Towards Partners Profiling in Human Robot Interaction Contexts

Salvatore M. Anzalone, Yuichiro Yoshikawa, Hiroshi Ishiguro[1],
Emanuele Menegatti, Enrico Pagello[2], and Rosario Sorbello[3]

[1] Intelligent Robotics Laboratory, Dept. of Systems Innovation,
Graduate School of Engineering Science, Osaka University
[2] Intelligent Autonomous Systems Laboratory, Dept. of Information Engineering,
Faculty of Engineering, University of Padua
[3] Dept. of Chemical, Management, Computer and Mechanical Engineering,
Faculty of Engineering, University of Palermo

Abstract. Individuality is one of the most important qualities of humans. Social robots should be able to model the individuality of the human partners and to modify their behaviours accordingly.This paper proposes a profiling system for social robots to be able to learn the individuality of human partners in social contexts. Profiles are expressed in terms of of identities and preferences bound together. In particular, people's identity is captured by the use of facial features, while preferences are extracted from the discussion between the partners. Both are bound using an Hebb network. Experiments show the feasibility and the performances of the approach presented.

Keywords: profiling, personal robots, human robot interaction.

1 Introduction

Human beings are social animals. People are individuals but are also members of a group. Our behaviours, our actions are not only highly influenced by our society, but are also capable of influencing the society itself, creating a strictly, deep connection between each single individual and the others. Moreover, social capabilities have an extremely significant role in our evolution, in our development and education. Sociability influences our deep human qualities, like identity, friendship, empathy and also emotion. It is not possible to understand human nature without considering our social capabilities. Classic robotics has been focused on making mobile robots completely autonomous, capable of exploring, moving and accomplishing missions in a safe way, inside real environments. According to this approach, robots are something like a tool or a sort of "intelligent instrument" employed to achieve missions that are too hazardous for humans to carry out, dangerous tasks in a remote, unreachable environment. This use of robots does not catch the sense of a real partnership between robots and humans because they always identify a master-slave relationship between them [1]. Despite of this, robots can collaborate in a strict way with humans as a social, cooperative and capable partner: not only as systems able to perceiving and acting in order to extend human productivity, but also as interactive and communicative subjects, reliable working

I. Noda et al. (Eds.): SIMPAR 2012, LNAI 7628, pp. 4–15, 2012.

partners [2]. However, going social is not enough [3]. Humans consider themselves individuals with an unique personality, with personal preferences, and with a social role, so they expect to be treated likewise. Robots should interact with people in a "personal" way. To be perceived as active and effective partners, truly accepted by humans, social robots need to learn how to relate with the individuality of single persons: sociable robots must be able to identify and represent human partners according to their physical features, their preferences and their social relations to adapt their behaviours, vocabulary, and social rules according to the individuals that are involved in the interaction.

2 System Overview

The system proposed in this paper tries to model the profiles of human partners. As depicted in figure 1, a robot is involved in the interactions between two human users: while people discuss, the robot is able to capture the information related to their identity and the data about the current discussion. Profile of each partner is modelled by bounding together this information. A person's profile can be described as a conjugation of several kinds of different features. A characterization of a face can be used to depict the identity of a human partner, but this can be improved using other related descriptors such as the voice. Using different kind of features is possible to obtain a more reliable model of the appearance of the partners, according to their physical cues. Through this characterization the robot is able to recognize the identity of its human partners. In the work presented on this paper only facial cues are considered. Then, the profile is completed by joining to this physical characterization the preferences of the partners, by analysing the conversations between them and by finding their topics: preferences can be seen as the most recurrent topics discussed by each partner. Finally, the robot can be enabled to adapt its behaviour according to this information perceived. Several software modules have been built in order to accomplish the main task, as shown in figure 2. Data from a camera is collected by a face recognition system and processed to extract faces, if any, and, from them, facial features. Then, such kind of features are classified together in order to obtain identity claims of the partners that share the environment with

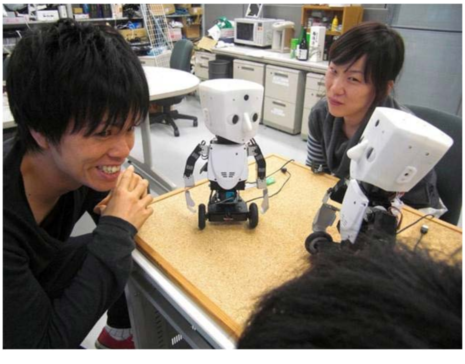

Fig. 1. A typical setup of the system

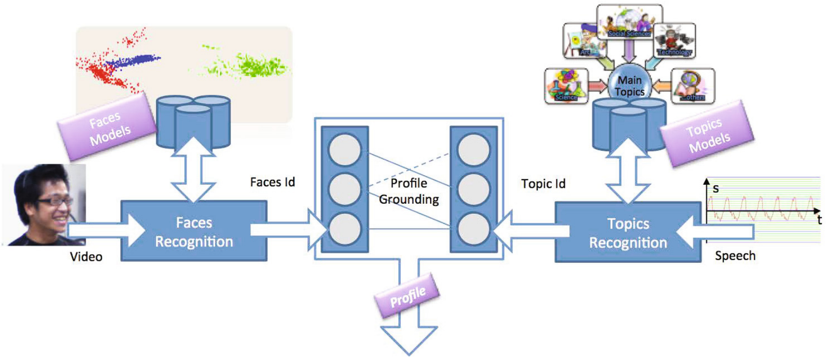

Fig. 2. An overview of the system

the robot. On the other side, through a speech recognition system, sound is processed to extract the utterances of the conversation [4]. Then, the topic recognition system tries to deduce the topic of the current conversation through a statistical characterization of these sentences. Finally, a profiling system will bound identities and topics together and will store this information as a model of the profile of each person.

3 People Identification

Identification of partners in a human-robot interaction context can be achieved using different ways, such as voice identification, face recognition, and so on [5]. While this can be an easy task in controlled environments, it becomes a challenging problem in daily life applications in unstructured environments. Focusing on vision based systems, features retrieved by a camera are usually very noisy, can change if the person in front of the camera changes pose, and accordingly to the light conditions of the environment. Such circumstances will strongly affect the recognition results. Furthermore, recognition results are also affected by non verbal communication, such as showing emotional states or social behaviours. But this is not enough; in long term interactions these problems become bigger because people can alter their appearance. tis possible to think about a robot that should interact with the same woman with and without make up, or with the same man with and without a beard. The approach here presented achieves partner identification using visual information. As shown in figure 3, data from the environment is perceived through the camera of the robot. This raw information, is analysed to detect humans features that will be extracted and collected. Facial features are retrieved through the use of Eigenfaces applied to the visual information. The features collected represent a biological signature of each person, so they are opportunely classified in a supervised way to obtain the identification claims. n detail, the face features extraction process is subdivided in two main steps: in a first phase the camera data is processed to find faces. The Viola-Jones classifier is an efficient detection system that tries to find features, called Haar-like features, that encode the existence of oriented contrast in different regions of the image [6]. A set of this kind of features has been chosen using different pictures of faces taken under the same lighting conditions and normalized to line up the eyes and mouths: these features encoded the contrasts and the special relationships showed by human faces. The Haar classifier is trained to detect and localize faces inside the images taken in the same conditions of the training set. Faces found by the Viola-Jones detector are processed in order to find a model capable of describing the identities of people, by extracting the most relevant information contained in them. The Principal Component Analysis offers theory concepts to achieve this, in particular using the technique of the Eigenfaces [7]. According to this approach, a small set of pictures is used to calculate some vectors, called Eigenfaces, that define a space that best encodes the variations of between faces. This space is called Eigenspace. In this project, the eigenspace has been built using a standard database, YaleFaces, to represent a generic space capable of describing the features of many kinds of faces [8]. The features of the faces seen are the eigenvalues calculated from this space by projecting on it the images found by the Viola-Jones detector. As a final step, vectors of faces are classified by a SVM properly trained to classify the frontal images of human users faces giving them a face claim identifier [9]. A statistical refinement of the output can be performed in this stage, forcing the system to return as a result of the current face identity the most recurrent one of the last 5 identities claimed.

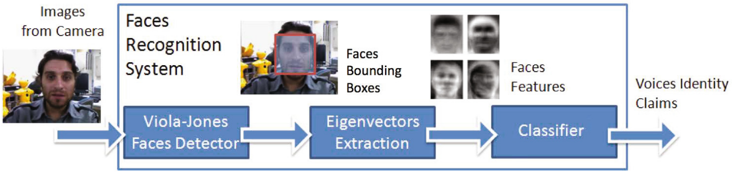

Fig. 3. The faces recognition system

4 Topics Identification

The human speech is a natural and intuitive way of communication with a robot [10]. However, the usage of the auditory channel in a human robot interaction context becomes very difficult due to its huge and noisy informative content, and due to the incompleteness and ambiguity of the natural languages. On one side, the auditive flow can encode one or more overlapped speeches, with echoes and environmental noise; on the other side, the natural language seems unable to describe them only in terms of syntax, semantics or phonetics rules, making its deep understanding a very hard task [11]. In order to avoid such problems, the approach used in this work tries to overcome the low recognition rate on the accuracy of the speech recognition system by grounding conversation between people to their topic, using only some relevant words. According to this approach, each word is weighed using a modified version of the classical "Term Frequency - Inverse Document Frequency" ranking function, a statistical measure often used in text mining and information retrieval [12]. Given a corpus of documents, the TF-IDF evaluates the importance of a word in a document. In the work here presented the same idea has been applied to evaluate the relevance of a word in a given topic,

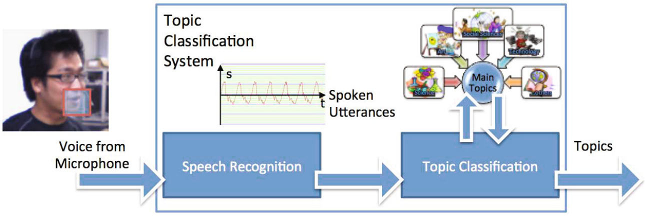

Fig. 4. The topics recognition system

performing a “Term Frequency - Inverse Topic Frequency” ranking function [13]. The TF-ITF approach in particular gives more weight to the terms often used in few topics and gives a low weight to the terms used in all the topics considered. In this way, it is possible to discard by thresholding all the negligible terms of the vocabulary, such as verbs, conjunctions, adjectives, by considering only the meaningful terms for each topic. The algorithm relies only on words frequencies, then any syntactical and semantical considerations are not taken in account: because of this, the system will not understand the details of the sentences, and, in particular, it will not be able to distinguishing affirmations and negations, likes and dislikes, and so on.

However, in order to approach free context conversations the system should be able to deal with thousands of topics. It is not possible to directly apply TF-ITF to such a huge amount of categories. A convenient way to approach this problem is by using a hierarchical categorization of the topics. Wikipedia, one of the biggest existing encyclopedias,can be seen as a huge repository of sentences categorized by thousand of topics. Moreover, each Wikipedia topic is itself categorized according to one or more parent topics [14]. From this point of view, Wikipedia offers an unique set of knowledge base that can be used to process natural language, categorized in a hierarchical graph, from the most abstract topics to the most detailed ones. Top level nodes, all children of a main root node, are the most general nodes, such as “Science”, “Art”, “Social Sciences”, “Technology” or “Society”; descending deeper to the bottom, topics start to become more and more concrete, arriving to the leaf topic nodes that represents topics related to very specific categories. In this hierarchy TF-ITF is calculated between nodes with a common parent: it is possible to evaluate for each node which is the most relevant topic between its children. Then, starting from the root of the hierarchy, it is possible to categorize words and sentences by finding a branch of the tree inside the Wikipedia based hierarchy that is coherent to its topic. In particular, Tf-Itf for each word of the sentence is normalized among the children nodes, then words with a high level of entropy are discarded. Using the remaining words, a probability of the sentence to belong to each child node is calculated. The child node with the highest probability is chosen and its topic is selected as being coherent with the sentence. A reference of the coherence of the topic with the sentence has been also retrieved by calculating the entropy of the sentence among the different childs. Through the recursive use of this algorithm, it is possible to explore the whole topic tree in order to find a path of the most related topics to a given sentence, from the abstract categories to the most detailed ones. This recursive algorithm can stop in the case of the reaching of a leaf, or in several other

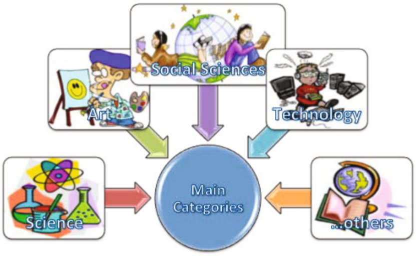

Fig. 5. Some of the top level categories in the hierarchy used by the topic recognition system

situations, such as the complete discarding of all the words in the sentence, due to high entropy. In the categorization of conversations for application in the real world, there are several aspects to be considered. First of all, according to the applications, the classification of the topic should be executed in realtime, or in several seconds, or in several minutes or, also, offline. A significant amount of data may involve a lot of calculations, and this may decrease the performance of the system. Moreover, the quality of these clusters should be settled in an accurate way. Given a specific number of clusters, a high degree of separation between them implies the use of highly-reliable, abstract topics, but which are not capable of informing about the details of the conversations. On the opposite side, a low degree of separation implies the use of more detailed, but less-reliables, topics. The choice of the amount of details for the topic classification should be carefully chosen according to the application, taking in account their reliability performances.

5 People Profile Grounding

Claims of faces coming from the SVM classifier are referred to clusters that better encode the similarities between features. However, they do not inform in an explicit way about the persons profile. The system should learn about the preferences by linking in some way the information about the identity with the information about the topic. This can be seen as a "symbol grounding" problem, in which features of the same person extracted from different modalities will converge to the same high level mental concept [15]. From this point of view, the symbol, or the model, of the human identity will "emerge" by binding between them all with the information about each partner, that is, his face, his preferences. An anchoring system, capable of linking in a correct way the features from each modality to their high level symbols representing the model of each human identity is needed. The solution here presented follows a statistical approach: while humans interact in front of the robot, claims of identities and topics recognized will be perceived at the same time. Then, the system will join them by learning this binding. According to this approach, identities and topics are bound together to form multi modal clusters that represent an unique complete model of each identity. This idea has been implemented through a Hebb network capable of representing the connection between the clusters from each modality [16]. In an Hebb network, if two connected neurons are repeatedly activated at the same time, they will strengthen their connection, tending to become more and more associated. According to this, while users converse in front of the robot, the system will learn and improve its bindings between the topics recognized and identities perceived. The system updates the weights of the most active links using the formula:

$$\triangle w_{i^* j^*} = \eta ({}^{l}d_{i^* j^*} a_{i^*} \cdot {}^{r}d_{i^* j^*} \cdot a_{j^*} - w_{i^* j^*}) \tag{1}$$

in which: w_{ij} are the weights; i^* and j^* are the indexes of the most active link; a_i and a_j are the activation level of input and output; η is a coefficient about the speed of the learning and ${}^{l}d_{i^* j^*}$ and ${}^{r}d_{i^* j^*}$ are coefficients calculated as:

$${}^{l}d_{ij} = \exp^{-\frac{\underset{k,k \neq j}{\Sigma w_{ik}}}{\sigma^2}} \tag{2}$$

$${}^{r}d_{ij} = \exp^{-\frac{\underset{k,k \neq i}{\Sigma w_{kj}}}{\sigma^2}} \tag{3}$$

where σ is a variance weight. According to the mutual exclusivity, the other links are inhibited:

$$w_{ij^*}(t+1) = w_{ij^*}(t) - \eta_l \cdot (1 - {}^{l}d_{ij^*} \triangle w_{i^* j^*}) w_{ij^*}(t) \tag{4}$$

$$w_{i^* j}(t+1) = w_{i^* j}(t) - \eta_r \cdot (1 - {}^{r}d_{i^* j} \triangle w_{i^* j^*}) w_{i^* j}(t) \tag{5}$$

where η_l and η_r are coefficients regarding the speed of the lateral inhibition. Following this approach, the system learns the relationship between the most active couples and it is also capable ofrecovering when something wrong has been learned. The lateral inhibition grants the system the possibility of forgeting and learning the correct connection. The dimension of the η coefficients, η, η_l and η_r, are important parameters: if the value of η is big the system will learn quickly, and if the value of η_l and η_r are low it will forget slowly. During the development, a careful evaluation of these parameters is needed. While people are discussing in front of the robot, face claims and topics from the conversation are calculated. The activation level of each of the two sides of the Hebb network will rely on this information. In particular, the probability of the current face to belong to the set of trained identities, calculated by the faces recognition system, will be used as activation level on one side of the network. The topics belonging to the branch of the tree calculated using the current discussion data, by the topic recognition system, are used for the second side of the network. In particular, the topics selected will be all together active using an activation level that will be dependent to their respective degree of detail, to their significance and to their entropy associated with the conversation considered. In this way, more abstract levels, with less informative content will give a small contribute on the identity modelling, than the others that will bring more informative content. According to the presented approach, as shown in figure 6, during different conversations, the Hebb network will store the models of all the human partners of the robot in terms of strong connections between its nodes. Favourite topics for each partners will be the most discussed and detailed topics. In this case, the network will model strong connections between the identity claims and their respective favourite topics and will be able to store the models of all the human partners of the robot.

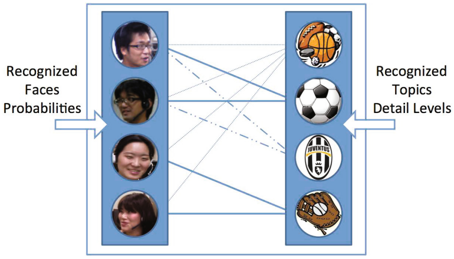

Fig. 6. The Hebb network models the profiles of the human partners as connections between identities and conversation topics

6 Experimental Results

The system has been tested in different scenarios in order to evaluate the performances of each single component. Then, an evaluation of the profiling system is performed.

6.1 People Identification Evaluation

To evaluate the physical identification of humans, the system was tested 10 times using a population of 5 people. In each experiment, a person reads, for about 15 seconds, a text in front of the robot. As shown in figure 7, the system has been tested by incrementing the number of the people in the training set. With a small number of identities in the training sets, performances of the system are very good, but while incrementing them discriminating identities in real environment, it becomes more and more difficult. To preserver the performances of the partners recognition a biggest number of features can be used, in order to rely on a more detailed description. The figure 7 shows also a comparison of the results obtained using a different number of the features. There are several reasons for the failures. The most relevant are the changes in the light conditions

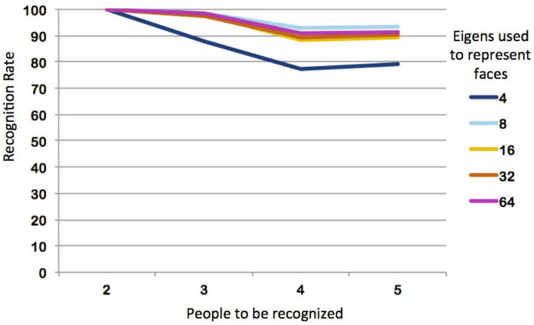

Fig. 7. Performances of the people identification by varying the number of faces to be recognized, and among different lengths of the features vector

of the environment and the change of the gaze direction of the persons, that can also bring a variation on the shadows on the faces. These conditions introduce important differences in respect to the training set, so the system is not able to identify human users in a correct way. Also, emotions can be the reason for a wrong identification: a big smile can alter considerably not only the features of the face, but also the gaze direction and, consequently, its shadows. Recognition problems will be overcome using more adaptive techniques of clustering instead of SVM or by using a more reliable multi modal approach, as using voices claims.

6.2 Topics Identification Evaluation

The Wikipedia based hierarchical topic recognition system has been evaluated using a set of 45 sentences coherent to 15 common topics (3 sentences per topic) such as “soccer”, “travel”, “recipes”, “manga”, “music” an other everyday topics of conversation. From the whole test set of conversations, the system was not able to recognize in 40% of the topics, because they were not included in the Wikipedia based training set. Despite of this, the 78% of the remaining 60% of the conversations considered, was successfully classified by the system into coherent topics. In order to judge the quality of the topics recognized, more tests have been performed to understand to what extent the results were acceptable for humans. We asked one volunteer cooperator to score the acceptability of them using points from 1 (“very unrelated”) to 5 (“very coherent”), comparing original sentences and recognition results. The collected data was normalized among the deepness of the tree explored, depicting with this their detail degree, and then statistically interpolated, in order to find a trend. As shown in Figure 8 by the blue line that represents the trend extracted from the coherence results, the most abstract topics are felt as less unrelated. Going through the most detailed nodes, the coherence grows up until a maximum, then it decades. This can be explained by the difficulty that the system finds on separating strongly tighten nodes, due to their low amount of information differences. It is interesting to explain something more about this curve: the two maximum points are both located to high coherent topics with a different level of detail, such as “sport” and “soccer”, according to the figure. This can be explained by the search in the graph, that explores among different levels of abstraction of the topics. Finally, the trend of the information entropy associated to the sentences, amongst the detail level of the tree explored, was calculated. A statistical interpolation of this data has been conduced in order to find the trend depicted in the Figure 8 by the red line. In this case, the closer it is to zero the entropy, the higher is the informative content of the detail level. As it is possible to see, the coherence results are highly correlated to the entropy, showing that this can be used as a criteria to judge the quality of the solution given. Despite of this important result achieved, several are the limitations of this system. The use of a huge amount of data elaborated during the search in the hierarchy of the topics is reflected on the long processing time. At this stage, the hierarchical topic recognition system cannot perform in real-time. Furthermore, the lack of information of the training set depicted in the results shows that, though the Japanese Wikipedia provides a vast and expansive set of topics, it is not able of including all the possible topics that can be discussed during daily life conversations. Lastly, the system at this stage is

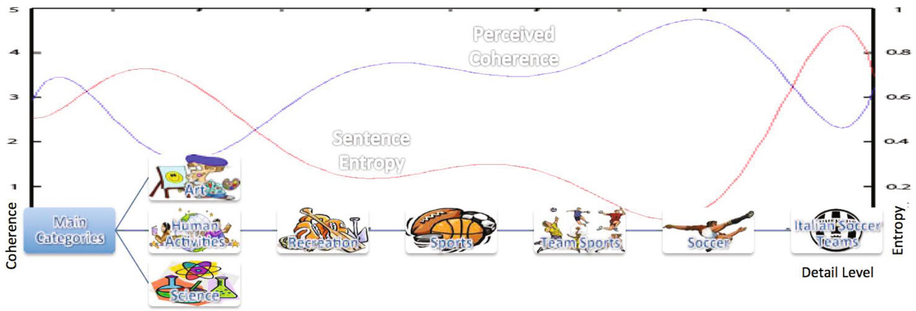

Fig. 8. Coherence trend of proposed topics, in blue, among their detail level, and their entropy

not able of infering any kind of information, other than what is explicitly described by the tree, but that can be obvious to an human listener.

6.3 Profiling Evaluation

Three couples of people were considered for the experiments. For each experiment, a person in the couple is chosen randomly to watch a video related to some sport. Then, as shown in figure 9, the couple is encouraged to discuss about it in front of a robot able of nodding and pointing to the current speaker. The experiment is performed three times for each couple, proposing three different videos related to two different categories. During each experiment, facial information and speech data is recorded and used offline to extract people profiles, according to the algorithm described. In particular, according to the previously obtained results, a set of 16 features was used for the face recognition system. Moreover, topics activation levels for the grounding system have been weighed according to their detail level using the trend function previously found. Analysis of the connections learnt by the profile grounding system revealed connections between people and topics coherent to the conversation in question. In

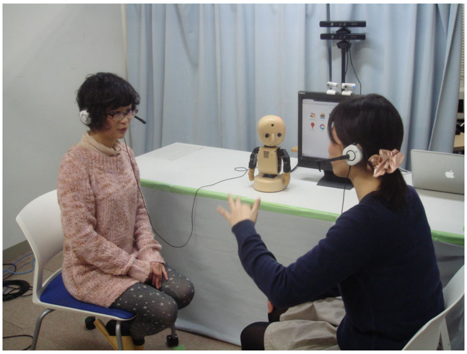

Fig. 9. People conversates in front of the robot during the experiment

particular, the strongest connections were found between people and the most coherent topics of the conversations in which they where involved, in the 77% of experiments. Moreover, the strongest connection between people and topics occurred with the most frequent topics discussed, coherently with the topic proposed in the conversations achieved, in 66% of experiments. Errors come mainly from the topic recognition system that is not always able to recognize coherent topics in all the different conversations in which human partners can be involved in. Moreover, despite the system being able to find coherent topics, it is not able to classify within identical topic conversations, that humans will easily recognize as belonging to the same topic, due to the ambiguity hiden in the natural language itself.

7 Conclusions

A profiling system for robots involved in human conversations was presented. Identities of human partners and topics discussed during the conversations were bound together in order to model their own profiles. Identities are recognized in a closed set by the classification of faces using eigenfaces technique, while topic recognition was achieved using an hierarchical approach based on "Term Frequency - Inverse Topic Frequency" ranking function. Profile modelling was exploited using a Hebb network. Experiments performed show the potential and the deficiencies of the system. Despite these problems that showed how the system is very far from its use in real, long-term, daily life contexts, the results obtained encourage us to pursuit its development and experimentation. In particular, more efforts should be focused on real time capabilities of the system and on obtaining stronger recognition identities, by relying on more stable features of the faces and by introducing a multi modal characterization, using other channels, such as the voices. Moreover, topic recognition system should be improved to have a better understanding of the conversation. In particular, ontologies [17] can be used in combination with the presented hierarchy of topics in order to improve the recognition itself, as a way to infer other connections between conversation topics. Furthermore, the system should be able to cope with dynamic situations by recognizing and incorporating in its set of known acquaintances new, unknown, people. Lastly, future experiments will focus on the effect of customized behaviours according to the profiling results during human robot interactions. Many are the real world applications that can take advantage from the idea of profiling. However, it is important to underline that here only a first approach using faces and conversation topics has been presented. It is possible to imagine more complex profiles, able to rely on more features, that can be used in several applications such as robotic companions, elderly assistants, human-robot gaming systems, and all the applications that need to rely in a strong characterization of the behaviours related to the individuality of the human partners.

Acknowledgment. Authors would like to thank Prof. M. Chetouani, Dr. S. Livieri, Dr. F. Dalla Libera, Prof. A. Chella, Prof. G. Vassallo for their support during the development of this project. This work has been supported by a Grant-in-Aid for scientific research fellowships from Japan Society for the Promotion of Science (JSPS). This project was partially founded by Regione Veneto Cod. progetto: 2105/201/5/2215/2009 approved with DGR n. 2215, 21/07/2009.

References

1. Breazeal, C.: Toward sociable robots. Robotics and Autonomous Systems 42(3-4) (2003)
2. Breazeal, C.L.: Designing sociable robots. The MIT Press (2004)
3. Anzalone, S., Nuzzo, A., Patti, N., Sorbello, R., Chella, A.: Emo-dramatic robotic stewards. Social Robotics, 382–391 (2010)
4. Lee, A., Kawahara, T., Shikano, K.: Julius—an open source real-time large vocabulary recognition engine. In: Seventh European Conference on Speech Communication and Technology (2001)
5. Anzalone, S.M., Menegatti, E., Pagello, E., Yoshikawa, Y., Ishiguro, H., Chella, A.: Audio-video people recognition system for an intelligent environment. In: 2011 4th International Conference on Human System Interactions (HSI), pp. 237–244. IEEE (2011)
6. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision 57(2), 137–154 (2002)
7. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 591, pp. 586–591 (1991)
8. Hurkens, C., Van Iersel, L., Keijsper, J., Kelk, S., Stougie, L., Tromp, J., Dolech, D., Eindhoven, A.: Face Image Database, publicly available for non-commercial use (2008), `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`
9. Steinwart, I., Christmann, A.: Support vector machines. Springer (2008)
10. Kraft, F., Kilgour, K., Saam, R., Stuker, S., Wolfel, M., Asfour, T., Waibel, A.: Towards social integration of humanoid robots by conversational concept learning. In: 2010 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 352–357. IEEE (2010)
11. Anzalone, S.M., Cinquegrani, F., Sorbello, R., Chella, A.: An emotional humanoid partner. Linguistic and Cognitive Approaches To Dialog Agents (LaCATODA 2010) At AISB (2010)
12. Jackson, P., Moulinier, I.: Natural language processing for online applications: Text retrieval, extraction and categorization, vol. 5. John Benjamins Pub. Co. (2007)
13. Anzalone, S.M., Yoshikawa, Y., Menegatti, E., Pagello, E., Sorbello, R., Ishiguro, H.: A topic recognition system for real world human-robot conversations. In: IAS 2012, 12th International Conference on Intelligent Autonomous Systems (2012)
14. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)
15. Coradeschi, S., Saffiotti, A.: An introduction to the anchoring problem. Robotics and Autonomous Systems 43(2-3), 85–96 (2003)
16. Yoshikawa, Y., Hosoda, K., Asada, M.: Unique association between self-occlusion and double-touching towards binding vision and touch. Neurocomputing 70(13-15), 2234–2244 (2007)
17. Kobayashi, S., Tamagawa, S., Morita, T., Yamaguchi, T.: Intelligent humanoid robot with japanese wikipedia ontology and robot action ontology. In: Proceedings of the 6th International Conference on Human-Robot Interaction, pp. 417–424. ACM (2011)