Multimodal High Level Data Integration

Olga Vybornova, Hildeberto Mendonça (1), Daniel Neiberg (2), David Antonio Gomez Jauregui (3), Ao Shen (4)
(1) UCI-TELE, Louvain-la-Neuve, Belgium, (2) TMH/CTT, KTH Royal Institute of Technology, Sweden, (3) TELECOM and Management SudParis, France, (4) University of Birmingham, UK

Abstract
A method of multimodal multi-level fusion integrating contextual information obtained from spoken input and visual scene analysis is defined and developed in the context of this research. The resulting experimental application poses a few major challenges: the system has to process unrestricted natural language, free human behaviour, and to manage data from two persons in their home/office environment. The application employs various components integrated into a single system: speech recognition, person identification, syntactic parsing, natural language semantic analysis, video analysis, human behavior analysis, a cognitive architecture that serves as the context-aware controller of all the processes, manages the domain ontology and provides the decision-making mechanism.

Index Terms—multimodal data integration, semantic representations, cognitive architecture, spoken input, video input, decision-making.

I. PRESENTATION OF THE PROJECT
A. Motivation
ince Bolt's groundbreaking work on speech- and gesture-based interaction [2], multimodal systems and fusion engines have progressively evolved towards more robust semantic interpretation and they all include some fusion mechanism to combine the separate data streams by reducing uncertainty. The interest to jointly process audio and visual information related to human activities, and to extend the technological developments in individual modalities for human-computer interaction has been increasing. To improve naturalness and robustness of user interfaces they should be able to automatically recognize human identity, intentions, speech, actions in computing environments, define the person localization, perform source separation, media synthesis and media content mining. A critical factor that allows to make multimodal processing in speech-based interactions effective is the robust integration or fusion of information from multiple modalities.

Multimodal fusion is a central question to solve and provide effective and advanced human-computer interaction. To understand and formalize the coordination between modalities involved in the same multimodal interface it is required to extract relevant features from signal representations and thereafter to proceed to high-level (semantic-level) fusion. Signal- and semantic-level processing is deemed tightly interrelated since signals are seen as bearers of the meaning. To provide natural and robust interaction with the user(s) the system must be able to interpret as a whole the various modalities used by a human to interact, i.e., to extract information arriving simultaneously from different sources and to combine them into one or several unified and coherent representations of the user’s behavior. A challenge is to fuse these different input modalities effectively to augment their natural strengths and use redundancy to increase robustness. Moreover, it is critical to decide what information should be fused and what should not.

We see multimodality as a set of variables describing states of the world (user’s input, an object, an event, behavior, etc.) represented in different media and through different information channels. In this respect the goal of data fusion (merging semantic content from multiple streams) should give an efficient joint interpretation of the multimodal behavior of the user(s) – to provide effective and advanced interaction.

We have been working with multimodal fusion considering two main modalities, which are speech recognition and human behavior analysis through image processing tools. We have conducted experiments in a hypothetical context, with two or more people interacting with each other and with objects in the scene. We observed that when intending to perform an action, the user(s) might:
• speak about it describing their actions or intentions (the ideal case, then the linguistic and action recognition data are complementary);
• be doing things, but speaking about something absolutely irrelevant to those actions (in this case the data from the two modalities should be analysed separately, without merging);
• the users’ words might contradict the actions performed (then the actions have priority, since “actions speak louder than words”);
• be just silent when performing actions (in this case the unimodal operation mode is required and no data fusion is needed).

Everything what is said or done is meaningful only in the particular context. To accomplish the task of semantic fusion we should take into account the information obtained at least in the following three types of context [7]:
- domain context: meaning prior knowledge of the domain, semantic frames with predefined action patterns, user profiles, situation modelling, a priori developed and dynamically updated ontology defining subjects, objects, activities and relations between them for a particular person;
During the early fusion, some data received from each modality, which meaning is redundant in two or more modalities, should be synchronized in order to strengthen the link between data and optimize the late fusion. An example of early fusion done in our experiment is the person identification from the speaker and from the image features. This singular meaning helps us to complement a recognized intention from the speech with the behaviour of the speaker, predicting what decision is more appropriated for the situation. When the early fusion finish, the search for semantics continues with the application trying to identify plurality of meanings. This is the input of the late fusion, which deals with higher symbolic elements abstracted by the modality recognizing. These elements are first mapped between the different modalities and then integrated during the final decision stage to resolve uncertainty and produce the user's intention interpretation. One example of plurality of meanings on speech recognition is to detect the human intention linking the person (subject) with a concrete or abstract things (object) through a action or relation (predicate). In the sentence “I want to call Nick” there is a clear intention represented by links between “I”, “want to call” and “Nick”, which are subject, predicate and object respectively. These meanings will be combined with other meanings during the late fusion. In addition to the early and late fusion, we are contributing with the reinforcement learning, which is a cyclical process that uses all found meanings to reinforce or redefine them selves, producing better inputs for the late fusion. The need for reinforcement came from a reflection about how a existent modality can help the analysis of other modalities preserving the modality cohesion, where a modality must not be aware of the presence of other modalities in order to avoid some level of coupling. We could observe this need in our experiment, when we implemented n-best reranking functions for the speech recognition analysis to find the best hypothesis about what was said. These functions need direct access to the list of concepts from the ontology and what was analysed from other modalities, in order to probabilistically rank the best hypothesis possible. The reinforcement fusion is more robust than early fusion, less conclusive than late fusion, and it is not mandatory, but recommended when there are clear possibilities of improvement.

Developing optimal strategies for fusion is still a crucial issue as none of the known approaches has so far emerged as the outstanding solution. Various different formalisms have been proposed and mainly consist of deterministic fusion schemes:  
- rule-based implementations: including variants such as unification grammars used in the multidimensional parsing of typed feature structures, finite state transducers supporting tighter coupling of parsing and input recognition, finite state machine for biologically inspired rhythmical synchronization;  
- statistical methods: such as integration based on maximum entropy;  
- mixed approaches: probabilistic agent-based resolution and integration performed with simple voting mechanism based on confidence values of resolved concepts, probabilistic salience weighting of action-based concepts that are bound to speech concepts according to a predefined set of rules.

II. APPLICATION, SCENARIOS AND MODALITIES OF INTEREST

In order to develop and validate the whole experiment, we defined five different scenarios where two people interact mutually in a room, talking in a natural way and behaving without restrictions. Each scenario tries to explore specific combinations of speech and behavior to increase the robustness of the system. This system is able to track people, analyze their behavior, movements, speech, and takes decisions about how to prompt them necessary information when required or provide any other assistance.

The main challenges that we face in this application are unrestricted human language and free natural behavior within home/office environment. In order to analyze behavior efficiently, the system has to correctly process, interpret and create joint meaning of the data coming from speech analysis and video scene analysis. We consider that human behavior is goal-oriented, so our main aim is to recognize the users’ plan, to see what they want to achieve. The system manages the data streams arriving from two sources – video scene and speech. In particular, we show a technique distinguishing between the data from different modalities that should be fused and the data that should not be fused but analyzed separately.

III. PROJECT ARCHITECTURE AND IMPLEMENTATION

The application employs various components integrated into a single system: speech recognition, speaker identification, syntactic parsing, natural language semantic analysis, video analysis, human behavior analysis component and a cognitive architecture that serves as the context-aware controller of all the processes, manages the domain ontology and provides the decision-making mechanism.
The next step is to analyse the human behaviour, comparing the movements of the user with a set of rules. The behaviour is relative to fixed objects in the scene, which are defined in the context domain and are directly associated with the aid to be given by the system. The rules define the boundaries of what is near of far from a certain object. Then the result of the rule processing at this stage is: [person] is near to [telephone], [person] is far from [computer] or [person] is moving to [library]. This result is produced for each person in each frame of the video. Individually, these results are not significant enough for fusion. We have to analyze the movements in many frames in order to have final conclusions. For instance: if in the last 80 frames the rule engine produced "[person] is moving to [library]" then we can conclude that there is a real intention to achieve the library, considering some variables of the environment, such as area of the room. The late fusion occurs when we identify a person moving to the library and we also detect the intention to find a book in a sentence like this: “I can find a book about it in the library.”. Therefore, we can conclude the person is moving to the library because he wants to find a certain book there. Previous sentences, already analysed, indicates that the article “it” in the current sentence actually means “French wines”, an object identified through the semantic analysis.

Many existent components were just reused, such as Sphinx, C & C Parser, C & C Boxer, Open CV, Protegé and Soar. Other components were developed from the beginning, which was the case of the Human Behavior Analyzer, the Fusion Mechanism and the interoperability between all components, which is implemented using sockets, a simple network communication strategy.

Speech Recognition

1) Speech Data

The speech was recorded by two non-native subjects, one 23 year Chinese male and one 32 year Swedish male. The data was recorded using 16kHz, 16 bit audio. The 5 scenarios consisted of a total of 72 sentences and 148 seconds. For development of speaker identification, we used ten phonetically rich sentences for training, and for parameter tuning we used another ten phonetically rich sentences and ten words of different length.

2) Speech Recognition

For speech recognition we used Sphinx 4, which is an open source, Java-based speech recognizer [29]. For acoustic modeling, we used the 8 Gaussian triphone models, trained on the Wall Street Journal Corpus, which are supplied along with Sphinx. Since we wanted to allow the system to monitor a discussion between two or more people, we want to have a large vocabulary language model. For this purpose, 3-grams with a maximum of ~5000 words were trained using the orthographic transcriptions from the Wall Street Journal Corpus. The 5000 words were selected as the most common ones plus the ones that are present in the scenarios.

In order to enhance understanding of concepts, we wanted to evaluate different methodologies to pick hypotheses from the N-best list which Sphinx can produce. For this purpose, Sphinx was configured to produce as long N-best lists as possible given a memory constraint of 1 Gb. This resulted in list lengths of ~100 hypotheses for short utterances with good accuracy and ~10000 list lengths for long utterances with poor accuracy. Two different approaches may be used. The first is to estimate a re-scoring function on data, the second is to apply re-ranking heuristics. The first approach is likely to be most efficient, but it requires enough data to be done.
optimally. The second approach is easier to apply and is likely to be more robust to less knowledge about the users may say.

By making an assumption about the domain, a word list consisting of all nameable things in the ontology was queried. Three kinds of re-ranking heuristics were tried out. In each of these, a score was assigned to each hypothesis and then N-best list was re-sorted using this score in such way that the initial ranking was preserved if two or more hypotheses had equal score.

Method A: The score is set to one if one or more words from the ontology are found.

Method B: One scoring point is added for each matching word in the ontology word list.

Method C: A score is added, which is equal to the word length in characters, for each matching word in the ontology word list.

In Tables 1 and 2, the performance gain by this approach is compared to selecting the 1st top hypothesis and by letting an oracle select the hypothesis which maximizes Accuracy. These results where obtained using a preliminary version of the ontology. While Accuracy takes substitutions, insertions and deletions into account, Correctness does only take substitutions and deletions into account. In table Y, Accuracy and Correctness is computed by only using the words from the ontology as a reference, forming Concept Accuracy and Concept Correctness. The idea is to examine to what extent information, which the cognitive modules can process, is present.

### Table 1: Word accuracy and correctness for different hypothesis selection methods

<table>
<thead>
<tr>
<th>Hypothesis selection method</th>
<th>Accuracy</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>56,30</td>
<td>64,52</td>
</tr>
<tr>
<td>Method A</td>
<td>50,13</td>
<td>62,98</td>
</tr>
<tr>
<td>Method B</td>
<td>46,53</td>
<td>63,50</td>
</tr>
<tr>
<td>Method C</td>
<td>47,81</td>
<td>64,01</td>
</tr>
<tr>
<td>Oracle</td>
<td>74,81</td>
<td>78,15</td>
</tr>
</tbody>
</table>

### Table 2. Concept accuracy and correctness for different hypothesis selection methods

<table>
<thead>
<tr>
<th>Hypothesis selection method</th>
<th>Accuracy</th>
<th>Correctness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>54,12</td>
<td>67,06</td>
</tr>
<tr>
<td>Method A</td>
<td>27,06</td>
<td>68,24</td>
</tr>
<tr>
<td>Method B</td>
<td>10,59</td>
<td>78,82</td>
</tr>
<tr>
<td>Method C</td>
<td>12,94</td>
<td>81,18</td>
</tr>
<tr>
<td>Oracle</td>
<td>78,82</td>
<td>74,12</td>
</tr>
</tbody>
</table>

From the results presented in Table X, it is clear that the methods don’t improve over the baseline (1st hypothesis). It should also be noted, that these results are obtained using a preliminary version of the ontology. From the results in Table Y, it should be noted that while there is a large drop in Accuracy, there is a clear increase in Correctness, which means that insertions are present. This means that the information content in terms of words known to the ontology is increased on the expense of word insertions.

3) **Speaker Identification**

Speaker Identification is the task to determine who is speaking. For the application is described in this report, a standard speaker identification system was considered. It is based on Mel-Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs), as in Renynold et. al. We used 28 log-Mel filters between 300 and 8000 Hz, cosine projected to 24 dimensions, where the first 12 with their delta where used. A simple data driven procedure for speech detection was tried: The MFCCs where clustered using the k-means algorithm with euclidean distance for two clusters. Then the cluster which had the highest energy was marked as “speech” and the other one as “non-speech”. Then these two clusters where used for frame based segmentation. Visual inspection showed that the approach seemed reasonable. No channel compensation was used. This system was implemented in Matlab.

The two speakers in the scenarios where enrolled, using ten phonetically rich sentences. For parameter tuning, another ten phonetically rich sentences and ten words were used. The performance measured in classification accuracy for various numbers of Gaussians are shown in Table 3. Further analysis showed that the errors occurred for test utterances containing very short words, such as “no” and “hi”. Using speech detection didn’t improve accuracy. Based on these results, 16 Gaussians where chosen for evaluation using the recorded data for the scenarios. The accuracy for the evaluation data was 94%.

### Table 3. Speaker identification accuracy, with and without speech detection (SD)

<table>
<thead>
<tr>
<th>Gaussians</th>
<th>No SD</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>85,00%</td>
<td>92,50%</td>
</tr>
<tr>
<td>4</td>
<td>90,00%</td>
<td>92,50%</td>
</tr>
<tr>
<td>8</td>
<td>97,50%</td>
<td>92,50%</td>
</tr>
<tr>
<td>12</td>
<td>97,50%</td>
<td>92,50%</td>
</tr>
<tr>
<td>16</td>
<td>97,50%</td>
<td>90,00%</td>
</tr>
<tr>
<td>20</td>
<td>97,50%</td>
<td>95,00%</td>
</tr>
<tr>
<td>24</td>
<td>92,50%</td>
<td>95,00%</td>
</tr>
<tr>
<td>32</td>
<td>97,50%</td>
<td>92,50%</td>
</tr>
</tbody>
</table>

4) **System integration of the speech components**

A speech module, consisting of Sphinx 4 and the Speaker Identification system described above was fully implemented
in a TPC/IP client. Also on the server side, the re-ranking method A was implemented, in way which makes it easy to extend to Method B or C.

B. Syntactic and Semantic analysis of spoken input

The next stage after speech recognition is syntactic and semantic analysis of the discourse. For our purposes we use the CCG (Combinatory Categorial Grammar) parser, Release 0.96, developed by S. Clark and J. Curran [9]. The grammar used by the parser is taken from CCGbank developed by J. Hockenmaier and M. Steedman [3].

CCGbank is a treebank containing phrase-structure trees in the Penn Treebank (WSJ texts) converted into CCG derivations. It allows easy recovery of long-range dependencies, provides a transparent interface between surface syntax and underlying semantic representation, including predicate-argument structure. The grammar is based on 'real' texts, and that is why it has wide-coverage, thus making parsing efficient and robust.

The CCG parser has Boxer [8], as add-on to generate semantic representations - Discourse Representation Structures (DRSs), the box representations of Discourse Representation Theory (DRT) [13]. DRSs consist of a set of discourse referents (representatives of objects introduced in the discourse) and a set of conditions for these referents (properties of the objects). Our initial experiments with the CCG parser together with Boxer showed that it suits well for our application in parsing speech of elderly people which is on one hand somewhat restricted to their environment, background and social relationships (the domain model and user profile help us cope with the challenge), but on the other hand is naturally broad enough to need wide-coverage tools for processing. DRSs can be generated in different output formats in Prolog or XML. To link the DRSs output with the ontology we use the XML format.

C. Visual scene analysis

1) Introduction

The video information provides a description of what happens in the environment at a given time. In this project a video sequence is made for each possible scenario. The video sequences were recorded using a distributed 8-camera voxelised visual hull [22]. The description of the environment is obtained by processing each image of the videos using computer vision algorithms.

2) Problem

In these video sequences, there are 3 types of fixed objects (a telephone, some books and a computer) located in different positions inside the scene, there are also two persons that are moving and interacting with these fixed objects. In order to have a good description of the environment for each one of these scenarios is necessary to extract the information of the position of each fixed object, the position of each person at any moment and also the motion direction of each person. Using this information it’s possible to know if a person is near or far from each object or if a person is moving to an object. By this way, the system will have good information about the human behavior in order to make better decisions.

3) Procedure

Extracting all this information of the video sequences are common problems of computer vision field. These problems are mainly:

- Object detection (to find the position of each fixed object)
- People detection and tracking (to find the position of each person)
- Motion analysis (to find the motion direction of each person)

The computer vision system to extract this information was implemented in C++ using OpenCV library [12] developed by Intel. In the next sub-sections, will be described the procedure and algorithms to resolve the problems already mentioned.

a) Finding the position of each object

The objects (the telephone, the books and the computer) in these video sequences are always fixed (in the same position), also they don’t change size or rotate because the camera’s viewpoint is always the same. Hence, the detection of these objects can be easily done by using a template matching algorithm [7]. These algorithms compare a template with a region of an image in order to determinate a similarity measure, wherein the similarity measure is determined using a statistical measure.

To obtain these templates, a sample picture of each object is captured from any image of the video sequence. The OpenCV operator cvMatchTemplate was used, this operator returns the probable positions in the image where the template can be located. By this way, the most probable position corresponds to the area where the object is detected. The similarity measure that provides better results for this problem was the correlation coefficient normalized. Because each object never changes position, this template matching step is only done once in the first captured image of the video sequence. Then, these positions are used for all the images in the video sequence.

b) Finding the position of each person

There has been much work in the area of people detection and tracking using computer vision techniques. This problem has been a big challenge since many real-time vision systems require robustness in different environments. In the video sequence of this project, two persons are talking with each other and also moving randomly inside the scenario in order to interact with these objects. The problems here are mainly:

1. The shape and size of each person can change over time because the persons can move far or close to the viewpoint in different parts of the scenario.
2. The persons can communicate too much from each other in the video, making difficult the identification of each one.
3. One person can be partially occluded by the other person.
4. Some body parts of each person can be outside the scenario because of the viewpoint of the camera.
5. Each person moves in a random way.

In order to resolve this problem, a color-based tracking was used. Principally the color of the clothes of each person was used, assuming that both people in the video will have different color clothes. However, in order to make this people detection robust to the cluttered environment, a background subtraction technique and also blob detection was implemented in order to discriminate noise. The procedure can be described in this way:

1. Firstly, the first frame in the image was used in order to estimate the background image. This, of course, assumes that nobody is going to be present in the first frames of the video sequence. In order to detect the foreground, it was calculated the absolute difference of the pixel values between image with the people present and the background image (using operator cvAbsDiff from OpenCV) [9]. Then, a threshold in the absolute difference image was applied. By this way, the foreground detected will be the silhouettes of each person. This technique, although is easy to implement and it has a very fast processing time, have the disadvantage of not discriminating the shadows of people.
2. Secondly, a Single-Gaussian Model [5] of each color was applied to the silhouettes in order to obtain the probability that the pixel values corresponds to a color. This Gaussian Model (mean and covariance matrix) was learned in an initialization step using samples manually taken of the clothes color of each person. The result of this process is a binary image where the pixels that are above a threshold are marked as belonging to each color.
3. After, a morphological opening operation was applied to the result image of the previous step. By applying this, we can obtain a more regular shape in the binary image. This was implemented using OpenCV operator cvMorphologyEx with a 3 x 3 square structuring element.
4. In this step, a blob detection and analysis was implemented in order to find the region of interest (ROI) that corresponds to each person. The analysis is made by using the size information of each blob in order to discriminate noise or false detections. After this analysis is done, we can estimate the bounding box of each person using the body human proportions.
5. In order to follow each person, the size of the bounding box of each person is increased for the next frame and the person is searched inside this area. This is a very simple tracking, however, it gives good results since the persons are not far from the camera and also they don’t change position too fast.

This procedure, although, it works very well for all the video sequences of this project maybe it can not give the same result for other types of video sequence. Consequently, some experiments were done using features tracking from each person, by this way, the idea is to identify and track using optical flow (pyramid-based KLT feature tracking) [15] a group of features conserving the distance properties between each feature of each group [14]. In these experiments some features were lost because of the occlusions, making it difficult to identify each person. However, this can be a very good approach for future research in order to improve the robustness of this system.

c) Finding the motion direction of each person

In order to find the motion direction of each person, Motion Templates algorithms were used based on papers by Davis and Bobick [10] and Bradski and Davis [4]. These algorithms are very fast and robust. The implementation was done using OpenCV Motion Template functions. This functions can determine where a motion occurred, how it occurred, and in which direction it occurred. To calculate the motion direction of each person, the silhouettes (obtained by background subtraction as described in section 3.2) are updated in time using cvUpdateMotionHistory operator, after the motion gradient is calculated using the information of the temporal silhouettes (applying cvCalcMotionGradient), finally connected regions of Motion History pixels are found using OpenCV operator cvSegmentMotion.

With this result, we have region of motions with their gradient directions in the foreground image. In order to find what region of motion corresponds to each person, it was used the result of the procedure described in section 3.2. By this way, the gradient direction of the biggest region of motion inside the bounding box of each person corresponds to the motion direction of the person.

4) Human behavior analysis

Using the information described in the section 3, it’s easy to know the human behavior at any time in the video sequence. In this step, we need to know principally:

1. If every object is near or far from each person.
2. The object toward which each person moves.

In order to find this information we only apply some rules and define thresholds using the result of the computer vision system (position of each fixed object and each person and motion direction of the persons). This human behavior information is sent by a socket TCP/IP connection to the multimodal high level data integration system in order to be fused with the speech recognition information (speech modality).

5) Results and conclusions

The computer vision system was executed in a laptop with an AMD Turion 64 2000 MHz processor, 1024 MB of RAM and a graphic card NVIDIA GForce Go 6150 (256MB). The computer vision system is capable of providing the required
information of all the 5 scenarios in real-time processing (25 frames by second). The system is robust to partial occlusions and it can detect and track both persons in any part of the scenario providing their position in all frames. In Figure 2, we can see the result of the computer vision system. The red circles with the red line inside is the motion direction (0 – 360 degrees) of each person, the blue pixels represent the Motion History of the human silhouettes, the white blob is the estimated area where each person is identified inside the scenario and the red small squares is the position of each fixed object (the computer, the books and the telephone).

In order to improve the robustness of the system in different illumination conditions, an approach of features tracking with optical flow can be used, by this way, the identification of each person is not going to rely only on the color and the size, also we’re assuming that each person has different color clothes, this is not always true in a real scenario. Maybe for future research it can be used also particles filter approach in order to estimate the position of each person in the next frame improving the robustness of the tracking.

D. Ontology design and reasoning

In order to have a database for semantic reasoning, an ontology with comprehensive modeling pattern of multimodal actions and prior knowledge about the objects and user is necessary. We used open source tool Protégé [20] to create the ontology in classes, properties and individuals. Figure 4 is the looking of a developed ontology in Protégé 4.0 alpha environment.

We use Soar [24] cognitive architecture in order to create and apply rules to query semantically the modality triples in the ontology. The ontology is re-implemented for adapting the specific working memory structure in Soar environment. To a certain range, class, property and individual in Protégé 4.0 alpha correspond to identifier, attribute and constant. Figure 5 shows an example of graph structure of the knowledge base. In order to have this knowledge base for reasoning, the ontology is created as states when the agent is initialized. After receiving the input from external application, an operator will be proposed and applied to map the input triple to the knowledge base then give out the result to the output link.
For the interacting between Soar and external world, we use SML (Soar Markup Language) to send and receive commands packages as XML packets. The agent created by soar is loaded by the SML codes and run till there is any outcome or run for a certain number of circles which is set by codes too.

The outcomes are checked by SML code as well and kept for further actions. See Figure 6 for the execution cycle.

In our project, when we detect there is an input triple in the input-link, we will try to map it to ontology. Different operators will be proposed corresponding to the conditions. Then in the application stage of the operator, output will be generated and retrieved by the SML code. Further actions are taken according to the outputs.

IV. EXPERIMENTS AND RESULTS

Let us consider one of the annotated example scenarios:

(1) [Beto] Hi Ronald! How is the life going?
(2) [Ronald] I am fine.
(3) [Ronald] I want to call Nick.
(4) [Beto] What for?
(5) [Ronald] He mentioned that he attended a wine tasting course.
(6) [Beto] It sounds interesting, I like wine.
(7) [Ronald] Actually I plan to join the next class. He also mentioned a book about French wines, but I cannot recall the name of the author.
(8) [Beto] Why don't you send a mail to Nick?
(9) [Ronald] Maybe I can find a book about it in the library.
(10) [Beto] Yes, you are right.
(11) [Beto] Did you find it?
(12) [Ronald] Yes, I did.

In our experimental work we used scenarios only with natural human language – we did not work with isolated words, commands, restricted language or something like that. My goal was to experiment with normal complete utterances expressed in a natural way. It should be noted here that we did not do spontaneous language like with pauses, interjections, hesitations, overlaps in speech, etc., because the wide-coverage CCG parser [8] that we used, is trained on newspaper texts, cannot analyze fragments of sentences, it accepts only well-formed complete utterances.

For us, to make multimodal data fusion means to interpret human behavior, to identify the users’ plan, infer their intentions. Having understood what the people in the scene want to do, the system takes the decision about how to assist them in the given case. Looking at our challenging example scenario, we can see that there are 4 points in this dialog where the speakers express their plan to do something. In (3) Beto wants to call Nick, in (7) Ronald plans to attend the next class, in (8) there is a possible plan to send a mail and then this possible path of decision is changed for another route - in (9) there is an intention to find a book in the library. How shall the system realize what plans to take into account and what not? When to react and when not? This is why we employ multimodal information – when the plan is identified from the user’s words, we look at the other modality data to see, if the person is going to “confirm” his words with the corresponding actions or not. In (3), (7) and (8) the persons expressing the said intentions were standing still, just continuing the talk.

And only in the (9) utterance Ronald moved to the bookshelves. That is why only in this last case of plan expression the system reacted and prompted where to find the desired book. By the way, in the phrase “Maybe I can find a book about it in the library” we have to resolve ambiguity between the library in the room and a library on the web. We again do that using information from the other modality – we look if the person is moving to the books in the room or if he is moving to the computer.

Schematic examples of decision making process:

(3) If (Ronald) [wants to send] {email to Nick} & (Ronald [is moving to] {the computer} | He [is close to] {the computer}) then

>> open the mail client with the “to” field filled with nick@uclouvain.be

(9) If (Ronald) [can find] {book} [about] {it} [in] {the library} & (Ronald [is moving to] {the library}) then

>> There is a book about French wines on the first shelf.

(9) If (Ronald) [can find] {book} [about] {it} [in] {the library} & (Ronald [is moving to] {the computer}) then

>> open a web search website and put the keyword in the search field

When identifying the person’s plan from speech, we basically rely on the linguistic semantic analysis as described in Section III B, but we certainly take advantage of the obvious lexical
signs of plan and intentions expression. For example, such verbs and phrases like “want”, “wish”, “plan”, “going to”, etc. (we defined 19 expressions like this in total) in the certain syntactic context and in the present or future tense clearly point at the person’s intention to do something. And vice versa, negative forms of verbs like “I don’t want”, “I have no wish to...”, “You don’t want to...” as well as verbs in the past tense serve as stop-words, and signal that this plan should be discarded and not taken into account, because no system response is needed.

V. CONCLUSIONS AND FUTURE WORK

During this challenging project we have posed and solved several difficult problems. We have:

• managed to deal with spatial relationships (based on the fixed “anchor” objects in the room)
• made semantic fusion of events not coinciding in time
• achieved good results in speaker identification - synchronisation between image and speech identification
• created an open framework to manage fusion between two (in our case) or more modalities (in enhanced future work)
• designed the system so that each component can run in a separated machine thanks to the distribution mechanism interchanging data through a TCP/IP network.

However, we are not going to stop at this point, but we have even more problems to solve in our future work. To name just a few, we should:

• implement effective learning mechanism
• perform efficient decision making even from information fragments
• handle spatial relationships relatively to moving people
• perform 3D video analysis
• identify detection of orientation of the people in the scene
• add at least one more modality - eye gaze tracking
• recognize various types of gestures
• learn to deal with natural language redundancy (repeating the same idea in different words).

VI. ACKNOWLEDGEMENT

The research described herein was started under the European FP6 SIMILAR Network of Excellence (www.similar.cc) and continues by the European FP6-35182 OpenInterface project (www.oi-project.org).

We thank Diego Ruiz and Ronald Moncarey (UCL-TELE, Belgium) for precious help in the project preparation.

VII. REFERENCES

[4] Bradski G., Davis J. "Motion Segmentation and Pose Recognition with Motion History Gradients" IEEE WACV'00, 2000
He received a BSc in Computer Systems Engineering in 2002 and a Master of Science in the field of electrical engineering, KTH (Royal Institute of Technology), Stockholm, Sweden, in 2004 with a specialization in Intelligent Systems and Knowledge Recognition and Speech Production. From 2002 to 2006 he worked as a Research engineer at Centre of Speech Technology (CTT) at KTH, where the main field of research is methods for automatic speech verification and emotion recognition. From 2006 to 2007 he attended a Trainee Internship at Asahi Kasei, where the working tasks included Verification and Emotion Recognition. From 2006 to 2007 he enrolled as PhD student at Centre of Speech Technology (CTT) at KTH, were he was doing work in Speaker Verification and Emotion Recognition. Now he has finished his first year of PhD study in Electronic Electrical & Computer Engineering Department, University of Birmingham. His main interests multimodal systems and semantic level applications, knowledge base and ontology creation.

Olga Vybornova received a PhD degree in applied and mathematical linguistics from Moscow State Linguistic University, Russia, in October, 2002. From May 2005 she was a post-doctoral researcher and from January 2008 to the present time a research assistant in Université Catholique de Louvain at the Communications and Remote Sensing Lab (UCL-TELE) in Louvain-la-Neuve, Belgium. Her main research interests are in the fields of multimodal data integration, assistive technology supporting active ageing and social cohesion, semantics-driven multimedia presentation generation, context-based inference, meaning representation, goals, intentions and commitments in communication, dialog modeling; cognitive linguistics; memory and attention, computational semantics.

Hildeberto Mendonça has a bachelor's degree in Computer Science at University of Fortaleza, Brazil. During his undergraduate studies he worked in the Laboratory of Human Computer Interaction with distance learning and digital TV projects. Afterwards, he worked in the Laboratory of Knowledge Management with public security and knowledge bases projects. Currently, he has finished his Diploma of Extended Studies in Human Computer Interaction Meta Representation as a requirement for the PhD in Applied Science in the field of Multimodal User Interaction at Catholic University of Louvain, Belgium. He has a rich professional experience in web application development, collaborative systems, and software engineering and software architecture projects. His technical skills are in the Java platform and its main extensions, C++, XML, relational databases, distributed systems, web services, web and desktop user interfaces, and creation and maintenance of ontology.

Daniel Neiberg received an Master of Science in field of electrical engineering, KTH (Royal Institute of Technology), Stockholm, Sweden, in 2003. From 2002 to 2006 he worked as a Research engineer at Centre of Speech Technology (CTT) at KTH, were he was doing work in Speaker Verification and Emotion Recognition. From 2006 - 2007 he attended a Trainee Internship at Asahi Kasei, where the working tasks included development and evaluation of Automatic Speech Recognition of Swedish. In 2007, he enrolled as PhD student at Centre of Speech Technology (CTT) at KTH, where the main field of research is methods for automatic speech recognition and speech production.

David Antonio Gómez Jáuregui was born in Mexico City on June 13, 1980. He received a BSc in Computer Systems Engineering in 2002 and a Msc. in Computer Science in 2004 with a specialization in Intelligent Systems and Computer Vision. He obtained these grades from Tec de Monterrey Campus Cuernavaca, Mexico. In 2003 he participated in the Robotic Soccer tournament (FIRA 2003) celebrated in Vienna, Austria in representation of Tec de Monterrey. During 2004 - 2007 he worked as a software developer in several companies (Real Time Services, Infomedia, Banco Azteca, Cognosite) learning and using different software programming technologies. In 2007, he obtained a scholarship from the Mexican Government to study a PhD in Computer Vision at Télécom SudParis (Evry, France). He is currently studying his second year of his PhD with the thesis subject 3D Human Gesture Acquisition by Computer Vision and Virtual Rendering? He has big research interest in Computer Vision, Artificial Intelligence and Computer Graphics. Other interests include videogames, learning new languages, eating pizza, listening music and meeting people.

Ao Shen has a bachelor's degree in Communication Science & Engineering at Fudan University, Shanghai, China and a bachelor's degree in Electronic Communication Engineering at Birmingham University, UK. He is experienced in coding mobile application and won the second prize in Vodafone Betavine Student Competition. Now he has finished his first year of PhD study in Electronic Electrical & Computer Engineering Department, University of Birmingham.