaired performance. In the audio testing scenario however, we did not observe a decrease in performance for classes with a low number of samples, supporting our reasoning that a larger feature space of the helper modality might have less detrimental effects.

Our class-specific performance analysis in Table 3 also revealed that both our proposed gaze-enhanced approach as well as our no gaze baseline implementation of EmoBed [26] achieved an F1 score of 0 on fear and surprise. These low performances are clearly related to the low number of training examples available for these classes. The consensus performance measurement on the OMG dataset is micro F1 [9, 26]. We choose this measurement in our evaluations for comparability with previous work. To improve performance on fear and surprise, a different evaluation metric (e.g. macro F1), and correspondingly, a different loss formulation giving more weight to those rare classes would be needed.

5.2 Limitations and Future Work

While our work showed the importance of gaze integration in flexible emotion recognition systems, it is still subject to several limitations that should be addressed in future work.

A key prerequisite for every gaze integration method are the available gaze estimates. While our system was able to clearly improve over no-gaze baselines already with the imperfect estimates provided by OpenFace [8], it potentially missed subtle gaze cues that are not well represented in these gaze estimates. More accurate gaze estimates obtained from dedicated eye tracking systems [20, 33] should be examined in future work as they have the potential to improve performance even further. Another limitation lies in the dataset we used for training and evaluation. The OMG dataset [9] consists of monologue videos collected from YouTube. While this covers a wide variety of different participants and personalities, it poses constraints on the recording situation (i.e. a single person in front of the camera). Future work should investigate how our approach can be extended to handle less constrained scenarios. One example is the recognition of emotions embedded in interactions of freely moving people [48]. This poses several additional challenges. For example, gaze estimates might not always be available due to occlusions.

While we used recent state-of-the-art gaze features for emotion recognition [50, 62], these featuresets do not represent the target of a persons' gaze. An important avenue for future work on gaze-based emotion recognition will be to investigate ways to integrate information on the gaze target. This is especially relevant in scenarios with multiple possible gaze targets, e.g. in group interactions [45]. Such an approach would require eye contact detection [25, 60] during pre-processing.

Finally, the principle of gaze integration presented is also applicable to tasks beyond emotion recognition in which gaze plays a prominent role. Examples for such tasks include emergent leadership detection [13, 44], prediction of speaking turns [11], rapport estimation [46], or personality prediction [27].

5.3 Applications

Our method can be applied in any HCI scenario where an RGB video of adequate quality to perform gaze estimation is available. OpenFace 2.0 [8] was successfully evaluated in comparison with state-of-the-art gaze estimation approaches and subsequent work has made use of it in a variety of settings, including group interactions recorded from ambient cameras [25] and interviews conducted via teleconferencing [47]. As a result, we expect our method to be applicable to a wide variety of situations in which emotion recognition is desired. Emotion recognition is key in many applications, including telemedicine [32], educational technology [6], and artificial mediators [45]. For example, the emotional responsiveness of a psychotherapy patient to her therapist can be used to estimate therapeutic alliance and potentially predict therapy outcome [39]. Gaze-enhanced cross-modal emotion embeddings are promising in this application, especially when unstable network connections result in low-quality or dropped video. In addition a large number of health-related interactions take place on the phone, without any video availability [38, 61]. This is an ideal application scenario where speech-based emotion recognition at test time can profit from gaze-enhanced cross-modal emotion embedding training. In an educational setting, emotion recognition can be valuable to access students' evaluation or learning material [15]. As students are often silent during learning or may be in a location with high ambient noise, gaze-enhanced cross-modal emotion embeddings can potentially be used to create video-only emotion recognition systems that can still profit from speech data during training.

6 CONCLUSION

In this work, we presented the first approach for gaze integration into a crossmodal training framework that allows single-modality emotion recognition at test time. We introduced a new module visual stream network to the crossmodal emotion embedding (EmoBed) framework. This visual stream network combines gaze features with a facial feature representation. With a model-level fusion approach of gaze and facial features, we outperformed the state of the art for video only testing on the popular One-Minute Gradual Emotional (OMG) Behavior dataset. Using an early fusion approach we also improved over the previous state of the art for audio-only testing on the same dataset. In ablation experiments, we showed that crossmodal training with the gaze-enhanced visual stream network leads to clear improvements over baselines without gaze integration. Taken together, our results underline the importance of gaze integration in practical emotion recognition systems.

ACKNOWLEDGMENTS

A. Bulling were funded by the European Research Council (ERC; grant agreement 801708). E. Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. P. Müller was funded by the German Ministry for Education and Research (BMBF), grant number 01IS20075. We would like to thank Dominike Thomas for her support and encouragement during the course of the project. We Thank The German University in Cairo (GUC) for the initial support to A. Abdou.

REFERENCES

[1] Reginald B Adams Jr and Robert E Kleck. 2003. Perceived gaze direction and the processing of facial displays of emotion. *Psychological science* 14, 6 (2003), 644–647.

[2] Reginald B Adams Jr and Robert E Kleck. 2005. Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion* 5, 1 (2005), 3.

[3] Ashwaq Alhargan, Neil Cooke, and Tareq Binjammaz. 2017. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 479–486.

[4] Suzan A Anwar. 2019. *Real Time Facial Expression Recognition and Eye Gaze Estimation System.* Ph. D. Dissertation. University of Arkansas at Little Rock.

[5] Claudio Aracena, Sebastián Basterrech, Václav Snáel, and Juan Velásquez. 2015. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2632–2637.

[6] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2016. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments* 24, 3 (2016), 590–605.

[7] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.

[8] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[9] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. 2018. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[10] Markus Bindemann, A Mike Burton, and Stephen RH Langton. 2008. How do eye gaze and facial expression interact? *Visual Cognition* 16, 6 (2008), 708–733.

[11] Chris Birmingham, Kalin Stefanov, and Maja J Mataric. 2021. Group-Level Focus of Visual Attention for Improved Next Speaker Prediction. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4838–4842.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[13] Francesca Capozzi, Cigdem Beyan, Antonio Pierro, Atesh Koul, Vittorio Murino, Stefano Livi, Andrew P Bayliss, Jelena Ristic, and Cristina Becchio. 2019. Tracking the leader: Gaze behavior in group interactions. *Iscience* 16 (2019), 242–249.

[14] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 375–388.

[15] Chih-Ming Chen and Hui-Ping Wang. 2011. Using emotion recognition technology to assess the effects of different multimedia materials on learning emotion and performance. *Library & Information Science Research* 33, 3 (2011), 244–255.

[16] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2014. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 508–513.

[17] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 19–26.

[18] Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology* 3, 02 (2005), 185–205.

[19] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)* 47, 3 (2015), 1–36.

[20] Stefan Dowiasch, Peter Wolf, and Frank Bremmer. 2020. Quantitative comparison of a mobile and a stationary video-based eye-tracker. *Behavior research methods* 52, 2 (2020), 667–680.

[21] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[22] Nathan J Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews* 24, 6 (2000), 581–604.

[23] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[25] Eugene Yujun Fu and Michael W Ngai. 2021. Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4873–4877.

[26] Jing Han, Zixing Zhang, Zhao Ren, and Bjoern W Schuller. 2019. Emobed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings. *IEEE Transactions on Affective Computing* (2019).

[27] Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. 2018. Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience* 12 (2018), 105.

[28] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 553–560.

[29] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, Phu Le, Vidhyasaharan Sethu, and Julien Epps. 2015. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. 41–48.

[30] Roxane J Itier and Magali Batty. 2009. Neural bases of eye and gaze processing: the core of social cognition. *Neuroscience & Biobehavioral Reviews* 33, 6 (2009), 843–863.

[31] Natasha Jaques, Cristina Conati, Jason M Harley, and Roger Azevedo. 2014. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International conference on intelligent tutoring systems*. Springer, 29–38.

[32] Athanasios Kallipolitis, Michael Galliakis, Andreas Menychtas, and Ilias Maglogiannis. 2021. Speech Based Affective Analysis of Patients Embedded in Telemedicine Platforms. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1857–1860.

[33] Moritz Kassner, William Patera, and Andreas Bulling. 2014. 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*. 1151–1160.

[34] Dacher Keltner. 1995. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of personality and social psychology* 68, 3 (1995), 441.

[35] Jung-Hwa Kim, Seung-June Choi, and Jin-Woo Jeong. 2019. Watch & Do: A smart iot interaction system with object detection and gaze estimation. *IEEE Transactions on Consumer Electronics* 65, 2 (2019), 195–204.

[36] Charles E Kimble and Donald A Olszewski. 1980. Gaze and emotional expression: The effects of message positivity-negativity and emotional intensity. *Journal of Research in Personality* 14, 1 (1980), 60–69.

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

[38] Alexandra König, Kevin Riviere, Nicklas Linz, Hali Lindsay, Julia Elbaum, Roxane Fabre, Alexandre Derreumaux, and Philippe Robert. 2021. Measuring Stress in Health Professionals Over the Phone Using Automatic Speech Analysis During the COVID-19 Pandemic: Observational Pilot Study. *Journal of Medical Internet Research* 23, 4 (2021), e24191.

[39] Sander L Koole and Wolfgang Tschacher. 2016. Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance. *Frontiers in psychology* 7 (2016), 862.

[40] Jing Liang, Yu-Qing Zou, Si-Yi Liang, Yu-Wei Wu, and Wen-Jing Yan. 2021. Emotional gaze: The effects of gaze direction on the perception of facial emotions. *Frontiers in Psychology* 12 (2021).

[41] C Neil Macrae, Bruce M Hood, Alan B Milne, Angela C Rowe, and Malia F Mason. 2002. Are you looking at me? Eye gaze and person perception. *Psychological science* 13, 5 (2002), 460–464.

[42] Maarten Milders, Jari K Hietanen, Jukka M Leppänen, and Marc Braun. 2011. Detection of emotional faces is modulated by the direction of eye gaze. *Emotion* 11, 6 (2011), 1456.

[43] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2227–2231.

[44] Philipp Müller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *2019 International Conference on Multimodal Interaction*. 274–278.

[45] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4878–4882. https://doi.org/10.1145/3474085.3479219

[46] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behavior. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*. 153–164. https://doi.org/10.1145/3172944.3172969

[47] Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating averted gaze in dyadic interactions. In *ACM Symposium on Eye Tracking Research and Applications*. 1–10.

[48] Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 663–669.

[49] Jonny O'Dwyer, Niall Murray, and Ronan Flynn. 2018. Affective computing using speech and eye gaze: a review and bimodal system proposal for continuous affect prediction. *arXiv preprint arXiv:1805.06652* (2018).

[50] Jonny O'Dwyer, Niall Murray, and Ronan Flynn. 2019. Eye-based Continuous Affect Prediction. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 137–143.

[51] Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*. 669–676.

[52] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, Article 41, 12 pages. https://doi.org/10.5244/C.29.41

[53] Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. 2021. Robust Latent Representations Via Cross-Modal Translation and Alignment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4315–4319.

[54] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (2015), 22–30.

[55] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[56] Philipp V Rouast, Marc Adam, and Raymond Chiong. 2019. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing* (2019).

[57] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, Aravind Namandi Vembu, and Rohit Prasad. 2012. Emotion recognition using acoustic and lexical features. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[58] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. 2018. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8615–8621.

[59] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. 2021. Leveraging recent advances in deep learning for audio-Visual emotion recognition. *Pattern Recognition Letters* (2021).

[60] Rémy Siegfried and Jean-Marc Odobez. 2021. Visual Focus of Attention Estimation in 3D Scene with an Arbitrary Number of Targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3153–3161.

[61] Johannes Tröger, Nicklas Linz, Alexandra König, Philippe Robert, and Jan Alexandersson. 2018. Telephone-based dementia screening I: automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 59–66.

[62] Thong Van Huynh, Hyung-Jeong Yang, Guee-Sang Lee, Soo-Hyung Kim, and In-Seop Na. 2019. Emotion recognition by integrating eye movement analysis and facial expression model. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*. 166–169.

[63] Laura Jean Wells, Steven Mark Gillespie, and Pia Rotshtein. 2016. Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PloS one* 11, 12 (2016), e0168307.

[64] Matthias J Wieser and Tobias Brosch. 2012. Faces in context: a review and systematization of contextual influences on affective face processing. *Frontiers in psychology* 3 (2012), 471.

[65] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.

[66] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. 2014. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing* 3 (2014).

[67] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 1 (2008), 39–58.

[68] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.

Received November 2021; revised January 2022; accepted April 2022