Resolving Object References in Multimodal Dialogues for Immersive Virtual Environments

Thies Pfeiffer
SFB 360
University of Bielefeld
33594 Bielefeld, Germany
Thies.Pfeiffer@Uni-Bielefeld.de

Marc Erich Latoschik
AI & VR Lab
University of Bielefeld
33594 Bielefeld, Germany
marcl@techfak.uni-bielefeld.de

Abstract

This paper describes the underlying concepts and the technical implementation of a system for resolving multimodal references in Virtual Reality (VR). In this system the temporal and semantic relations intrinsic to referential utterances are expressed as a constraint satisfaction problem, where the propositional value of each referential unit during a multimodal dialogue updates incrementally the active set of constraints. As the system is based on findings of human cognition research it also regards, e.g., constraints implicitly assumed by human communicators. The implementation takes VR related real-time and immersive conditions into account and adapts its architecture to well known scene-graph based design patterns by introducing a so-called reference resolution engine. Regarding the conceptual work as well as regarding the implementation, special care has been taken to allow further refinements and modifications to the underlying resolving processes on a high level basis.

1. Introduction

Resolving references plays a central role during the analysis of natural human utterances. In communication, references to objects are based on both perceptual properties such as shape, color or orientation, and on inferred conceptual properties such as function or linguistic identifier. A referential expression is a set of feature descriptions discriminating certain objects. During our research on multimodal interaction for interactive graphics applications, we have been continuously faced with the conceptual and technical problems of finding the correct mapping between referential expressions and entities in our graphics applications, the so-called referent-objects. This has led to several partly application tailored approaches where the applied concepts and techniques were deeply bound to the systems' internal architectures and algorithms. Their reuse, adaption and refinement was and is an error-prone work since some achievements were lost due to the fact that implementations often included heuristics and tweaks not documented and which could not easily be reversely engineered. On the other hand, human cognition oriented projects focussing on concepts of reference processing led to persistent results explicitly described or formally developed. But they often lacked a technical implementation at all. To provide a means of adjustment and refinement of the reference resolving process, a framework which makes its internal conceptual basis and heuristics explicit and accessible from the outside is required.

Since nowadays all of our current and upcoming work is centered around VR applications, resolving references under immersive conditions demanded for an approach even closer related to human perception and cognition capabilities. Figure 1 in our color plate shows a typical multimodal interaction with one of our systems where the relative position between the head tracked user and the scene's possibly moving objects might constantly change. This makes it necessary to process referential utterances in a way that their propositional values are valid with respect to the specific utterance's times and the given relative scene representations or view of the user at those times. This close temporal binding of the utterance's semantic is deeply complicating a technical implementation which will always be running –at least mildly– behind. Hence, VR based applications that handle human referential utterances for interaction purposes are faced with external conditions very similar to those found in natural human-to-human communication.

Scene-graph based approaches encapsulate functionality for input handling and simple animation in dedicated modules like sensors, manipulators or even engines for more complex tasks. Our architecture for the processing of multimodal utterances follows this scheme and offers a variety of

modules for gesture recognition or multimodal parsing and integration. One of these modules is the *Reference Resolution Engine* (RRE) described in this paper.

2. Related Work

Speech and multimodal driven interactions with graphics applications can be dated back to Bolts "Put that there" system [4] and later work which explores the usefulness of deictic expressions for specifying object references in computer graphics systems. Many approaches did utilize these methods for 2D non-immersive large screen display systems [1][19][20][13]. The ICONIC system [12] favored iconic gestures describing shape or orientation of objects in conjunction with speech to identify and manipulate objects. Cavazza et al. [5] make the related problem of finding the right mapping between deictic percepts and referent-objects explicit by introducing the term *extended pointing* when combining pointing gestures with speech. Recent work either references objects only unimodally using well known ray-casting approaches [26] or, like Arangarasan et al. [2], describes the referencing problem implicitly as an abstract task (object selection) in multimodal VR interactions but does not discuss any conceptual solution. Such a conceptual model of multimodal referring based on *reference domains* is found in [14], but it lacks any details about concrete algorithms or methods for a technical realization and only uses examples based on a static 2D domain.
In the desktop based design application VIENA [18], objects and places could be referenced by typing and simple word by word speech input accompanied by mouse clicks. The system already distinguished three different frames of reference during the analysis process: 1) egocentric: based on the user's view (static during interaction), 2) intrinsic: based on certain object features, e.g., a dedicated *front* area of a desk and 3) extrinsic: based on a communication partner's view, in VIENA an anthropomorphic agent called Hamilton. References could be made using noun phrases accompanied by mouse clicks to specify object types, colors and relative positions. The CODY system [9] additionally incorporated application specific functional context information as a constraint when finding the referent-objects. Its reference analysis process searched for objects for which all of the requested attributes were true and which could perform the requested operation, e.g., the instantiation of a connection between two CAD parts. Both systems already included references to objects from former interactions during their reference analysis ("*...and now take* **it**..."). The SGIM system [15] was faced with the implications of an immersive environment which introduced the problem of temporal dependencies between the user's referential utterances and the perceived scene. SGIM's reference analysis process was conceptually handled as a set theory problem where each additional constraint would be used to generate a new intersection with an existing set of possible objects based initially on all the objects the user had in her view during a given time span. This work is now extended to communicate with MAX, an anthropomorphic agent who interacts in a Virtual Environment (VE) with the user and who introduces his own dynamic frame of reference. The results from these systems have nowadays been utilized for a variety of current projects located in our VE, e.g., the Virtual Workplace [3][16].

A cognitive adequate reference resolution system has to account for results from linguistics and psychology. However, empirical findings from Natural Language Processing (NLP) are mainly based on experimental settings of a static nature, such as understanding written text, at least with respect to the time frame of a single utterance. This we will call the *static world assumption*. Interactive immersive VR systems lead to highly dynamic scenes where a static world assumption no longer holds. The spatial reference systems which anchor referential utterances and the virtual world itself may change even during a simple utterance. This situation is covered by our *dynamic world assumption* first incorporated in the SGIM project. Our research is associated and partly embedded in an interdisciplinary Collaborative Research Center (SFB360) at the University of Bielefeld where technical work has been accompanied by simulative exploration and evaluation of communication taking place in our application scenarios. Relevant aspects of this work are consistent to other results, e.g. in [21], and can be summarized as follows:

Naming: Naming objects is in general an important mechanism for their discrimination. Regarding generic objects, naming objects by proper names or type names has only a reduced selectivity. Also, aggregates of objects might not have a proper name or type at all. This has been addressed to some extent by the development of COAR [10] which, e.g., allows referring to well known aggregates with their functional description (e.g. "propeller") or to generic objects with their functional role within an aggregate (e.g. a bar might be a "rotor-blade" if it is part of a "propeller").

Features: Besides names, experimental studies showed that humans tend to choose object features which are easily perceptually available. In Computer Graphics and VR based application scenarios the most important features are color, shape, position, orientation and size. In experiments German test subjects showed a preference of the order size - color - shape or sometimes color - size - shape when combining several features [7]. Features are not only bound to the modality of speech, as orientation, position or shape are quite naturally expressed using coverbal gestures.

Spatial References: To express spatial references, humans generally specify the position of the intended referent-object(s) in relation to other referent-object(s). To allow for a linguistic encoding, space has to be categorized based on a certain perspective, such as to the left or to the right, or a referent-object such as in "behind the chair". It can be assumed that the spatial categories have blurred boundaries and are by no means distinct but have overlapping regions [8]. Vorwerg et al. have developed a computational model for adequately categorizing spatial areas into these categories [25]. A combination of several spatial references might include a shift in referent-objects ("the block behind the wheel to the left of the plane") or even the perspective. Perspectives are not always made explicit which might lead to ambiguities as in ("in front of the plane") where at least the egocentric and intrinsic perspectives are competing.

Common Ground: In natural communication participants align their utterances to a common ground which the communicating partners share [6], at least over a subset of the referential expression. This is a dynamic process for the common ground might also change, e.g., as the result of negotiations after violations (here misunderstandings) are encountered. The common ground between two individuals is influenced, e.g., by their physical being, their social and cultural background, their individual long-term experiences and the situative context they are sharing.

2.1. Discussion and Requirements

Generally, results from related work often emphasize what a system was or is capable of but only few insights are given into the applied specific concepts and methods to analyze references in the multimodal input streams. If internals are discussed, often the focus is on the principle integration method applied, e.g., if it is frame based or using a TAG (Tree-Adjoining-Grammar) formalism. Even own work often did not explicitly describe the reference resolution process but consisted technically of an unmanageable amount of deeply nested `if-then-else` style conditions sometimes dependent on internal side-effects. But combining the experiences from existing technical solutions with the conceptual work of experimental studies leads to significant aspects which a system that processes multimodal references for VR applications has to take into account:

Reference Types: To reference objects, the user chooses criteria with a high selectivity for the intended objects to identify or distinguish them from the surroundings. Names, features and spatial relations are such important criteria.

Utterance Complexity: Referential utterances can become fairly complex. In language terms, they can consist of several nested relative clauses which should be processed at least to a certain level. Gestures can occur simultaneously at each time during the multimodal utterance regarding a referential speech unit.

Competing Frames of Reference: Projective spatial relations are anchored in a frame of reference (egocentric, intrinsic and extrinsic). The speaker can make the frame of reference explicit by using an extrinsic frame of reference, even if she refers to herself ("*Looking from my side, the propeller is in front of the plane.*"). But even if the anchor of a frame of reference is given, there is still room for different interpretations of the intrinsic orientation of an object [11].

Dynamic Frames of References: Temporal dependencies between a referential utterance's units and the represented scene plays a crucial role and has to be handled in an immersive setting. The propositional value of a reference is only valid during a limited time span.

Common Ground: A technical system designed to understand natural communication has to share a common ground with its communicating partners - at least at a phenomenological level. While the processes from perception to interpretation might be substantially different from that of humans, they have to produce representations of a similar quality to provide a solid ground for reasoning steps leading to a successful participation in natural communication.

Uncertainty: Most approaches seem to handle multiple object references using boolean logic. But gestures are inherently vague and speech is highly ambiguous and mostly uses qualifying expressions like *"..left of.."*. We seldomly see the world as a collection of objects with selfcontained properties but we compare these properties with other objects and their respective properties. Qualifying expressions are inherently relative and hence can only be handled partially by using best guess sorted lists for constraining properties. This problem is of an intrinsic nature and should be represented in principle when analyzing references, e.g., using certainty values.

Technical Concerns: For interactive VR systems, responses 1) must be generated as fast as possible and 2) must provide delayed access. As a central module in the analysis of multimodal utterances, the RRE must consider those real-time constraints. The engine should be designed to be compatible with existing VR frameworks.

Administration of Heuristics: The engine's underlying methods and concepts should be made explicit and provide a high-level interface for administration and modification. This would greatly enhance the process of integration of individual experimental results, which might induce further refinements of the system. Support for easy modification of the heuristics is crucial for application specific experimental studies accompanied with evaluations to ensure a robust and effective system interaction [18].

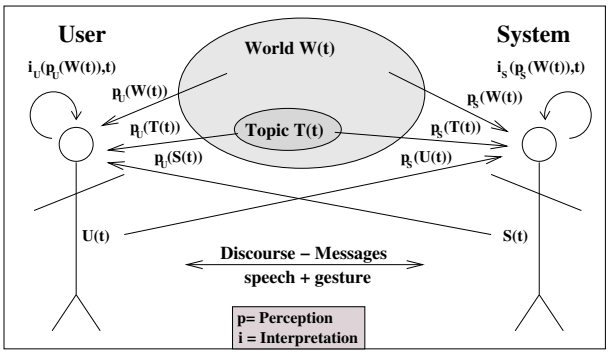

Figure 1. A model for multimodal discourse in a typical VR setting. The user and the system are fully immersed and perceive $p_{U|S}$ each other $U(t)|S(t)$ and the world $W(t)$ in real-time. They are permanently interpreting $i_{U|S}$ their percepts according to long- and short-term knowledge.

3. Concepts

We start with the presentation of a basic model for reference communication in human-computer interactions found in typical VR settings. Based on this model we will show crucial aspects of a system for understanding references in human utterances and we will use the model to position the reference resolution subprocess.

3.1. Model for Reference Use

Popescu-Belis presented a generic model for reference use in human communication [22] which we believe to be a suitable abstraction of the basic situation of reference use in multimodal discourse in VEs. For his model is tailored to application scenarios of a static nature, such as simple desktop applications or text understanding systems, it has to be refined to meet the extended requirements of today's immersive VR systems.

The model considers two communicating partners which are exchanging multimodal discourse messages over a topic T in the world W. While this is still true in VR discourse, we concentrate on a user U communicating with a technical system S (see Figure 1), which can work transparently in the background or, e.g., can be represented by an embodied conversational agent (ECA).

Popescu-Belis does not explicitly regard the aspect of time in his model, although he mentions that this could be done in principle. To meet our dynamic world assumption, our refined model therefore extends the original model by parameterizing the representations with the time of their validity or perception. Further, the immersive character of VR systems rises a need for the integration of the communicating partners in the world model, for otherwise egocentric or extrinsic references could not be handled. This is reflected by the percepts $p_U(S(t))$ and $p_S(U(t))$.

In discourse the turn might shift between the communicating partners. For presentational reasons we will concentrate on the situation of a single instruction which is uttered by the user and perceived and interpreted by the system. By an instruction we mean a semantical coherent expression which is assumed to fully specify an intentioned action. If this assumption does not hold, e.g., due to underspecified utterances or recognition errors, the reference resolution process can be used to identify objectives for reparation or clarification.

An instruction may be uttered using different (interacting) modalities and may contain references to entities in the world $W(t)$, which together build the topic subset $T(t)$ - explicitly including the communicating partners.

3.2. The Reference Resolution Process

In our model the world $W(t)$ subsumes the topic $T(t)$, the user $U(t)$ and the system $S(t)$. The system perceives $p_S(W(t))$ which includes $p_S(U(t))$ and thereby the utterances of the user over time. These percepts are interpreted using classifiers based on long- and short-term knowledge (discourse history), resulting in a conceptual representation of the world. Within this process of interpretation the concepts of the recognized references from the perceived utterances are matched against the updated conceptual representation of the world. This subprocess of the interpretation $i_S(p_S(W(t)), t)$ we call the *reference resolution process*. It can be said that the system has correctly resolved the references, if it has fully identified the topic $T(t)$ - which not necessarily implies that all of the references uttered by the user have been perceived or interpreted correctly.

4. System Design

The design of our reference resolution engine which processes $i_S(p_S(W(t)), t)$ is guided by the idea of an expert system with different shells. We thereby found the distinction of three nested shells useful, as indicated in Figure 2.

4.1. Phases of Reference Resolution

The process of reference resolution can be logically divided into at least five phases, which are distributed over the three shells:

1. Preprocessing: referential expressions are transformed into queries to the RRE.

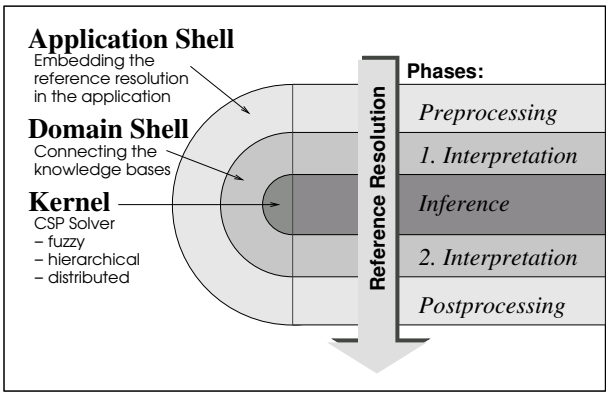

Figure 2. The reference resolution engine (RRE) is an expert system with three shells. The Kernel holds a constraint satisfaction problem solver. The Domain shell provides access to knowledge bases. The Application shell mediates between the embedding application and the RRE.

2. First interpretation: queries are analyzed and interpreted according to conceptual and operational knowledge.

3. Inference: the query is resolved and a set of solutions is generated.

4. Second interpretation: solutions are reinterpreted according to domain specific knowledge.

5. Postprocessing: solutions are transformed into application specific structures and returned.

4.2. The Kernel

Referential expressions (REs) may contain several constituents, each with a specific selectivity, which in conjunction are expected to enable an exact identification of the referent-object(s). This motivates us to handle such an expression as a constraint satisfaction problem (CSP). A RE is thereby represented as a variable and the constituents of the RE are mapped to one or more constraints. If the RE makes use of relative statements, such as "*the screw to the left of the block*", additional variables are introduced for each of the referent-objects. To meet the requirements for a reference resolution system, our Kernel combines several mechanisms for solving CSPs:

Fuzziness: To handle ambiguous expressions, we chose a fuzzy-based approach instead of pure boolean logic. For each constraint we evaluate a saliency expressed by a number in the interval $[0..1]$. The saliency of a combination $\cap$ of several constraints is computed by the T-norm. There are several of such T-norms, with min being the most well-known. Using the min T-norm ensures that the saliency of the whole expression is determined by the weakest constraint. While this might be useful for applications in the area of process control or quality management, the constraints in reference resolution are of a different quality. Here each constraint contributes to the discrimination of the solution. Therefore a T-norm is needed where the result of the operation is influenced by each of the saliencies of the single constraints. A simple norm which satisfies this criteria is the quasilinear T-norm which is computed by $max(0, a + b - 1)$.

Using a fuzzy-based approach also enables us to differentiate between strong and weak constraints. This is a different quality than constraints having a strong or weak saliency. The later is the result of the evaluation while the first influences the evaluation and therefore enables a fine grained adjustment of the interplay of the constraints. There are several of such modifiers known in fuzzy literature. A very simple one is $\mu_{not}(x) = 1 - \mu(x)$, which negates the original constraint $\mu(x)$. Stressing a constraint can be done, e.g., using $\mu_{very}(x) = \mu(x)^2$ and diminishing with $\mu_{\text{more-or-less}}(x) = \sqrt{\mu(x)}$.

Hierarchies: As we have seen in the Related Work, humans have certain preferences about the order of the constraints. There are several factors that might induce such a preference: selectivity of the constraint, perceptual availability, user specific preferences or, e.g., cultural conventions. When evaluating REs the selectivity is quite important. The sooner the set of possible interpretations of an expression can be reduced, the less evaluations have to be computed.

Also the order of the interpretation of the constraints is not independent, as, e.g., in "*the left car*", which is assumed to be evaluated as "*from the cars, the left one*" and not as "*the left object which is also a car*". Although "*left*" precedes "*car*", evaluating the constraint for selecting the objects of a specific type before the spatial constraint for the leftness of the object gives significantly better results.

The constraints in our CSP-system are therefore hierarchically ordered with respect to selectivity, preference and salient evaluation order. The default hierarchy can always be overwritten by single constraints to ensure that exceptions are always possible.

Refinement: Besides the constraints explicitly given in the utterance there can be others which are only implicitly assumed. For example, experiments showed that humans tend to prefer objects in the direction of their preferred cultural reading direction, if none other spatial information is given. Also objects in grasp-space may be preferred before objects far away. When two objects are combined, one would possibly prefer pairs of objects that are close together before others. These assumptions can be modeled by

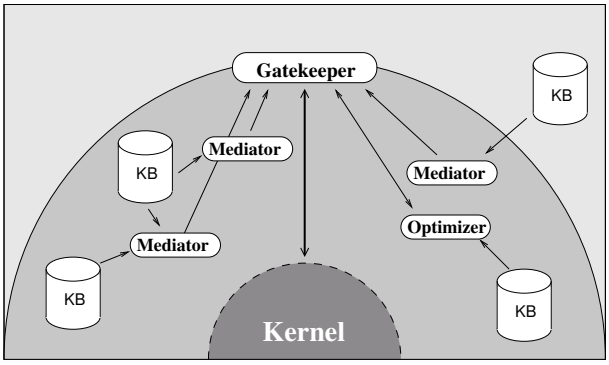

Figure 3. Queries are handled by the gatekeeper of the domain shell. He maintains a common ontology of concepts and operations, supported by several mediators for accessing different knowledge bases (KB). Optimizers are using domain specific knowledge to improve the quality of the queries before they are evaluated by the Kernel.

a combination of the fuzzy and the hierarchical capabilities of the system, e.g., by using a *more-or-less* predicate and by giving them a low priority in the hierarchy. The constraint solver is able to refine the solutions of the evaluation process by incrementally adding constraints of lower priority. By doing so, these additional assumptions might only be considered for discrimination in cases of ambiguous REs.

Performance: Compared to boolean CSPs, fuzzy CSPs lead to an increased search space. A simple way to approach this problem is to introduce a threshold for the minimal saliency we accept for a solution. This allows the CSP solver to prune all non-promising paths in the search tree at an early stage.

Instead of a hard threshold which may lead to an empty set of solutions when the optimal solution is below the threshold, we define the maximal number of solutions we would like to receive. If the CSP solver finds a solution better than one in the current list of solutions, the worst solution in the list is dropped and the new solution is inserted. The saliency of the worst solution in the list thereby defines the actual dynamic threshold.

4.3. The Domain Shell

The basis of the domain shell is a common ontology which is used to map queries to operations on the knowledge bases (KB). Several modules are involved in this process, as shown in Figure 3.

The **Gatekeeper** manages the querying process and administers access to all KBs using a dynamic common ontology about the conceptual and operational knowledge found in the different special purpose KBs.

The **Mediators** are the interfaces between the Gatekeeper and the special purpose KBs. They describe the conceptual knowledge of the underlaying KB and provide operations for exploration and filtering. These informations are incorporated into the common ontology by the Gatekeeper.

The **Optimizers** use domain specific knowledge to improve the quality of the solutions and increase the performance of the search process. This also includes the enrichment of the original query by additional constraints that can be assumed implicitly. For example, to successfully connect a screw with a block, the block has to have an empty opening matching the screws thread and the thread of the screw has to be exposed to some extent. By doing so, the search space can be significantly reduced.

4.4. The Application Shell

The application shell decouples the RRE from the embedding application by providing a special propositional **query language** to express the reference problems.

This query language is somewhat oriented on the idea of conceptual graphs of Sowa [23], with a syntax oriented on the functional programming language Scheme. For example, one query language representation of the utterance "*Take* [pointing] *this red part here and connect it to the* [pointing] *left front chassis.*" from Figure 1 in our color plate is:

```
(inst ?object-1 (time t2) (type THING))
(color ?object-1 (time t1) (color RED))
(more-or-less (pointed-at ?object-1 (time t1 t2)
              (agent USER_1)))
(inst ?object-2 (time t6) (type THING))
(more-or-less (pointed-at ?object-2 (time t4 t5 t6)
              (agent USER_1)))
(inst ?role-1 (type CHASSIS))
(has-role ?object-2 ?role-1)
(position ?object-2 (time t4)
          (qualifier LEFT)
          (perspective USER_1))
(position ?object-2 (time t5)
          (qualifier FRONT)
          (perspective USER_1))
(connectable ?object-1 ?object-2 (time t3))
```

For dealing with the dynamic structure of VR scenarios, all propositions have a timestamp. The query language also allows for fuzzy expressions, such as *not* or *more-or-less*, to modify the stringency of certain propositions, as can be seen in the example, where the fuzziness of the pointing gesture is stressed by the *more-or-less* modifier.

Also, the default hierarchical priority (see Kernel/Hierarchies) can be modified to model exceptions. This feature is intensively used by the Optimizers of the domain shell for adding implicit constraints. In our example the following constraints might be added to allow for a more distinct discrimination of the target objects:

```
(position ?object-1 (qualifier NEAR)
          (perspective USER_1) (priority LOW))
(position ?object-2 (qualifier NEAR)
          (perspective USER_1) (priority LOW))
(position ?object-2 (qualifier NEAR)
          (perspective ?object-1) (priority LOW))
```

Results: In general more than one solution will be generated by the fuzzy based CSP solver, which can be distinguished by their saliency. The differences of the alternative solutions may be exploited to initiate a subdialogue for clarification.

5. Implementation and Usage

The described concepts have been realized in an implementation for our Virtual Reality platform, which is driven by Avango [24], a high-level software framework on top of OpenGL Performer. We are using Scheme for rapid prototyping of conceptual changes and new constraints. After evaluation, performance critical parts are reimplemented in C++ in a second iteration. The composition of the system is done by Scheme scripts, which allows a dynamic reconfiguration and rapid prototyping of the complete system using a high-level language.

5.1. The Application Scenario

In one application scenario a user is instructing the system, represented by the ECA Max, to assemble models out of toy building blocks (see Figure 2 in our color plate). By introducing Max into the scene, spatial references can be specified using all alternative perspectives (see Related Work). Max is able to interact with the user, e.g., in that he reflects an interpreted referential expression by pointing at objects in the virtual world and uttering an appropriate sentence.

5.2. Internal Architecture

The RRE in the system is coupled with several other engines (see Figure 4). The PrOSA (Patterns On Sequences of Attributes) [15] engine realizes $p_S(U(t))$ and aspects of $i_S(p_S(U(t)), t)$, providing a high-level access to the recognized instruction. Besides gesture recognition tasks, PrOSA preprocesses spatial object relations regarding important body anchored reference systems (views, hands, ...) by continuously sorting objects into so-called *spacemaps* for delayed evaluation. The interpretation process is managed by a tATN (temporal Augmented Transition Network) [16] which analyzes the output of PrOSA and feeds the reference resolution process incrementally. The COAR engine (not shown in the figure) realizes $p_S(W(t))$ and the conceptualization of objects and aggregates in $i_S(p_S(W(t)), t)$. Semantic Entity nodes provide scene-graph based traversal

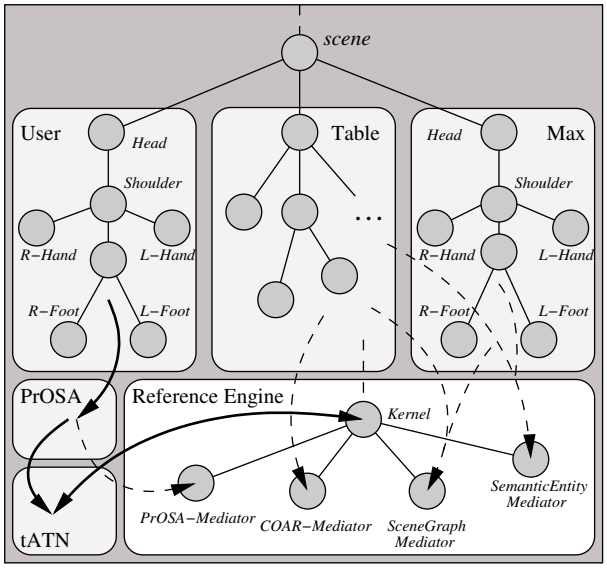

Figure 4. The RRE is embedded in the scene-graph of our VR setting. Several Mediators are connecting the engine to different knowledge sources. The PrOSA/tATN engines analyze and interpret the users utterances, incrementally feeding the RRE.

type access of external knowledge represented in a semantic net [17], and they are therefore also part of $i_S(p_S(W(t)), t)$. Several Mediators are binding these engines to the RRE, providing access to their domain specific knowledge.

6. Conclusion

Resolving references is a central part in task oriented dialogues. In the past, mechanisms for resolving references have only played a minor role in the VR society. Today scenarios with an evermore increasing complexity are available for virtual investigation and manipulation. Natural behavior in such environments results in complex user interactions utilizing several modalities.

These increasing demands can only be met by a reference resolution engine which is able to grow with the evolving systems. To understand natural user interactions, the interplay of both explicit user utterances and implicit assumptions that are being made by human recipients and which are inherently expected by the speaker in natural dialogues is essentially. Here, an open framework is needed which allows for an easy combination and evaluation of reference resolution heuristics.

The presented RRE is targeted at these requirements. Its architecture is designed for an integration in todays scene-graph based VR systems. The underlying algorithms are

tailored for fast evaluations and incremental processing. By using fuzzy techniques coupled with hierarchical structuring, the interplay of different constraints can be modelled down to a very fine-grained level. Enhancements are easily made by designing new constraints or incorporating new knowledge bases by adding new mediators as needed. A pool of established constraints can be used as a toolbox for the design and rapid prototyping of new applications.

References

[1] M. Andre, V. G. Popescu, A. Shaikh, A. Medl, I. Marsic, C. Kulikowsky, and J. Flanagan. Integration of Speech and Gesture for Multimodal Human-Computer Interaction. pages 20–27, Tilburg, Jan. 1998.

[2] R. Arangarasan and G. N. J. Phillips. Modular Approach of Multimodal Integration in a Virtual Environment. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 331–336. IEEE, 2002.

[3] P. Biermann and I. Wachsmuth. An Implemented Approach for a Visual Programming Environment in VR. In *Proceedings Fifth Virtual Reality International Conference (VRIC 2003), Laval, France*, pages 229–234, may.

[4] R. A. Bolt. Put-That-There: Voice and gesture at the Graphics Interface. In *ACM SIGGRAPHComputer Graphics*, New York, 1980. ACM Press.

[5] M. Cavazza, X. Pouteau, and D. Pernel. Multimodal Communication in Virtual Environments. In *Symbiosis of Human and Artifact*, pages 597–604. Elsevier Science B. V., 1995.

[6] H. H. Clark. Dogmas of Understanding. *Discourse Processing*, pages 567–598, 1997.

[7] H. Eikmeyer, U. Schade, and M. Kupietz. Ein konnektionistisches Modell für die Produktion von Objektbenennungen. Technical Report 94/5, SFB 360 - Universität Bielefeld, 1994.

[8] W. Hayward and M. Tarr. Spatial Language and Spatial Representation. *Cognition*, 55:39–84, 1995.

[9] B. Jung, M. E. Latoschik, and I. Wachsmuth. Knowledge-Based Assembly Simulation for Virtual Prototype Modeling. In *IECON'98: Proceedings of the 24th annual Conference of the IEEE Industrial Electronics Society*, pages 2152–2157, Aachen, Sept. 1998. IEEE, IEEE Computer Society Press.

[10] B. Jung and I. Wachsmuth. Integration of Geometric and Conceptual Reasoning for Interacting with Virtual Environments. In *Proceedings of the 98'AAAI Spring Symposium on Multimodal Reasoning*, pages 22–27, 1998.

[11] T. Jörding and I. Wachsmuth. An Anthropomorphic Agent for the Use of Spatial Language. In K. Coventry and P. Olivier, editors, *Spatial Language: Cognitive and Computational Aspects*, chapter 4. Kluwer, Dordrecht, 2001.

[12] D. Koons, C. Sparrel, and K. Thorisson. Intergrating Simultaneous Input from Speech, Gaze and Hand Gestures. In *Intelligent Multimedia Interfaces*. AAAI Press, 1993.

[13] N. Krahnstoever, S. Kettebekov, M. Yeasin, and R. Sharma. A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 349–354. IEEE, 2002.

[14] F. Landragin, N. Bellalem, and L. Romary. Referring to Objects with Spoken and Haptic Modalities. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 99–104. IEEE, 2002.

[15] M. E. Latoschik. *Multimodale Interaktion in Virtueller Realität am Beispiel der virtuellen Konstruktion*. PhD thesis, University of Bielefeld, 2001.

[16] M. E. Latoschik. Designing Transition Networks for Multimodal VR-Interactions Using a Markup Language. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 411–416. IEEE, IEEE, 2002.

[17] M. E. Latoschik and M. Schilling. Incorporating VR Databases into AI Knowledge Representations: A Framework for Intelligent Graphics Applications. In *Proceedings of the Sixth IASTED International Conference on Computer Graphics and Imaging*. IASTED, ACTA Press, 2003.

[18] B. Lenzmann. *Benutzeradaptive und multimodale Interface-Agenten*. PhD thesis, Technische Fakultät, Universität Bielefeld, 1998.

[19] M. Lucente, G.-J. Zwart, and A. D. George. Visualization Space: A Testbed for Deviceless Multimodal User Interface. In *Intelligent Environments Symposium*, American Assoc. for Artificial Intelligence Spring Symposium Series, Mar. 1998.

[20] B. Myers, R. Malkin, M. Bett, A. Waibel, B. Bostwick, R. C. Miller, J. Yang, M. Denecke, E. Seemann, J. Zhu, C. H. Peck, D. Kong, J. Nichols, and B. Scherlis. Flexi-modal and Multi-Machine User Interfaces. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 343–348. IEEE, 2002.

[21] D. Petrelli, A. D. Angeli, W. Gerbino, and G. Cassano. Referring in Multimodal Systems: The Importance of User Expertise and System Features, 1997.

[22] A. Popescu-Belis, I. Robba, and G. Sabah. Reference Resolution beyond Coreference: a Conceptual Frame and its Application. In *Coling-ACL'98 (International Conference on Computational Linguistics - Meeting of the Association for Computational Linguisitics)*, pages 1046–1052, Montreal, Canada, 1998.

[23] J. F. Sowa, editor. *Knowledge-Based Systems: Special Issue on Conceptual Graphs*, volume 5/3. September 1992.

[24] H. Tramberend. Avango: A Distributed Virtual Reality Framework. In *Proceedings of Afrigraph '01*. ACM, 2001.

[25] C. Vorwerg, G. Socher, T. Fuhr, G. Sagerer, and G. Rickheit. Projective Relations for 3D Space: Computational Model, Application, and Psychological Evaluation. In *AAAI/IAAI*, pages 159–164, 1997.

[26] E. Zudilova, P. Sloot, and R. Belleman. A Multi-modal Interface for an Interactive Simulated Vascular Reconstruction System. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces ICMI'02, Pittsburgh, Pennsylvania*, pages 313–318. IEEE, 2002.