Human-Robot greeting: A model based on social studies and Hidden Markov Models

Manuel Picão Fernandes Campos de Carvalho
manuel.carvalho@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

January 2021

Abstract

Social mobile robots should be capable of effectively open interaction with people. However, greeting someone is a complex task. Adam Kendon modeled greetings as a set of six phases: Initiation of Approach, Distance Salutation, Head Dip, Approach, Final Approach, and Close Salutation. These are valuable for a social robot to infer people's greeting intentions and comply with them.

This work proposes a system for mobile social robots that estimates the greeting phase through an HMM (*Hidden Markov Model*) by extracting observable features, and follows it with the appropriate behaviors using BTs (*Behavior Trees*).

We used publicly available datasets to train the HMM, through the EM (*Expectation-Maximization*) algorithm, extracting and labeling the necessary observable greeting features. Later, we tested the state estimation with sequences from the same datasets, and obtained an average accuracy of 80,9%.

To test the system we used a mobile humanoid robot from the Institute for Systems and Robotics, Vizzy. We conducted experiments on a simulator, obtaining an accuracy around 92% while predicting states seen by the robot, in different greeting situations. When connecting BTs to the state prediction, we confirmed that every state was properly replicated and natural greetings were achieved, confirming the system's applicability for HRI (*Human-Robot Interaction*).
Keywords: social robots, greetings, Hidden Markov Model, Behavior Trees

1. Introduction

Although most of the time people do not notice it, at the beginning of every interaction between humans, the two parties tend to follow a greeting ritual. Kendon [13] proposed a model for this ritual, composed of several steps, starting on the moment people sight each other and finishing, generally, with a salutation.

Greeting may be a struggling and unnatural behavior even for humans, since there are plenty of different approaches that can be taken. These may vary, for instance, according to the social relationship between the two parties, cultures, or education.

Despite the difficulties, the greeting ritual emerges with major social importance, both for humans and social robots. Proper and natural greetings can be the beginning of a good interaction, as unnatural and strange behaviors may bring a lack of comfort, or make the other abandon the interaction. Social robots have been growing substantially in the last few years, already serving human jobs such as receptionist [19, 20, 22] or companion of people in need [15, 17]. Therefore, to approximate the greeting behavior to the humans' is crucial for their Human-Robot Interaction (HRI).

For this, we will be based on Kendon's greeting model [13], which consists of six distinct phases: *Initiation of Approach*, *Distance Salutation*, *Head Dip*, *Approach*, *Final Approach*, and *Close Salutation*. These phases, described in the following section, do not happen always nor necessarily by this order. Thus, a social robot needs to be able to estimate the current phase from observable human social signals.

To keep track of the greeting ritual, we will model the phase estimation problem as a Hidden Markov Model (HMM). HMMs are probabilistic models defined by a set of states which are not directly observable (hidden), and a set of possible observations that the states depend on. For this problem, the hidden states will represent the six greeting phases. An HMM bases its ideas on the Markov property, which assumes that, at each moment, the decision of the next state depends entirely on the present state. Since these states are hidden, an HMM will also depend on which observations are found. In our case, observations can be seen as characteristics of the phases, as detailed later.

Our robot will then use a Behavior Tree (BT) as a control mechanism to react to the predicted state. BTs consist of flexible sequences of tasks (actions or movements, for example), that are performed according to conditions. A BT can be divided into several smaller trees, allowing us to create a sub-tree for each greeting phase, with its distinctive movements.

The content of this article is described as follows. Section 2 presents background content and related works; section 3 contains the implementation with an HMM; section 4 describes the robot reaction model using BTs; section 5 draws some conclusions and proposes future works.

2. Related Work

As stated earlier, Kendon [13] created a greeting model composed of six phases, based on a deep video analysis of a birthday party.

These phases are described as follows: i) *Initiation of Approach* (IA), where people sight a target person, make the decision to greet and start to prepare for the approach by orienting their body and looking directly; ii) *Distance Salutation* (DS), where people display a long-distance salutation, without physical contact. This salutation can vary from a head movement (head tossing, head lower or nod, for instance) to an arm movement (such as waving) and is usually accompanied by a smile and direct gaze; iii) *Head Dip* (HD), a subtle head lower movement, that commonly follows a DS; iv) *Approach* (APP), where people begin to move toward the target, sometimes without looking directly to him/her; v) *Final Approach* (FA), the final moments of the approach, where the greeters start to prepare for a close interaction, by adjusting their head position, looking directly to the other and, usually, smiling; and vi) *Close Salutation* (CS), where the sequence generally ends, by performing a salutation, such as a handshake, kisses, embracing, or a subtle head movement.

To the best of our knowledge, Heenan et al. [9] built the only solution to directly predict Kendon's phases and replicate them with a social robot, however, the authors opted for a Finite State-Machine to control the state changing. This model always performed the same phase sequence, which did not bring much flexibility. Also, it could not change state until the movement had ended, and ignored many of the social signals involved, as well as its uncertainty, in opposition to an HMM.

Other social robots [3, 24, 26] computed characteristics such as the person's position, orientation and availability to interact. However, these were usually used to change between their own greeting phases, and never to estimate them in a person greeting.

Regarding the phases implementation, several projects [3, 9, 23, 24, 26] implemented an approach movement toward a target person on a social robot, even though only a few [9, 24, 26] could distinguish an *Approach* and a *Final Approach* phase, where the robot would begin to prepare for an interaction. Only two projects [9, 23] implemented a waving movement as a *Distance Salutation*, while [3, 7, 9, 24] displayed a *Close Salutation*, despite only [9] produced a contact salutation (a handshake). The IA was the phase most commonly replicated [3, 7, 9, 23, 24, 26], given that most HRIs call for it, while the *Head Dip* did not seem to have any public reproduction.

3. Greeting Model using a Hidden Markov Model

3.1. Overview

As stated before, we model Kendon's greeting model as a Hidden Markov Model (HMM), where each phase of the greeting corresponds to one hidden state. For this, we chose to adopt a Gaussian HMM, assuming that our observation features follow a Gaussian distribution. This was adequate since the features selected were mostly physical features that are always present, though with different values, and not events that either happen or do not, as the general HMM observations. However, this brings a few differences, when comparing to the common HMM.

Firstly, our HMM contains a sequence of observations, $\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_V$, where each one is a vector with length M, with each value being the numerical value of one of the M features.

Secondly, the emission probability matrix $\boldsymbol{B}$ is split into: i) A matrix $\boldsymbol{M}$, with the mean values for each observation in each state; ii) A covariance matrix $\boldsymbol{C}$ for each state, with the covariance values between observations, or variances, in the diagonal values.

As the usual HMMs, our model will also have a matrix $\boldsymbol{A}$ with the transition probabilities between states and a vector $\boldsymbol{\pi}$ with the probabilities for the initial state.

3.2. Observations

To have the model predicting the states accurately, there was a need to choose observable features that characterized only some phases or could allow to distinguish them. We chose an observation rate of 5 per second to be high enough for the model to change states without noticeable time gaps, but also low enough to not exceed any other robot connection rate, provoking errors. The five chosen observation features are described as follows.

Distance. In a greeting sequence, people commonly start far away from the target and end up

close. Therefore, the robot-person distance is a key factor for predicting the state. For the calculations, we used OpenFace [2], which returns the face information according to the camera, including the 3D position and orientation. As we are only interested in a 2D distance, we discarded the height coordinate. After a transformation from the camera referential to the robot's base frame (see figure 1), distance is given, in millimeters by the following equation, where $p'(x, y)$ is the position of the person in the robot's XY plane:

$$Distance = \|p'\| \tag{1}$$

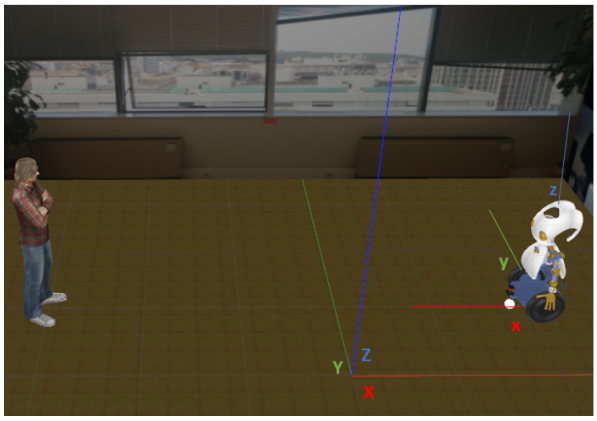

Figure 1: Representation of the robot (x,y,z) and world (X,Y,Z) coordinate frames

Speed. The speed of a target person allows distinguishing static phases of the greeting model (usually IA and CS) and moving phases (APP and FA, at least). For a simple computation of this feature, we used the distance that a person moves in the robot's direction within two observations, and later divided it by the time interval between observations, 0.2 seconds. Having the person-robot distance in the last observation, $PreviousDistance$, and the distance from the person to the last observation's robot position, $Distance'$, speed can be given, in millimeters per second, by:

$$Speed = \frac{(PreviousDistance - Distance')}{0.2} \tag{2}$$

Gaze. The direction of a person's gaze is another meaningful factor in this greeting model, as already described. Therefore, we developed two methods in order to compute if the person appears to be looking at the robot, or not. The first method uses eye gaze direction vectors, extracted from OpenFace (green lines starting at the eyes in figure 2) to estimate a gaze point, that is, a point in the robot's YZ plane to which the gaze direction of the person is pointing. Having the 3D position of both eyes (also returned from OpenFace) in robot coordinates, e_0, e_1 and the direction vectors, $\boldsymbol{g}_0, \boldsymbol{g}_1$, the gaze point gp is computed by:

$$gp = \frac{(\boldsymbol{g}_0 + e_0) + (\boldsymbol{g}_1 + e_1)}{2} \tag{3}$$

In the above equation, we assume the gaze point to be estimated by the center of both left and right eye gaze points.

To determine the likelihood of direct gaze, we used the cone model for the field of view of a human [1, 28]. Larger opening angles of the cone are associated with images progressively more blurred, while smaller angles bring more details in the view. Our visual attention is commonly associated with an angle smaller than 60^{o}. Thus, we pictured a cone with the vertex on the person's eye and a 60^{o} angle opening which represented the person's view. We consider the person is looking directly if the robot's face is inside the cone's base, i.e., $\|gp - f\| < r$, being f the center of the robot face, in its coordinates and r the radius of the base. The gaze feature is given by the following equation, where smaller values correspond to a higher gaze likelihood.

$$Gaze = \frac{\|gp - f\|}{r} \tag{4}$$

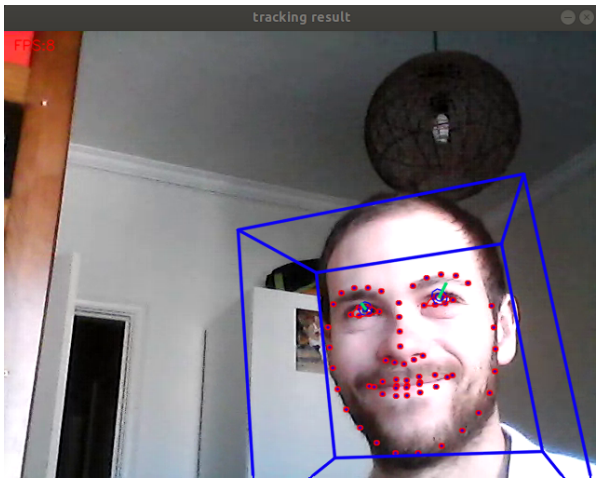

Figure 2: Output of OpenFace: face landmarks, face orientation and gaze direction

As the direction vectors proved to be inaccurate at long distances, we also built a gaze direction detector based on the face's orientation (blue structure in Figure 2, which Kendon [13] stated could be a good indicator in most cases.

To estimate the gaze point, $gp(gp_y, gp_z)$ we used the following equations:

$$gp_y = |t_x| \; tan(\lambda) - t_y \tag{5}$$

$$gp_z = |t_x| \; tan(\beta) + t_z \tag{6}$$

Where β and λ are the rotations of the face around the Y and Z axes of the robot's referential, and $t = (t_x, t_y, t_z)$ is the face's position. After estimating GP, the method is identical to the first one, and $Gaze$ values are calculated.

Smile. The intensity of smiling is yet another very important characteristic of some phases of Kendon's greeting model and it will also function as an observation feature. To create a reliable smile detector, we used the Action Units (AUs) detector from OpenFace. AUs are defined by the Facial Action Coding System as the contraction or relaxation of one or more facial muscles and are commonly used in various fields to detect emotions, for instance. Several studies, for instance [21, 25], mention that the feeling of happiness can be physically displayed by the combination of AU6 and AU12, i.e., the combination of a raise of the cheeks and a pull of the lip corners. With the presumable assumption that both AUs have similar importance, the smile feature is given as follows:

$$Smile = \frac{AU6 + AU12}{2} \tag{7}$$

Here, the value of $Smile$ depends on the OpenFace AU scale, in which we considered 1 to be approximately a neutral face, and 2 to be a common smile.

Movements. The *Distance Salutation*, *Head Dip* and *Close Salutation* phases are characterized by a typical head or arms movement. Therefore, the last observation feature would be a detector that could distinguish between these three kinds of movement and return the probability for each one.

Due to its implementation complexity, the detector implemented rested on a non-automated process of a user, external to the greeting, pressing a key, whether a movement from one of the three kinds was performed, and returning a probability of 1, depending on the kind detected. We separated these 3 probabilities into $HDip$, $DSal$ and $CSal$, therefore, every observation vector has the format $\boldsymbol{O} = [Distance \;\; Speed \;\; Gaze \;\; Smile \;\; DSal \;\; CSal \;\; HDip]$

3.3. Model training

The training of our HMM was performed using information from videos of real greeting sequences, extracted from the AVDIAR Dataset [8] and the UoL 3D Social Interaction Dataset [5].

This information would serve as input for the Expectation-Maximization (EM) algorithm [11], which, given a sequence of observations $\boldsymbol{O}$, and the set of possible states for the HMM, should return an estimate for the matrix parameters. This algorithm starts with an initial estimation of the HMM parameters. Then, two steps run iteratively until the algorithm reaches a convergence point: The E-Step and the M-Step. The E-Step uses the matrices from the last iteration to estimate an expected state transition count and the expected amount of transitions between each pair of states, through the entire given sequence. On the M-Step, these estimations are used to compute new probabilities for the HMM and to make an estimation of the four matrix parameters, for a Gaussian HMM.

To use the EM algorithm, we had to extract several sequences of observations from the abovementioned datasets. The AVDIAR dataset, apart from the videos, provided a file with the 2D head position from all the participants in the video and the calibration information from the stereo cameras. Given this, the observations' computation process for this dataset was the following, for each video:

1. Extraction of the left and right camera's rectified images for each frame;
2. Calculation of the disparity map for these images, using the Semi-Global Block Matching algorithm [10] and the values of depth, using the disparity-to-depth matrix provided;
3. Computation of the distance and speed features for each frame of the greeting sequence;
4. Correction of some inconsistent values by a median filter and interpolation and ensuring a correct interval between observations;
5. Analysis and labeling of the remaining 3 features, using adequate scales.

For the second dataset, the job was easier, as we were already given the 3D position of both greeters at each frame. The computation of gaze direction was also possible, due to having head orientation values. Thus, after computing distance and speed similarly to the previous dataset, we used the second gaze detector described earlier and computed this observation feature. Following this, we made some corrections to inconsistent gaze values and labeled the smile and movements features, as in the AVDIAR Dataset. In Table 1 we summarize the two approaches for the observations' extraction.

With the processes above described, we managed to extract 33 complete greeting sequences with their observations, separated by 0.2 seconds. From this 33, we decided around 75% would serve for the model's training and the other 25% would belong to a test set, where each observation was labeled with the seeming greeting phase of the respective moment. Firstly, the sequences were manually split,

	AVDIAR	UoL
Format of available data	2D	3D
2D-3D conversion	YES	NO
Distance/Speed extraction	Using 3D position	Using 3D position
Filter to ensure 0.2 seconds	Median	Average
Outlier interpolation	YES	NO
Gaze extraction	Labeling	Using head orientation
Smile/movements extraction	Labeling	Labeling

Table 1: Comparison of the observations' extraction approaches for the 2 Datasets

choosing 25 sequences ($\sim 75.8\%$) that were long and contained most of the phases, providing more information for the model. These were the input for the EM algorithm, together with the desired number of states (6). The algorithm was set to stop when it considered convergence (log-likelihood lower than 0.01) or divergence (iteration number higher than 100) had been achieved, returning the matrices for our HMM: the transition probabilities ($\boldsymbol{A}$), the initial state probabilities ($\boldsymbol{\pi}$), the mean values ($\boldsymbol{M}$), and the covariances ($\boldsymbol{C}$) for the observations. The first three are found below for further analysis.

$$\mathbf{A} = \begin{bmatrix} 0.608 & 0.144 & 0 & 0.030 & 0.218 & 0 \\ 0 & 0.625 & 0.075 & 0 & 0.300 & 0 \\ 0 & 0 & 0.400 & 0.198 & 0.402 & 0 \\ 0.066 & 0 & 0 & 0.631 & 0.303 & 0 \\ 0 & 0.051 & 0 & 0 & 0.673 & 0.277 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\pi} = \begin{bmatrix} 0.929 & 0 & 0 & 0.071 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{M} = \begin{bmatrix} 2068 & 99 & 1.47 & 1.33 & 0 & 0 & 0 \\ 1617 & 142 & 0.80 & 1.89 & 1 & 0 & 0 \\ 1635 & 256 & 1.26 & 1.40 & 0 & 0 & 1 \\ 2771 & 1896 & 0.97 & 1.77 & 0 & 0 & 0 \\ 1569 & 389 & 0.68 & 1.73 & 0 & 0 & 0 \\ 1083 & 131 & 0.57 & 1.91 & 0 & 1 & 0 \end{bmatrix}$$

As the EM algorithm is unsupervised, its output represents a model with six states clustered according to the information provided, and not necessarily the six phases we wish. Therefore, the matrices above were previously organized so that each generated state is connected to its most similar phase, in the following order: IA, DS, HD, APP, FA, CS.

This HMM, onward mentioned as the Trained Model, has a few differences comparing to Kendon's greeting description. Firstly, almost every distance and speed mean value was shorter than expected, arguably due to the two models being based on highly different environments: an outdoors, large, and crowded party, and an indoors, small room generally with two people. This also provoked a scarcity of the *Approach* phases, since smaller distances resulted in people starting the approach already preparing for the salutation (*Final Approach* phase). Despite these limitations and a few other small details, the Data-Driven Model still estimated six states very similar to Kendon's description of the six original phases.

3.4. Vizzy

Vizzy [18] is a humanoid-like robot developed by the Institute for Systems and Robotics (ISR) for assistive robotics, whose appearance can be seen in Figure 3. Vizzy was already part of multiple initiatives, which have implemented several skills on it, including reaching and grasping for simple shape objects, 3D face detection (position and orientation), localization and autonomous navigation in a known map, arm gestures as a handshake, waving, arm stretching, pick objects and drop objects, head control for a 3D fixation point (gaze) and speaking.

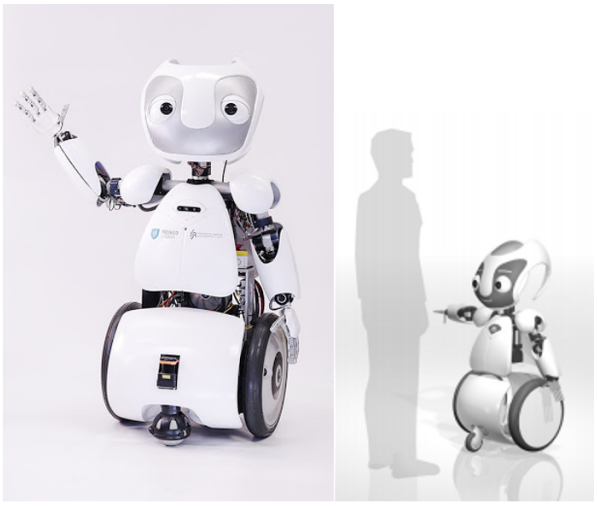

Figure 3: Left: Vizzy waving; Right: Vizzy's size comparing with a 1,75 m person

3.5. Model testing and results

After a brief comparison with Kendon's notes, we used the test set, with sequences extracted from the greeting videos, to compute two testing metrics for the model. These were the accuracy of the model, i.e., the percentage of state labels it can predict, comparing to the previous manual labeling; and the

confusion matrix, to provide a superior comprehension of the errors.

For the accuracy, two prediction algorithms were used: the Viterbi algorithm [6] and the forward algorithm [11]. The main difference between these two predictive algorithms is that, while the forward algorithm receives a sequence of observations and calculates, iteratively, the probability of each state for each index on the sequence, the Viterbi algorithm computes directly the most probable state path that corresponds to the sequence of observations. This difference makes the forward most capable to predict the states in real-time situations, as we will not obtain entire sequences but, instead, one observation at a time. The Trained Model's accuracy results are in Table 2. As a means of comparison of these results, we manually created an HMM, Kendon Model, entirely based on Kendon's greeting notes, and whose results are in the bottom row of the Table referred to.

	Forward Algorithm	Viterbi Algorithm
Data-Driven Model	0.839	0.893
Kendon Model	0.785	0.774

Table 2: Accuracy of the two models on the test set

The model achieved an accuracy over 83% with both predictive algorithms, which was higher than both Kendon Model's results, arguably due to the mentioned differences between the environments. The Viterbi algorithm was the most successful in the Trained Model case, which was expected, given that the sequences provided were complete.

Table 3 contains the second analyzed metric, the confusion matrix. Here, we could confirm the *Approach* and *Final Approach* limitation, with 50% of the errors coming from confounding these two states. However, the IA, DS, and HD were not properly tested, since the smaller amount of these states required their presence on the testing set. Figures 4 and 5 also provide us the two sequences with the most incorrectly predicted labels, confirming that errors did not escalate easily, neither were mostly present in one sequence.

	IA	DS	HD	APP	FA	CS	Real Total
IA	1	0	0	0	0	0	1
DS	0	0	0	2	0	0	2
HD	0	0	0	0	0	0	0
APP	1	0	0	8	3	0	12
FA	2	0	0	2	47	0	51
CS	0	0	0	0	0	27	27
Predicted Total	4	0	0	12	50	27	**93**

Table 3: Confusion Matrix of the chosen model on the test set, using the Viterbi algorithm

To evaluate the robustness of the model, we later created 15 train-test splits and calculated the accuracy values, given that each case generated an HMM and had different test sequences. The results can be found (mean +/- standard deviation format) in Table 4, with the comparison of the Kendon Model on the same test set.

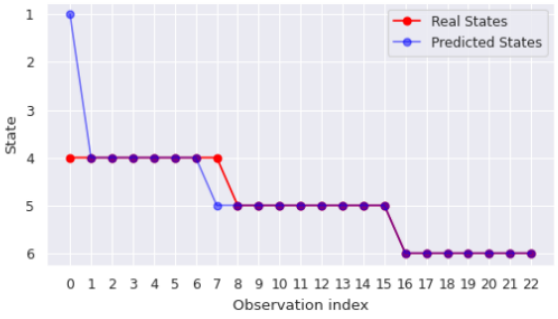

Figure 4: Instance of a sequence predicted with the Viterbi algorithm

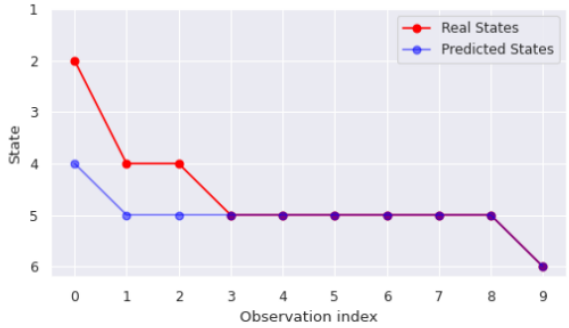

Figure 5: Instance of a sequence predicted with the Viterbi algorithm

The average values of accuracy were slightly lower than with the Data-Driven Model, as expected. As the size of the training set is small, randomly choosing the sequences resulted in some cases where the training missed important information to identify the six desired states. However, this also created a high standard deviation for the accuracy, opposing to the Kendon Model, since there is no training. With a higher quantity of greeting data, we predict the model would stabilize similarly to the first trained model, providing a smaller standard deviation and an average accuracy that should exceed the Kendon Model.

	Forward Algorithm	Viterbi Algorithm
Data-Driven Model	0.780 +/- 0.132	0.801 +/- 0.115
Kendon Model	0.810 +/- 0.058	0.823 +/- 0.051

Table 4: Accuracy of the two models on 15 different train and test sets

Finally, we implemented a few complementary tests using real-time observation sequences, in a simulator built for Vizzy using Rviz [12] and Gazebo

[14] as visualization and execution tools for the two middlewares used with Vizzy: Yet Another Robot Platform (YARP) [16] and Robot Operating System (ROS) [27]. Several types of greeting situations were simulated, using Vizzy in simulation, and a model of a fake person, whose face information could be extracted.

In each experiment, the model received observation features from the person and predicted the most probable state every time stamp, using the forward algorithm. Here, the gaze and smile features were kept constant at regular values for simplicity. In Table 5) there is a small description about the six situations experimented, which led to an average accuracy of 91.8 %.

Description	**Accuracy**
Normal greeting with every state	0.957
Greeting without DS and HD	0.889
Greeting with a DS only close to person	1
Greeting starting at short distance	1
Smiling greeting (smile=2.5)	1
Gazing greeting (gaze=0.4)	0.733
Global Accuracy	0.918

Table 5: Accuracy of the model on different sequences

The global structure of the BT that was developed contained, as can be seen in Figure 6, a node subscribing the state being predicted and a block containing a switch function, running one sub-tree, according to the prediction. The reactiveness of BTs permits an almost constant checking on the predicted state and immediate change of sub-tree, if necessary. The checking rate was chosen to be 10 times/second, to ensure that every change of state (updated 5 times/second) had a reaction on the BT.

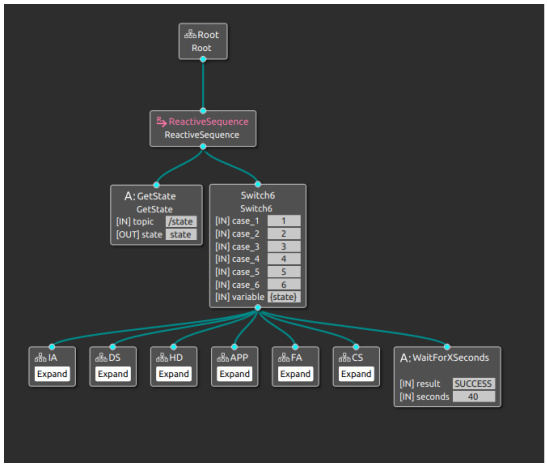

Figure 6: Global Behavior Tree of the system

4. Greeting Model using Behavior Trees

4.1. Overview

After many validations of our Hidden Markov Model's capacity to correctly predict states in real greeting sequences, we implemented the robot's reaction to each one of these states. With the usage of Behavior Trees (BTs), we created a Control Architecture for our system, that could read the predicted state, published in a specific ROS topic by the HMM, and command the robot to perform the respective sequence of movements, given a specific prediction.

4.2. Behavior Trees

Behavior Trees (BTs) [4] are a Control Architecture, whose function is to structure the switching between different tasks (represented as nodes) in an autonomous agent, such as a robot. We chose BTs to control our robot's reaction model, firstly, because of the reactiveness they provide. A BT allows good handling of unexpected changes and errors by being able to check every condition and roll back to a previous task of the sequence, quickly and efficiently. Secondly, their modularity allows the components to be developed and tested separately, which was highly beneficial in our system, as we could create a BT containing six smaller sub-trees, one for each state.

Initiation of Approach. As a response to the usual first phase of a greeting sequence, the robot performs two actions described by Kendon: frontal orientation at the target person, following by direct looking. This orientation changing is computed by four steps: i) identifying target face's position and robot position in world referential $r = (r_x, r_y, r_z)$; ii) changing the face position to the world referential, $p = (p_x, p_y, p_z)$; iii) calculating the goal orientation using equation 8, by considering only the rotation around the Z-axis (λ), and where the function $atan2$ is the 2-argument arctangent; iv) rotating robot to the target orientation, keeping its position.

$$\lambda = atan2(y = p_y - r_y, x = p_x - r_x) \quad (8)$$

After it, the robot moves its head in the direction of a point that corresponds to the target's face in the robot coordinate frame.

Distance Salutation. Reacting to this phase, Vizzy was set to produce a waving movement with its right arm and ensure a direct looking at the person, as it could have faded due to another phase.

Head Dip. As it detects this phase, the robot reacts accordingly, by performing a similar movement that consists of setting a head orientation lower than the target's face.

Approach. As soon as the person starts its approach movement and the HMM predicts the APP state, the robot starts two parallel branches of the BT that run in parallel. On one side, Vizzy receives and executes approach plans according to the person's position, by the following method: i) obtaining person's position and orientation through OpenFace; ii) converting the obtained values to the world referential; iii) calculating the goal position as the point at a specified distance from the person (1 meter in our case); iv) calculating the goal orientation as the opposite of the person's, to prepare for a frontal interaction; v) moving the robot the goal position and orientation; vi) Steps i) to vi) are repeated at the BT's rate, ensuring that any change in the person's position is noted and updates the approach plan.

On the other branch, the non-direct gaze identified by Kendon in this phase is provided by looking at a position lower than the target's face.

Final Approach. When the greeter starts to prepare for its interaction with the robot, the HMM should start to predict the FA phase. As the reaction, Vizzy continues using the same approaching logic as in the previous phase, however, it changes its gaze display to look directly at the target.

Close Salutation. To replicate the usual final phase of the greeting, the robot was programmed to display a handshake movement with its right arm, while combining it with a direct gaze and a verbal greeting: "Muito prazer", which translates to "Pleasure to meet you!".

4.3. System Testing

Our final experiments consisted of testing the phase prediction model combined with the reactions from the Behavior Trees.

For this, we set a situation similar to the previous experiments, using the Vizzy simulator. The difference was that the state would be continuously published in a ROS topic that would activate the Behavior Tree and trigger the robot to move, according to the phase predicted by the HMM. The greeting sequence tested started with the robot turning to the person, by detecting the *Initiation of Approach* and passed through the six states, finishing with a *Close Salutation*. Three different parts of the sequence are present in Figure 7 to 9.

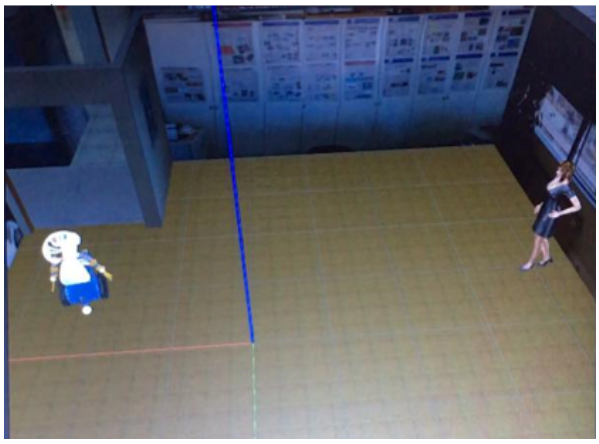

Figure 7: Simulation: Initial positions

Figure 8: Simulation: Approaching movements

5. Conclusions and Future Work

To build a Hidden Markov Model trained with real data that could represent Kendon's greeting model in a manner that no other project had accomplished was, undoubtedly, a major challenge in this work. Considering the low quantity and questionable quality of the sequences found, to manage to train a model that could identify six states particularly similar to the ones Kendon had described was a significant achievement. Our HMM also returned positive results when experimented on a test set, predicting 89.3% and 83.9% of the state labels with the Viterbi and the forward algorithm, respectively; an average accuracy of 78% and 80.1% using several combinations of train and test sets; and almost 92% while testing in the simulator with different situations.

When testing Vizzy reacting to the state prediction of the HMM, the implementation using Behavior Trees also responded as predicted. All states could be correctly displayed in our simulator and the time gaps for the reactions were not too long, allowing natural greeting sequences.

For the future, it might be interesting to discover

Figure 9: Simulation: *Close Salutation*

how the model would adapt to having more training information from greetings, preferably from more than one experiment environment. Certain features could have also enhanced the quality of the robot's greeting and were not implemented, such as a smiling display, a few subtle arms and body movements, or other salutations, since people can expect to be greeted differently. Another valuable extension to this work would be an adaptation to group greeting, by changing the HMM for continuous greetings, ensuring it does not repeat any target.

References

[1] Patricia L. Alfano and George F. Michel. Restricting the field of view: Perceptual and performance effects. *Perceptual and Motor Skills*, 70(1):35–45, 1990. doi: 10.2466/pms.1990.70.1.35. URL `https://doi.org/10.2466/pms.1990.70.1.35`. PMID: 2326136.

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. doi: 10.1109/FG.2018.00019.

[3] Drazen Brscic, Tetsushi Ikeda, and Takayuki Kanda. Do you need help? a robot providing information to people who behave atypically. *IEEE Transactions on Robotics*, PP:1–7, 01 2017. doi: 10.1109/TRO.2016.2645206.

[4] Michele Colledanchise and Petter Ogren. *Behavior Trees in Robotics and AI: An Introduction*. 07 2018. ISBN 9781138593732. doi: 10.1201/9780429489105.

[5] C. Coppola, S. Cosar, D. Faria, and N. Bellotto. Automatic detection of human interactions from rgb-d data for social activity classification. In *IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 871–876, 2017.

[6] G. D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[7] Mary Foster, Rachid Alami, Olli Gestranius, Oliver Lemon, Marketta Niemelä, Jean-Marc Odobez, and Amit Kumar Pandey. The mummer project: Engaging human-robot interaction in real-world public spaces. volume 9979, pages 753–763, 11 2016. ISBN 978-3-319-47436-6. doi: 10.1007/978-3-319-47437-3_74.

[8] Israel D. Gebru, Silèye Ba, Xiaofei Li, and Radu Horaud. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099, 2018. doi: 10.1109/TPAMI.2017.2648793.

[9] Brandon Heenan, Saul Greenberg, Setareh Aghel Manesh, and Ehud Sharlin. Designing social greetings in human robot interaction, 2014. doi:10.1145/2598510.2598513.

[10] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814 vol. 2, 2005. doi: 10.1109/CVPR.2005.56.

[11] Daniel Jurafsky and James H. Martin. Hidden markov models. In *Speech and Language Processing*, pages 548–563. Prentice Hall PTR, USA, 1st edition, 2000. ISBN 0130950696.

[12] Hyeong Ryeol Kam, Sung-Ho Lee, Taejung Park, and Chang-Hun Kim. Rviz: a toolkit for real domain data visualization. *Telecommunication Systems*, 60:337–345, 2015.

[13] Adam Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990. ISBN:978-0521389389.

[14] N. Koenig and A. Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2149–2154 vol.3, 2004. doi: 10.1109/IROS.2004.1389727.

[15] Jenna Lebersfeld, Caleb Brasher, Christian Clesi, Carl Stevens Jr, Fred Biasini, and Maria Hopkins. 2157 the socially animated machine (sam) robot: A social skills intervention for

children with autism spectrum disorder. *Journal of Clinical and Translational Science*, 2:49–49, 06 2018. doi: 10.1017/cts.2018.190.

[16] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. Yarp: Yet another robot platform. *International Journal of Advanced Robotic Systems*, 3(1):8, 2006. doi: 10.5772/5761. URL `https://doi.org/10.5772/5761`.

[17] Grégoire Milliez. Buddy: A companion robot for the whole family. pages 40–40, 03 2018. ISBN 978-1-4503-5615-2. doi: 10.1145/3173386.3177839.

[18] Plinio Moreno, Ricardo Nunes, Rui Figueiredo, Ricardo Ferreira, Alexandre Bernardino, José Santos-Victor, Ricardo Beira, Luís Vargas, Duarte Aragão, and Miguel Aragão. Vizzy: A humanoid on wheels for assistive robotics. In *Robot 2015: Second Iberian Robotics Conference*, pages 17–28. Springer, Cham, 2015. doi: 10.1007/978-3-319-27146-0_2.

[19] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Dilip Kumar Limbu, Swee Lan See, and Alvin Hong Yee Wong. Socializing with olivia, the youngest robot receptionist outside the lab. In Shuzhi Sam Ge, Haizhou Li, John-John Cabibihan, and Yeow Kee Tan, editors, *Social Robotics*, pages 50–62, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-17248-9.

[20] Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, PP:1–1, 07 2018. doi: 10.1109/MRA.2018.2833157.

[21] Seho Park, Kunyoung Lee, Jae-A Lim, Hyunwoong Ko, Taehoon Kim, Jung-In Lee, Hakrim Kim, Seong-Jae Han, Jeong-Shim Kim, Soowon Park, Jun-Young Lee, and Eui Chul Lee. Differences in facial expressions between spontaneous and posed smiles: Automated method by action units and three-dimensional facial landmarks. *Sensors*, 20: 1199, 02 2020. doi: 10.3390/s20041199.

[22] Manoj Ramanathan, Nidhi Mishra, and Nadia Thalmann. *Nadine Humanoid Social Robotics Platform*, pages 490–496. 06 2019. ISBN 978-3-030-22513-1. doi: 10.1007/978-3-030-22514-8_49.

[23] E. Saad, J. Broekens, M. A. Neerincx, and K. V. Hindriks. Enthusiastic robots make better contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1094–1100, 2019. doi: 10.1109/IROS40897.2019.8967950.

[24] Satoru Satake, Takayuki Kanda, Dylan Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. A robot that approaches pedestrians. *Robotics, IEEE Transactions on*, 29:508–524, 04 2013. doi: 10.1109/TRO.2012.2226387.

[25] Karen Schmidt and Jeffrey Cohn. Dynamics of facial expression: Normative characteristics and individual differences. volume 0, 01 2001. doi: 10.1109/ICME.2001.1237778.

[26] Chao Shi, Satoru Satake, Takayuki Kanda, and Hiroshi Ishiguro. A robot that distributes flyers to pedestrians in a shopping mall. *International Journal of Social Robotics*, nov 2017. doi: 10.1007/s12369-017-0442-7.

[27] Stanford Artificial Intelligence Laboratory et al. Robotic operating system. URL `https://www.ros.org`.

[28] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13–13, 12 2011. ISSN 1534-7362. doi: 10.1167/11.5.13. URL `https://doi.org/10.1167/11.5.13`.