MuDiS – A Multimodal Dialogue System for Human-Robot Interaction

Manuel Giuliani, Michael Kaßecker
Robotics and Embedded Systems Group
Dep. of Informatics, Technische Universität München
Boltzmannstraße 3, D-85748 Garching bei München, Germany
Email: {giuliani, kassecke}@in.tum.de

Stefan Schwärzler, Alexander Bannat,
Jürgen Gast, Frank Wallhoff
Lehrstuhl für Mensch-Maschine-Kommunikation
Dep. of Elec. Engineering, Technische Universität München
Arcisstraße 21, 80333 München, Germany
Email: {s,bannat, gast, wallhoff}@tum.de

Christoph Mayer, Matthias Wimmer
Image Understanding & Knowledge-Based Systems
Dep. of Informatics, Technische Universität München
Boltzmannstraße 3, D-85748 Garching bei München, Germany
Email: {mayerc, matthias.wimmer}@in.tum.de

Cornelia Wendt, Sabrina Schmidt
Institut für Arbeitswissenschaft
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, D-85577 Neubiberg, Germany
Email: {cornelia.wendt, sa.schmidt}@unibw.de

***Abstract*—We present the MuDiS project. The main goal of MuDiS is to develop a Multimodal Dialogue System that can be adapted quickly to a wide range of various scenarios. In this interdisciplinary project, we unite researchers from diverse areas, including computational linguistics, computer science, electrical engineering, and psychology. The different research lines of MuDiS reflect the interdisciplinary character of the project. In this publication, we show how MuDiS collects data from human-human experiments to get new insights in multimodal human interaction. Furthermore, we present the first version of the MuDiS system architecture, which contains new components for classification of head movements, multimodal fusion, and dialogue management. Finally, we describe the application of the MuDiS system in a human-robot interaction scenario to prove that MuDiS can be implemented in different domains.**

I. Introduction

Human communication is multimodal. Humans use their whole body, their hands, their head, gazes or their face to express complex information, including their current emotional status, their intentions about their next actions, or simply their agreement or disagreement. But humans are not only able to express themselves in a multimodal way, in addition they are experts in interpreting the multimodal utterances of other humans, to understand their emotions and intentions.

In this publication, we present the project MuDiS, which develops a new kind of *Multimodal Dialogue System* that enables an artificial agent—e.g. a computer or a robot—to interact with a human in a natural and multimodal way. The main feature of this dialogue system is its generality. It can be adapted to a wide range of applications and domains in a short time. To accomplish this ambitious goal, the project unites researchers from such diverse research areas as computational linguistics, computer science, electrical engineering, and psychology, which explore the aspects of a multimodal dialogue system from different directions. The new feature of MuDiS besides its generality, in comparison to other multimodal dialogue systems, will be that it will interpret input of new input channels that have never been integrated in a multimodal dialogue system before. Namely, we are planning to integrate components for emotion recognition, multi person tracking, and focus of attention detection.

The remainder of this publication is organised as follows: in Section II we review shortly some important multimodal systems that influenced the design of the MuDiS system architecture. The psychologist partners of MuDiS are executing human-human experiments, to find out more about how humans use multimodality. The setup for these experiments is described in Section III. The results of the experiments will directly feed into new models for multimodal fusion and dialogue management. After that, Section IV gives a general overview of the MuDiS system architecture and highlights some of the specific system parts, including components for head movement recognition, multimodal fusion and a dialogue manager. Finally, Section V shows how the MuDiS approach can be applied to a human-robot interaction scenario, before Section VI concludes this publication.

II. Related Work

Many approaches for multimodal dialogue systems have been described in literature, since Bolt introduced the "put-that-there" system [**?**]. These systems can be roughly separated into those systems that fuse the input from several input channels already on feature level ([1], [**?**]), and those systems that integrate the input after an initial unimodal interpretation in a later stage ([2], [**?**], [**?**]).

Nearly all of these systems have in common that they claim to be multimodal, but most of them combine only two

From: 1st Intern. Cotesys Workshop 2008, Technische Universität München

modalities, for example speech and gestures or speech and pen input. A second disadvantage of many multimodal systems is that they are developed for a certain domain and can only be ported to applications with changing substantial parts of the system. There are some exceptions, including Johnston et al. [?], who present a framework for rapid prototyping of information systems with speech and pen input, or the talk project [?], who describe a grammatical framework for development of multimodal grammars. The work from Landragin et al. [?] is also very interesting, as they are showing how they port MMIL, the MultiModal Interface Language that is used for multimodal meaning representation, to a new domain.

The goal of MuDiS is to overcome these obstacles by providing a software architecture that is extensible and applicable to a wide range of applications. Additionally, we plan to integrate information from new input channels, including emotion recognition and person tracking, which, to our knowledge, has not been done before.

The next section describes the human-human interaction experiments we are conducting to get more insights about human behaviour during a collaboration task. The results of these studies will be directly implemented in the MuDiS system architecture, by applying the findings in new models for multimodal fusion and dialogue management.

III. Human-Human Experiments

A. *Experimental Setup*

In order to improve human-robot interaction, a near to naturalistic interaction between two humans was observed. One of the humans represented the robot, the other one the human co-worker in an industrial setting. The "robot" was an instructed experimenter whereas the human co-worker was a test person. "Robot" and participant sat face to face at a table and were asked to solve a joint construction task with the LEGO Mindstorms system (see Figure 1).

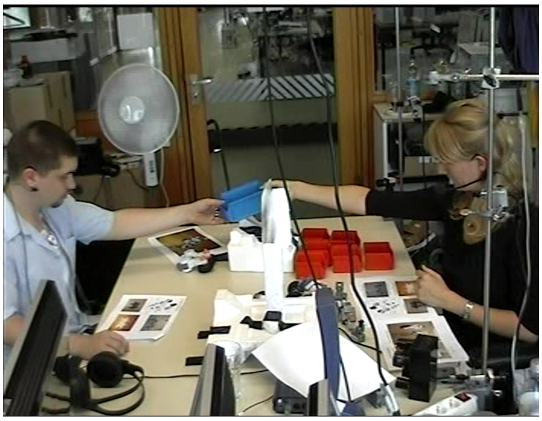


Fig. 1. Experimental setup for the MuDiS human-human experiments: the experimenter on the right enacts a "robot" that is able to hand over LEGO pieces to the experiment participant, who assembles the pieces according to a given assembly plan.

The "robot" had been instructed to act like an ideal robot in terms of being supportive, not getting impatient, handing over LEGO parts just in time, explaining difficult construction steps, or trying to cheer up the co-worker. "It" was able to point, to speak, to hear and to react like a human, apart from two constraints: like a real robot, it could not put together LEGO parts, and it's sight was artificially impaired by placing a semi-transparent foil between the two actors. The foil was about 30 cm high so that LEGO parts could be exchanged, and the faces of the interaction partners were still visible for both of them. Thus, mimic information could still be used, for example to identify the co-worker's emotional state.

B. *Procedure*

To make sure that every participant had a comparable level of experience with the LEGO Mindstorms parts, there was a practice phase before the experimental phase. In this practice task the participants had to construct the word "TEST" from available LEGO bricks under supervision of the "robot". Thus, the experiment participants could make themselves acquainted with the instructions, the different LEGO parts, and the overall setting.

Before the test phase, participants were asked to complete a questionnaire surveying their former experience with LEGO, their affinity to technology in general, and their current emotional state. After the experiment, a second questionnaire had to be completed concerning the affective state again, the experienced quality of the interaction, and ratings of the task and the support by the robot.

C. *Data Collection*

During the experiment, recorded data included speech of participant and robot, physiological data (heart rate, skin conductance, pulse) and camera views from four different angles. For the analysis of facial expressions, one camera was pointed directly on the face of the participant. Another camera perspective showed the table from the top for gesture recognition purposes and for coding the task progress. Additionally, there was also a camera aiming at the "robot" to analyse the (appropriate) reactions of it. The fourth camera recorded the whole interaction scene from a more distant point of view. This allows for an identification of important dynamic events occurring between the interacting partners.

D. *Data Analysis*

With regard to the video data, a qualitative data analysis is planned, comparable to those suggested by Kahn et al. [3], Zara et al. [4], or Dautenhahn and Werry [5]. As evaluation categories are highly context-dependent, we will develop a system of our own with diverse dimensions. Coding the different modalities independently will be the most basic step. This comprises not only facial expressions and speech, but also gestures, certain movements, as well as changes in physiological parameters. Beyond those micro events, we are also interested in their psychologically relevant interplay or certain timely orders (e.g. "event *a* always shortly happens before event *b*"). With regard to speech, the complete dialogue structure of the interaction will be analysed. On this meta level, the dynamics of initiative are another interesting aspect.

Such a categorisation will help on the one hand to find principles and structures in human-human interaction that have to be implemented for an improvement of human-robot interaction. On the other hand, it helps to give tags to the different behaviours, actions, and events that have to be recognised in the interaction process. With them it shall be possible to train automatic recognition algorithms with suitable parameters.

We plan to implement the findings from the human-human experiments in some components of the MuDiS system architecture, which will be described in the following section.

IV. MuDiS System Architecture

In this section, we provide an overview for the MuDiS system architecture and describe parts of the system in more detail in the following subsections. Figure 2 shows a schematic overview of the MuDiS system architecture. The figure shows that the MuDiS approach for a multimodal system is based on the concepts of modularity and generality. The interfaces between the single system parts are strictly defined. This way, parts of the system can be exchanged without the need for adjusting the rest of the system.

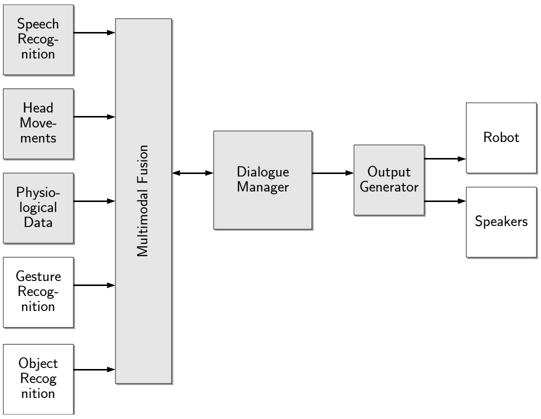


Fig. 2. The MuDiS system architecture. The figure shows the minimal system configuration. The system parts highlighted in grey are developed by MuDiS researchers. For the remaining parts MuDiS defines general interfaces so that the components can be exchanged according to a given scenario.

Currently, the MuDiS team is working on the system parts that are highlighted in grey in Figure 2. The *multimodal fusion* component gets input by *speech recognition*, a component that classifies human *head movements* and *physiological data* from a component that measures the skin conductance response, heart rate, and blood volume pressure. Fusion takes the speech and head movement data (physiological data can not be interpreted yet), produces a semantic translation of the data and sends it to the *dialogue manager*. The dialogue manager keeps track of the current state of the conversation, such that together with the input by the multimodal fusion, it can infer actions that need to be executed next. These actions are then sent to an *output generation* component, which translates the actions into commands for a connected output device—Figure 2 shows for example a robot and speakers to play speech messages—to give multimodal feedback to a human, who is working with the system.

A. Head Movement Recognition

The *head movement recognition* component we are describing here is the first step to a more sophisticated emotion recognition module, which we plan to implement in a later stage of the MuDiS project. Recent progress in the field of computer vision allows a robust and accurate classification of human faces with a high runtime. This renders our component for face interpretation inevitable to realize the paradigm of intuitive human-machine interaction. It is able to robustly recognise facial features independent of the visible person's ethnic group or culture. It allows to determine facial expressions, gaze direction, gender, and age and much more. For the MuDiS project we currently classify two head movements that are important for multimodal communication: head nodding and shaking. These movements indicate or emphasize an affirmative or a rejecting attitude.

Our approach uses model-based image interpretation, which allows to accurately infer these highly semantic interpretation results by exploiting a priori knowledge of human heads and faces, such as shape and skin colour information. These techniques reduce the large amount of image data to a small set of model parameters that describe the current pose of the head, which facilitates and accelerates subsequent interpretation. Fitting the face model is the computational challenge of finding the model parameters that best describes the face within a given image. This section describes the main components of model-based techniques, compare to [6].

The face model contains a parameter vector $\mathbf{p}$ that represents its configurations. We integrate a deformable 3D wire frame model of a human face (Candide-3), which is introduced and explained in detail in [7]. The model consists of 116 anatomical landmarks and its parameter vector $\mathbf{p} = (r_x, r_y, r_z, s, t_x, t_y, \sigma, \alpha)^T$ describes the affine transformation $(r_x, r_y, r_z, s, t_x, t_y)$ and the deformation (σ, α). The deformation parameters indicate the shape and animation units such as state of the mouth, roundness of the eyes, raising of the eye brows, etc., see Figure 3.

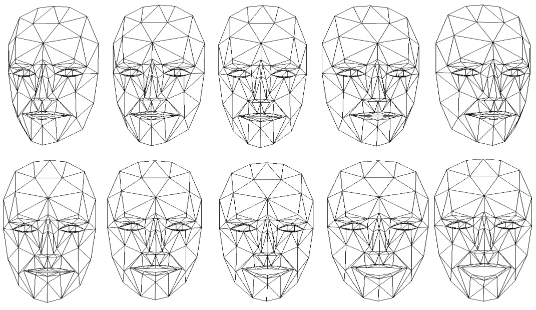


Fig. 3. The Candide-3 face model is able to reflect various face shapes and facial expressions.

The Candide-3 face model is inspired by the biological constitution of a head, it represents the three-dimensional structure of the head and the muscular deformation of the facial components. This makes it highly suited to head and face interpretation scenarios and outperforms earlier approaches, such as two-dimensional active shape models [**?**].

The localization algorithm computes a rough estimate of the most important model parameters by investigating the global position and the size of the visible face. These initial model parameters are further refined by the subsequent fitting algorithm. Our system integrates the approach of [8], which detects the model's affine transformation in case the image shows a frontal view face.

The objective function $f(I, \mathbf{p})$ computes a comparable value that specifies how accurately a parametrised model $\mathbf{p}$ matches an image I. It is also known as the likelihood, similarity, energy, cost, goodness, or quality function. Traditional approaches attempt to figure out good computation rules of this function in a manual procedure which is laborious and erroneous. In contrast, our earlier publication [6] proposes a method that automatically learns the best computation rules of this function based on objective information theoretic measures. This approach considers a large set of training images and the quality of different model parameterizations is known for each image. Using this information, it is able to learn a good objective function and it is also able to objectively evaluate its performance via previously unseen test images with annotations. The advantages of this approach are that it does not rely on expert knowledge in the field of computer vision and it yields more robust and accurate objective functions, which facilitate the task of the associated fitting algorithms. Furthermore, it is not restricted to face model fitting but applicable independently of the domain.

The fitting algorithm searches for the model parameters that best describe the visible face. These parameters correspond to the global minimum of the objective function. Fitting algorithms have been subject of intensive research and evaluation. We refer to [9] for a recent overview and categorization. Since we adapt the objective function rather than the fitting algorithm to the specifics of the face interpretation scenario, we are able to use any standard fitting technique. Because facial expression recognition requires real-time capabilities, we chose a quick hill climbing algorithm. Note, that the reasonable specification of the objective function makes this local optimisation method nearly as accurate as global optimization strategies, such as genetic algorithms.

Inferring the head action, such as nodding or shaking, relies on the correctly fitted face model. We record several short image sequences of persons displaying a head gesture (nodding, shaking). Additionally, image sequences were recorded that show the person's head in a neutral position. The model is tracked through these short image sequences and the model parameters are exploited to calculate the motional and transitional speed from the affine transformation parameters. From this data a classifier is trained to infer head gestures. We rely on Hidden Markov Models [**?**] because they specifically consider temporal dependencies within the presented training data.

Estimating Facial Expressions is tackled by computing a large number of features that describe the structure and the muscular activity of the face. Our approach considers both, structural and temporal features of the face. For each image within the image sequence of the live video stream, the model's deformation parameters represent structural information and the motion of model's landmark points yields temporal features. This large amount of feature data is assembled into a feature vector that describes one individual image, whereas a stream of feature vectors describes the visible activity within a video stream. From this vast amount of data, a classifier is generated with machine learning techniques that infers the facial expression. Our approach implements decision trees [**?**] as a quick and robust classifier.

B. Multimodal Fusion

In this section, we describe the *multimodal fusion* component, which is used to integrate the input by speech and head movement recognition to a combined meaning representation. MuDiS is a system that applies a late fusion approach. Accordingly, on an abstract level, multimodal content is considered as sequences of discrete data objects. The general approach for the multimodal fusion looks as follows: every input modality connects to the fusion component over a dedicated one-way channel to send information about characterised events. A typical example is a speech recogniser sending text strings of the recognised speech. The multimodal fusion module generates abstract semantic representations for the single input streams, for example by using a parser that processes the input string by the speech recogniser and analyses the structure of the recognised sentence. The semantic representations for each input channel are then combined to produce an integrated representation of multimodal events, which are afterwards passed on to the dialogue manager.

One of the main goals in MuDiS is to keep the system architecture general and easy to adapt to various domains. For this reason, we developed a new algorithm for speech processing, in which sentences are broken up into tokens that can be translated to unique semantic primitives (actions or events). These primitives are stored in a database so that they can be easily exchanged when MuDiS is applied to a new domain. This new approach solves some of the typical problems that are related to speech, for example ambiguity or the use of different word structures to express similar content. To illustrate this, consider the following example: a human working with a robot might order the robot to bring a certain object, for which she uses one of the sentences in (1) (other examples are also thinkable).

(1) a. Bring me object X.
 b. Get me object X.

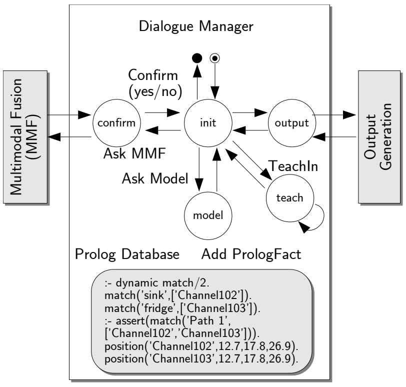
The MuDiS multimodal fusion breaks both of these sentences into two tokens "*bring me*" and "*object X*" or "*get me*" and "*object X*", respectively. Then it looks up the semantic primitives that correspond to the tokens in its internal database. The database maps both tokens "*bring me*" and "*get me*" to the same semantic primitive "*GET*". This way, different verbal utterances can be linked to a set of defined semantic primitives just as needed for a given scenario.

C. Dialogue Manager

After processing of the input coming from the different input modalities, multimodal fusion sends its interpretation to the *dialogue manager*, which is the central component of the MuDiS architecture that keeps track of the current state in the dialogue and initiates the next actions and utterances.

Currently, the MuDiS dialogue manager is based on a state machine and a knowledge base implemented in Prolog. Prolog is a declarative logic programming language and we use it for checks, building temporarily queries, and as a knowledge database in the dialogue steps. Therefore, relations are defined as clauses. For example, if an instantiation of a *match* is found, Prolog makes the union of clauses and translates them into a query, e.g. *exists('Channel105')*. If the clause returns false, Prolog tries to backtrack alternative query rules. Complex query rules are implemented by means of recursive predicates.

When the dialogue manager gets input by the multimodal fusion component, it queries the knowledge base to infer the next dialogue steps. Afterwards, it generates the commands and speech output to the human. The underlying finite state machine represents general dialogue nodes, which are independent from the application and can be reused. The state machine is based on the work by Mealy [12]. Finally, the dialogue manager sends the commands to the robot and the speech output to a simple text-to-speech component that replaces a more sophisticated output generation module, which will be implemented in a later stage of the project. The robot additionally gives feedback to the dialogue manager, which is also used to infer the next actions. Figure 4 shows a schematic overview of the dialogue manager, it represents the general dialogue modes, which are application-independent and thus reusable. The dialogue manager runs into dialogue modes by *events* generated from the multimodal fusion. The *events* and the outputs are analogue to [12] in the transitions of the finite state machine.



```
:- dynamic match/2.
match('sink',['Channel102']).
match('fridge',['Channel103']).
:- assert(match('Path 1',
['Channel102','Channel103'])).
position('Channel102',12.7,17.8,26.9).
position('Channel103',12.7,17.8,26.9).
```

Fig. 4. The figure shows the MuDiS dialogue manager, which is a combination of a finite state machine and a Prolog database. The dialogue manager is triggered by events from the multimodal fusion component, and the Prolog knowledge database.

V. MuDiS in a Human-Robot Scenario

To prove the general applicability of the MuDiS system, we implemented MuDiS in a human-robot interaction scenario. In this scenario a mobile robot, which is equipped with two 7-DOF arms, moves around in a kitchen environment, and serves coffee and other refreshments to a human that communicates with the robot over two modalities. On the one hand the human can speak to the robot via a head-mounted microphone, on the other hand the robot is able to detect the humans head movements and to recognise if the human nods or shakes the head to signal agreement or disagreement. Figure 5 shows a schematic overview of the kitchen environment. The whole environment is monitored by a infrared tracking system, which tracks infrared emitters that are mounted on certain objects and locations in the kitchen. The tracked objects include a *cup*, a *glass*, and a *plate*, while the marked locations of the kitchen are the *fridge*, the *sink*, and the *oven*. These objects and locations, as well as the commands a human can give to the robot, are stored as tokens in a database, as we described in Section IV-B.

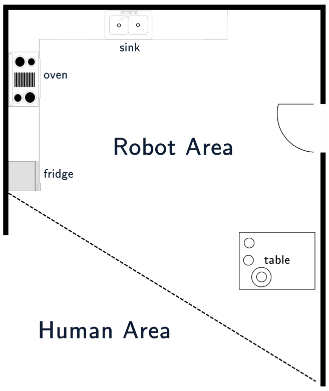


Fig. 5. MuDiS in a human-robot interaction scenario: the figure shows a schematic overview in which robot and human have separated work areas; the robot can bring objects from a table and can be ordered to move to certain locations (table, fridge, oven, sink) in the kitchen.

The two input modalities for this scenario—speech and head movements—were already presented in Section IV-A. Table I shows an example dialogue between the human and the robot. The MuDiS system enables the robot to understand basic sentences uttered by the human. In this simple scenario, the human mainly gives the robot commands it has to execute. However, the robot is already able to learn new labels and associate them to objects and locations in the kitchen area, which is also presented in the dialogue in Table I. In addition, the robot has to reason over the orders it gets by the human. If it cannot execute a given order it has to react appropriately by generating an adequate spoken remark, for example when it has to move to a certain location and does not know the position yet, the robot has to ask for the according infrared marker that is associated with the location.

Speaker	Dialogue and Actions
Human	Hello.
Robot	Hello, how are you today? What can I do for you?
Human	Please, bring me a glass of water.
Robot	Already on my way.
	Robot goes to the table, grasps the glass of water, returns to the human, and hands over the glass.
Robot	Here you are. Is this what you wanted?
	Human nods. Robot recognises the nodding and displays it on its display.
Human	Thank you. Now, bring me the cookie plate.
Robot	No problem.
	Robot goes to the table, grasps plate with cookies, returns to the human, and hands over the plate.
Robot	Here you are.
	[. . .]
Human	Go to the sink.
Robot	I don't know where the sink is.
Human	It's at marker X1.
	Robot moves to marker X1/the sink.
Robot	Okay.
	[. . .]

TABLE I

Example dialogue between human and robot that comprises input from speech and head movements.

VI. Conclusion and Future Work

We presented MuDiS, an interdisciplinary project that unites researchers from computational linguistics, computer science, electrical engineering, and psychology, which develop a multimodal dialogue system. The main feature of this new dialogue system is that it has been designed to be applicable to various domains already from early developmental phases.

In this publication, we showed the experimental setup for the human-human interaction experiments we are currently conducting to get more insights in human behaviour during a collaboration task. The results from these experiments will be applied to define aspects of the dialogue systems in a later stage of the project, e.g. to determine timing relations of modalities in the multimodal fusion component. Furthermore, we introduced the general structure of the MuDiS system architecture and highlighted three system components: head movement classifier, multimodal fusion, and dialogue manager. Finally, we showed how MuDiS was integrated in a human-robot interaction scenario.

In the future, we plan to extend the MuDiS system architecture and apply it to other domains to show the generality of our approach. Especially, we plan to implement a sophisticated emotion recognition module and versions of multimodal fusion and dialogue manager that are based on findings of the human-human experiments and can handle more complex dialogue situations.

Acknowledgement

This work was supported by the CoTeSys cluster of excellence

References

[1] R. A. Bolt, "”put-that-there": Voice and gesture at the graphics interface," in *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1980, pp. 262–270.

[2] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "Quickset: multimodal interaction for distributed applications," in *MULTIMEDIA '97: Proceedings of the fifth ACM international conference on Multimedia*. New York, NY, USA: ACM Press, 1997, pp. 31–40.

[3] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "Match: An architecture for multimodal dialogue systems," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 376–383, 2002.

[4] W. Wahlster, *SmartKomFoundations of Multimodal Dialogue Systems*. Heidelberg, Germany: Springer, 2006.

[5] L. Boves, A. Neumann, L. Vuurpijl, L. Bosch, S. Rossignol, R. Engel, and N. Pfleger, "Multimodal interaction in architectural design applications," *Lecture Notes In Computer Science*, pp. 384–390, 2004.

[6] M. Rickert, M. E. Foster, M. Giuliani, T. By, G. Panin, and A. Knoll, "Integrating language, vision and action for human robot dialog systems," in *Proceedings of the 4th International Conference on Universal Access in Human-Computer Interaction, HCI International, Part II*, ser. Lecture Notes in Computer Science, C. Stephanidis, Ed., vol. 4555. Beijing: Springer, Jul. 2007, pp. 987–995. [Online]. Available: http://www.springerlink.com/content/d038121x21629077/fulltext.pdf

[7] P. Ljunglöf, G. Amores, R. Cooper, D. Hjelm, O. Lemon, P. Manchón, G. Pérez, and A. Ranta, "Talk: Multimodal grammar library," *Deliverable D1. 2b, TALK Project*, 2006.

[8] F. Landragin, A. Denis, A. Ricci, and L. Romary, "Multimodal meaning representation for generic dialogue systems architectures," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 2004, pp. 521–524.

[9] P. H. Kahn, B. Friedman, N. Freier, and R. Severson, "Coding manual for children's interactions with aibo, the robotic dog – the preschool study (uw cse technical report 03-04-03)," Seattle: University of Washington, Department of Computer Science and Engineering., Tech. Rep., 2003.

[10] A. Zara, V. Maffiolo, J. Martin, and L. Devillers, "Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics," *Lecture Notes in Computer Science*, vol. 4738, p. 464, 2007.

[11] K. Dautenhahn and I. Werry, "A quantitative technique for analysing robot-human interactions," in *International Conference on Intelligent Robots and Systems*, 2002.

[12] M. Wimmer, S. Pietzsch, F. Stulp, and B. Radig, "Learning robust objective functions with application to face model fitting," in *Proceedings of the 29th DAGM Symposium*, vol. 1, Heidelberg, Germany, September 2007, pp. 486–496.

[13] J. Ahlberg, "Candide-3 – an updated parameterized face," Linköping University, Sweden, Tech. Rep. LiTH-ISY-R-2326, 2001.

[14] T. F. Cootes and C. J. Taylor, "Active shape models – smart snakes," in *Proceedings of the 3^{rd} British Machine Vision Conference*. Springer Verlag, 1992, pp. 266 – 275.

[15] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[16] R. Hanek, "Fitting parametric curve models to images using ocal self-adapting seperation criteria," Ph.D. dissertation, Department of Informatics, Technische Universität München, 2004. [Online]. Available: http://tumb1.biblio.tu-muenchen.de/publ/diss/in/2004/hanek.html

[17] L. Rabiner, "A tutorial on hidden markov models and selected applications inspeech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[18] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann, 1993.

[19] G. Mealy, "A method to synthesizing sequential circuits," in *Bell System Technical Journal*, 1955, pp. 1045–1079.