MAKING THE UNKNOWN MORE CERTAIN: A STACKED ENSEMBLE CLASSIFIER FOR OPEN GESTURE RECOGNITION WITH A SOCIAL ROBOT

Heike Brock and Randy Gomez

Honda Research Institute Japan Co., Ltd
8-1 Honcho, Wako-shi
Saitama 351-0188, Japan

ABSTRACT

We introduce a novel stacked ensemble classifier for the unconstrained recognition of known and unknown gestural input data in nonverbal communication with a social robot. The architecture utilizes three separate CNNs of different expected data input size and combines their output predictions to a unified estimate. Analysis shows that in comparison to a single CNN architecture, the combined estimate reduces prediction confidence values for unknown gestural movement segments, making the system able to identify unknown data input with higher certainty under both laboratory and real environment conditions. In a human-robot interaction experiment, we are able to improve unknown class detection accuracy by up to 40% under maintained or equal known class recognition performance, and hence considerably enhance the overall robustness of the recognition system.

Index Terms— open gesture recognition, ensemble classifier, nonverbal communication, human-robot interaction, social robots

1. INTRODUCTION

Gestural communication is a central component of nonverbal interaction among two or more humans [1, 2, 3]. As such, it also constitutes a natural and straight-forward way of communicating with machines. Communication through gestures can complement and add to the information provided by verbal utterances, as well as provide information on its own whenever necessary. For example, a vertical finger in front of one's mouth will tell others that they should be quiet. However, the way how individuals use gestures to express their thoughts and intentions differs based on their cultural background, individual character, age or technical expertise. For this reason, it is essential to develop systems that are able to handle diverse and variably performed gestures, and adapt to individual styles of gestural communication.

In recent years, various machine learning models were introduced that reliably recognize activities and gestures from video input [4, 5, 6]. Gesture communication has also been studied in human-robot interaction [7, 8, 9]. However, the majority of these architectures operate under a distorted perspective that impedes their usability for actual human-centered perception tasks: presupposing that a user will only communicate via pre-defined gestures, a system cannot react to any other tracked movements that were not included during model training.

For social robots that are designed to co-habitate with humans [10], it is important to address the previous issues to enhance the quality of interaction and maintain long term user interest. The aim of our work is therefore to build a real-time recognition system that is capable to handle unknown data input, while stile being accurate and reliable. Specifically, we target perception capabilities for the online understanding of gestural indications towards a social robot companion [11]. To develop a system that is able to respond appropriately to both known and unknown user expressions, we propose a novel ensemble architecture built of three stacked CNNs. The system is evaluated on target-specific data and compared to a baseline consisting of a conventional single CNN. Showing a clear superiority of the ensemble approach, its integration onto the robot system can then provide a more flexible, natural and robust gestural communication system in the future.

2. BACKGROUND

The problem to identify unknown, and simultaneously recognize known classes, is commonly referred to as open set recognition. It extends and generalizes the problem of zero-shot learning of unseen classes based on data attributes, which has been a popular research topic for object recognition within the last years [12] and has also been introduced to the gesture recognition domain [13]. The main idea of open set recognition is to identify and reject movements that do not belong to any of the previously learned gesture classes. One possibility to enforce such classifier decision-making is the avoidance of low-confidence predictions [14]. This means input data is labeled as unknown once the class of maximum likelihood does not surpass a certain threshold. In a previous work [15], we developed a basic architecture

based on a single CNN classifier to address this task within a communication setting with a social robot. The architecture is complemented by a gesture detection process [16] and a verbal speech interaction pipeline, which can provide additional knowledge on the performed unknown gestures for labeling and future reference. In the current work, we now propose an enhanced network constituting of an ensemble of three CNN classifiers to improve reliability and robustness of the unknown class detection step.

3. SYSTEM SETUP

The system's perception is based on the Microsoft Kinect Azure, which provides three-dimensional body joint positions of users directing nonverbal gestural communication towards our social robot Haru [10]. Since we focus on hand gestural communication, we do not consider the lower extremities and track 24 joints of interest per person body (4 spinal joints, 2x4 arm joints, 2x3 hand joints, and 6 head joints). Under run time operation, the processing of the relevant body joint positions is as described in [15]. First, a selection of joints trajectories is used to detect the beginning and end of gestural movements for gesture segmentation. Next, the gesture segment is transformed into a standardized data structure of angular and distal features between related body segments following the procedure introduced in [17]. We then normalize the feature segment and resize it to a standardized size using cubic spline interpolation [18]. After evaluation of the resulting feature matrix by a CNN classifier, the final open set recognition constitutes of a simple sanity check. Here, the classifier's label estimates are rejected when the class of maximal prediction confidence does not surpass a certain confidence threshold τ observed during the training process, and the respective movement segments are assigned an 'unknown' label.

3.1. Single CNN Architecture

Our baseline builds the previously evaluated single **CNN** architecture shown to be superior to other learning approaches under the given scenario [15]. It consists of three stacks of two convolutional layers with kernel size (3x3) and a (2x2) pooling layer each, followed by two fully connected layers. All convolutional and the first dense layers are batch normalized, followed by an ELU activation. The top layer contains a sigmoid activation for classification. The CNN accepts input data (i.e. standardized feature matrices) of size $[96 \times 64]$.

3.2. Ensemble CNN Architecture

We speculate that the standardization of a feature segment's length might introduce intra-class variability based on the execution speed and duration of every gestural expression. Therefore, we propose a stacked ensemble classifier **ENS** utilizing the prediction estimates of three distinct CNNs. The

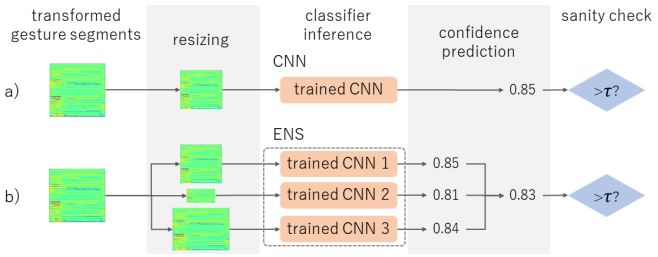

Fig. 1. Open recognition flow. a) Previously introduced baseline architecture (CNN), b) Proposed ensemble architecture (ENS).

CNNs all constitute of the basic architecture described above and only differ by their expected data input size. The chosen input data sizes are $[96 \times 64]$ as well as $[48 \times 32]$ (matrix squeezing) and $[128 \times 96]$ (matrix stretching). To obtain an ensemble evaluation of incoming feature segments, we combine the predictions of all three separately trained models into a joint estimate by simply averaging their class-wise confidence values (Figure 1).

4. METHODOLOGY

The aim of our work is to propose a classifier that identifies unknown gesture segments with higher reliability than the baseline, while maintaining accuracy in the classification of known gesture segments.

4.1. Basic Performance Evaluation

To obtain a genuine assessment of performance differences between the two architectures, we run an evaluation in a near-perfect, laboratory environment on the data set introduced in [15]. The collection comprises manually annotated gesture expressions of 18 target gestures for human-robot communication (representing known classes) and 9 gesture variations (representing unknown classes) performed by 15 diverse gesture actors. In total, the data set contains 5902 gesture segments labeled as one of the 18 known classes and 756 gesture segments labeled as unknown. To increase the amount of available data, we furthermore augment every data sample once by shifting its start and end frames with a random value between [-15, 15] and the superposition of random noise.

Next, we organize our data under a 4-fold group-based cross-validation, whereas the original gesture segment and its augmentation build one group. We split the known class gesture segments with a ratio of $60 : 20 : 20$ for training, validation and testing. We then combine the testing subset with the gesture segments of unknown class, and evaluate the performances of both architectures under the open test set. For every cross-validation cycle, we determine performance as average over 4 runs. For simplification, we only report the averaged metrics of all open test data in the following.

Evaluation Metrics To evaluate a classifier's performance in unknown class detection, we perform binary evaluations. Here, all samples from known classes in an investigated test set constitute one group, and all samples of unknown classes constitute another group. We determine the accuracy as A_k for the set of known samples and A_u for the set of unknown samples. Additionally, we compute the Youdens index J [19] defined by the binary recall R and the specificity S as

$$J = R + S - 1, \tag{1}$$

whereas S is defined by the number of true negatives TN and the number of false positives FP as $S = \frac{TN}{TN+FP}$. J is defined in the range $[-1, \ldots, 1]$, with -1 meaning that all data was incorrectly classify, 0 constituting an uninformative classifier, and 1 representing a perfect classification. To measure the classifier's potential to distinguish between known classes (while still taking into account incorrectly rejected data), we furthermore compute the macro-averaged F1-score of all known classes, whereas every sample of known class label that is classified as unknown is counted as false negative. We compare the classifiers' performance for a range of threshold values $\tau \in [0.55, 0.6, 0.65, \ldots, 0.9, 0.95]$.

4.2. Interaction Performance Evaluation

We deploy both architectures within a more realistic setting under the overall interaction pipeline. For this, we train the two architectures on the complete set of 5902 known gesture segments from the previous data collection. We utilize the trained models to determine the content of 24 consecutive freestyle gesture expressions; respectively to predict the labels of their sequence of automatically segmented gesture segments. The gesture expressions were enacted by 6 volunteers that did not participate in the training data collection. Participants were instructed to perform sequences of 5 to 8 gestures for communication with the robot to their personal thinking, leading to very unconstrained and unstructured data. Results of the gesture segmentation were recorded, and the resulting gesture segments manually labeled from video. This sequence annotation was then utilized as ground truth and compared to the predictions obtained with CNN and ENS. Errors in segmentation (e.g. short gesture segments caused by movement noise or missing segment end detections) were not further addressed, and the erroneous gesture segments were labeled as unknown. This reflects the strategy applied to actual interactions under run time environments, where treating perceived information as unknown is preferred over generating inappropriate robot behavior caused by false positives in recognition [20].

Evaluation Metrics We report performance of the sequence predictions in terms of element-wise macro-averaged accuracy, precision, recall and $F1$ score, whereas all unknown gestures are counted as one class. For unknown class rejection, we choose $\tau = 0.75$.

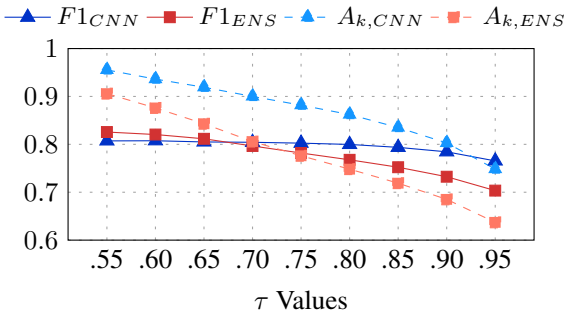

Fig. 2. Change in F1 and A_k scores with increasing threshold values.

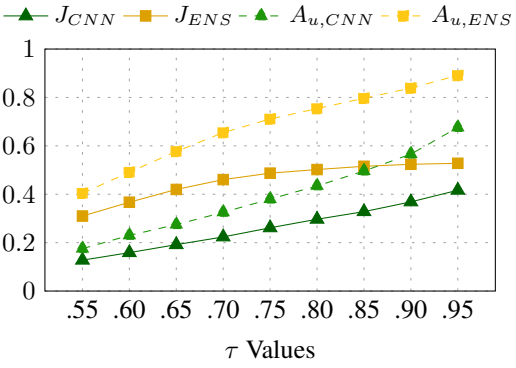

Fig. 3. Change in J and A_u scores with increasing threshold values.

5. RESULTS AND ANALYSIS

We consider the basic performance metrics under increasing τ values for both architectures. As can be expected, we observe a trade-off between classifier performance relative to the chosen threshold value: while a higher threshold value makes it easier to identify and reject unknown classes, it can also lead to erroneous rejections of known classes that could have been correctly classified otherwise. In concrete, we see that ENS achieves higher or similar $F1$ scores than CNN for lower τ values, but suffers from a stronger drop in performance with increasing τ (Figure 2). A similar loss of performance as a function of τ can also be found for A_k, whereas in this case CNN consistently outperforms ENS by 0.05 to 0.10 points. This correlation is reversed and amplified for the correct identification of unknown classes (Figure 3). Here, J and A_u scores of ENS are consistently higher than scores obtained with CNN, whereas the performance score differences range from approximately 0.2 to 0.4 points, and are hence considerably larger than for $F1$ and A_k.

We note that with $A_u < 0.3$, CNN is rarely successful in detecting unknown classes under small threshold values. ENS on the other hand shows higher reliability ($A_u \geq 0.4$). For further analysis, we have a separate look at the J and A_u

Table 1. Unknown class detection scores for low threshold values separately evaluated for all classifier.

	CNN1 [96 × 64]		CNN2 [48 × 32]		CNN3 [128 × 96]		ENS	
τ	J	A_u	J	A_u	J	A_u	J	A_u
.55	0.13	0.17	0.12	0.19	0.13	0.16	0.31	0.40
.60	0.16	0.23	0.16	0.24	0.16	0.21	0.37	0.49
.65	0.19	0.27	0.19	0.29	0.19	0.26	0.42	0.58

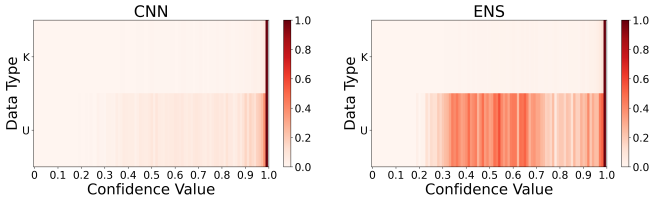

Fig. 4. Normalized intensity mapping of known (K, top) and unknown class (U, bottom) confidence values.

scores of all trained CNNs within ENS for lower τ (Table 1). Although CNN2 (input size $[48 \times 32]$) performs slightly better than CNN1 and CNN3 (input size $[128 \times 96]$), none of the separate CNNs is able to reach performance values of ENS. This suggests that the ensemble of classifier models allows for a better separation of unknown and previously learned gestural expressions. The distribution of confidence values further supports this assumption (Figure 4, Figure 5). Whereas both architectures evaluate the majority of known gesture segments consistently with high confidence values close to 1.0, ENS achieves to move its predictions of unknown gestural input data into the lower confidence value range. This ultimately helps to make a more informed decision about novel incoming data.

Lastly, to determine whether the ensemble classifier can make actual interaction with our social robot more robust, we analyse the interaction performance evaluation. Here, ENS clearly outperforms CNN in all common metrics (Table 2). We can also confirm the previous conclusion that ENS helps to identify unknown gesture input more accurately (Figure 6), with correct recognition of unknown segments rising from

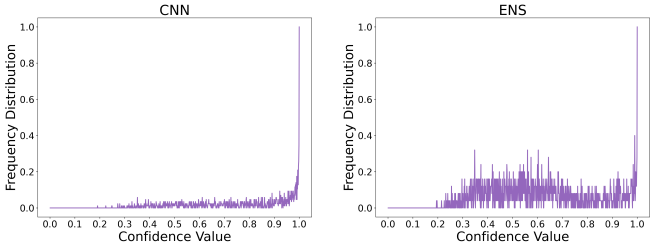

Fig. 5. Normalized frequency distribution of confidence values for unknown classes.

Table 2. Classifier performance averaged over all sequences.

	accuracy	*precision*	*recall*	$F1$
CNN	0.410	0.475	0.676	0.400
ENS	0.707	0.676	0.689	0.564

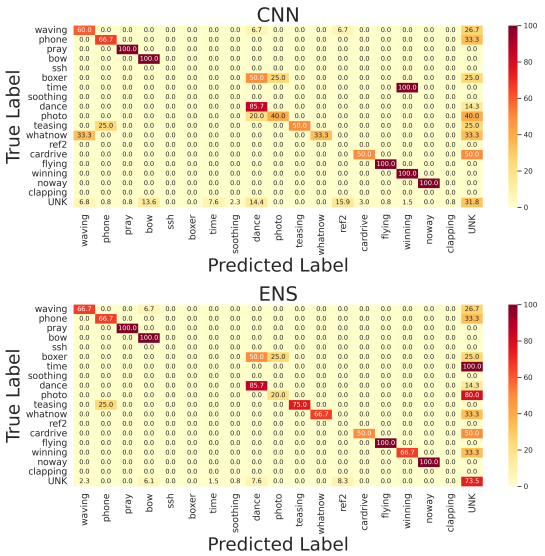

Fig. 6. Normalized confusion matrices for the sequential test data.

31.8% to 73.5%. Although few gestural segments could only be recognized with CNN (see e.g. 'time' and 'photo'), we conclude that ENS is better suited for the actual target setting of human-robot nonverbal interactions in real environments. In the following, we will now work to enhance the gesture detection scheme, to further reduce the number of inappropriate robot interactions caused by false positives.

6. CONCLUSION

We discuss the benefit of an ensemble CNN architecture built from the stacked predictions of three independent CNN models for different data input size under an open gesture communication setting in human-robot interaction. We performed an extensive evaluation of the proposed architecture under both laboratory and realistic conditions, and compare results to the baseline of a single CNN. Our analysis shows that a single classifier might achieve slightly better accuracy in recognition of known classes, however, advantages of the ensemble architecture prevail. In particular, the ensemble classifier helps to make unknown gesture class predictions more distinct to known ones, improving the robustness and reliability of unknown class rejection. Integration of the proposed ensemble architecture into a social robot framework could hence enhance its gesture recognition skills and make nonverbal communication more natural in the future.

7. REFERENCES

[1] Susan Goldin-Meadow, "The role of gesture in communication and thinking," *Trends in cognitive sciences*, vol. 3, no. 11, pp. 419–429, 1999.

[2] Adam Kendon, *Gesture: Visible action as utterance*, Cambridge University Press, 2004.

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

[4] Jesus Suarez and Robin R Murphy, "Hand gesture recognition with depth images: A review," in *2012 IEEE RO-MAN: the 21st IEEE international symposium on robot and human interactive communication*. IEEE, 2012, pp. 411–417.

[5] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[6] Siddharth S Rautaray and Anupam Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[7] Jochen Triesch and Christoph Von Der Malsburg, "A gesture interface for human-robot-interaction," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 546–551.

[8] Hee-Deok Yang, A-Yeon Park, and Seong-Whan Lee, "Gesture spotting and recognition for human–robot interaction," *IEEE Transactions on robotics*, vol. 23, no. 2, pp. 256–270, 2007.

[9] Hongyi Liu and Lihui Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.

[10] Randy Gomez, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura, "Haru: Hardware design of an experimental tabletop robot assistant," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 233–240.

[11] Heike Brock, Selma Sabanovic, Keisuke Nakamura, and Randy Gomez, "Robust real-time hand gestural recognition for non-verbal communication with tabletop robot haru," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 891–898.

[12] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[13] Naveen Madapana, "Zero-shot learning for gesture recognition," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 754–757, Association for Computing Machinery.

[14] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[15] Heike Brock and Randy Gomez, "Personalization of human-robot gestural communication through voice interaction grounding," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2021)*. IEEE, 2021.

[16] Heike Brock, Iva Farag, and Kazuhiro Nakadai, "Recognition of non-manual content in continuous japanese sign language," *Sensors*, vol. 20, no. 19, pp. 5621, 2020.

[17] Iva Farag and Heike Brock, "Learning motion disfluencies for automatic sign language segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7360–7364.

[18] E Kiran Kumar, PVV Kishore, ASCS Sastry, M Teja Kiran Kumar, and D Anil Kumar, "Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 645–649, 2018.

[19] Matheus Gutoski, Andre Eugenio Lazzaretti, and Heitor Silvério Lopes, "Deep metric learning for open-set human action recognition in videos," *Neural Computing and Applications*, pp. 1–14, 2020.

[20] David Cameron, Stevienna de Saille, Emily C Collins, Jonathan M Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, pp. 106561, 2021.