Towards Visual Behavior Detection in Human-Machine Conversations

1st Oliver Roesler
Modality.AI Inc.
San Francisco, CA, USA
oliver.roesler@modality.ai

2nd David Suendermann-Oeft
Modality.AI Inc.
San Francisco, CA, USA
david.suendermann-oeft@modality.ai

***Abstract*—In this paper, we investigate how multiple conversational behaviors can be detected by automatically analyzing facial expressions in video recordings of users talking to a dialog system. To this end, we recorded a video corpus of human-machine interactions containing distances between facial landmarks as well as a manually annotated behavior labels for each recorded video frame. We evaluated the difficulty of defining unambiguous conversational behaviors and used a deep neural network to predict conversational behaviors on a frame-by-frame basis that, after extracting facial landmarks of detected persons, produced an F1-score of up to 0.86.**

***Index Terms*—visual conversational behavior detection, human-machine interaction, facial landmarks, deep neural networks**

I. Introduction

For many applications, facial expression detection can be of substantial benefit. For instance, facial expressions can be used to detect student engagement and emotions in game-based learning environments and during interactions with intelligent tutoring systems to adjust the environment appropriately to support learning [15]. In general, positive emotions have been shown to benefit learning [14], but also confusion can play a supporting role [6], while anger or anxiety are usually harmful for learning [14]. Detecting emotions is also beneficial in dialog systems to allow the system to react accordingly, i.e. to adjust the dialog flow to respond in real time to emotional conditions [3]. For example, a system might provide additional information which rebuts a wrong user assumption that caused a negative emotion, or it might show sympathy. In cases of positive emotions it might encourage the user to express its emotional state or provide feedback [3]. However, dialog systems are usually unimodal, i.e. they use only speech communicating to the user and no visual information. Thus, previous studies that investigated behavior detection in dialogs did neither consider facial expressions to recognize emotions nor did they consider other important non-emotional behaviors that occur during natural conversations and are used by humans to influence the the flow of a conversation. Example non-emotional behaviors are *distraction* and *thoughtfulness* both of which are good candidates for a dialog system to take real-time action. For instance, if someone is distracted, the system might repeat the previous sentence to bring the attention back to the conversation or explicitly comment on it, e.g. "please pay attention to our conversation.". In contrast, if a person appears thoughtful, the agent might give the user some more time to find an appropriate answer or might offer some additional help, if it seems to be likely that the user is hesitating due to missing information.

The literature shows several studies that investigate emotion recognition from facial expressions during human-robot interactions because robots are often equipped with cameras. For example, Liu et al. [12] detected emotions from facial expressions, while Alonso-Martín et al. [1] used also voice, i.e. they used two modalities. In contrast to these previous works, the present study focuses on detecting conversational behaviors, which include also non-emotional behaviors, such as *distraction* and *confusion*. The distinction between emotional and non-emotional behaviors is not always clear. Three of the six behaviors used in this study, i.e. *smiling*, *annoyed*, and *approving*, might also be considered emotional behaviors. For example, *smiling* might be seen as an indication of happiness. In the end, whether a conversational behavior, i.e. a behavior that suggests a change in the dialog flow, represents an emotion, is not important for the purpose of this study. *To the best of our knowledge*, conversational behavior detection has not been investigated before, which means that no baselines, corpora, or even guidelines of important conversational behaviors with a description of their characteristics are available. The latter makes it especially difficult to employ several annotators to create a large corpus to sufficiently train a classification algorithm because of the high likelihood of significant inter-annotator disagreement.

The rest of this paper is structured as follows: Section (II) describes the data collection, corpus, and annotator agreement. Classification results are described in Section (III). Finally, Section (IV) concludes the paper.

II. Materials and Methods

A. Data collection

The video data was collected from 19 participants via Amazon Mechanical Turk[1] [7]. Each video differed in the

Further contributions from Franziska Haller, Michael Neumann, Mike Merrill, David Pautler, Amanda Robinson, Andrew Cornish, Doug Habberstad, Jackson Liscombe, and Renko Geffarth from Modality.AI Inc., San Francisco, CA, USA.

[1]https://www.mturk.com/

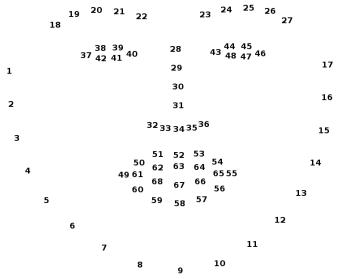

Fig. 1: Illustration of the 68 used facial landmarks.

distance of the participant to the camera, the background, as well as the light conditions. Most of the participants were sitting in front of the camera, while some were standing and one of them was lying. During the video the participants had a conversation with a conversational agent that was asking them different question, such as "What have you been doing for pleasure?" or "What has gotten you down lately?", as well as commenting on the provided answer in a general manner. The system is a multimodal cloud-based dialog system that is delivered through a web browser. Overall, 46 minutes and 19 seconds of conversational video were obtained.

All videos were manually analyzed and annotated by three different annotators. Except for a list of the six possible conversational behaviors {SMILING, DISTRACTED, THOUGHTFUL, ANNOYED, APPROVING, and CONFUSED}, the only additional instruction provided was that it is possible that not all behaviors are present and that they are not equally distributed. Thus, the annotators had to decide the specific characteristics of the behaviors themselves, which led to significant inter-rater disagreement, as described in Section (II-C).

B. Corpus

Each instance in the corpus consists of all 2,278 distances between all 68 facial landmarks according to MultiPIE[8] (Figure 1) and a behavior label (by definition, in the corpus only one behavior is present at a time). Facial landmark detection consists of two main steps: (1) Face detection, and (2) Detection of facial landmarks. The face detection algorithm takes an image as input and outputs the (x, y)-coordinates of faces in it. This study employs the Dlib[2] face detector, which uses 5 histograms of oriented gradients for face detection [5]. The employed method is fast and works well for frontal and slightly non-frontal faces, while it does not work for faces smaller than 80x80, side faces and faces that are looking up or down[3]. For facial landmark detection the Dlib facial landmark detector was employed, which uses an ensemble of regression trees proposed by Kazemi and Sullivan [10].

C. Inter-annotator agreement

In order to determine the range of valid interpretations and to validate annotation procedures, we calculated the inter-annotator agreement [2] for the corpus. To ensure that annotators were not influenced in their freedom of interpretation, they were only provided with minimal instructions, thereby ensuring that they would come up with their own sets of characteristics for all possible behaviors.

[2]http://dlib.net/

[3]In future work, we will use a Single Shot MultiBox[11] based face detector to detect also non-frontal faces across various scales.

TABLE I: Inter-annotator agreements between all annotators for all labels and only behavior labels, i.e. excluding neutral.

		A and B	A and C	B and C
Percent Agreement	**All**	24.94%	68.15%	30.22%
	Behavior	57.57%	89.16%	89.93%
Cohen's Kappa	**All**	0.11	0.28	0.09
	Behavior	0.41	0.85	0.81

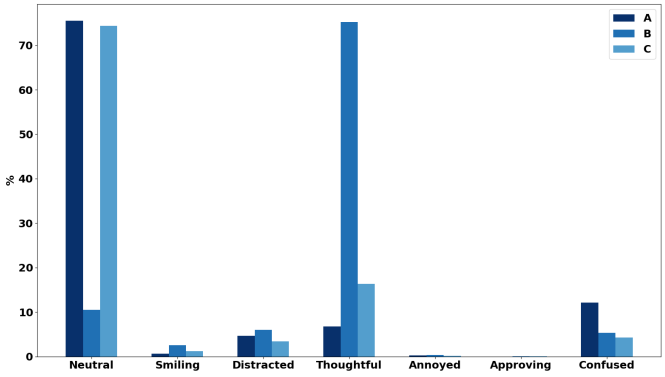

Fig. 2: Class distributions for all three datasets.

Inter-annotator or inter-rater agreements have been calculated using percent agreement and Cohens Kappa [4]. Table (I) shows the inter-annotator agreements between all three annotators considering only behavior labels, i.e. excluding *neutral*. Overall, the agreements are rather low, when considering all labels, while they are quite high, when only considering behavior labels. The results show a major difference between annotator **B** and the other two annotators because **B** labeled most instances as *thoughtful*, while the other two, i.e. **A** and **C**, labeled most instances as *neutral*, as shown in Figure (2). This clearly illustrates that one conversational behavior can be interpreted quite differently by different people making it hard to define a general annotation scheme, which again complicates the creation of larger corpora. More precise rubrics will have to be defined in the future to overcome this limitation.

D. Deep Neural Network

A deep neural network with ten fully connected layers, each with 500 nodes, was used for classification. The network had 2,278 input nodes, representing the distances between all facial landmark coordinates, and 6 output nodes, i.e. one for each behavior class. As activation functions rectified linear units were used for all layers [9, 13].

III. Results and Discussion

The network (Section III) was trained for 10 epochs with a learning rate of 1e-7, and a batch size of 32. Ten-fold cross-validation was carried out for nine different data set combinations {AA, AB, AC, BA, BB, BC, CA, CB, CC}. The

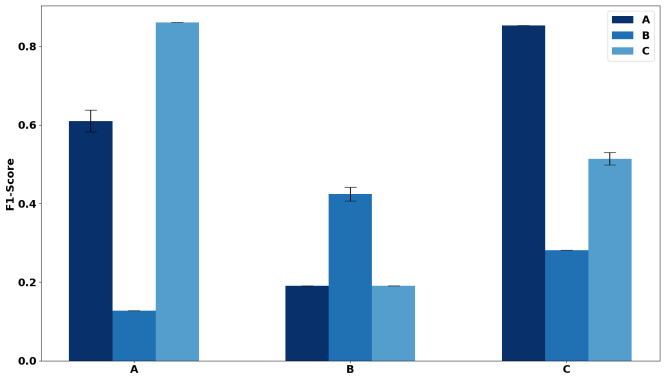

Fig. 3: Inter- and intra-dataset classification results. The training dataset is indicated by the x-axis labels, while the test dataset is represented by the colors of the bars.

best classification performance, with an F1-Score of 0.86, was achieved when the classifier was trained with data set **A** and tested with **C** or trained with **C** and tested with **A**. This is followed by the intra-dataset classifications, which achieved F1-scores between 0.4 and 0.6, while the worst performance was achieved when training or testing with dataset **B**.

These results can be explained by the obtained inter-annotator agreement (Table I). The Cohen's Kappa coefficient for **A** and **C** indicates fair agreement, while the coefficients for **B** are three times as low indicating only slight agreement. Therefore, the inter-annotator agreement shows a strong correlation with the inter-data-set classification results[4].

IV. Conclusions and Future Work

We investigated the use of a deep neural network to recognize non-emotional behaviors in human conversations with a dialog system. Furthermore, we evaluated the difficulty of defining unambigous non-emotional behaviors through the evaluation of inter-annotator agreement.

The classification performance of the employed network was significantly dependent on the used annotation set. When different annotations were used for the training and test set, the performance correlated with the inter-annotator agreement, i.e. higher agreement led to better classification results. The inter-annotator agreement was in general quite low, clearly illustrating that there are no clear and intuitive characteristics for all the behaviors considered in this study.

In future work, we will increase the number of used conversational videos and annotators to enhance generalizability and accuracy of the classifier. Furthermore, we will conduct a user study to determine whether there is a set of non-emotional behaviors with characteristics most people agree on. Finally, we will investigate to also use audio and textual information since previous research has shown that multi-modal input can significantly increase classification performance.

[4]In this case, we consider only the inter-annotator agreement for all labels because the classifier was trained and tested on all labels, i.e. including *neutral*.

V. Acknowledgments

Our gratitude goes to all involved participants of the study.

References

[1] F. Alonso-Martín, M. Malfaz, J. Sequeira, J. F. Gorostiza, and M. A. Salichs. A multimodal emotion detection system during humanrobot interaction. *Sensors*, 13: 15549–15581, November 2013.

[2] R. Artstein. Inter-annotator agreement. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer, Dordrecht, 2017.

[3] F. Burkhardt, K. P. Engelbrecht, M. van Ballegooy, T. Polzehl, and J. Stegmann. Emotion detection in dialog systems - usecases, strategies and challenges. In *3rd International Conference on Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, Netherlands, December 2009.

[4] J. Cohen. A coefcient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, USA, June 2005.

[6] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, February 2014.

[7] M. Eskenazi, G. Levow, H. Meng, G. Parent, and D. Suendermann. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. Wiley, Hoboken, USA, 2013.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, volume 28, pages 807–813, 2010.

[9] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung. Digital selection and analogueamplication coexist in acortex-inspired silicon circuit. *Nature*, 405:947–951, June 2000.

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, June 2014.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, October 2016.

[12] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676, October 2017.

[13] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the*

27th International Conference on International Conference on Machine Learning (ICML), Haifa, Israel, June 2010.

[14] R. Pekrun. Emotions as drivers of learning and cognitive development. In R. A. Calvo and S. K. D'Mello, editors, *New Perspectives on Affect and Learning Technologies*, pages 23–39. Springer New York, New York, USA, June 2011.

[15] R. Sawyer, A. Smith, J. Rowe, R. Azevedo, and J. Lester. Enhancing student models in game-based learning with facial expression recognition. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP)*, Bratislava, Slovakia, July 2017.