Nonverbal Communication Cue Recognition: A Pathway to More Accessible Communication

Zoya Shafique, Haiyan Wang, Yingli Tian
Department of Electrical Engineering, The City College of New York
{zshafiq001,hwang005}@citymail.cuny.edu, ytian@ccny.cuny.edu

Abstract

Nonverbal communication, such as body language, facial expressions, and hand gestures, is crucial to human communication as it conveys more information about emotions and attitudes than spoken words. However, individuals who are blind or have low-vision (BLV) may not have access to this method of communication, leading to asymmetry in conversations. Developing systems to recognize nonverbal communication cues (NVCs) for the BLV community would enhance communication and understanding for both parties. This paper focuses on developing a multimodal computer vision system to recognize and detect NVCs. To accomplish our objective, we are collecting a dataset focused on nonverbal communication cues. Here, we propose a baseline model for recognizing NVCs and present initial results on the Aff-Wild2 dataset. Our baseline model achieved an accuracy of 68% and a F1-Score of 64% on the Aff-Wild2 validation set, making it comparable with previous state of the art results. Furthermore, we discuss the various challenges associated with NVC recognition as well as the limitations of our current work.

1. Introduction

Nonverbal communication is a fundamental part of human interaction that involves relaying information through channels other than words, such as facial expressions, body language, hand gestures, etc. These nonverbal communication cues (NVCs) can convey a wealth of information about a person's emotions, attitudes, and intentions and can even contradict or emphasize spoken words. Nonverbal communication is a complex process that affects how people interact with others and plays an essential role in building social relationships, establishing trust, and conveying meaning in a variety of contexts. Understanding nonverbal communication is crucial for effective communication, especially in situations where language barriers exist or in situations where words may be unclear or misleading [29]. Although an essential part of how humans interact with each other, nonverbal communication is largely inaccessible for those who are blind or have low-vision (BLV).

According to the World Health Organization, 2.2 billion people worldwide have some form of vision impairment; 217 million people have moderate to severe vision impairment, and 36 million are blind [1]. Research studies show that the BLV community may understand other people's intentions, feelings and beliefs differently than sighted people mainly because they have limited access to the information about others' mental states during communication [32]. Such conversational asymmetry is contributed to by the inaccessibility of nonverbal communication, as studies have shown that nonverbal communication cues make up at least 55% of the emotional information conveyed during a conversation [13]. As such, software which can accurately classify nonverbal communication cues is a critical step towards building accessible NVC recognition systems.

NVC recognition is a challenging task due to the large variety of nonverbal cues with subtle differences. A nonverbal cue can be a head nod, shrugging shoulders, or arms folded across one's chest. In different contexts, each of the aforementioned cues can convey different emotions. Despite the large variety of nonverbal communication cues, they can be broken down into three basic tasks: facial expression recognition (FER), body pose estimation, and hand gesture recognition. Work combining all three tasks is relatively sparse compared to studies focusing on FER for emotion recognition. Additionally, FER datasets [15, 26–28] are limited to basic emotions and neglect more common NVCs such as if the conversation partner is thinking, confused, or agreeing with something that was said.

The challenges of accurately recognizing NVCs for emotion classification include: (i) large variation in temporal duration of actions, (ii) large intraclass variance, and (iii) accumulating a uniform distribution of emotions conveyed by NVCs. Specifically, nonverbal communication cues can range from extremely short actions to very long actions. Recent studies in action detection have shown that classifying and localizing very short actions in videos is a challeng-

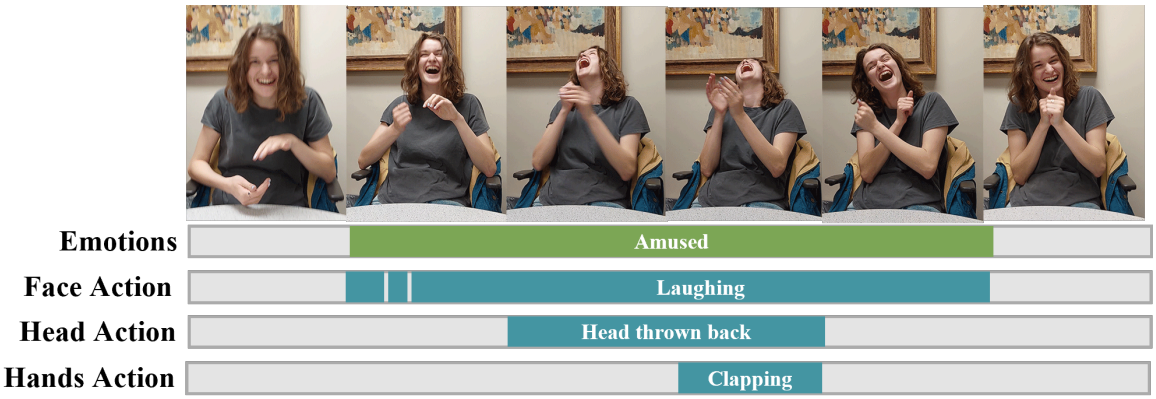

Figure 1. Our CCNY NVC Dataset contains two levels of annotations for ten different emotion classes along with temporal annotations of the start and end time for all actions. The high-level label represents the emotion conveyed by nonverbal cues whereas the fine-grained label represents the action cues themselves. The fine-grained actions are labeled for three modalities: facial expressions, head movement, and hand movements.

Table 1. Distribution of labels in common FER datasets as a percentage of the total annotations where the emotions are abbreviated as follows: Neutral: Ntrl, Anger: Agr, Disgust: Dgst, Happy Hpy, Surprised: Sprs. Instances from classes such as disgust, fear, and sadness are relatively sparse compared to classes. Note that, CK+ has another class that is not listed here: *contempt*, which is approximately 5% of the total dataset [26].

Dataset	Ntrl	Agr	Dgst	Fear	Hpy	Sad	Sprs.
Aff-Wild2 [15]	41	4	2	2	30	13	8
CK+ [26]	–	14	18	8	21	9	25

ing task for current methods [38]. Therefore, current methods are limited for NVC recognition, considering that some NVCs, such as nodding, can last for less than ten seconds and occur very frequently in conversations. In addition, there is also a large intra-class variance as people tend to express nonverbal communication cues differently depending on the situation [13, 29]. For instance, tilting one's head to the side, looking to the side without tilting one's head, scratching one's chin and pursing one's lips can all be seen as indicators of someone who is thinking. Lastly, ensuring a uniform distribution of NVCs in a dataset presents a challenge as some NVCs occur less frequently than others. NVCs which indicate anger and sadness are not as common in casual conversations as NVCs expressing thought, agreement, or amusement. Many commonly used FER datasets, two of which are shown in Table 1, have a great class imbalance. Such imbalance can pose a problem not only for network training but may also negatively impact the robustness of a NVC recognition aid in real-world scenarios.

To combat the above-mentioned challenges, we first collect a NVC dataset, which we name the CCNY NVC Dataset, by conducting casual interviews and capturing videos with a wide range of NVCs. During the video recording, participants were shown videos and asked riddles and various questions to elicit responses (e.g. amusement, thought, confusion, and sadness, among others). A subset of the questions is shown in Table 2 The collected videos are labeled at two levels: a coarse emotion category related to the NVCs and a fine-action category consisting of the NVCs themselves. The fine-action category is labeled for multiple modalities, as shown in Figure 1. We individually label fine-grained NVCs to study whether learning fine-grained NVCs directly can aid deep neural networks in better extrapolating emotions. One example of such a framework is a network which takes in as input the NVC cues and predicts the high level emotion. We also label multiple modalities as NVC cues can take the form of facial expressions, hand movements, or body gestures. An ideal model for NVC cue recognition would take into account the multiple modalities, yet very few multimodal emotion datasets currently exist.

To fill this gap, we are both building the CCNY NVC Dataset and constructing a multimodal baseline for NVC cue and emotion recognition. We choose a 3D-ResNet [8] as our baseline as they can easily be stacked together to capture multimodal information as shown by Vahdani et al. [10, 34]. To showcase the potential of our baseline, we present initial results for the seven basic expressions classification task using the Aff-Wild2 dataset [14–22, 35], which focuses on facial expression recognition as a means of emotion recognition. Our main contributions can be described as

- A study of the challenges present in NVC recognition as compared to FER recognition.
- An analysis of current FER datasets and their limitations for NVC recognition.

- Design of a multimodal baseline model for NVC recognition.

- Demonstrating comparable performance of our baseline with current state-of-the-art methods on the Aff-Wild2 dataset [14–22, 35] expression classification task.

The rest of this paper is structured as follows: Section 2 discusses previous attempts at building an assistive aid to recognize NVCs for the BLV community and the scope of multimodal models for NVC recognition. Section 3 describes current datasets which have a potential for NVC recognition and discusses their limitations to highlight how the CCNY NVC Dataset fills a gap in the current domain. Section 4 introduces the baseline model for NVC recognition. Section 5 discusses the experimental results for a branch of the proposed baseline on the Aff-Wild2 dataset.

2. Related Work

Facial Expression Recognition as a Means of NVC Recognition Previous studies [2, 3, 23, 33] have focused on facial expression recognition as a means of NVC recognition. Real-time systems using FER have been developed to aid those who are BLV perceive NVCs in video-calls [33] and in casual conversations [2, 23]. Specifically, Shi et al. [33] developed an accessible video calling prototype that detects visual conversation cues in a video call (i.e., attention, agreement, disagreement, happiness, thinking, and surprise) and uses audio cues to convey them to a user who is blind or low vision. However, this system is impractical for other scenarios where NVC recognition is needed (e.g., a blind user in a meeting room). Furthermore, some users found the audio feedback to be distracting from the conversation. With their Expression system, Anam et al. [2] used Google Glass to record videos of the conversation partner which were then sent to a server which detects facial features to classify the NVC and relays the information to the user through speech feedback. As with [33], the speech feedback of Expression can be distracting and obtrusive in a conversation. Additionally, having a remote server for classification is not scalable in real-world scenarios. VibroGlove [23] is proposed as another assistive technology which relays facial expression information to users through vibrations from sensors mounted on a glove where each emotion is correlated with a specific vibration pattern. Although less obtrusive than audio feedback, this method is not scalable as adding common NVC classes such as paying attention, thinking, or confused to the seven basic emotions tested in [23] would increase the number of vibration patterns and may prove confusing for the user. The usability study for VibroGlove is also limited as studies were carried out with only one participant who was blind. Therefore, this is a large gap in the field of assistive devices for nonverbal cue recognition.

Multimodal Action Recognition NVC recognition is a multimodal task as a NVC can consist of a facial expression, a head movement, body posture, hand movements or some combination. Most previous research in the NVC domain have not used these modalities and have relied solely on FER recognition or micro-gesture recognition [25]. However, other modalities such as spoken words, speech signals, heart-rate and other physiological signals [30, 36] have been used in various FER recognition tasks such as action unit (AU) detection and expression recognition. A drawback to such modalities in the NVC domain is that spoken words and speech patterns may communicate different information than NVC cues as there are many instances where verbal and nonverbal communication cues may contradict each other. Furthermore, for the BLV community, audio cues are readily accessible and may not contribute as much to conversational symmetry as augmenting the conversation with NVCs.

Multimodal studies using body poses, facial expressions, and hand gestures can be found in the sign language recognition domain as many signs, for example in American sign language (ASL), are composed of several body movements in addition to the hand gestures. Previous studies in this domain [10, 34] have used an ensemble of 3D residual networks, with each network corresponding to a different modality from hand gestures, facial expressions, and head movements. The results from each network are fused together and post-processed using a majority voting algorithm to determine the final action class. This structure is highly adaptable for NVC recognition as it allows for the seamless fusion of multiple modalities.

3. NVC Datasets

3.1. Existing Datasets for Emotion Recognition

There are many datasets for FER, gesture and emotion recognition. Here we briefly summarize several commonly used datasets. The iMiGUE dataset [25] consists of 359 videos of press conferences with athletes participating in the Grand Slam tournament after a match. It is the first dataset of its kind, with labels for micro-gestures, which the authors define as subconscious actions which reveal underlying emotions. In other words, the iMiGUE dataset is analogous to nonverbal cues for the head, hands, and body modalities. Although iMiGUE is a spontaneous emotion recognition dataset based on micro-gestures, the emotion classification is binary; emotions are labeled as positive and negative. For NVC recognition, this binary labelling is too coarse. Many other emotional gesture datasets [6, 12, 31] are made from posed actions and therefore are limited for spon-

Table 2. A subset of questions asked during nonverbal communication cue capturing.

Question	Intended Emotion
What time is it when an elephant sits on a fence?	Amused, Confused, Thinking
Can you tell me about what you do?	Neutral
When was the last time you were really frustrated?	Agitated, Upset
When was the last time you laughed so hard your stomach hurt?	Happy
The more you take, the more you leave behind. What are they?	Confusion, Thinking
Interviewer gave random compliments	Happy
How has your day been?	Neutral

(a) (b)

Figure 2. (a) Two examples from the surprised class from the CCNY NVC Dataset. Even though both examples are of the same person, the surprised emotion is expressed differently in both instances. (b) Two instances of the thinking class from the CCNY NVC Dataset. Looking to the side and rubbing ones face are common nonverbal cues used to express thought [29]. Because there are a large variety of nonverbal cues for each emotion, there is a large intra-class variance when relating NVCs to emotional states.

taneous NVC recognition. FABO [7] is perhaps the closest approximation for a multimodal dataset for NVC recognition as it is a bimodal dataset with annotations for facial and body gestures. This dataset focuses on basic emotions such as uncertainty, anger, surprise, fear, anxiety, happiness, disgust, boredom, and sadness. Although FABO [7] has a more extensive list of emotions as compared to other datasets, it is a posed dataset and does not include annotations for very common nonverbal communication cues such as thinking and agreement.

On the hand, there are many FER datasets [26] which feature labelling for seven basic expressions in both spontaneous and posed environments. The Aff-Wild2 dataset [14–22, 35] is a large-scale dataset containing 548 videos labelled for the recognition of seven basic expressions, with an additional other category. Specifically, the emotions are neutral, anger, disgust, fear, happiness, sadness, and surprise.

Such a large scale dataset could be useful for NVC recognition as the the large number of instances (almost 2 million annotated frames) would help learn a wide variety of cues used to relate emotions. Furthermore, the Aff-Wild2 dataset [14–22, 35] contains many real-life situations, such as when there is a large glare due to sunlight through a window and low-lighting conditions, which could help make a NVC recognition aid robust under different conditions. However, the emotions for which the Aff-Wild2 dataset is annotated are limited and more common NVCs such as thinking, confusion, and paying attention, are not included. As a FER dataset for emotion recognition (among other tasks), Aff-Wild2 [14–22, 35] is a large step forward but still falls short of the requirements for a NVC recognition dataset. Other FER datasets, such as CK+ [26], DIFSA/DIFSA+ [27, 28], and BP4D+ [5], have similar emotions as Aff-Wild2 or focus on action unit detection instead of emotion recognition and therefore also have limited applicability for NVC recognition. Additionally, these datasets contain posed expressions which cannot directly be used to recognize a variety of spontaneous nonverbal cues. Currently there is no available dataset which combines common nonverbal communication with the basic emotions found in emotion recognition datasets, whether they be for FER or gesture recognition. Furthermore, there is no multimodal dataset for emotion recognition that combines facial expressions with hand gestures and body pose.

Such datasets do however exist for ASL recognition. One such dataset is the ASL-HW-RGBD dataset [34], which is proposed for the task of error recognition in ASL gesturing. As an ASL gesture consists of both a manual sign, or the actual hand sign, and a nonmanual sign (head movements or facial expressions), the authors propose to recognize errors in gesturing by matching the manual and nonmanual signs. As such, there are multiple levels of an-

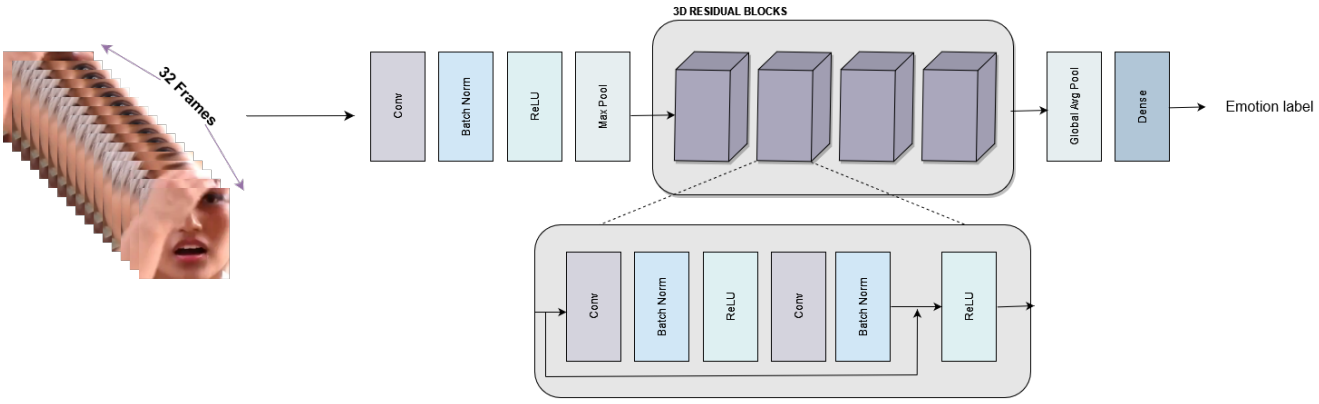

Figure 3. An overview of our baseline for NVC recognition. OpenPose [4] is employed to crop faces out from all frames in the input videos into 134 x 134 pixel images. After preprocessing, the raw RGB frames are input into a 3D-ResNet with 34 layers and 5 convolutional blocks in groups of 32 frames. Final predictions are made by processing the 3D-ResNet output with a fully connected layer. Our network is adapted from [8, 10, 34]. For the Aff-Wild2 dataset, we use the provided cropped and aligned images for training.

notation, with each level relating to a different aspect of ASL grammar. Facial expressions are annotated separately for ASL grammar. For example, asking questions in ASL is associated with a set of facial expressions and head movements that complete the gesture. With these multilevel annotations, Vahdani et al. train three separate networks for head movement recognition, facial expression recognition, and hand gesture recognition, respectively, and combine the output of each network to recognize errors in signing ASL gestures. Such dataset and network architectures can be extended to the NVC recognition domain for multimodal detection of cues.

3.2. CCNY NVC Dataset

Dataset Design and Annotation: The CCNY Nonverbal Cue (NVC) Dataset is a human emotion detection dataset, featuring 128 videos with multilevel class annotations and temporal boundary annotations. To the best of our knowledge, this dataset is the first of its kind, with videos of casual conversations from the first-person point of view. Although various facial expression recognition datasets exist, they do not provide temporal boundaries as our NVC dataset does nor do they annotate common nonverbal cues and emotions such as thinking, paying attention and confused.

In the CCNY-NVC Dataset, NVCs are labeled at two levels. The first level classifies the high-level emotion represented by the NVC while the second level labels the fine-grained action. For example, an instance of the speaker nodding in a video is labeled as nodding at an action level and also as "agreement/understanding" at a higher level. The high-level semantic NVC labels feature 10 categories: agreement/understanding, amused, happy, confused, thinking, upset, disagreeing, dislike, exasperated, and surprised. Our fine-grained action labels are further

Table 3. The ten classes in the current CCNY-NVC dataset.

NVC Classes	
Thinking	Amused
Agreement/Understanding	Confused
Surprised	Upset
Happy	Exasperated
Dislike	Disagreement

divided into multiple categories. We provide annotations for facial expressions, gaze, head movements, and hand gestures. We take multiple modalities into consideration as a NVC can consist of more than just facial expressions. As such, multimodal annotations can be used to accurately represent the components of complex nonverbal communication cues. Table 3 lists the classes currently available in the CCNY-NVC dataset. The dataset will be extended to include more NVCs from more scenarios including group conversations.

Collection Methodology: To capture a wide range of NVCs, participants were asked an initial question to start the conversation and the conversations were allowed to progress naturally. At some points in the conversations, participants were randomly asked riddles and shown videos in order to capture uncommon NVCs such as anger or sadness. All video clips of NVCs were captured on a Samsung Galaxy Tab S7 FE 12.4" to test the portability of the device for real-world deployment. Consent for the release of media was obtained from all participants.

Key Challenges: Capturing a balanced NVC dataset (i.e. with a uniform distribution of NVCs) presents a great challenge as NVC classes such as anger and sadness are not as common as thinking, agreement/understanding, or happiness in casual conversations. Furthermore, many NVCs,

such as nodding, have a very short temporal duration (< 10 seconds) leading to noisy temporal annotations. Many NVC classes also have a large intra-class variance as shown in Figure 2. In Figure 2(a), the top can be interpreted as genuine surprise whereas, based on the facial expression, the bottom image is a mixture of disbelief and surprise. Subtle differences between the two instances change the meaning of the emotion in context even though the same emotion is being represented in both cases.

4. NVC Recognition Baseline Network

As shown in Figure 3, we propose a 3D-ResNet as the baseline [8, 10, 34] for NVC recognition due to its ability to effectively model spatio-temporal features in videos in a straight-forward manner. Rather than classifying individual frames, we believe using both spatial and temporal features will help enhance NVC learning as the meaning of nonverbal cues depends heavily on context. The proposed network has a total of 34 layers over five convolutional blocks, four of which are 3D residual network connections. The first blocks in the network consists of a convolutional layer with 64 kernels, batch normalization, ReLU activation, and max-pooling layers. This block is followed by four 3D-ResNet blocks, with 64, 128, 256, and 512 kernels, respectively. After the last residual block, the output of the network is processed by global average pooling and dense layers to produce the final prediction. As input, the network takes in groups of 32 frames to represent a clip from the video.

After testing the feasibility of our design on the facial expression recognition task, we aim to extend the baseline into a multimodal network as shown in Figure 4, where each branch of the network aims to predict emotional states based on one of the following modalities: facial expression, hand gestures, head/body pose. We propose to crop out the face and hands from video frames as input to the face and hand networks, respectively. For the head/body recognition network, we aim to input skeleton keypoints obtained from OpenPose [4] along with the RGB frame. The outputs of the networks would be fused together to make the final prediction.

5. Experiments

5.1. Datasets and Settings

Datasets: As the CCNY NVC Dataset is in its early stages of production, it is of a small scale and suffers from a significant class imbalance. Therefore, for our preliminary tests, we use the Aff-Wild2 dataset [14–22, 35]. Presented in the Affective Behavior in-the-wild (ABAW2) competition held alongside ICCV 2021, the dataset contains 548 videos scraped from YouTube with approximately 2.8 million annotated frames. Each frame is annotated and videos range from 0.04 to 26.22 minutes. This dataset is split into

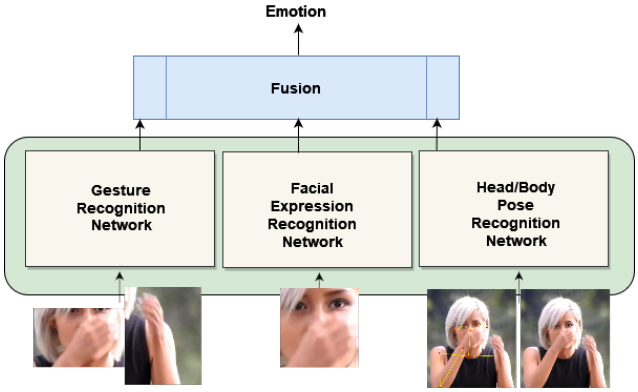

Figure 4. Our proposed pipeline for NVC recognition. We propose that a multimodal network would be able to best model nonverbal cues as they are complex actions involving different parts of they body.

three tasks: Valence-Arousal Estimation (VAE), Seven Basic Expression Classification (EXPR) and Twelve Action Unit Detection (AU). For our purposes, we use the EXPR split of the dataset, which contains 248 videos for training, 70 videos for validation and 228 videos for testing. As the dataset originates from a competition, we only had the training and validation data available. These videos are labeled for recognition of seven basic emotions: neutral, anger, disgust, fear, happiness, sadness, and surprise. Most importantly, these videos are of spontaneous behavior in the wild, making the Aff-Wild2 dataset [14–22, 35] a close approximation for naturally occurring NVCs. Furthermore, as mentioned in Section 3, the Aff-Wild2 dataset [14–22, 35] contains videos featuring many real-life scenarios such as low lighting, sun glare, and a shaky camera frame. A variety of real-world scenarios is necessary for training a robust NVC recognition aid for the BLV community.

Implementation Details To prepare data for training, faces from the videos must be cropped out and aligned. For our preliminary experiments, we use the cropped and aligned faces provided in the Aff-Wild2 dataset. Missing frames were interpolated using neighboring frames. The original Aff-Wild2 annotations were restructured to represent instances of each emotion. The original annotations provided labels for each frame in the video. To input into our 3D-ResNet and adapt the dataset to the video action recognition domain, we concatenated consecutive labels belonging to the same emotion to create action level annotations. In other words, if frames 31 to 64 were labeled as 'Neutral', a new annotation was made with the starting frame, ending frame, and expression category as "Neutral/31/64" to replace the original frame-level annotations. These clips were passed into a 3D-ResNet with 34 layers.

We trained our network with an initial learning rate of 0.001 and a batch size of 128. To combat over fitting and

Table 4. Results for our baseline method on the official validation set of the Aff-Wild2 Expression Classification task. Our method shows promise as it is comparable to previous state of the art results. [37] reported the best results on the validation set out of all competing teams and came in first on the test set. [9] placed second on the official test set.

Method	F1 Score	Accuracy	ABAW2 Metric
Ours	64.3	68.2	65.6
Netease Fuxi Virtual Human [37]	75.7	85.6	79
CPIC-DIR2021 [9]	40.2	63	47.7
Aff-Wild2 Baseline [22]	30	50	36.6

class imbalance, we implemented weight decay as regularization and used focal loss [24] and stochastic gradient descent for optimization. We also implemented a weighted sampling to combat the class imbalance alongside focal loss [24] and ensure that the network saw a more even distribution of classes in each batch. To aid our training, we used pretrained weights for the ResNet-34 from the Kinetics [11] dataset. Lastly, we used data augmentation techniques such as random cropping, random horizontal flip, and random rotation.

5.2. Evaluation Metrics

To measure the performance of our model, we report the F_1 score and total accuracy. The F_1 score for emotion and NVC recognition is computed on a per-frame basis, i.e., were all frames classified correctly for a given emotion class. We also report the total accuracy as the ratio between the number of correct predictions and the number of total predictions. To ensure a fair comparison, we calculate a weighted average between the F1 score and the total accuracy, which was the main evaluation criterion for the ABAW2 competition [16]. The exact formulation is presented in Equation (1).

$$\epsilon_{total} = 0.67 \times F_1 + 0.33 \times TAcc \tag{1}$$

5.3. Results

As shown in Table 4, we achieved a F_1 score of 64.3 and an accuracy of 68.2 on the Aff-Wild2 dataset. Our results are comparable to [9, 37], which placed second and first on the official test set, respectively. As the test set was not available to us, we report our performance on the official validation set and compare with the performance reported in [9, 37] on the official validation set. Both these methods used either additional datasets, pseudo-labelling techniques or prior architectures on top of which their models were built, however, we achieve comparable results with no additional data, pseudo-labelling, or prior models. As such, our method shows great promise for emotion recognition.

6. Conclusion

Although many emotion recognition datasets and models exist, they are limited in their applicability to nonverbal cue communication, which is an essential part of how we communicate with each other. Such nonverbal communication however is largely inaccessible to those in the blind or low vision community, leading to conversational imbalance between speakers. Furthermore, current accessibility aids fall short in terms of ease of use and scalability. To combat these issues and work towards a practical model for nonverbal cue recognition, we are building the CCNY NVC Dataset. Such a task is nontrivial due to large class imbalances and noisy labels. The CCNY NVC Dataset is a multimodal dataset with both emotion annotations and fine-grained nonverbal cue annotations. We also propose a multimodal baseline for the NVC cue recognition task. As our dataset is still in production, we present preliminary results on the AFF-Wild2 dataset; our results show promise for our proposed baseline method. In future work, we will aim to refine our CCNY NVC Dataset and apply our baseline model to detect nonverbal cues and emotions. We will further develop more advanced methods to recognize NVCs in long, untrimmed videos by incorporating other modalities and temporal localization.

References

[1] Blindness and vision impairment. https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment. 1

[2] Asm Iftekhar Anam, Shahinur Alam, and Mohammed Yeasin. Expression: A dyadic conversation aid using google glass for people with visual impairments. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pages 211–214, 2014. 3

[3] Douglas Astler, Harrison Chau, Kailin Hsu, Alvin Hua, Andrew Kannan, Lydia Lei, Melissa Nathanson, Esmaeel Paryavi, Michelle Rosen, Hayato Unno, et al. Increased accessibility to nonverbal communication through facial and expression recognition technologies for blind/visually impaired subjects. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 259–260, 2011. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5, 6

[5] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 4

[6] Donald Glowinski, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus Scherer. Technique for automatic emotion recognition by body gesture analysis. In *2008 IEEE Computer society conference on computer vision and pattern recognition workshops*, pages 1–6. IEEE, 2008. 3

[7] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International conference on pattern recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE, 2006. 4

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 2, 5, 6

[9] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition. *arXiv preprint arXiv:2107.04187*, 2021. 7

[10] Longlong Jing, Elahe Vahdani, Matt Huenerfauth, and Yingli Tian. Recognizing american sign language manual signs from rgb-d videos. *arXiv preprint arXiv:1906.02851*, 2019. 2, 3, 5, 6

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[12] Michael Kipp and Jean-Claude Martin. Gesture and emotion: Can basic gestural form features discriminate emotions? In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–8. IEEE, 2009. 3

[13] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013. 1, 2

[14] Dimitrios Kollias. Abaw: Learning from synthetic data & multi-task learning challenges. *arXiv preprint arXiv:2207.01138*, 2022. 2, 3, 4, 6

[15] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 1, 2, 3, 4, 6

[16] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 2, 3, 4, 6, 7

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2, 3, 4, 6

[18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2, 3, 4, 6

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 2, 3, 4, 6

[20] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2, 3, 4, 6

[21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2, 3, 4, 6

[22] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 2, 3, 4, 6, 7

[23] Sreekar Krishna, Shantanu Bala, Troy McDaniel, Stephen McGuire, and Sethuraman Panchanathan. Vibroglove: an assistive technology aid for conveying facial expressions. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3637–3642. 2010. 3

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[25] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10631–10642, 2021. 3

[26] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 1, 2, 4

[27] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016. 1, 4

[28] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 1, 4

[29] Joe Navarro. *The dictionary of body language: a field guide to human behavior*. HarperCollins, 2018. 1, 2, 4

[30] Yassine Ouzar, Frédéric Bousefsaf, Djamaleddine Djeldjli, and Choubeila Maaoui. Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2460–2469, 2022. 3

[31] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 3

[32] Jolanta Sak-Wernicka. Exploring theory of mind use in blind adults during natural communication. *Journal of psycholinguistic research*, 45:857–869, 2016. 1

[33] Lei Shi, Brianna J Tomlinson, John Tang, Edward Cutrell, Daniel McDuff, Gina Venolia, Paul Johns, and Kael Rowan.

Accessible video calling: Enabling nonvisual perception of visual conversation cues. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–22, 2019. 3

[34] Elahe Vahdani, Longlong Jing, Yingli Tian, and Matt Huenerfauth. Recognizing american sign language nonmanual signal grammar errors in continuous videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1–8. IEEE, 2021. 2, 3, 4, 5, 6

[35] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 2, 3, 4, 6

[36] Ren Zhang, Ning He, Shengjie Liu, Ying Wu, Kang Yan, Yuzhe He, and Ke Lu. Your heart rate betrays you: multimodal learning with spatio-temporal fusion networks for micro-expression recognition. *International Journal of Multimedia Information Retrieval*, 11(4):553–566, 2022. 3

[37] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 7

[38] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021. 2