Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Detection of human activities and human intentions to support people with disabilities

Rapporto Tecnico N.:
RT-ICAR-PA-11-03

novembre 2011

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sede di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sede di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Detection of human activities and human intentions to support people with disabilities

S. Gaglio[12], I. Infantino[1], C. Lodato[1]

Rapporto Tecnico N.:
RT-ICAR-PA-11-03

Data:
novembre 2011

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sede di Palermo, Viale delle Scienze edificio 11, 90128 Palermo.

[2] Università degli Studi di Palermo, Dipartimento di Ingegneria Chimica Gestionale Informatica e Meccanica, Viale delle Scienze, 90128 Palermo.

I rapporti tecnici dell'ICAR-CNR sono pubblicati dall'Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche. Tali rapporti, approntati sotto l'esclusiva responsabilità scientifica degli autori, descrivono attività di ricerca del personale e dei collaboratori dell'ICAR, in alcuni casi in un formato preliminare prima della pubblicazione definitiva in altra sede.

Detection of human activities and human intentions to support people with disabilities

Salvatore Gaglio, Ignazio Infantino, Carmelo Lodato

Technical Report n. RT-ICAR-PA-11-03, November 2011

1. Introduction

In the wider context of capturing and understanding human behavior (Pantic et al., 2006), it is important to perceive (detect) signals such as facial expressions, body posture, and movements while being able to identify objects and interactions with other components of the environment. The techniques of computer vision and machine learning methodologies enable the gathering and processing of such data in an increasingly accurate and robust way (Kelley et al., 2010). If the system captures the temporal extent of these signals, then it can make predictions and create expectations of their evolution. In this sense, we speak of detecting human intentions, and in a simplified manner, they are related to elementary actions of a human agent (Kelley et al., 2008).

Over the last few years has changed the approach pursued in the field of HCI, shifting the focus on human-centered design for HCI, namely the creation of systems of interaction made for humans and based on models of human behavior (Pantic et al., 2006). The Human centered design, however, requires thorough analysis and correct processing of all that flows into man-machine communication: the linguistic message, non-linguistic signals of conversation, emotions, attitudes, modes by which information are transmitted, i.e. facial expressions, head movements, non-linguistic vocalizations, movements of hands and body posture, and finally must recognize the context in which information is transmitted.

In general, the modeling of human behavior is a challenging task and is based on the various behavioral signals: affective and attitudinal states (e.g. fear, joy, inattention, stress); manipulative behavior (actions used to act on objects environment or self-manipulative actions like biting lips), culture-specific symbols (conventional signs as a wink or a thumbs30 up); illustrators actions accompanying the speech, regulators and conversational mediators as who nods the head and smiles. Systems for the automatic analysis of human behavior should treat all human interaction channels (audio, visual, and tactile), and should analyze both verbal and non verbal signals (words, body gestures, facial expressions and voice, and also physiological reactions). In fact, the human behavioral signals are closely related to affective states, which are conducted by both physiological and using expressions. Due to physiological mechanisms, emotional arousal affects somatic properties such as the size of the pupil, heart rate, sweating, body temperature, respiration rate. These parameters can be easily detected and are objective measures, but often require that the person wearing specific sensors. Such devices in future may be low-cost and miniaturized, distributed in clothing and environment, but which are now unusable on a large scale and in non structured situations.

The visual channel that takes into account facial expressions and gestures of the body seems to be relatively more important to human judgment that recognizes and classifies behavioral states. The human judgment on the observed behavior seems to be more accurate if you consider the face and body as elements of analysis. A given set of behavioral signals usually does not transmit only one type of message, but can transmit different depending on the context. The context can be completely defined if you find the answers to the following questions: Who, Where, What, How, When and Why (Pantic et al., 2006). These responses disambiguating the situation in which there are both artificial agent that observes and the human being observed. In the case of human-robot interaction, one of the most important aspects to be explored in the detection of human behavior is the recognition of the intent (Kelley et al., 2008): the problem is to predict the intentions of a person by direct observation of his actions and behaviors. In practice we try to infer the result of a goal-directed mental activity that is not 0 observable, and characterizing precisely the intent. Humans recognize, or otherwise seek to predict the intentions of others, using the result of an innate mechanism to represent, interpret and predict the actions of the other. This mechanism probably is based on taking the perspective of others (Gopnick & Moore, 1994), allowing you to watch and think with eyes and mind of the other. The interpretation of intentions can anticipate the evolution of the action, and thus capture its temporal dynamic evolution. An approach widely used in statistical classification of systems that evolve over time, is what uses Hidden Markov Model (Duda et al., 2000). The use of HMM in the recognition of intent (emphasizing the prediction) has been suggested in (Tavakkoli et al., 2007), that draws a link between the HMM approach and the theory of the mind.

The recognition of the intent intersects with the recognition of human activity and human behavior. It differs from the recognition of the activity as a predictive component: determining the intentions of an agent, we can actually give an opinion on what we believe are the most likely actions that the agent will perform in the immediate future. The intent can also be clarified or better defined if we recognize the behavior. Again the context is important and how it may serve to disambiguate (Kelley et al., 2008). There are a pairs of actions that may appear identical in every aspect but have different explanations depending on their underlying intentions and the context in which they occur. Both to understand the behaviors and the intentions, some of the tools necessary to address these problems are developed for the analysis of video sequences and images (Turaga et al., 2008). The aspects of security, monitoring, indexing of archives, led the development of algorithms oriented to the recognition of human activities that can form the basis for the recognition of intentions and behaviors. Starting from the bottom level of processing, the first step is to identify the movements in the scene, to distinguish the background from the rest, to limit the objects of interest, and to monitor changes in time and space. We use then, techniques based on optical flow, segmentation, blob detection, and application of space37 time filters on certain features extracted from the scene. When viewing a scene, the man is able to distinguish the background from the rest, that is, instant by instant, automatically rejects unnecessary information. In this context, a model of attention is necessary to select the relevant parts of the scene correctly. One problem may be, however, that in these regions labeled as background is contained the information that allows for example the recognition of context that allows the disambiguation. Moreover, considering a temporal evolution, what is considered as background in a given instant, may be at the center of attention in successive time instants. Identified objects in the scene, as well as being associated with a certain spatial location (either 2D, 2D and 1/2, or 3D) and an area or volume of interest, have relations between them and with the background. So the analysis of the temporal evolution of the scene, should be accompanied with a recognition of relationships (spatial, and semantic) between the various entities involved (the robot itself, humans, actions, objects 1 of interest, components of the background) for the correct interpretation of the context of action.

But defining the context in this way, how can we bind the contexts and intentions? There are two possible approaches: the intentions are aware of the contexts, or vice versa the intentions are aware of the contexts (Kelley et al., 2008). In the first case, we ranked every intention carries with it all possible contexts in which it applies, and real-time scenario is not applicable. The second approach, given a context, we should define all the intentions that it may have held (or in a deterministic or probabilistic way). The same kind of reasoning can be done with the behaviors and habits, so think of binding (in the sense of action or sequence of actions to be carried out prototype) with the behaviors. A model of intention should be composed of two parts (Kelley et al, 2008): a model of activity, which is given for example by a particular HMM, and an associated label. This is the minimum amount of information required to enable a robot to perform disambiguation of context. One could better define the intent, noting a particular sequence of hidden states from the model of activity, and specifying an action to be taken in response. A context model, at a minimum, shall consist of a name or other identifier to distinguish it from other possible contexts in the system, as well as a method to discriminate between intentions. This method may take the form of a set of deterministic rules, or may be a discrete probability distribution defined on the intentions which the context is aware. There are many sources of contextual information that may be useful to infer the intentions, and perhaps one of the most attractive is to consider the so-called affordances of the object, indicating the actions you can perform on it. It is possible then build a representation from probabilities of all actions that can be performed on that object. For example, you can use an approach based on natural language (Kelley et al., 2008), building a graph whose vertices are words and a label is the weighed connecting arc indicating the existence of some kind of grammatical relationship. The label indicates the nature of the relationship, and the weight can be proportional to the frequency with which the pair of words exist in that particular relationship. From such a graph, we can calculate the probability to determine the necessary context to interpret an activity. Natural language is a very effective vehicle for expressing the facts of the world, including the affordances of the objects. If the scene is complex, performance and accuracy can be very poor when you consider all the entities involved. then, can be introduced for example the abstraction of the interaction space, where each agent or object in the scene is represented as a point in a space with a defined distance on it related to the degree of interaction (Kelley et al, 2008). In this case, then consider the physical artificial agent (in our case the humanoid) and its relationship with the space around it, giving more importance to neighboring entities to it and ignore those far away.

2. Developing an Intentional System[1]

In the following, we describe a cognitive architecture developed with the aim of detecting human movements and perceiving actions and intents (see Infantino et al. 2008) and the design of a

[1] I. Infantino, C. Lodato, S. Lopes, F. Vella (2008). An Intentional System based on a Knowledge Base of Visual Perception. AIxIA

semantic structure linked to visual data. In particular, we implemented an “intentional” vision system, that is, a system that “looks at people” and automatically perceives information relevant to interpret the human behavior (see for example Kuno et al. 1999), distinguishing between unintentional human movements, movement for manipulating objects, and gestures used for communicating. The use of word "intentional" in this context concerns the purpose of generating a stream of pre-processed data useful for reasoning, recognition, reacting, and interacting when a human and his activity are objects of observation from the artificial system. The raw data coming from multiple sources of images and videos are filtered and processed in order to retain information useful to understand the human will, state and condition.

In order to model, recognize, and interpret human behavior, several tasks must be addressed (see for example Turk 2004): face detection, location, tracking and recognition; facial expression analysis, and human emotion recognition; audiovisual speech recognition; eye-gaze tracking; body tracking; hand tracking; gait recognition; recognition of postures, gestures, and activity in general (see for example Moeslund et al. 2001).

The complexity of the “*intentional*” analysis can be managed by a semantic approach (see for example Gruber 1991): an ontology is the semantic structure which encodes the implicit rules constraining the structure of a piece of reality (Guarino 1995). We propose an ontology that describes the logical structure of a “intentional system” domain, its concepts and the relations between them. This conceptualization is associated to visual data and instances of classes, forming a knowledge base where the values and their relationship are stored in the same information structure (for example see the approach used in Garcıa-Rojas 2006).

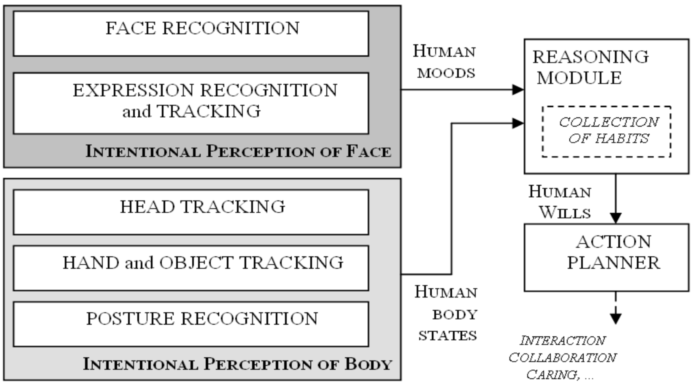


Fig. 1. SeARCH In: Sensing-Acting-Reasoning: Computer understands Human Intentions. Intentional vision framework scheme.

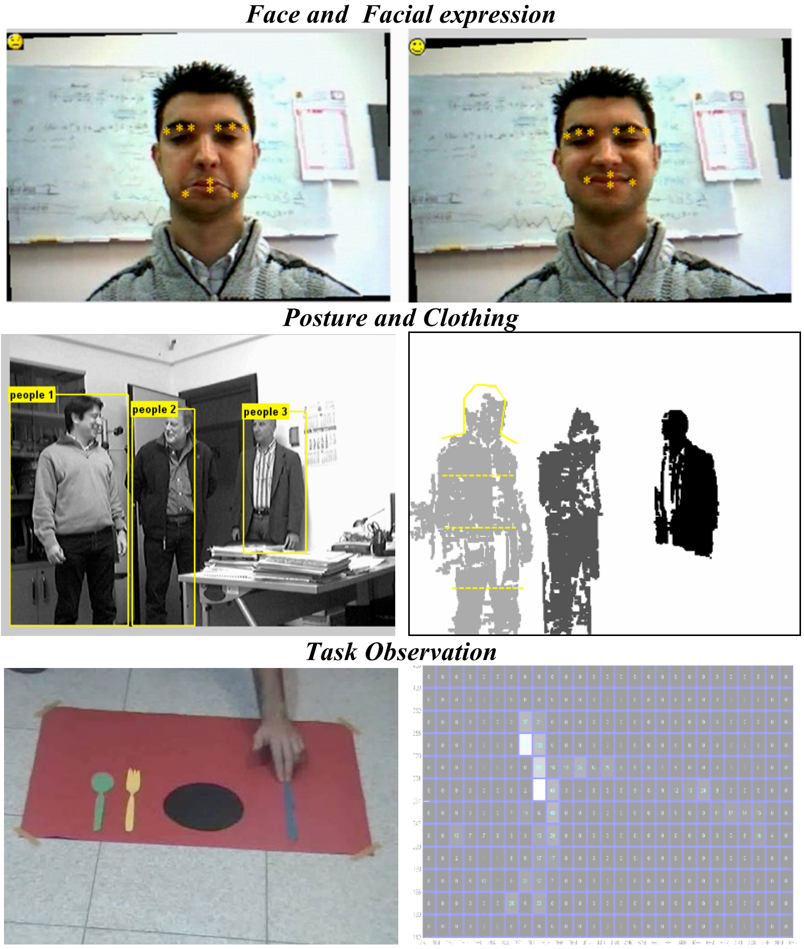
2.1 The Intentional Framework

The proposed framework is named SeARCH In (Sensing-Acting-Reasoning: Computer understands Human Intentions). The relevant modules of the proposed architecture and their functional interconnections are depicted in figure 1. The core of intentional vision system is composed by two specialized modules: Intentional Perception of Body (IPB) and Intentional Perception of Face (IPF). The first module (IPB) deals with the detection and tracking of human bodies. In particular, it tries to locate silhouette, head, and hands of the people detected in the scene and performs their posture recognition. Furthermore, IPB detects and tracks also relevant objects moved by the hands. The output of this module consists in sequences of positions, and shape descriptors corresponding to all the detected entities. The second module (IPF) performs the recognition of the human detected in the scene and his face expression analysis. The main output of IPF module is a temporal sequence of recognized facial expressions characterizing the human mood.

The sequences coming from both modules are linked to the relevant human states (hungry, sleeping, and so on) by the Reasoning Module (RM). RM outputs the interpreted human wills (to eat, to sleep, etc.) on the basis of IPF, and IPM data stream. Its effectiveness is improved on if knowledge of the individual is stored in the Collection of Habits (CH) that represents the memory of RM. Finally, the Action Planner module (AP) decides if and how the system has to interact, collaborate, or assist the human.

Actually, RM has been implemented as a simple rule based algorithm, and it employs queries to consult the knowledge data base described by OWL. This database includes the Collection Habit that has been built by means of a supervised learning phase. Finally, an imitation based approach has been used to record in the Action Planner all the operations necessary to accomplish the task just as the human is used to do it.

The IPF module is devoted to recognize the people detected in the scene and to analyze their facial expression. Its main output is a temporal sequence of recognized facial expressions (see first row of figure 2) that is sent to the reasoning module. The technique described in Viola and Jones (2004) is used for detecting the presence of faces in the scene, and for detecting relevant facial feature points. The facial expression recognition is performed by the tracking (using a Particle Filter) of eyebrows and lips movements. We have implemented a Facial Action Coding System (FACS) (Ekman and Friesen 1978) classifying simple expressions: anger, disgust, fear, joy, sadness, surprise. The sequences of these elementary emotions are recorded to the aim of building a sort of signature representative of a particular human condition in the scene: he/she is hungry, he/she is bored, and so on. In a learning phase, relevant sequences are recorded, and they will be included in the Collection of Habits of a particular person.


Face and Facial expression

Posture and Clothing

Task Observation

Fig. 2. Results of various processed visual inputs. First row: facial expression tracking and recognition example. A rule-based recognition algorithm classifies simple expressions (anger, disgust, etc.) that are indicated by the yellow small icon depicted on upper-left corner of image; sequences of elementary emotions are considered to recognize human mood. Second row: Perception of human movements. Example of multiple people tracking and silhouette extraction. Third row: robot observation of a task shown by the human user: the task of “to lay the table” is simulated by the placing of simple planar objects (cutlery and plate) on red area (the table). Trajectories of object movements are recorded in order to build an occurrence matrix (see left side of the figure related to “fork”). This matrix is used for finding positional relation of objects in the table, and for calculating paths to place an object.

The Intentional Perception of Body module has been designed for the detection of human presence and activity. It accomplishes three main tasks: human silhouette localization and posture recognition (see second row of figure 2), and hands tracking. The input to IPB module comes from at least one fixed camera observing an indoor environment whose map is known. Two more cameras placed on a mobile robotic platform are then used to take close views of human standing in an area “of interest” monitored by fixed cameras. We use a Condensation algorithm to track the people and PCA technique for recognize postures (see Chella et al. 2006). When robot is observing an area “of interest”, the system attention will be focalized on hands and objects movements. The aim is to learn a human task by means of observation, or to collaborate/interact with the person if the task is known. The approach used for implementing both capabilities is inspired to the work of Rao et al. (2007), where a simple statistical analysis is employed. The robot camera view is rectified in respect to the plane of the working area, where human hands manipulate several objects. The detected features are: position of centre of mass, color, and shape (by Fourier descriptors) of objects, and hands position. An occurrences matrix for each entity records the number of times that the object is detected in a particular location of the working plane (see third row of figure 2). When a new working area is observed, or an object never seen before is noticed, occurrences matrices and features values are built or updated. If the knowledge of the working area is already acquired and known objects are detected, the intentional system could execute the task replacing the human actions or collaborate with him/her to place some objects until the final configuration is obtained.

Previous described modules could be composed in a flexible and dynamical way. The aim is to have an adaptable “intentional vision software system” which is capable to act in different situations or scenarios. Generally, we could suppose to have a series of color cameras and video streams accessible by a large band network. A coordinator software agent named CA will collect all visual information resulting from various sources, and will provide an aggregate view of the whole scenario, making available this data to the cognitive architecture for a further analysis. We perform simple experiments, where CA agent includes reasoning module, action planner module, and collection of habits, allowing to have a simplified cognitive architecture. The knowledge managed by intentional vision subsystem, updated at regular time interval, will record the following data: label of identified person, his/her localization, state of motion, posture, and facial expression, positions of his/her visible body parts, behavior pattern. Other information could be considered in order to completely satisfy the requirements of the cognitive architecture.

2.2 Semantic Structure of Visual Perceptual Data

The semantic structure takes in account an experimental scenario that has two special places indicated as working areas. The red working area has been used for showing how “to lay the table” using spoon, fork, knife, plate, and glass; the blue working area has been used for showing the task “to tidy”, using book, pencil, stapler, and eraser. In these areas, the robot observes actions in order to learn tasks by means of examples given to it by humans. Details about each module of proposed architecture are reported in Infantino et al. (2008). Table 1 reports relevant entity definitions and recognition performances where applicable.

	Entity	**Quantity**	**Recognition rate**	**Definition or range**
f_i	*Face*	15 persons / 750 faces	95%	0000_2-1111_2
e_i	*Expression*	7 elementary emotions	63%	000_2-111_2
m	*Mood*	15 clusters	-	$[e_1,e_2,\ldots,e_{50}]$

k				
p i	*Posture*	7 body postures	95%	000_2-111_2
d i	*Clothing*	10	-	[r_1, g_1, b_1, v_{r1}, v_{g1}, v_{b1}, …, r_4, g_4, b_4, v_{r4}, v_{g4}, v_{b4}]
o i M i	*Object*	10	90%	{color, shape}
			-	Occurrence matrix
t_i	*Task*	7	-	000_2-111_2
h i	*Habit*	15	-	h_{ik}=[f_i,d_i,m_k,t_k]

Table 1. List of relevant entity definitions and recognition performances. For example, second row indicates that 750 faces of 15 persons has been processed, the face recognition rate was 95%. A binary value f i is associated to each person (code 00002 means that face is not recognized).

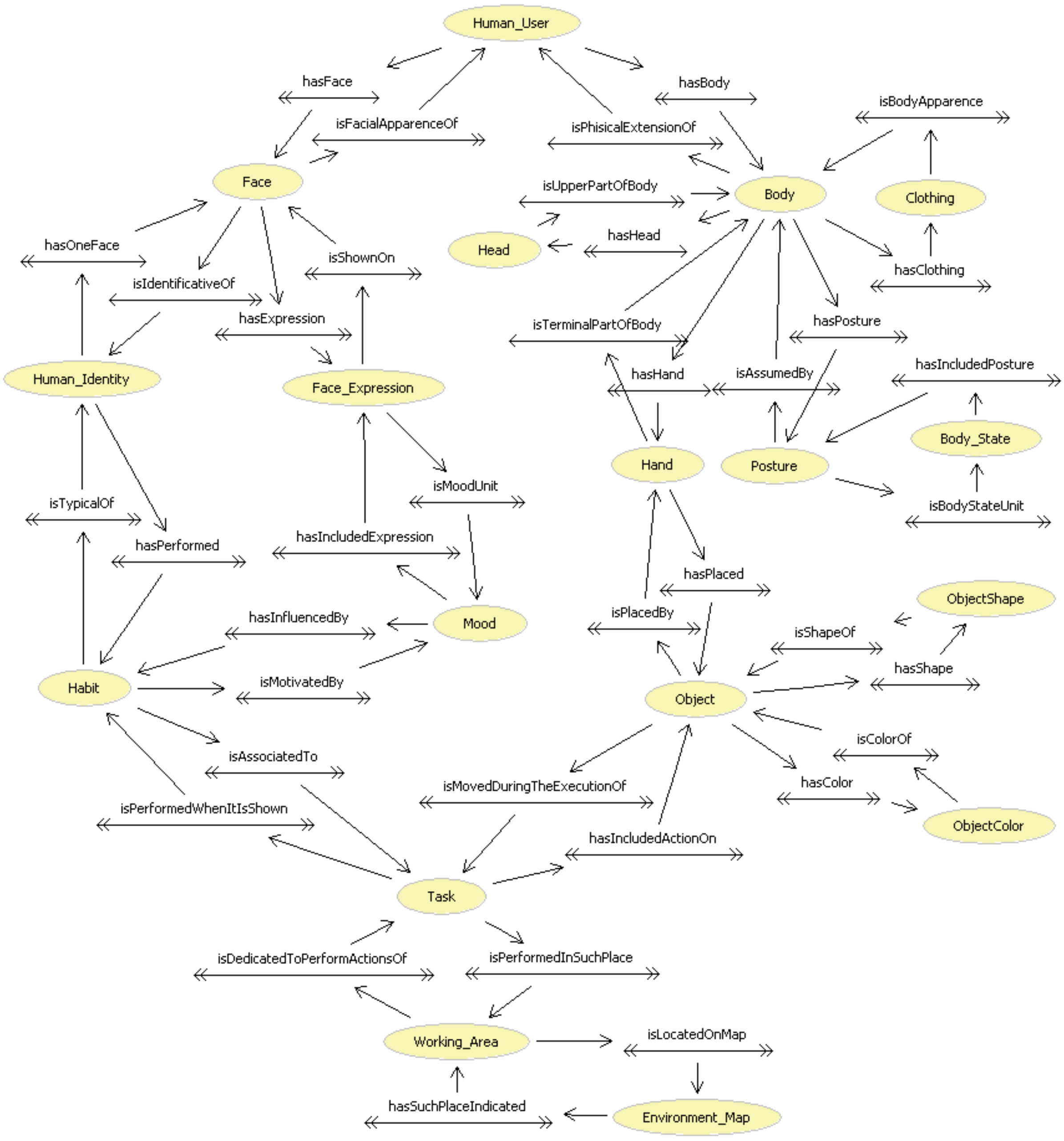


Fig. 3. Diagram of SearchIn Ontology: principal classes and attributes (direct and inverse relationship).

We designed an ontology related to *SearchIn* framework domain in order to manage extracted visual knowledge, and to process it by an inference engine. The ontology is implemented as OWL-DL model by using Protegè (see Protégé link in the bibliography). Fact++ reasoning engine has been used for checking ontology consistence. Some APIs has been used for performing queries, and data retrieving. Even if numeric data are related to the previous described scenario, the ontology can be adapted to other similar experiments. All data described in table 1 correspond to values of individuals specified in the ontology and are included in the following classes: Human_Identity, Face_Expression, Human_Mood, Body_State, Posture, Clothing, Working_Area, Environment_Map, Task, Object, Habit. A scheme of the principal defined classes, and attributes (direct and inverse relationship) are showed in figure 3 (by using GrOWL tool, see the bibliography). Moreover the following subclasses are defined: ProfileFace, and FrontalFace; LeftHand, and RightHand; TableObject, and DeskObject (they have more subclasses such as fork, spoon, knife, pen, eraser, and so on).

As example of expressiveness assured by the implemented semantic structure, some queries (using Protegè DL Query Tab) are reported in the following:

Who has at least one recorded habit?
Query: *Human_Identity and hasPerformed some Habit*
Results (Instances): *InoKnownIdentity*
DanielaKnownIdentity
IgnazioKnownIdentity
FilippoKnownIdentity
What are Filippo's habits?
Query: *Habit and isTypicalOf value FilippoKnownIdentity*
Results: *H4_Habit*
H1_Habit
(H1 and H4 Habit are data stored in knowledge base)
Which is the Filippo's habit when he is hungry?
Query: *Habit and isTypicalOf value FilippoKnownIdentity and isMotivatedBy value HungryMood*
Results: *H1_Habit*
Which task is executed by Ignazio when he is confused?
Query: *Task and isPerformedWhenItIsShown some (Habit and isTypicalOf value IgnazioKnownIdentity and isMotivatedBy value ConfusedMood)*
Results: *ToTidyUpTask*

2.3 Example of application

A list of objects and occurrences matrices $\{o_j, M_j\}_k$ corresponds to each task t_k. During the learning phase, when a human is near to a working area, the robot goes there to recognize him/her and observe actions. In normal activity, after the learning phase, the human-robot interaction is regulated by following set of simple rules:

- if the people tracking module detects a person close to a working area, and d_i is similar to a known one, the CA agent sends a command to make the robot approach such a place;
- if the face is recognized, then the robot observes the face expressions in order to determine his/her mood; else a new person is introduced in face database;

afterwards the robot searches and selects a task among the available "collection of his/her habits" given the recognized mood. This task represents the human will to satisfy. We have performed 10 experiments for each task ("to lay the table", and "to tidy"): 5 are related to the learning phase and 5 to the collaboration one. Even if this is a preliminary experimentation, we report only 3 failures: 2 are due to erroneous recognition of human moods, and the other to erroneous recognition of the human face.

2.4 Future works

The described framework aims to obtain a vision systems focused on the extraction of information useful to understand human wills. We have described a possible composition of several standard artificial vision algorithms for implementing an intentional vision system to insert in a cognitive architecture. Different applicative scenarios will be considered to have an exhaustive testing phase of the proposed architecture. Our intent is to exploit all the advantages of semantic structure, and to obtain more sophisticated reasoning and planning modules.

3. Bibliography

Chella, A., Dindo, H., and Infantino, I., (2006), "People Tracking and Posture Recognition for Human-Robot Interaction", in proc. of International Workshop on Vision Based Human-Robot Interaction, EUROS-2006.

Duda, R.; Hart, P. & Stork, D. (2000). Pattern Classification, Wiley-Interscience

Ekman, P., and Friesen, W.V., (1978), Manual for the Facial Action Coding System, Consulting Psychologists Press, Inc.

Garcıa-Rojas, A., Vexo F., Thalmann, D., Raouzaiou, A., Karpouzis, K., Kollias, S., Moccozet, L., and Magnenat-Thalmann, N., (2006), "Emotional face expression profiles supported by virtual human ontology", Comp. Anim. Virtual Worlds, vol. 17, pp. 259–269.

Gopnick, A. & Moore, A. (1994). "Changing your views: How understanding visual perception can lead to a new theory of mind," in Children's Early Understanding of Mind, eds. C. Lewis and P. Mitchell, 157-181. Lawrence Erlbaum

Guarino, N., 1995, "Formal Ontology, Conceptual Analysis and Knowledge Representation", International Journal of Human and Computer Studies, vol. 43(5/6), pp. 625-640.

GrOWL, Graphical editor of OWL-DL models, http://home.dei.polimi.it/arrigoni/GrOWL/

Gruber, TR., 1991, "The role of a common ontology in achieving sharable, reusable knowledge bases". In Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning, pp. 601–602.

Infantino, I., Lodato, C., Lopes, S., Vella, F., (2008), "Implementation of a Intentional Vision System to support Cognitive Architectures", International Workshop on Robotic Perception (VISAPP-RoboPerc08), 3rd International Conference on Computer Vision Theory and Applications VISAPP'08, Funchal, Madeira – Portugal.

Kelley, R.; Tavakkoli, A.; King, C.; Nicolescu, M.; Nicolescu, M. & Bebis, G.. (2008). Understanding human intentions via hidden markov models in autonomous mobile robots. In Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction (HRI '08). ACM, New York, NY, USA, 367-374. DOI=10.1145/1349822.1349870

Kelley, R.; Tavakkoli, A.; King, C.; Nicolescu, M. & Nicolescu, M. (2010). Understanding Activities and Intentions for Human-Robot Interaction, Human-Robot Interaction, Daisuke Chugo (Ed.), ISBN: 978-953-307-051-3, InTech, Available from: http://www.intechopen.com/articles/show/title/understanding-activities-and intentions-for-human-robot-interaction.

Kuno, Y., Ishiyama, T., et al., (1999), "Combining observations of intentional and unintentional behaviors for human-computer interaction", in proc. of the SIGCHI conference on Human factors in computing systems, Pittsburgh, Pennsylvania, USA, pp. 238-245.

Moeslund, T.B., and Granum, E., (2001), "A survey of computer vision-based human motion capture", Computer Vision and Image Understanding, vol. 18, pp. 231-268.

Pantic, M.; Pentland, A.; Nijholt, A. & Huang, T.S. (2006). Human Computing and Machine Understanding of Human Behavior: A Survey, in proc. Of Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06), pp. 239-248.

Protegè, Ontology editor, http://protege.stanford.edu/

Tavakkoli, A.; Kelley, R.; King, C.; Nicolescu, M.; Nicolescu, M. & Bebis, G. (2007). "A Vision-Based Architecture for Intent Recognition," Proc. of the International Symposium on Visual Computing, pp. 173-182

Turk, M., (2004), "Computer Vision in the Interface", Comm. of the ACM, vol. 47, no 1.

Viola, P., and Jones, M. J., (2004), “Robust Real-Time Face Detection”, International Journal of Computer Vision, vol. 57, no 2, pp. 137-154.