PALADYN Journal of Behavioral Robotics

VERSITA

Research article · DOI: 10.2478/s13230-011-0003-y · JBR · 1(3) · 20110 · 169-178

FaceBots: Steps Towards Enhanced Long-Term Human-Robot Interaction by Utilizing and Publishing Online Social Information

Nikolaos Mavridis*, Michael Petychakis, Alexandros Tsamakos, Panos Toulis, Shervin Emami, Wajahat Kazmi, Chandan Datta, Chiraz BenAbdelkader, Andry Tanoto

Interactive Robots and Media Lab, United Arab Emirates University, 17551 Al Ain, UAE

Received 30 August 2010

Accepted 28 January 2011

Abstract

The overarching goal of the FaceBots project is to support the achievement of sustainable long-term human-robot relationships through the creation of robots with face recognition and natural language capabilities, which exploit and publish online information, and especially social information available on Facebook, and which achieve two significant novelties. The underlying experimental hypothesis is that such relationships can be significantly enhanced if the human and the robot are gradually creating a pool of episodic memories that they can co-refer to ("shared memories"), and if they are both embedded in a social web of other humans and robots they mutually know ("shared friends"). We present a description of system architecture, as well as important concrete results regarding face recognition and transferability of training, with training and testing sets coming from either one or a combination of two sources: an onboard camera which can provide sequences of images, as well as facebook-derived photos. Furthermore, early interaction-related results are presented, and evaluation methodologies as well as interesting extensions are discussed.

Keywords

Human-Robot Interaction · social networks · conversational robots · face recognition · long-term HRI

1. Introduction

The overarching goal of the FaceBots project is that of the creation of sustainable and meaningful long-term human robot relationships. This is a most important goal towards the further-reaching ultimate goal of human-robot symbiosis, i.e. harmonious and beneficial integration of robots in everyday human life. Taking a narrower but more directly applicable view, the creation of such sustainable relationships could prove highly beneficial towards the successful application of robots to numerous areas: disabled and elderly assistance, companion robots, tutor-robots etc.

But why could sustainable relationships be beneficial for such applications? This is an important question (Q1), which is also logically followed by (Q2): *How can one achieve long-term sustainable relationships*?, i.e. what aspects of the human-robot interaction design, as well as robot behaviors and form, could help towards creating such relationships?

A plausible account of a line of argumentation answering question Q1 is provided in the Appendix of this paper, and some relevant empirical evidence follows. Regarding question Q2, let us start by considering the state-of-the-art, and then by introducing the main hypothesis of FaceBots towards answering this question and thus providing aid towards creating long-term human-robot relationships.

Thus, let us try to examine the following question (Q3): *How much have we advanced towards achieving such sustainable long-term relationships with robots?* So far, empirical investigations have shown that currently we have minimal advancement towards fulfilling general requirements for the creation of such relationships: Although existing robotic systems are interesting to interact with in the short term, and novelty effects are exhibited, it has been shown that after some weeks of quasi-regular encounters, humans gradually lose their interest, and meaningful long-term human-robot relationships are not established. A solid example of this phenomenon is the case of Robovie [1], for which there was a steady and significant decrease in the total time of interaction of the robot with humans over six months - interest had worn off. Few further long-term interaction studies exist; notable exceptions include Kidd and Breazeal [2], and Sung, Christensen, and Grinte [3].

In [2], weight-loss coaching conversational robots were given to subjects and interacted with them over a period of six weeks, and were actually shown to be effective towards aiding with weight loss, i.e. were able to elicit a behavioral change through their interactions. Apart from the much more limited time-horizon as compared to [1], there were two more important differences in this experiment: First, the robots were interacting with a single person, and not with a wider social circle, as was the case of Robovie and the employees in the offices of ATR in [1]. Second, there was a clear shared goal driving the human-robot interactions in [2], namely weight loss of the human, while in [1], the interactions were more casual, and without a predefined explicit goal.

In the case of [3], instead of a conversational mobile humanoid (as is the case of Robovie in [1]), the 6-month total duration study is concerned with Roomba floor-cleaning robots: non-conversational, and highly non-anthropomorphic in both appearance as well as, arguably, behavior. The paper is mainly concerned with methodological aspects of performing such studies in homes. It is interesting to note that even in this case of highly non-anthropomorphic robots, over long-term interactions, people gave human names to them and were often describing them in agentive, and not passive-object only terms.

Thus, having briefly mentioned the need for supporting the creation

*E-mail: nmav@alum.mit.edu

of sustainable long-term relationships between robots and humans, in a variety of application areas (with supporting argumentation in the Appendix), and discussed key papers on the state-of-the-art in long-term human robot relationship studies, it is time to revisit (Q2): What is required in order for such relationships to be achieved? According to the argumentation in the Appendix, one plausible starting point is to consider long-term human-human relationships. But then, the further question follows: from all the expectations, habits, and intricacies of long-term human-human interactions, where should one start from when endowing robots with such capabilities?

Our proposed solution to the problem of creating sustainable and meaningful long-term human robot relationships is based on the following underlying hypothesis which is central to the FaceBots projects: *Such relationships can be significantly enhanced if the human and the robot are gradually creating a pool of shared episodic memories that they can co-refer to ("shared memories"), and if they are both embedded in a social web of other humans and robots they both know and encounter ("shared friends").*

This is the starting point that we have chosen. Of course, the notion of "shared intersection" can be generalized, for example to goals and preferences; however, in FaceBots we are starting with shared memories and friends.

Thus, here we present a conversational mobile robot with face recognition, which apart from having an onboard social- as well as interaction-database, is also connected in real-time to Facebook, a highly successful online networking resource for humans, towards enhancing long-term human robot relationships, by helping to address the above two prerequisites. The system, apart from aiming towards enhancing long-term relationships, also achieves many tangential side-gains, elaborated in the discussion section.

Furthermore, and quite importantly regarding the contributions of this paper, a *lengthy description and evaluation of our attempts towards dual-source face recognition* in FaceBots cover a big part of the presented results, and include important insights transferrable to other similar situations, in which there is both an on-board camera on a robot providing multiple images as well as the possibility of using partially tagged images derived from the internet (in our case, from the facebook website) towards face recognition. The main question addressed here is that of transferring training from one source across to testing in the other, and of the potential benefits of combining training sets across sources, also for the most important practically case of small data sets with few snapshots available.

Most importantly, it is worth mentioning that the FaceBots social robot *achieves two important novelties*: being the first such robot that is embedded in a social web, and being the first robot that can purposefully exploit and create social information that is available online. Furthermore, it is expected to provide empirical support for our main driving hypothesis, that the formation of shared episodic memories within a social web can lead to more meaningful long-term human-robot relationships. The experience gained by the creation of such a system as well as the software created is invaluable towards providing similar capabilities to other robots, and as a starting point for further enhancements of robots truly embedded in a social web that use and create online social information. Finally, the exposure of the robot to Facebook, through the public availability of its own Facebook page containing its friends and experiences as well as photos, will create public interest that will further support endeavors to similar directions in the future.

The structure of the remaining paper that follows, proceeds along the following lines: Having briefly mentioned some key long-term human-robot interaction studies, in this introduction section, fundamental background is also provided for the wider supporting area of interactive social robots with conversational abilities, as well as for face-recognition and robots, and social networks and robots, in the next section. Then, in Section 3, we proceed with a thorough description of the system architecture, the software modules, and their functionality and implementation, and in Section 4 we discuss the operation of the robot through a real-world interaction example (videos of the robot in operation can be found in YouTube channel irmluaeu). Then, in Section 5, we discuss and present extensive concrete results regarding evaluation and dual-source face recognition with FaceBots, followed by a discussion and extensions in 6, and a concluding section.

2. Further Related Research

Although numerous attempts towards *interactive social robots* have taken place (such as MIT's Kismet [4] and Leonardo [5], the Maggie [6] robot, ATR's Robovies [7] and recently many more), no existing systems have utilized a connection between robots and Facebook.

However, face detection has a long history as a field, and furthermore, *face-detecting conversational robots* are not new; there are numerous projects built-around face-detecting robots [8, 9], which might even carry out conversations with multiple humans such as in [10].

Regarding the sustainability of long-term *human-robot relationships*, the technical and experimental difficulties involved in creating a larger-scale field trial have prohibited the realization of such experiments until quite recently. A key long-term (six month) study is [1], as discussed in the introduction section. Shorter field studies in other contexts have taken place in the past; for example, the 18-day field trial of conversational robots in a Japanese elementary school [11]; as well as a 9-week field study in a classroom is reported in [12], in which, in a similar fashion to [1], novelty effects were originally exhibited, but were quickly followed with a significant decrease in interest over time. Numerous other field trials are underway, including a possible massive deployment of humanoids in malls [Ishiguro, personal communication].

Finally, regarding the real-time *utilization of web resources* by robots, much has not been done yet, but exciting prospects exist; for example, "Peekaboom" [13], could serve as a real-time repository for object recognition. Regarding depositories for robot-related data primarily for offline use, a notable exception the EU project RoboEarth [14].

3. System Architecture

3.1. Hardware

Our robot is composed of an ActivMedia PeopleBot robot [15], augmented with a SICK laser range finder, a touch screen, and a stereo Bumblebee camera [16] on a pan-tilt base [17] that is at eye-level with humans

3.2. Software Architecture

We have created an expandable modular software architecture, with modules intercommunicating through the ICE IPC system [18]. The modules can be running on multiple CPUs or PCs which are part of a network, and are written in C++, Java, and Perl. Effectively, a callable-method API is exposed by each module towards the others. The modules we have created are:

(M1) *Vision Module with Face Detection & Recognition*, from camera- or Facebook- derived pictures. Includes real-time externally callable training set modification / new classifier generation capabilities, and pluggable face detectors / classifiers. (M2) *Natural Language Dialogue Module*, with real-time language model switching

capabilities. (M3) *Social Database Module*, which locally holds basic personal info / friendship relationship / simple event data / photos for the people the robot knows, and which connects and updates through Facebook for those that are members of it. (M4) *Navigation and Motion Module*, to build a map of its environment and drive to key social locations, and (M5) *Controller Module*, which issues calls to all other modules, and where high-level system operation routines can easily be scripted. A more detailed description of the modules follows.

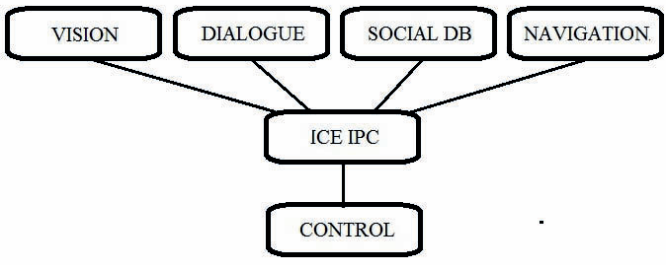

Figure 1. *Modules intercommunicating using ICE IPC.*

3.3. Vision Module

The purpose of the vision module is to detect human faces and recognize their identity, either using images fed by be the robot's onboard camera, or using Facebook-derived photos, which are passed to it through the social database module. The module is written in C++. Its API provides externally callable methods for doing the following online tasks:

- Create/add/remove face training set (for given person);
- Create/remove/retrain face classifier (for given person);
- Export a camera snapshot or detected face regions;
- Export recognized identities (score vector or hard decision);
- Tagged Image handling (for Facebook-imported photos);
- Set/query operational parameters.

1) *Capture:* At the front end of the module, there is a camera-capture connection that first looks for the onboard stereo camera, and if this is not found, looks for a remote stereo camera coming through an ICE feed.
2) *Face Detection & Preprocessing:* Face detection follows after the capture step. This module is based on the OpenCV implementation of [19]. The detected rectangular regions which are candidate faces, are then passed through a simple skin color detection algorithm. This algorithm is based on an RGB-triplet inequality, which classifies each pixel as either skin-color or non-skin. Then, the percentage of skin color pixels in the candidate face region is calculated, and if the percentage is less than the experimentally chosen threshold of 20%, the candidate face region is discarded. Once a candidate face region passes the skin filtering test, an elliptical mask is applied to the region in order to remove irrelevant background clutter, followed by brightness normalization. The latter consists of histogram equalization and normalizing pixel values to have a zero mean and unit standard deviation [20]. At the end of these multiple processing phases, an elliptical face region is obtained.
3) *Face Recognition & Temporal Evidence Accumulation:* The elliptical face region is then fed to an embedded-HMM-based array of classifiers [21], one classifier for each person in our face database. The output of each of the embedded-HMM classifiers is a scaled log-probability measure of how well the currently viewed face matches the classifiers model. This is furthermore accumulated through a moving-window-average running across multiple frames. The choice of the number of frames (window size) and the evidence accumulation method is discussed in the tuning and evaluation section of this paper. Then, the variance of the temporally-accumulated vector of scores is calculated, and if it is below a threshold (i.e. if all the classifiers are almost equally confident about the identity of the face), then the face is marked as unknown. The choice of this threshold will also be discussed later, in the tuning and evaluation section.
4) *Facebook Photo Handling:* Facebook photos often come pre-tagged. However, the tags supplied contain a name augmented with a rough estimated center location, and not with a rectangular bounding box. Therefore, for a pre-tagged face, face detection is still applied, and the detected face region whose center is closest to the rough estimated center reported by Facebook is chosen as the bounding rectangle. One further complication arises because the user-tagged faces in Facebook photos might be either in frontal or in profile poses. However, we currently cannot recognize profile faces but only frontal, and thus we need to make sure that a user-tagged face in profile pose will not be ignored while a nearby frontal face is detected. In this unfortunate but quite usual case, the bounding rectangle of a nearby face of another person might be reported from our module. Even worse, it might subsequently be used as a training set picture for the classifier of the user-tagged person, while it belongs to somebody else. Therefore, we perform face detection with two face detectors in parallel: a profile-tuned detector as well as a frontal-tuned detector (while, as we mentioned before, we only have frontal face, and not profile, face recognizers). If the nearest face to the user-supplied rough center is in profile pose, then it is discarded, and so even if the second-nearest might be frontal, it is not reported. However, if the nearest face to the user-supplied rough center is in frontal pose, then it is reported, and thus can be safely used as a training set picture.
5) *Training sets:* One of the interesting novelties of our system is that it has access to two groups of pictures - coming either from Facebook photos (fully tagged, partially tagged, or untagged), or from live or stored camera pictures. The important question of appropriate training set selection, pruning, and retraining is further discussed in the tuning and evaluation section.

3.4. Social Database Module

The purpose of the natural language dialogue-support module is to provide speech recognition, speech synthesis, as well as basic NLP services. Speech recognition is based on Sphinx 4 [22], and language models can be switched during operation. Speech synthesis is based on the Cepstral text-to- speech system that is part of the ARIA SDK [23] of the Peoplebot robot. The module is written in a mix of Java and C++ languages.

3.5. Natural Language Dialogue Module

The purpose of the social database module is to locally store relevant social information for the friends of the robot, and to perform acquisition and deposition of social information from Facebook. The module contains two databases: the social database and the interaction database. The social database contains entries both for people that the robot encounters which are on Facebook (Facebook friends), as well as for

people that the robot encounters which are not (non-Facebook friends). Encounters can be either physical (face-to-face) or virtual (online). Furthermore, the module also contains an interaction database, storing important past interactions that can be referred to. The ultimate purpose of the social information obtained and deposited is to enable meaningful and interesting interactions between the robot and its friends, and we will soon elaborate more on how this is achieved. The database is implemented in MySQL, and the module uses interoperable Java and Perl objects, and also utilizes the Facebook API.

1) *Social Database Structure*: This module is essentially a friend database, which includes a subset of the Facebook available information. It contains fields for general information (affiliation, current location, education, high school, hometown location, work history), a friend list (containing the list of friends of the robot's friend whose entry we are describing), an event list (as posted on Facebook), and a set of photos possibly with tags.

2) *Interaction Database Structure*: This module is essentially a form of an episodic memory of the robot, whose primary key is a timestamp. Every session of interactions with a specific friend has a unique ID. Also, other fields include an interaction type identifier and a description, as well as a number of boolean flags, and the userID of the friend the robot is interacting with.

3) *The Social Module API*: methods for accessing the social database (and Facebook) include general information reads / writes, friend-list derived relation queries (including some higher-order queries which are however constrained by privacy / visibility settings), event-list queries as well as methods for getting / depositing possibly tagged photos. On the other hand, there are also methods for direct messaging through Facebook, and for accessing the interaction database. The former include methods for messaging and chatting, while the latter include methods for depositing and selectively retrieving past interactions (including for example past responses of the user to the robot's questions)

3.6. Navigation and Motion Module

The purpose of this module is to provide basic navigation, mapping, as well as motion capabilities to our system. The module is written in C++, and is using the ARIA navigation library (ARNL). Its API currently provides externally callable methods for:

- Start/Stop spatial exploration;
- Report current position / Go To target position;
- Move camera pan-tilt unit.

3.7. Controller Module

The purpose of the central controller module is to orchestrate the operation of the Facebot by issuing calls to the other modules. Within it, high-level system routines can easily be scripted, in a form that hides most of the module-specific implementation details.

4. System Operation

The most important underlying question regarding the design of the high-level system operation (scripted within the main controller module) is: how can one appropriately utilize social information and interaction history information in order to create interesting interactions? Here, a number of possibilities were initially explored, starting with single initiative robot driven dialogue. An actual demo example follows:

Greetings / Recognition:
H:Hello! My name is Sarah. Welcome to the interactive robots and media laboratory.
R:Hi! Are you Panos Toulis?
H:Yes.
(Robot changes its online status to ''interacting with Panos Toulis'')

Query of overall state:
R: Hey, Panos Toulis, are you doing well today?
H: Yes, pretty well.

General news-based discussion
R: Let me tell you some interesting news that I've heard.
R: The latest windows release will be available to everyone after a surge in demand crashed the Microsoft website on the 9th of January.
(Item taken from BBC news online)
R: Have you heard about this?
H: Yes.
R: That's great!

Robot discusses Facebook status of human
R: I see you are chasing robots in the desert.

Robot gives news about a mutual friend, acquired through Facebook
R: Did you know that our friend Shervin Emami has posted a new photo on Facebook?
(Information acquired from Facebook mini-feed)
H: No.

Robot sends reminder to human
R: I am sending you a message about this to check it out.

Robot remembers past encounter with mutual friend
R: I saw Shervin yesterday evening. He was doing well.
(Information acquired from interaction database)

Robot gives news about a mutual friend
H: Did you know that our friend Chandan is ''waiting for Sarah to have that little sweet conversation with me''?
(Information acquired through Chandan's Facebook status)
R: Yes, of course.

Robot connects an online friend with the human
R: One of our friends, Nikolaos Mavridis, is online -- would you like me to send him a message?
H: Yes.
(`Nikolaos, Panos says hello'' is sent)
End of interaction
R: Hey, Panos, it was nice talking to you!
R: I have to go now.
R: See you later!

A generic framework systematizing the utilization of social information and structuring the dialogues is under development. The overall form of the actual demo session of which was analyzed above is the following: The robot initially wanders around our lab, avoiding obstacles, and occasionally speaking to itself. When a human face is detected through the vision system, an attempt towards recognition is made, and if there is enough confidence about identity, the robot greets the human and asks if he is indeed the person the robot has recognized. If not, the second choice is announced, and a verification question is given again.

Then, some pictures are taken, which are added to the training set of the appropriate classifier: either of the already known recognized person, or of a new classifier in case of a new person, who is also asked about his name. In case of a new person, there is an attempt to find social information about him/her through Facebook, if he/she already is a member. For example, if the new person according to Facebook is a friend of an already known friend of the robot, then this is announced and indirectly asked for confirmation. In the case of an already known person appearing before the robot and being saluted and recognized correctly, a mix of the basic dialogue steps exemplified by the above transcript is utilized. For example, status changes of mutual friends are discussed, news items announced, reference is made to possible meetings with common friends in the mean time or to previous meetings with the person, instant messaging to other online friends mediated by the robot takes place etc. During the interaction, information is also posted on the robot's Facebook page. Also, some pre-scripted segments of dialogue, containing announcements or jokes, can embellish the conversations. Finally, the robot says goodbye and continues its wandering. An earlier demo of the robot is already published as a video at the HRI 2009 conference.

5. Evaluation and Dual-Source Face Recognition

In such a complex system, there exist multiple parameters that need to be tuned as well as many discrete design choices to be taken. Furthermore, numerous types of evaluation can be carried out. Some of these are:

T1) *Module-level evaluations*: What is the performance of the vision system, of speech recognition etc. when viewed as isolated modules.

T2) *System-level evaluations*: Engineering metrics implicating more than one module; for example delay between appearance of face and announcement of greeting, crossing across the vision, main controller, and speech modules.

T3) *Task-centered evaluations*: Successful task completion rates, statistics of task duration, task-performance metrics, quantitative error analysis and error taxonomies etc.

T4) *User-centered evaluations*: Self-reported or externally measured, including user satisfaction, ease-of-use ratings, expectations etc.

T5) *Human-replacement-oriented evaluations*: How close were the actions (speech and motor) taken by the robot to those of a human when put in the same situational context.

T6) *Long-term field trials*: For example, measurements of frequency, duration, and content of interactions during a multi-month operational deployment.

Here we will present some first results belonging mainly to the first, second and third type - in decreasing order of extent. A long-term field trial is planned in the mid-term future, once some further extensions have taken place. The purpose of this field trial will be to provide concrete evidence for our main underlying experimental hypothesis: that human-robot relationships can be significantly enhanced if the human and the robot are gradually creating a pool of shared episodic memories that they can co-refer to (''shared memories''), and if they are both embedded in a social web of other humans and robots they both know and encounter (''shared friends''). Also, notice that for the case of an ongoing project where additions and modifications are still taking place, the evaluations carried out often also function as tuning sessions for parameters or design choices.

A considerable amount of effort was directed towards questions related to our vision system. A first choice that had to be made was *the choice of an appropriate threshold of variance across classifier scores in order to decide that a face should be classified as "unknown*'', as well as a minimally acceptable winning match score. The underlying assumption here is that an ''unknown'' face should not create a clear winner among our classifiers, and that even if it does, the corresponding score will be low. The appropriate variance value that was chosen was 1.2 (Figure 2).

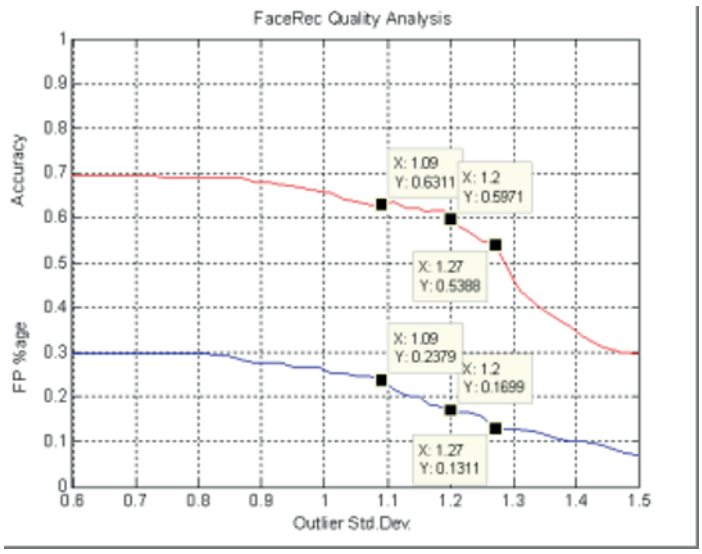

Figure 2. Effect of varying outlier standard deviation on recognition accuracy and false positive percentage.

A second very interesting question is: *"over how many frames (window size) should we accumulate evidence before deciding upon the identity of a face?"*, and also *"at what stage and through what accumulation process?"*. For the latter, after some initial experimentation we decided to accumulate at the level of the continuous scores of the classifiers (before choosing a discrete ''winner''), and to do so with a fixed-size window equal-weight averaging. For the former, we performed empirical tuning, by varying window sizes, and looking at the changes of recognition accuracy. Of course, there is a practical limit to the number of frames; the human does not wait for too long. In practice, a camera-derived training set for five individuals, with a duration of 100 frames was acquired within our lab, and also two testing sets, each of 100 frames, one within the lab (easier), and one outside, where lighting conditions and background differ (more difficult). Then, the window size was varied, and overall accuracy graphs plotted (Figure 3). From the results, it became clear that after 25 frames or so, corresponding to 5 seconds at 5 frames per second (an acceptable exposure time), there was no more significant increase in accuracy to be expected.

A third important question is concerned with the *size of the training set*. If there is no option for incremental retraining (which is the case

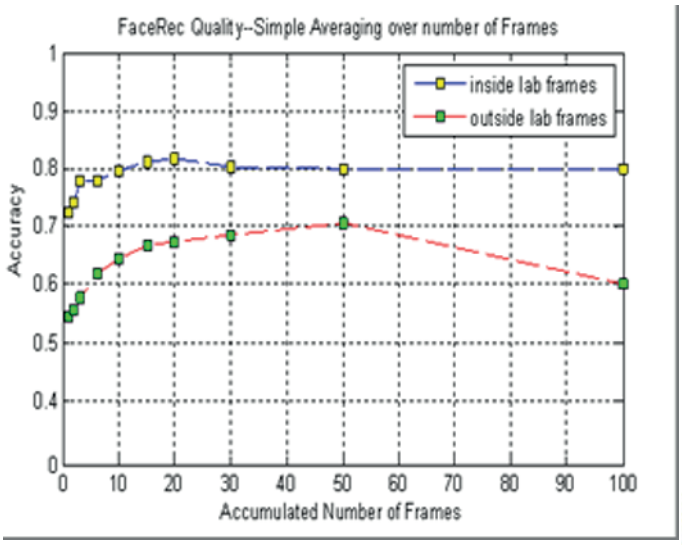

Figure 3. *Recognition Accuracy as a function of evidence accumulation over time (number of frames at 5fps).*

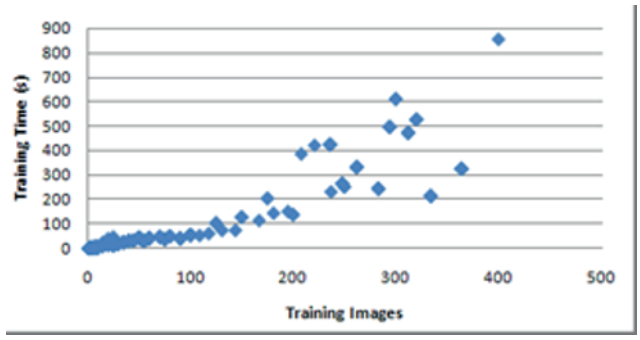

Figure 4. *Time taken for training a classifier.*

Training Set:	Testing Set: Camera (30)	Facebook (30)	Cam+FB (60)
Camera (30)	98.9	46.7	72.8
Facebook (30)	47.8	78.9	63.3
Cam+FB (30)	96.7	78.9	87.7
Cam+FB (60)	94.4	80	87.2

Figure 5. *Transferability of training from Facebook pictures to camera photos and vice-versa: Recognition Accuracy over different combinations of training and testing sets. Camera training set acquired across one month.*

in certain recognition methods), and if offline retraining does not take place during idle periods, then one needs to keep retraining time to reasonable lengths. Also, a larger training set is not necessarily a better one; the mid-term variations of human faces (facial hair, movement of light sources) would require more emphasis to be put on the recent shots of the face as compared to older ones. Furthermore, pruning of outliers is another idea we are exploring. Regarding the question of training time versus set size, empirical results are shown in Figure 4. Thus, 30 or so is the maximal tolerable size for quasi-real-time online retraining (20 seconds), while figures as high as 400 seem to be acceptable for offline (during idle periods), possibly also with a fixed CPU time-slice allocation for backgrounding (15 minutes at 100 percent CPU utilization, one hour at 25 percent - possibly multiplied by the number of people whose classifiers need to be updated).

Now, let us examine the fourth, and most intriguing question. This is more of a system-level question, as it requires interoperation of multiple modules. Having access to both live as well as stored camera pictures, but also to potentially partially or fully tagged Facebook photos, creates many interesting possibilities for using one or the other or a mix for training, and then transferring the knowledge to testing to any of the three species (for automating tagging for example- note that social information can also be utilized in addition towards that purpose, as discussed later in this paper, in the Discussion section). Thus, a crucial question that was asked is: *how well do either Facebook photos or camera pictures function as a training set, when they are tested on Facebook photos or camera pictures or across / in a mix?*

An evaluation was carried out with five persons, and using two 30-picture per person training sets (one from Facebook, and one from the robot vision system detected face regions), and two 30- picture per person disjoint testing sets (again one from each source). The results can be seen in the two across-set accuracy matrices: the difference between that two is that the first one, shown in Figure 5, uses a camera training set containing photos from five sessions spaced over a month (i.e. has been friend for a while), while the second (Figure 6) contains a training set containing frames from a single encounter (new friend, have just met once).

The results are quite interesting. *First,* as expected, the temporally spread camera training set (Figure 5) is much stronger (produces much more accurate recognition) than the single session-derived set (Figure 6). *Second,* in the case of the combined camera plus Facebook training sets, one can see that increasing the size of the set from 30 to 60 does not necessarily produce better accuracy - i.e. 30 is adequate (compare row 3 with row 4 in each of the two figures). *Third*, the cross-combinations (training by camera and testing on Facebook and vice-versa) produce unacceptable results. Camera to Facebook gives 46% and 31%, while Facebook to camera gives 48% and 49% - which is noticeably a little better. One could try to explain this asymmetry on the basis of the greater variance across Facebook pictures, which when used as a training set reduces over-fitting. On the other hand, Facebook to Facebook performs tolerably (80%), and so does camera to camera after a single session. At the top level, we have the camera to camera combination for the case of the spread-over-the-

Training Set:	Testing Set: Camera (30)	Facebook (30)	Cam+FB (60)
Camera (30)	76.7	31.1	53.8
Facebook (30)	48.9	78.9	63.8
Cam+FB (30)	75.6	77.8	76.6
Cam+FB (60)	73.3	74.4	73.8

Figure 6. *Transferability of training from Facebook pictures to camera photos and vice-versa: Recognition Accuracy over different combinations of training and testing sets. Camera training set acquired from a single session.*

month camera training set, with a high 98%. *Fourth,* it is worth noting that although on their own the Facebook-only and camera-only sets do not cross-generalize, if they are used in conjunction (i.e. the third and fourth rows), then they always produce better results than alone in the across-case, and do not significantly deteriorate recognition in the same-species case (98% falls to 97, and 76 to 75 etc.).

Finally, notice that at least for the case of real-time camera-shot recognition, errors can be corrected by appropriate verbal feedback by the user; and also, quick experimentation showed that in most cases, in the case of an error, the second-best choice of our system is correct. Thus: "R: Hello! Are you John? H: No R: Oh sorry! I misrecognized you. You are George, right?" is a viable option as it is not so disturbing for your friend to be misrecognized sometimes, as long as he is a new friend (case of Figure 6 - for a new friend, the robot only has a single session of face training data), and as long as your second guess is correct.

Finally, two *initial task-level evaluations* were carried out: In the first one, the robot physically interacted with five people, each one of which for four times. The interactions were videotaped, and automated logs were taken. The duration of the interactions is on the order of 125-145 seconds, during which 8-10 conversational turns took place. In the second, the robot did not interact physically, but through Facebook chat, with people that had befriended it, after an alpha-test opening to received invitations. During the week of the initial test, the robot had 197 Facebook friends. The average duration of the interactions was again on the order of 150 seconds, the robot was switched on for 116 hours (almost 5 full days), and 167 interactions took place. 70 out of the 197 friends interacted with the robot during this period (35%), the median number of interactions per user was 2. The interaction schedule of this initial week is shown below. Further data for a comparison across two versions (with and without shared memories and friends) will be collected in the near future, after accepting more Facebook friends, in parallel to physical interactions.

	W	R	F	S	U	M
00		4		3		4
01		5				2
02		6	4	1		2
03		3	2	1	3	
04		1	3			2
05		2	1			2
06		2				
07			1			1
08						X
09				1		
10				1		
11				1		
12	X				6	
13	4	4	4			
14	1	3	3			
15	3	2			3	
16	5	3		2		
17		2	2	2	5	
18		3		2	3	
19	3	1	1	4		
20	8	1	2	3		
21	8			1		
22		1	1	1	1	
23	7			2	2	

Figure 7. *Interactions through Facebook chat for 6 days.*

6. Discussion And Extensions

Multiple Extensions are currently underway:

E1) Extensions of the main controller-scripted basic cycle are undergoing testing and further development.
E2) Corresponding language models in order to support the various stages of the robot-driven dialogue are being created. Also, the possibility of supporting some human-initiative or mixed-initiative dialogue turns is considered.
E3) Different ways to utilize the available social information in the form of verbal interactions are being thought out, as well as ways to implement the acquisition of such information through questions.
E4) Increased exploitation of the Facebook messaging and chat channels for verbal interactions is underway, including the possibility of the robot sustaining a conversation with more than one friend at the same time: one face-to-face physical, and one possibly more than one over Facebook.
E5) The social database and interaction database are undergoing redesign aiming towards compactness and functionality enhancement.
E6) Experiments regarding the periodic retraining of classifiers in order to account for facial appearance changes and training set pruning / augmentations are taking place.
E7) A form of basic "active sampling" technique for acquisition of multiple face poses through intentional movement of the pan-tilt of the robot's camera and/or robot body movement is being examined.
E8) The utilization of other online resources apart from Facebook towards driving dialogues and enhancing interactions is being examined.
E9) The possibility of using social-information-driven dialogue as a mode of dialogue, existing alongside other modes (for example, dialogue about the physical situational context, along the lines of [9]), and better integration with situation model theory.
E10) The whole system is being moved to a different embodiment: IbnSina ([24, 25]), our humanlike humanoid robot (Figure [8]), which supports facial expression, hand gestures, and more, and is part of a unique interactive theatre installation supporting multiple forms of teleparticipation ([26, 27]) . Extension to a second language is also underway.

Figure 8. *IbnSina Android Robot at the IRML Lab, UAEU.*

Furthermore, another extension direction deserves special attention. There exists a possibility for utilizing friendship information, in order to enhance automated tagging in Facebook pictures. The underlying assumption is that friends are more likely to co-occur in photos - thus we can start biasing our recognition hypothesis set towards friends, once we know the identity of a person in a photo. The process goes as follows: suppose we know the identity of a person, either through recognition, or through pre-tagging, and that we are quite confident of it. Then, we acquire his circle of friends through the social database, and we bias our hypothesis space (bigger priors, larger score weight etc.) towards the circle of friends. Then, we recognize the other faces, and choose the one whose identity we are most confident of. Now, we have two circles of friends: the first face's friends (F1), and the second (F2). We also have their intersection: their mutual friends (F1&2). Thus, we can now bias with three levels of strength: small weight for non-friends (not belonging to either F1 or F2), large weight for mutual friends (F1&2), and intermediate weight for friends which are not mutual. Implementation details as well as results can be found at [28].

Finally, it is worth noticing that the correlation between friendship and co-occurrence in pictures can be utilized in the inverse way too; once we have seen two people co-occuring in multiple photos, it is quite likely that they might be Facebook friends. This idea has been partially utilized for example in the "click-expansion" option of Touchgraph [29], and can provide an indirect way for the robot to have a starting set of hypothesis in order to ask questions to people about their friends.

Now, after having discussed extensions that are underway, we will take a higher-level viewpoint, and discuss where this project fits within a bigger picture. First, this is arguably the first example of *utilization and publishing of online information deposited by non-expert humans by an interactive conversational robot*, and we foresee a wide array of prospects arising through this stance. Second, since multiple "FaceBots" can share social information among themselves, and can "switch embodiments", effectively creating a single identity with multiple distributed embodiments, this creates the prospect for an *ultra-social robotic being*, which might have a circle of friends much wider than a usual human, and such an entity could for example be used in order to increase social capital by creating new connections among humans, and interfering in multiple ways in the information diffusion as well as structure of the social networks which it bridges. For more details on the potential applications of physical or virtual agents within human social networks, the reader is referred to [30]. Third, again moving across the spectrum of mobility and physicality, although here we are presenting a mobile robot which can explore physical space and encounter humans, one could easily port a part of the system's functionality to an entity having *possibly a virtual body but connected to a physical camera and speech subsystems*, effectively remaining stationary in physical space, or anyway just being human-transported. This possibility would also enable a much wider deployment of the system in the near future, which would contribute towards the acquisition of larger training sets and cumulative experience of interactions for analysis.

7. Conclusion

Towards sustainable long-term human-robot relationships, a mobile robot with vision, a dialogue system, a social database and a Facebook connection was created, which achieves two important novelties: being the first such robot that is embedded in a social web, and being the first robot that can purposefully exploit and create social information that is available online. The main hypothesis towards achieving sustainable long-term human relationships which underlies the creation of our robot is that such relationships can be significantly enhanced if the human and the robot are gradually creating a pool of shared episodic memories that they can co-refer to ("shared memories"), and if they are both embedded in a social web of other humans and robots they both know and encounter ("shared friends").

We present an extensive description of the architecture of the robot, as well as important concrete results regarding face recognition and transferability of training for face recognition, with training and testing sets coming from either or a combination of two sources: an onboard camera which can provide sequences of images, as well as Facebook-derived photos, either pre-tagged or untagged. Furthermore, early interaction-related results were presented, and evaluation methodologies as well as multiple interesting extensions, also positioning the project in a wider context, were discussed.

Through the FaceBots project, by helping towards creating sustainable long-term human-robot relationships, we believe that the worthy ultimate goal of harmonious human-robot symbiosis has come a step closer to its realization.

Acknowledgment

We would like to thank Microsoft External Research for supporting this project through its Social Robotics CFP, as well as all the members, students, and friends of the Interactive Robots and Media Lab that have helped, each in their own way, towards the realization of the FaceBots project.

Appendix

A plausible account for a chain of argumentation supporting the claim that: "The creation of sustainable relationships between humans and robots could prove highly beneficial towards the successful application of robots to numerous areas, such as disabled and elderly assistance, companion robots, tutor-robots" is provided here. Apart from the logical argumentation, some relevant empirical evidence was also covered in the introduction and background section.

The basic premises of the argument are:

(S1): More natural and unobtrusive interaction with a robot, can aid towards diminishing human adaptation to the robot, and towards getting closer to transparency-in-use, which would subsequently lead to better performance towards the goal of the specific application the robot is aiming towards.

(S2): In the mind of humans, depending on the observed or expected capabilities of the device they are interacting with, different kinds of mental models of the device are construed.

(S3): If the device that the human is interacting with affords natural-language communication as well as demonstrates partially anthropomorphic appearance and behavior, then it will be construed with an agentive mental model (with beliefs, intentions, affect etc.), and not a passive-object mental model, and expectations of interactions with it will follow expectations of interactions with humans in a similar role as the robot is playing.

(S4): In applications where the encounter between the human and the robot is not one-off, but multiple encounters with a robot take place over a longer period of time, extra expectations arise for the robot, as well as requirements for the interactions it can afford, which must be partially fulfilled in order to achieve natural and unobtrusive interaction with the human in the long-term, and stronger sustainment.

(S5): These extra requirements in order to be able to achieve sustainable long-term interaction, are again based on expectations of long-term human-human relationships, for the case of robots that have been construed in the minds of their human partners through human-like agentive mental model.

Thus, the argument follows:
In applications where the robot is aiming towards companionship, assistantship of disabled or elderly etc., the natural preconceptions of humans regarding the role it must play is heavily shaped by the role that other humans would play, if they were in the position of the robot. Thus, for short-term interactions, it is to start with desirable to have (S3) hold; i.e. to have robots with natural-language and partial anthropomorphicity in behavior and form, so that they are construed with an agentive mental model by the humans they are interacting with, so that (S1) can naturally arise, and thus short-term performance for the application under investigation can be achieved.
Specifically, for the case of companionship, assistantship of disabled or elderly etc., there is usually a need not only for one-off, but for multiple encounters with a robot during a longer time period. Thus (S4) applies, and as the robots, as argued before, due to their role, would better support natural language and partial anthropomorphicity, they would probably be construed with mental models of agentive entities in the mind of humans. Thus, (S5) applies, and requirements arising from human-human long-term relationships become important for the robots. When such requirements are fulfilled, long-term relationships with the robot can be achieved, in a natural and unobtrusive manner; and thus (S1) holds not only for the short-term, but also for the long-term.
And thus, through the above chain of argumentation, arriving at (S1) for both the short- as well as the long-term, it logically follows that such sustainable relationships can be beneficial for such applications, thus providing a plausible explanation which answers questions (Q1).

References

[1] N. Mitsunaga, T. Miyashita, H. Ishiguro, K. Kogure, and N. Hagita, Robovie-IV: A Communication Robot Interacting with People Daily in an Office, In Proceedings of IROS 2006, p. 5066-5072.

[2] C.D. Kidd, C. Breazeal, Robots at home: Understanding long-term human-robot interaction, In proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 22-26 Sept. 2008, pp 3230-3235.

[3] J. Sung, H.I. Christensen, and R.E. Grinter, Robots in the Wild: Understanding Long-Term Use. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (HRI '09). San Jose, USA, 2009.

[4] C. Breazeal, Emotion and sociable humanoid robots, International Journal of Human-Computer Studies, Volume 59, Issues 1-2, Applications of Affective Computing in Human-Computer Interaction, July 2003, Pages 119-155, ISSN 1071-5819, DOI: 10.1016/S1071-5819(03)00018-1.

[5] A.G. Brooks, J. Gray, G. Hoffman, A. Lockerd, H. Lee, and C. Breazeal, Robot's play: interactive games with sociable machines. Computer Entertainment. 2, 3 (Jul. 2004), 10-10.

[6] M.A. Salichs, R. Barber, A.M. Khamis, M. Malfaz, J.F. Gorostiza, R. Pacheco; R. Rivas, A. Corrales, E. Delgado. Maggie: A Robotic Platform for Human-Robot Social Interaction. IEEE International Conference on Robotics, Automation and Mechatronics (RAM 2006). Bangkok. Thailand. Jun, 2006.

[7] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, R. Nakatsu, Robovie: an interactive humanoid robot. Industrial Robot: An International Journal, Volume 28, Number 6, 2001, pp. 498-504(7)

[8] H. Okuno, K. Nakadai, and H. Kitano, Social Interaction of Humanoid RobotBased on Audio-Visual Tracking, In Proceedings of the International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems IEA/AIE 2002, p.140-173.

[9] N. Mavridis,Grounded Situation Models for Situated Conversational Assistants, PhD thesis, Massachusetts Institute of Technology, available online at MIT DSpace Collections Archive at http://hdl.handle.net/1721.1/38523 and at http://www.dr-nikolaos-mavridis.com

[10] M. Bennewitz, F. Faber, J. Dominik, M. Schreiber, and S. Behnke, Multimodal Conversation between a Humanoid Robot and Multiple Persons. In Proc. MCHI ws at AAAI05.

[11] T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. In Human-Computer Interaction, 19(1&2):61-84.

[12] T. Kanda, R. Sato, N. Saiwaki, H. and Ishiguro, A Two-Month Field Trial in an Elementary School for Long-Term Human-Robot Interaction, IEEE Transactions on Robotics, 23, 5 (2007), pp. 962-971

[13] L. Ahn, R. Liu, M. Blum, Peekaboom: A Game for Locating Objects in Images. In ACM Conference on Human Factors in Computing Systems CHI 2006. p. 55-64.

[14] European Union RoboEarth project website, available online at : http://www.roboearth.org

[15] Pioneer 3 / PeopleBot Human-Robot Interaction Robot Operations Manual, ActivMedia Robotics, 44 Concord St., Peterborough NH, 03458

[16] Point Grey Research, Inc.,Vancouver, Canada, provider of pan-tilt-zoom camera motorization base units, web site available online at: http://www.ptgrey.com

[17] Directed Perception, Inc., 890C Cowan Road, Burlingame, CA 94010, provider of BumbleBee stereo cameras, web site available online at: http://www.dperception.com

[18] The Internet Communications Engine, ZeroC Inc., web site available online at: http://www.zeroc.com/ice.html

[19] P. Viola and M. Jones. Robust real-time object detection. International Journal of Computer Vision, 57(2):137-154, 2004.

[20] D. Bolme, R. Beveridge, M. Teixeira and B. Draper (2003) The CSU Face Identification Evaluation System: Its Purpose, Features and Structure, International Conference on Vision Systems, pp 304-311, Graz, Austria, April 1-3.

[21] A.V. Nefian and M.H. Hayes, Hidden markov models for face detection and recognition, in Proceedings of the International Conference on Image Processing, 1998.

[22] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, 2004.

[23] ActivMedia Robotics, Aria library reference manual, Technical Report (1.1.10), 2002.

[24] N. Mavridis, and D. Hanson, The IbnSina Center: An Augmented Reality Theater with Intelligent Robotic and Virtual Characters, in Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication Ro-MAN 2009, Toyama, Japan.

[25] N. Mavridis, A. Dhaheri, L. Dhaheri, M. Khanji, N. Darmaki, Transforming IbnSina into an Advanced Multilingual Interactive Android Robot, in Proceedings of the IEEE GCC Conference, 2011.

[26] N. Mavridis, E. Machado, et al.,Real-time Teleoperation of an Industrial Robotic Arm Through Human Arm Movement Imitation, in Proceedings of the International Symposium on Robotics and Intelligent Sensors (IRIS), 2010, Nagoya, Japan.
[27] C. Christoforou, N. Mavridis, et al.,Android tele-operation through Brain-Computer Interfacing: A real-world demo with non-expert users, in Proceedings of the International Symposium on Robotics and Intelligent Sensors (IRIS), 2010, Nagoya, Japan.
[28] N. Mavridis, W. Kazmi, P. and Toulis, Friends with Faces: How Social Networks Can Enhance Face Recognition and Vice Versa, in Book "Computational Social Networks Analysis: Trends, Tools and Research Advances", Springer Verlag, 2009.
[29] Touchgraph Limited Liability Company, web site available online at : http://www.touchgraph.com/
[30] N. Mavridis, "Robot Friends: Artificial Agents entering Human Social Networks", in Book "The Networked Self: Identity, Community and Culture on Social Network Sites", Routledge, New York, 2010.