

Classification of Visual Interest based on Gaze and Facial Features for Human-robot Interaction

Andreas Risskov Sørensen, Oskar Palinko^a and Norbert Krüger^b

The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Keywords: Human-robot Interaction, Visual Interest, Gaze, Classification.

Abstract: It is important for a social robot to know if a nearby human is showing interest in interacting with it. We approximate this interest with expressed visual interest. To find it, we train a number of classifiers with previously labeled data. The input features for these are facial features like head orientation, eye gaze and facial action units, which are provided by the OpenFace library. As training data, we use video footage collected during an in-the-wild human-robot interaction scenario, where a social robot was approaching people at a cafeteria to serve them water. The most successful classifier that we trained tested at a 94% accuracy for detecting interest on an unrelated testing dataset. This allows us to create an effective tool for our social robot, which enables it to start talking to people only when it is fairly certain that the addressed persons are interested in talking to it.

1 INTRODUCTION

Robots are becoming more and more commonplace in today's world and thus an effort is made on creating a common ground between them and humans. A large part of communication between humans is based on body-language, eye gaze and other subtle movements (Argyle, 1972), so for robots to truly understand our intentions they must be able to pick up on these non-verbal cues (Mavridis, 2015). Human-robot interaction can benefit a lot from a robots' ability to read humans or take commands that are not direct, in order to initiate communication faster and make it smoother.


Classifying whether a connection has been made using gaze and facial expressions of a person could enable a robot to quickly discern whether it should engage them or offer its services elsewhere, saving time for both itself and others.

Visual interest in this work will mean a human's expressed attention towards the robot and its behavior. This interest can partly be expressed by establishing mutual gaze, i.e. eye contact with the robot. This is a special communication situation when both agents become aware of each other's attention, which creates a dedicated communication channel between them. In addition to eye contact, other face

features, like emotion expression can also contribute to a stronger expressed interest.

This work is using the SMOOTH robot as a platform (Juel *et al*, 2020). It is a modular social robot for helping in care homes and other locations. It is designed to complete logistics (transporting laundry bins) and people-related tasks, e.g. guiding of elderly to the cafeteria, while navigating among them. An additional task is serving drinks to care home residents, as dehydration is a large problem, as some elderly people tend to forget to drink water.

In our current work, this task was generalized to the wider public, as the robot was tested at a university cafeteria which also caters a nearby concert hall (Palinko *et al*, 2020) (see Figure 1). The videos used in creating the interest classifier were recorded at this location during lunch breaks for students as well as preceding events at the concert hall. To detect face and eye features an open source library, OpenFace, was utilized in order to extract information from people's images, including not only gaze and head orientation but also details about different facial features (Baltrusaitis *et al*, 2016). The robot was operated at the cafeteria in a Wizard-of-Oz manner (i.e. controlled by human operators from a concealed location).

^a  <https://orcid.org/0000-0002-4513-5350>


^b  <https://orcid.org/0000-0002-3931-116X>



Figure 1: The setup of the SMOOTH robot.

Regarding classification of visual interest, the features which OpenFace provides will be inspected in order to make sure they have the desired relevance regarding interest. The process of how the used datasets were chosen in order to minimize bias will be explained, along with the labelling technique used to designate ground truth to said data.

These datasets will be used to train several classifiers belonging to three different categories: Decision Trees, k-Nearest Neighbors and Support Vector Machines. The classification models will be tested on parts of the human-robot interaction data in the pursuit of finding the one yielding the highest interest prediction accuracy.

Other ways to improve the classification models will be explored, including a look at the importance of the individual features and combinations thereof, along with the structure of the input parameters.

2 BACKGROUND

Recently, automatic face feature analysis has seen a great expansion. OpenFace is one of the most popular open source libraries which provides automatic face feature, head pose, facial action unit and eye gaze recognition (Baltrusaitis *et al*, 2016). OpenFace utilizes a set of Action Units developed by Swedish anatomist Carl-Herman Hjortsjö in 1969 (Hjortsjö, 1969). Action Units are defined by the Facial Action Coding System which classifies human facial movements by their appearance on the face and can be used in recognition of basic emotions.

Regarding visual interest, there is considerable background on visual attention between two agents expressed by mutual gaze (Argyle *et al*, 1976). Mutual gaze is an important tool not only in human-human but also in human-robot interaction (Palinko *et al*, 2015). A human and a robot sharing mutual gaze are evidently also sharing visual interest in each other. But this interest can have additional elements like face expressions caused by emotions: a smiling person might be more interested in communicating with the robot compared to a person with a neutral

expression. This type of interest has not yet been very well explored in the field of human-robot interaction. The only study which the authors found on this topic, (Munoz-Salinas *et al*, 2005), showed a system for detecting, tracking, and estimating the interest of people in mobile robots using fuzzy logic and stereo vision. A fuzzy system computes a level of probability of interest based on the person's position in relation to the robot and estimate of attention given by the orientation of the person's head. The orientation is determined by the amount of skin detected, as it follows the assumption that more of a person's skin is visible when facing the robot. However, fuzzy logic has in recent years not seen so many applications as compared to other machine learning approaches, which will be discussed in this study.

Mutual gaze can be found by determining the eye gaze angles of two agents. Researchers have in recent time chosen to replace eye gaze with its first proxy, head orientation, in situation where eye gaze was too inaccurate or impractical. (Palinko *et al*, 2016) describe how the richer information gained from eye gaze has a significant impact on the human-robot interaction compared to that of the head orientation alone.

Machine learning and classification algorithms are as popular as ever, and a wide variety of classifier types are available. Some of the more recent advances in this field are described in (Zhang, 2010).

3 APPROACH

This section describes the experiment from where the data was collected (Section 3.1), what the data consist of and how it was labelled (Section 3.2), and finally which features were extracted from it (Section 3.3).

3.1 Robot Experiment

An "in the wild" Wizard-of-Oz human-robot interaction experiment was performed in which the SMOOTH robot was serving glasses of water to a naïve audience. It was conducted at the University of Southern Denmark's Sønderborg location at the ground floor cafeteria, which is also co-located with a concert hall. The drink serving was conducted during lunch hours where many students were present as well as in the evening before events at the concert hall, when event-goers populated the sitting areas.

The Wizard-of-Oz setup meant that the robot was controlled from a remote location by human operators, but the interacting audience thought the

robot is acting autonomously. The operators had a visual contact with the robot and could hear what people were saying. One operator was driving the robot, while the other pushed buttons to select what and when the robot would say from a predefined set of sentences, which were designed to convince people to drink water. Results of this experiment were reported on in (Palinko *et al*, 2020).

In our current study we used video recordings of subjects interacting with the robot while getting water. The recording device was a GoPro Hero Session located on top of the robot's head, just above the simulated eyes, between the two loudspeakers, see Figure 1 left and center. The data used for training different classifiers was taken from videos recorded by the robot during this water serving experiment.

3.2 Training and Testing Data

The data consists of a wide range of people; some are interacting with the robot while others are not.

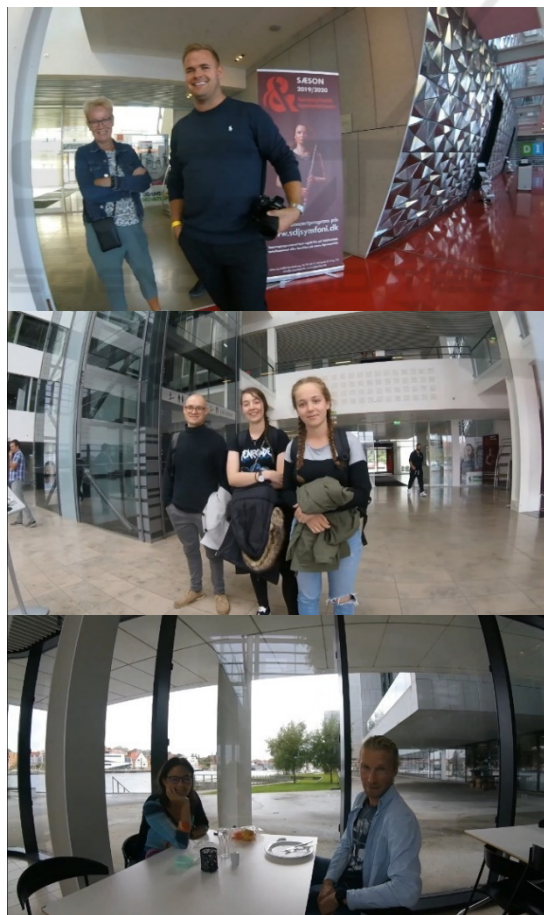


Figure 2: Examples of frames classified as 'Interest'.

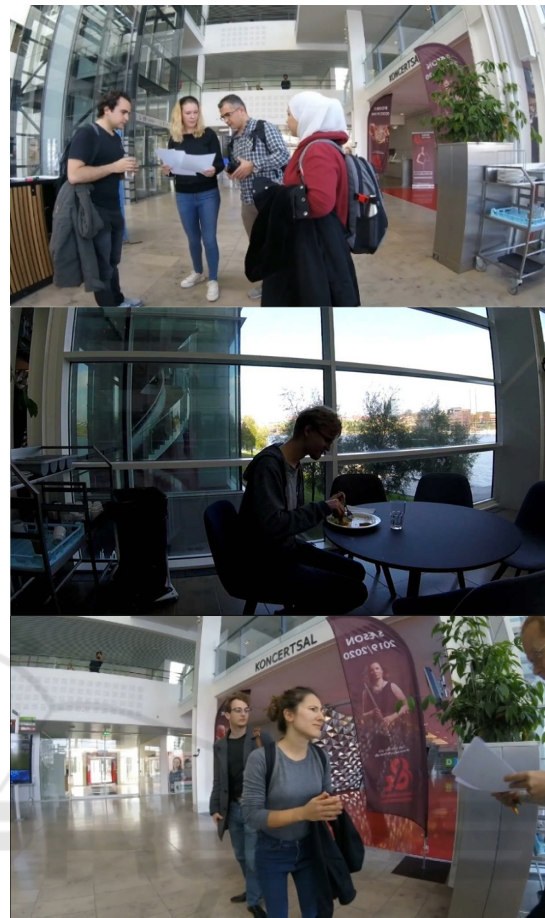


Figure 3: Examples of frames classified as 'No Interest'.

From this data several smaller snippets were extracted and put together to obtain two data sets of people showing interest and no interest respectively. Some examples of frames with each of the two classification are shown in Figure 2 and Figure 3.

We selected snippets from the overall video recordings and applied either the "interested" or "not interested" category to each clip. Because of technical difficulties in tracking each person in the video separately, only those clips were selected in which there was either only one person or where all people shared the same interest or non-interest in the robot. Clips with mixed interest were eliminated. The assignment of interest was up to the judgement of the authors and was based on people's direction of gaze, head orientation and face expressions.

A total of 48 video snippets were selected; 22 of people showing interest and 26 of people showing no interest. Each frame in the video snippets and each face in each frame was used for training the classifier separately. We had 22.604 frames labeled as

interested and 25.350 labeled as not interested in the training dataset.

3.3 Gaze and Face Features

OpenFace is an open source facial behavior analysis toolkit and is used in this work to extract facial features to be used for classification. The system is capable of performing Facial Landmark Detection, head orientation tracking, and gaze tracking, and is able to recognize certain Facial Action Units. Nearly any anatomically possible facial expression can be deconstructed using the Facial Action Coding System (FACS), into the specific Action Units (AU) that produced the expression (Ekman *et al*, 2012), which is a common standard to objectively describe facial expressions. OpenFace is able to recognize a subset of AUs and selection of these are shown in Figure 4.

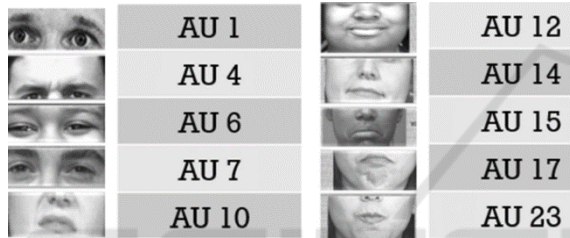


Figure 4: A selection of Action Units visualized: Outer Brow Raiser, Brow Lowerer, Cheek Raiser, Lid Tightener, Upper Lip Raiser, Lip Corner Puller, Dimpler, Lip Corner Depressor, Chin Raiser, Lip Tightener.

In the following we will discuss which features were used for training the classifiers:

Gaze Related. The gaze vectors for the eyes are given by the three coordinates $\mathbf{G}^L = (g_x^L, g_y^L, g_z^L)$ for the left eye in the image and $\mathbf{G}^R = (g_x^R, g_y^R, g_z^R)$ for the right eye. The vectors are normalized and given in world coordinates.

The eye gaze angles $\mathbf{A} = (a_x, a_y)$ are found by averaging the two gaze vectors ($\mathbf{G}^{avg} = 0.5(\mathbf{G}^L + \mathbf{G}^R)$) and taking the angles in x and y direction. These angles are given in radians and world coordinates. If a person's gaze is shifting left to right this will result in the change of a_x going from positive to negative, and if a person's gaze is shifting up to down this will result in a change of a_y going from negative to positive.



Figure 5: Visualization of eye gaze vectors as green lines.

Head Orientation. Rotation of the head pose in world coordinates with the camera as origin is given as $\mathbf{O} = (o_x, o_y, o_z)$. The rotation is in radians around x,y,z axes. This can be seen as pitch (o_x), yaw (o_y), and roll (o_z).

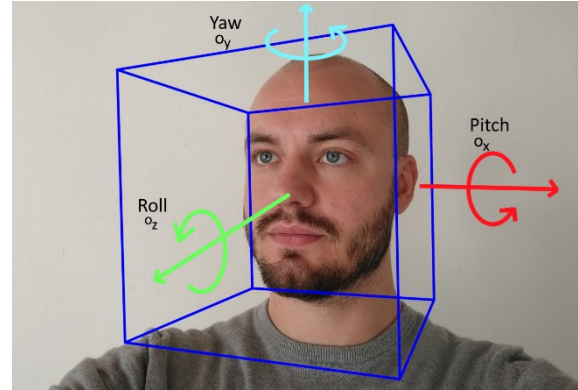


Figure 6: Visualization of head orientation.

Facial Action Units. Action units can be described in two ways. The first is a binary value determining if the Action Unit is present or not, and the second is the intensity of the Action Unit represented as a value between 0 and 5. OpenFace provides both of these parameters. It can detect the intensity and presence of the following 17 Action Units:

$$\mathbf{AU}_k = (\mathbf{AU}_k^{01}, \mathbf{AU}_k^{02}, \mathbf{AU}_k^{04} - \mathbf{AU}_k^{07}, \mathbf{AU}_k^{09}, \mathbf{AU}_k^{10}, \mathbf{AU}_k^{12}, \mathbf{AU}_k^{14}, \mathbf{AU}_k^{15}, \mathbf{AU}_k^{17}, \mathbf{AU}_k^{20}, \mathbf{AU}_k^{23}, \mathbf{AU}_k^{25}, \mathbf{AU}_k^{26}, \mathbf{AU}_k^{45})$$

where $k = i, p$ for intensity and presence, respectively. Additionally, the presence of Action Unit \mathbf{AU}_p^{28} can be detected.

The chosen parameters for use in the classification are then the 6 features for the two gaze vectors ($\mathbf{G}^L, \mathbf{G}^R$), the 2 features for the gaze angles (\mathbf{A}), the 3 rotation features for the head orientation (\mathbf{O}), the 17 intensity Action Units (\mathbf{AU}_i), and the 18 presence Action Units (\mathbf{AU}_p). This gives a total of 46 features.

3.4 Data Processing

Change of Coordinate System. Head and gaze angles in OpenFace are given in so-called world coordinates. Since interest shown in the robot is in relation to the camera, the angles were re-calculated with respect to camera coordinates to reflect that. Figure 7 shows two scenarios of the robot's field of view seen from above, with two people in the frame. On the left image the orientations of the two people are parallel and in the world coordinate system they have the angle value of 0. On the right they are both facing the camera of the

robot which means that both their angles with respect to the camera are equal to zero.

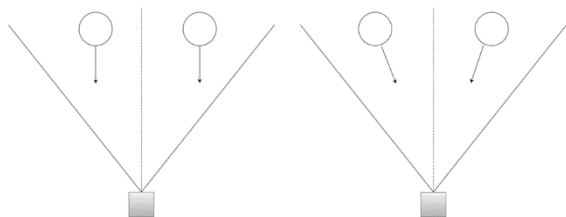


Figure 7: Left: the two people have the same orientation if using world coordinates. Right: the two people have the same orientation with respect to the camera.

Absolute Values. Gaze and head pose angles (**A** and **O**) range from negative to positive depending on their direction relative to the camera. The absolute values of the horizontal angles ($|a_x|$) were used instead to avoid potential bias of one data set having significantly more people looking a certain direction away from the camera. The same was not chosen for the vertical angles (angles that change when looking up-down) since the screen on the robot was located below the camera, so direction of angle away from zero could be related to interest.

Test Data. A separate dataset was created for use in testing the classification models. This dataset consisted of relatively short video snippets showing the first author having varying gazes and head orientations. The reason for not using parts of the other source videos, as was done for the training data, was to have an independent test set. In addition, using average parameter values of several frames is significantly easier with a single face compared to the multiple faces in the source videos, which would have to be tracked individually.

Training was done using 5-folds cross-validation to avoid overfitting, and the accuracies of the models were then found by using the models on the test dataset.

An additional observation made during these experiments was that sometimes the outputs given by OpenFace could be unreliable or incorrect for a short period. That is, the gaze vectors were, during some recordings, seen to be stable for most of the time, but occasionally flicker in seemingly arbitrary directions in short bursts. Therefore, it was chosen to use averages of video clips for the testing data instead of relying on single frames.

4 RESULTS

The dataset made from the data collected in Sønderborg was used to train several classification

models, some only including a limited set of features, to see which are most useful in classifying interest.

The classifier type was chosen to be Fine k-Nearest Neighbours (meaning $k = 1$) as it in initial tests showed good accuracy combined with fast training speed. The accuracies of the models were found by evaluating the models on the separate testing dataset by calculating the percentage of correct classifications. One set of models were trained using the camera coordinates, and another using the world coordinates to determine whether there is an improvement in accuracy as expected.

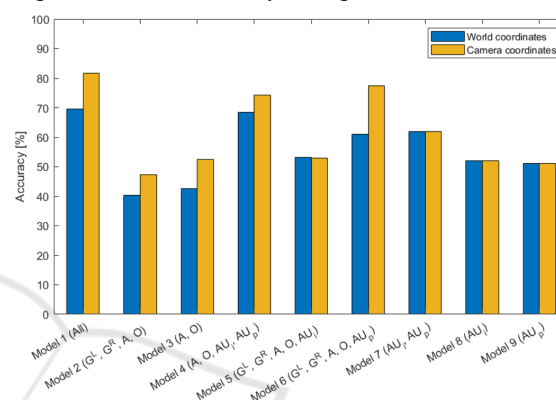


Figure 8: Comparison of accuracies between different models using world coordinates and camera coordinates, respectively. The used features are listed in parentheses.

Figure 8 shows that using all the features achieved the highest accuracy. Using the camera coordinates resulted in a significant improvement in the accuracy compared to the world coordinates by almost 12 percentage points in the model with all features. Using only features related to gaze and head orientation (Figure 8, Model 2) reduced performance to around 50 %, which is no better than guessing, and is significantly worse than just the Action Units by themselves. By comparing Model 5 to Model 6 it also becomes clear that the presence Action Units are more important than the intensity Action Units.

It was expected that the gaze and pose features would be more relevant than the Action Units, so to understand the poor performance of Model 2, the prediction results for that model was inspected closer. What proved to be the reason was that this model gave a prediction of 'No Interest' for every point in the test dataset. To see which of these 11 features were the cause for this behaviour, further models were made using different combinations of just these gaze and head orientation features.

Table 1: Models using combinations of the 11 features.

Model	Features	Accuracy [%]
10	g_x^L, g_x^R, a_x, o_y	74
11	$g_x^L, g_y^L, g_x^R, g_y^R, a_x, a_y, o_x, o_y$	61
12	$g_x^L, g_z^L, g_x^R, g_z^R, a_x, o_y$	65
13	$g_x^L, g_x^R, a_x, o_y, AU_i, AU_p$	73
14	$G^L, G^R, a_x, o_y, AU_i^{02}, AU_i^{04}, AU_i^{12}, AU_p$	82

Table 1 shows that the features describing the horizontal angles of gaze and head orientation (Model 10) by themselves give a relatively good accuracy but combining them with the features for the vertical angles (g_y^L, g_y^R, a_y, o_x) results in a poorer performance. The explanation for this could be the difference between the training and testing datasets, where the test data has less variation and possibly some bias in what camera angles were used. Keeping these horizontal features and adding the Action Units back in (Model 13), however, still gives a lower accuracy than using all the 46 features. Using this knowledge and trying to exclude the various intensity Action Units a final set of features with the highest accuracy was derived and is shown as Model 14. The 29 features included are the two gaze vectors, the horizontal gaze and head orientation angles, all the presence Action Units and three intensity Action Units.

In order to find the optimal classifier type additional models were trained using the above feature set. The k-Nearest Neighbours models and the Tree models all performed relatively well, while the Support Vector Machine showed more varied accuracies, with Fine SVM having the lowest score and Linear SVM having the second highest. The highest accuracy achieved was 90.24 % and was reached by the Fine Tree model.

As discussed in section 3.4, the outputs of OpenFace can sometimes be unreliable when looking at individual images. Therefore, the testing data was taken temporally by averaging over 30 data points at a time, which is equivalent to one second. The models using the various classifier types were tested using this temporal averaging and the comparison to the previous static classifiers can be seen in Figure 9.

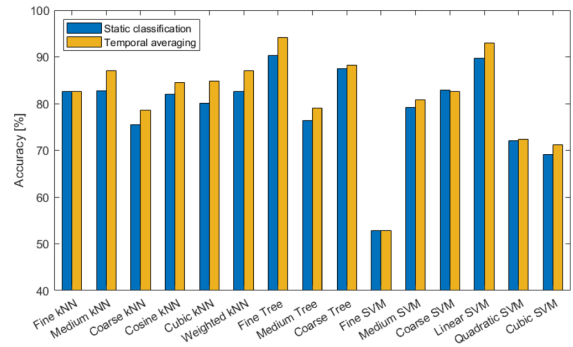


Figure 9: Comparison of various classifier models using static and temporal data, respectively.

Almost all of the classifiers had improved performance by applying temporal averaging. The best performing model was still of the Fine Tree type where the accuracy was increased to 94.10 %. A final comparison was made by retraining the models using all 46 features and applying temporal averaging. Figure 10 shows that the set of 29 features still results in the highest performing models, but for some of the k-Nearest Neighbours models using the full feature set gave slightly higher accuracies. The performance of the best model and its features are shown in Table 2.

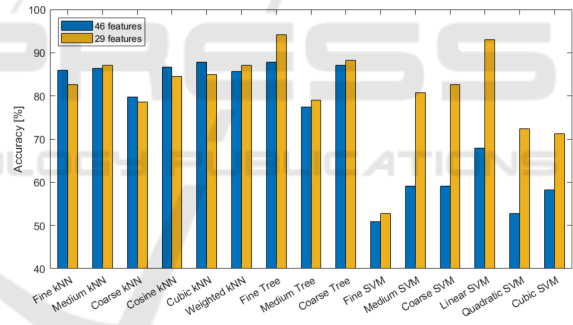


Figure 10: Comparison of various classifier models using the set of all features and the set of 29 features, respectively.

Table 2: Fine Tree model with the highest achieved accuracy.

Classifier	Features	Accuracy [%]
Fine Tree	$G^L, G^R, a_x, o_y, AU_i^{02}, AU_i^{04}, AU_i^{12}, AU_p$	94.10

5 DISCUSSION AND CONCLUSION

An interest detection system was developed by training classifiers of various types and the best performing model achieved an accuracy of 94 % on

the test dataset. Fifteen classifier types were tested, and their performances evaluated using subsets of the features. The model achieving the highest accuracy was of the Fine Tree type and used a subset of 29 features. Applying a temporal averaging to the test data in order to remove noise showed an increased performance for most of the classifiers.

The three intensity Action Units that proved to be relevant for the classification model were AU_i^{02} , AU_i^{04} , AU_i^{12} . They refer to 'Outer Brow Raiser', 'Brow Lowerer' and 'Lip Corner Puller' respectively. These are facial features that relate to frowning and smiling, which could have significance when evaluating interest. It should be noted, however, that the Action Units might not be completely reliable, especially in many of the 'No Interest' cases where the person is not facing the camera directly. Also, the reliability of the Action Unit values are noted to possibly be lower when using the feature extraction method on sequences containing multiple faces, which was used for this work, but despite this the achieved performance of the classifiers was good.

The only previous work found discussing classification of human interest in a robot (Munoz-Salinas *et al.*, 2005) used the detection of skin area based on colour to determine how much interest a person was showing. This method has two major limitations. Firstly, the usage of skin colour as a determining factor is not ideal as discerning in the case of bald people and people with low contrast between hair and skin colour can be problematic. Secondly, face orientation does not necessarily signify an interest in the robot. Both of these limitations are addressed with our suggested method. Including gaze provides a more reliable estimate a person's focus and thereby their point of interest.

The interest detection system described above can have different applications, and it will be primarily used on our robot for detecting which person in the robot's environment is interested in interacting with it and taking a cup of water which the robot carries. This detection algorithm will be especially useful in high traffic noisy situations where we cannot rely on the verbal communication channel, i.e. speech recognition, to gauge people's interest in having a cup of water.

ACKNOWLEDGEMENTS

This work has been supported by InnovationsFonden Danmark in the context of the Project "Seamless huMan-robot interactiOn fOr THE support of elderly people" (SMOOTH).

REFERENCES

- Argyle, M. (1972). Non-verbal communication in human social interaction.
- Mavridis, N. (2015). A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, 63, 22-35.
- Juel, W. K., Haarslev, F., Ramirez, E. R., Marchetti, E., Fischer, K., Shaikh, D., ... & Krüger, N. (2020). SMOOTH Robot: Design for a novel modular welfare robot. *Journal of Intelligent & Robotic Systems*, 98(1), 19-37.
- Palinko, O., Fischer, K., Ruiz Ramirez, E., Damsgaard Nissen, L., & Langedijk, R. M. (2020, March). A Drink-Serving Mobile Social Robot Selects who to Interact with Using Gaze. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 384-385).
- Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1-10). IEEE.
- Hjortsjö, C. H. (1969). *Man's face and mimic language*. Studentlitteratur.
- Argyle, M., & Cook, M. (1976). Gaze and mutual gaze.
- Palinko, O., Rea, F., Sandini, G., & Sciutti, A. (2015, November). Eye gaze tracking for a humanoid robot. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)* (pp. 318-324). IEEE.
- Munoz-Salinas, R., Aguirre, E., García-Silvente, M., & González, A. (2005). A fuzzy system for visual detection of interest in human-robot interaction. In *2nd International Conference on Machine Intelligence (ACIDCA-ICMI'2005)* (pp. 574-581).
- Palinko, O., Rea, F., Sandini, G., & Sciutti, A. (2016, October). Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5048-5054). IEEE.
- Zhang, Y. (Ed.). (2010). *New advances in machine learning*. BoD-Books on Demand.
- Ekman, R. (2012). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*.