Dublin Institute of Technology
ARROW@DIT

Conference papers

School of Computing

2006-01-01

Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue.

Geert-Jan Kruijff
DFKI, gj@dfki.de

John D. Kelleher
Dublin Institute of Technology, john.d.kelleher@dit.ie

Nick Hawes
University of Birmingham, n.a.hawes@cs.bham.ac.uk

Follow this and additional works at: http://arrow.dit.ie/scschcomcon

Part of the Computational Engineering Commons

Recommended Citation

Kruijff, G.,Kelleher, J., Hawes, N.:Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue.PIT 06: Proceedings of Perception and Interactive Technologies, 2006, Kloster Irsee, Germany.

This Conference Paper is brought to you for free and open access by the School of Computing at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License

School of Computing

Articles

Dublin Institute of Technology *Year* 2006

Information Fusion For Visual Refer ence Resolution In Dynamic Situated Dialogue.

Geert-Jan Kruijff* John D. Kelleher† Nick Hawes‡

*DFKI, gj@dfki.de

†Dublin Institute of Technology, john.d.kelleher@dit.ie

‡University of Birmingham, n.a.hawes@cs.bham.ac.uk

This paper is posted at ARROW@DIT.

http://arrow.dit.ie/scschcomart/1

— Use Licence —

Attribution-NonCommercial-ShareAlike 1.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution.
  You must give the original author credit.
- Non-Commercial.
  You may not use this work for commercial purposes.
- Share Alike.
  If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

For any reuse or distribution, you must make clear to others the license terms of this work. Any of these conditions can be waived if you get permission from the author.

Your fair use and other rights are in no way affected by the above.

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit:

- URL (human-readable summary):
  http://creativecommons.org/licenses/by-nc-sa/1.0/
- URL (legal code):
  http://creativecommons.org/worldwide/uk/translated-license

Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher2, and Nick Hawes[3]

[1] Language Technology Lab, DFKI GmbH
gj@dfki.de,
WWW home page: http://www.dfki.de/~gj
[2] Dublin Institute of Technology
John.Kelleher@comp.dit.ie,
WWW home page: www.comp.dit.ie/jkelleher
[3] School of Computer Science, University of Birmingham
n.a.hawes@cs.bham.ac.uk,
WWW home page: http://www.cs.bham.ac.uk/~nah

Abstract. Human-Robot Interaction (HRI) invariably involves dialogue about objects in the environment in which the agents are situated. The paper focuses on the issue of resolving discourse references to such visual objects. The paper addresses the problem using strategies for *intra-modal fusion* (identifying that different occurrences concern the same object), and *inter-modal fusion*, (relating object references across different modalities). Core to these strategies are sensori-motoric coordination, and ontology-based mediation between content in different modalities. The approach has been fully implemented, and is illustrated with several working examples.

1 Introduction

The context of this work is the development of dialog systems for human-robot collaboration. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution. Reference resolution in situated dialog is a particular instance of the anchoring problem [Coradeschi and Saffiotti, 2003]: how can an artificial system create and maintain correspondences between the symbols and sensor data that refer to the same physical object?

In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Each referring expression used in the dialog introduces a representation into the semantics of its utterance. This representation must be bound to an element in the context model in order for the utterance's semantics to be fully resolved. Referring expressions that access a representation in the context are called *anaphoric*. In a *situated* dialog, human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. This introduces a form of reference, called *exophoric* reference. Exophoric references denote objects that have entered the dialog context through a non-linguistic modality (such as vision), but have not been previously evoked into the context. Consequently, for a robot to participate in a situated dialog,

the framework it uses for reference resolution must support the integration of different forms of perceptual knowledge with the context models it constructs through dialog. The importance of anaphoric and exphoric references to situated discourse is evidenced by the frequency with which they occur. For example, the two most common cases of definite descriptions in the TRAINS corpus on situated dialogue were anaphoric and exophoric definites [Poesio, 1994].

The dynamics of environmental interaction make exophoric reference resolution a genuine problem. The movement of objects and agents in the environment may result in changes to how objects are perceived, and thus how these objects may be referred to. For example, you may be talking to a robot about an orange juice carton, and then rotate it so the robot now faces the side of the carton. Regardless of this change, the robot should understand that it is still the same orange juice carton which you talked about earlier. To manage environmental change, we must consider how different sightings of an object may be identified as being one and the same object. Furthermore, we should ensure that this identification enables the robot to construct a more complete understanding of the object. This will allow us to address the uncertainty and partial coverage of the robot's perception. For example, the robot should be able to combine the perceptual information it gets from the front and side sightings of the same orange juice carton to construct a more complete representation of it.

In this paper, we present an approach that addresses the problem of resolving references in situated dialog. The approach uses a combination of two fusion strategies. To establish whether different object occurrences in a single modality concern one and the same object, we use an *intra-modal fusion* strategy. The results of this strategy are *equivalence classes* which store different occurrences of the same object within a single modality, and describe an object at a conceptual level. To establish relations between the equivalence classes that have been established in different modalities (i.e. establishing cross-modal bindings) we use an *inter-modal fusion* strategy. For this, we use ontology-based mediation, exploiting the conceptual character of equivalence classes. Inter-modal fusion provides the basis for exophoric reference resolution, and can help the further completion and disambiguation of content within modalities. The resulting approach has been fully implemented, and used on a Pioneer PeopleBot mobile robot in HRI scenarios dealing with human-augmented mapping, and visual object learning. We will discuss several working examples.

Contributions This paper presents a novel approach to the resolution of contextual references to visual objects in dynamic environments. We base reference resolution on establishing relations between equivalence classes, rather than individual instances. This makes it possible to handle changes in sightings of objects, without losing the conceptual integrity of the object. Maintaining this integrity means we can refer to the object as still "the same" as before, even though the situation has changed.

Overview §2 highlights some of the factors that affect situated reference resolution and reviews previous approaches. In §3 we describe our approach, and in §4 we describe how the approach has been implemented. Following this, in §5, we provide various worked examples to illustrate the implementation. The paper finishes with conclusions and future work.

2 Data and Previous Work

In §1 we introduced the concepts of anaphoric and exophoric references and noted how the dynamics of the environment and the agents in the environment can make the resolution of these types of reference difficult. In this section we discuss one particular difficulty, namely the temporal aspect of situated discourse reference induced by these dynamics, and how it affects reference resolution. We also review some of the previous approaches to reference resolution against this background.

The types of referring expressions we are concerned with denote physical objects that persist in time and space. Some of the properties of these objects are preserved across time while others change. One consequence of this environmental variability is that connecting object representations across different modalities must take this temporal dimension, and its attendant variability, into account. [Coradeschi and Saffiotti, 2003] observe that we cannot model this binding as a one-shot process. However, previous approaches to intra-modal fusion, for example [Loutfi et al., 2005,Gurevych et al., 2003] and [Alexandersson and Becker, 2003] focus on fusing only the most recent perceptual representation of an object with the representation of a linguistic referent that denotes the object. For example, [Loutfi et al., 2005] propose a *track* function that ensures that the cross-modal binding points to "the most recent and adequate representation of the object". [Gurevych et al., 2003] use the *overlay* operator of [Alexandersson and Becker, 2003] which computes "the maximally compatible combination of new information and old information". Adopting such an approach results in the loss of information relating to the previous perceptual states of objects. If a robot does not maintain knowledge about the previous state of an object related to the discourse, it will not be able to answer questions relating to that object's prior state. The following example dialog highlights this problem:[4]

1. **scene**: A ball is placed to the right of a box.

H1 "*This*$_i$ is *a box*$_i$"
R1 "OK"
H2 "*This*$_j$ is *a ball*$_j$."
R2 "OK"

2. **scene change**: The ball is moved to the left of the box.

H3 "Where is *the ball*$_j$?"
R3 "*It*$_j$ is to the left of *the box*$_i$."
H4 "Where was *it*$_j$ when you first saw *it*$_j$?"

In our framework, the resolution of an anaphoric reference, for example the pronominal references in R3 and H4, represents one type of intra-modal fusion. When a representation is first evoked in a dialog, an equivalence class is created to hold it. Subsequent linguistic references to this representation are then all fused into the same (linguistic)

[4] In this example the indices i and j indicate that all the referring expressions marked by a particular index refer to the same physical entity, and that the representations of these references are intra-modally bound within the linguistic context model.

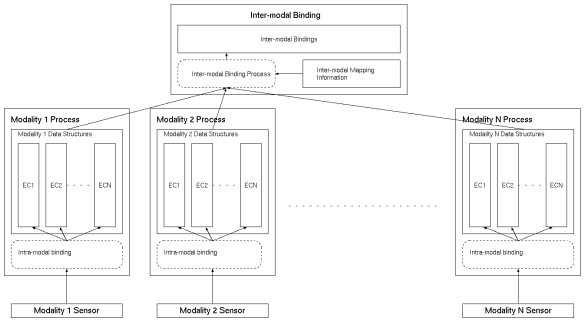

Fig. 1. Binding processes

equivalence class to denote that they are all about the same representation. The details of how we do this are presented in §3.1.

Our framework supports resolving exophoric references, for example the deictic references in H1 and H2, through inter-modal fusion. The reference to the object from the dialog is first entered into a linguistic equivalence class. It is then fused inter-modally (i.e. across modalities), with an equivalence class from another modality that represents the referent in that modality (e.g. the visual equivalence class that stores multiple sightings of the object being spoken about). We present the details of this in §3.2.

Given the above context, unless the robot retains information regarding the original location of the ball, it will not be able to answer the question posed in H4. This interaction points to the necessity of maintaining a perceptual history, similar to a discourse history. Indeed, the requirement for such a history has been noted elsewhere, see [Byron, 2003]. Finally, as noted in §1 the partial coverage and uncertainty inherent in robotic perception provides an extra layer of difficulty to reference resolution. An interesting effect of maintaining such a perceptual history is that it provides a mechanism for dealing with noise in real world sensor data. For example, it is possible that a vision sensor may not detect either the ball or the box at the time question H3 is posed to the robot, even though they are in the robot's visual field. However, if the robot maintains a history of an object's perceptual states (including where it appears in the world), it can use the last known location of the object to construct an answer to the question.

3 Approach

In the previous sections we briefly described our notion of an equivalence class, and indicated how these are created through intra-modal fusion. We then discussed how equivalence classes are associated through a process of inter-modal fusion to support references across modalities (e.g. the resolution of exophoric references). Figure 1 gives an illustration of these processes, which we discuss in more detail below.

3.1 The formation of equivalence classes through intra-modal fusion

An *equivalence class* (EC) represents an object at two levels: a concept-level (i.e. intensional) characterization of the properties of an object, and an instance-level (i.e. ex-

tensional) characterization of different sightings of the object. By separating these two levels of representation we can deal with the issue of change, and information completion. While sightings track change, the conceptual level provides a constant (though extensible) representation of the identity of the object.

We establish equivalence classes for objects within individual modalities, through *intra-modal fusion*. Intra-modal fusion represents a level of modality-specific information fusion. The input to an intra-modal binding process is formed by sensor events. These events are sets of entity descriptions from the input modality. Each of these structures contain information related to a particular sensed entity. For example, a sensor event from a visual sensor would consist of a set of structures each describing the visual properties of an entity in a scene.

The function of the intra-modal binding process is to bind each structure in a sensor event to an equivalence class that represents the entity the sensor data describes. For example, in visual fusion we use visual categorization to establish whether we are still looking at the same kind of object, based on aspects of visual appearance. Sensorimotoric coordination, i.e. the alignment between coordinate systems in perceptual and motoric modalities, subsequently enables us to determine whether the spatial positioning of an object has remained the same despite movement of the robot, or manipulation of the object. In our dialogue model, we use anaphoric binding to relate different mentions of (or references to) a discourse object to an equivalence class for that object.

Currently, we adopt a conservative approach to intra-modal binding, by trying to minimise the number of equivalence classes that are created within each modality. As the examples above illustrate, the particular process used for intra-modal binding is specific to the type of data being processed. Regardless of modality specific details, we propose a general strategy for intra-modality binding. The steps are as follows:

1. For each structure in a sensor event, retrieve the set of equivalence classes whose previously bound structures do not conflict with the input structure.
2. If the number of retrieved ECs $== 0$, create a new EC and bind the structure to it.
3. If the number of retrieved ECs $== 1$, bind the structure to the retrieved EC.
4. If the number of retrieved ECs > 1 trigger a conflict resolution strategy.

Step 4 of the above description introduces the notion of *conflict resolution*. The conflict occurs because the input sensor structure could possibly be bound to more than one EC. The purpose of conflict resolution to to determine which, if any, of these equivalence classes is the appropriate one. This issue is pervasive in any binding process where there is the possibility of ambiguity.

In the implementation of the architecture (described in §4), we resolve conflicts by binding the input structure to the EC with the largest amount of agreement between the sensor data in the structure and the sensor data in the structures already bound to the EC. In situations where this overlap heuristic still does not provide a single EC to bind to we can use a simple first-come first-served approach, or have the robot initiate a clarification dialogue as described in [Kruijff et al., 2006a]. Conflict resolution is one of the primary focuses for our future work.

Another point of note relating to the intra-modal binding process is that this process may trigger the inter-modal binding process. In the next section we describe inter-modal binding in more detail, and how it interacts with intra-modal binding.

3.2 Reference resolution through inter-modal fusion

The purpose of inter-modal fusion is to establish relations between equivalence classes across different modalities. Figure 1 shows the three components involved in inter-modal binding: inter-modal mapping information, the actual inter-modal binding process, and the resulting inter-modal bindings. Inter-modal mapping information addresses the problem of establishing a relation between the *content* in different modalities. For this, we exploit the conceptual characterization that an equivalence class provides for an object. We use ontology-based mediation (cf. [Wache et al., 2001] and [Gurevych et al., 2003]) to provide an ontological mapping between (modality-specific) conceptual systems to establish whether we can relate the content of the equivalence classes.

The inter-modal mapping information informs the inter-modal binding process about any correspondences between possible sensor inputs in different modalities. For example, this inter-modal mapping information helps us to bind the linguistic token *red* (specified conceptually as a *color*) to a cluster of visual sensor readings around the value *rgb(255,0,0)* (which we can also conceptualize as a color).

The inter-modal bindings component contains a set of structures that represent the inter-modal bindings that have been previously created by the binding process. Each structure is referenced by an id (e.g. b_1). This binding id is a unique identifier for the binding, and is used for indexing. Each binding structure contains at most one equivalence class from each modality, and at least one equivalence class in total. All equivalence classes bound together in this way can be assumed to refer to the same conceptual entity. The existence of an inter-modal binding structure with only one bound equivalence class indicates that there is an equivalence class in a particular modality that has not been bound to a representation from another modality. This may occur if, for example, the agent has seen an object that has not yet been talked about.

The inter-modal binding process is triggered by the intra-modal binding process. This happens when a new EC is created, and when an addition to an EC extends the range of sensor data associated with it (i.e. when a new attribute is used to help distinguish the additional entity from other entities, rather than when an existing attribute changes in value). The latter case is important as the addition of new information to an EC may cause a conflict between the extended EC and one or more of the ECs it may be inter-modally bound to.

When a new equivalence class EC_m is created in modality m, the inter-modal binding process must execute the following steps:

1. Retrieve the set of inter-modal binding structures that (1) do not have an EC from m bound to them, and (2) are already bound to ECs that can be aligned with EC_m. Whether alignment is possible is determined by applying the inter-modal binding information.
2. If the number of retrieved inter-modal binding structures $== 0$, create a new inter-modal binding structure containing EC_m.
3. If the number of retrieved inter-modal binding structures $== 1$, bind EC_m to the returned inter-modal binding structure.
4. If the number of retrieved inter-modal binding structures > 1, trigger a conflict resolution strategy.

As with the intra-modal binding algorithm, the issue of ambiguity introduces the need for conflict resolution. We adopt a similar resolution strategy here as was outlined for intra-modal binding: the candidate binding structures are ordered by the degree of possible alignment between the ECs bound to them and the input EC. Again, in situations where this overlap heuristic does not provide a single inter-modal binding structure, we use a simple first-come first-served approach or initiate a clarification dialogue.

4 Implementation

We have implemented the approach of §3 in a distributed architecture which integrates different sensorimotoric and cognitive modalities. The architecture enables a robot to move about in an indoor environment, and have a dialogue with a human about visual and spatial aspects of the situation. We have used this system in scenarios for human-augmented mapping and simple visual object manipulation, using a Pioneer PeopleBot.

Figure 2 shows the relevant aspects of the architecture. We have subsystems for communication, spatial localization & mapping, and visual processing. We use a BDI-based process to mediate between the different subsystems. We use beliefs to provide a common ground between different modalities, rather than being a layer on top of the different modalities. Beliefs thus provide a means for cross-modal information fusion, in its minimal form by co-indexing references to information in individual modalities [Gurevych et al., 2003]. Below we describe the communication subsystem and the visual subsystem in greater detail.

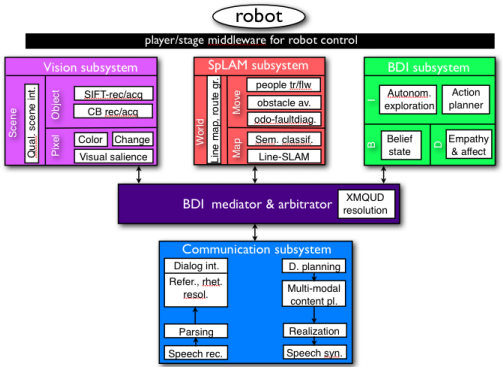

Fig. 2. The Implemented Architecture

The communication subsystem consists of several components for the analysis and production of natural language. It has been implemented as a distributed architecture using the Open Agent Architecture [Cheyer and Martin, 2001], following the idea of agent-based models for multi-modal dialogue systems [Allen et al., 2000].

For analysis we use the Sphinx4 speech recognition engine[5] with a domain-specific JSAPI speech grammar. The string-based output of Sphinx4 is then parsed with OpenCCG[6]. OpenCCG employs a combinatory categorial grammar [Baldridge and Kruijff, 2003] to yield a representation of the linguistic meaning

[5] http://cmusphinx.sourceforge.net/sphinx4/

[6] http://openccg.sf.net

that the string (i.e. the utterance) represents. We represent linguistic meaning as an ontologically rich, relational structure in a description logic-like formalism [Baldridge and Kruijff, 2002]. In structural dialogue analysis we relate the linguistic meaning of an utterance to the current dialogue context, in terms of how it rhetorically and referentially relates to preceding utterances. This yields an updated model of the situated dialogue context [Asher and Lascarides, 2003,Bos et al., 2003].

In the communication subsystem we use dialogue planning to produce flexible, contextually appropriate interaction. Based on a need to communicate, arising from the current dialogue flow or from another modality, the dialogue planner establishes a communicative goal. In turn, we plan the content to express this communicative goal, possibly in a multi-modal way using non-verbal (pose, head moves) and verbal means. In these planning steps, we can inquire the models of the situated context (e.g. dialogue context, visually scene) to ensure that the content we plan is contextually appropriate [Kruijff, 2005]. We realize verbal content using the OpenCCG realizer, which generates a string for the utterance, and then synthesize this string using text-to-speech[7].

In the vision subsystem, we have implemented visual scene understanding based on three cues: identity, color, and size of objects in the scene.[8] We use SIFT (Scale Invariant Feature Transform) features [Lowe, 2004] to recognize object identity. Each SIFT feature is a vector representing a particular arrangement of pixels centered on a particular point. When learning an object, SIFT features are extracted from the object (segmented by a fixed-size bounding box) and stored in a database along with a description of the object. During object recognition, SIFT features are again extracted from the image, and these are compared to the features associated with the previously learned objects. If the number of feature matches for an object is over a given threshold, the affine transformation of the object is estimated based on affinities between matched features. We obtain the pose of the object by applying this affine transformation to the model's segmentation mask. The robot calculates the color histogram over the segmented region of the IHS color space. The peak of smoothed histogram indicates the color. The vision subsystem consists of several CORBA[9] servers. We use an OAA agent to serve as a mediator between the communication subsystem and the vision subsystem.

5 Examples

In this section we provide various working examples to illustrate the implementation of our approach (§3) within our larger architecture for human-robot interaction (§4).

Figure 3 gives a flow diagram for processing simple dialogue describing a new visual object and some of its properties. We analyze the utterance "This is a box" in terms of its *linguistic meaning* and a (complex) characterization of its *dialogue act*. We model linguistic meaning as a relational structure over ontologically sorted content [Baldridge and Kruijff, 2002], and determine dialogue acts from the content and mood of the utterance. For "This is a box" this yields an assertion stating that an observed

[7] http://mary.dfki.de

[8] The vision subsystem was primarily implemented by Gregor Berginc (University of Ljubljana) and Bastian Leibe (TU Darmstadt).

[9] http://www.corba.org/

endurant (physical object) is an instance of a given type (a box). We inform BDI mediation of the linguistic meaning, its dialogue act, and the ECs for its discourse referents. Based on this, BDI mediation then decides to trigger a learning event in vision.

In vision we use a bounding box-method to determine the region of interest in the image for which we should learn a SIFT-based model. We create a visual referent id for the resulting model, and store this id in a new visual EC for the object. We provide the EC with a structural description of the object ("box") based on what was said [Kruijff et al., 2006b]. We then return the identifiers of the sighting and its visual EC to BDI mediation. BDI mediation creates a belief in which the dialogue and visual ECs are connected, and informs the communication subsystem that a visual model has been successfully acquired for the robot to provide feedback.

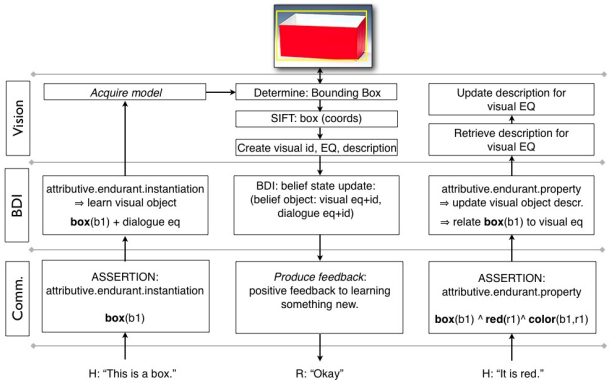

Fig. 3. Incremental update

When we next say, "It is red", reference resolution links the "it" to the discourse referent for "box". We provide the linguistic meaning (with the resolved reference) to the BDI mediator, as before, with a characterization that it expresses an assertion attributing a property to an object. Based on the dialogue act, BDI mediation retrieves the visual object EC to which the discourse referent for "box" corresponds, and informs the vision subsystem to update its description for the visual EC with the property "red".

The mechanisms for incrementally updating structural descriptions of visual object ECs thus rely on our ability to use the identifiers of ECs for co-indexing across utterances, relating the content we attribute to an object over the course of a dialogue. The same mechanism we can use in relating structural descriptions across different visual object models. For example, assume we say "This is an orange juice carton," rotate the carton and then say "This is its side." We can resolve the possessive pronoun to refer to the earlier mentioned carton. This provides the basis on which we can relate the (newly created) visual EC for "side" to the earlier created visual EC for "orange juice carton." The possessive expresses we have a *part-of* relation between "side" and the visual object that corresponds to the discourse referent to which we have resolved the pronoun. Using this information, we subsequently create a structural description for "side" which expresses this part-of relation between side, and the EC for "orange juice carton."

An interesting challenge is presented by "This is the side": We do not have a possessive pronoun, only the definite determiner. In this case, we use reasoning over hierarchical models to establish that a *side* is a part of an *object*. The definite determiner

leads us to check whether we can consider it a part of a visually salient object. In the current scene, the carton is salient, and it is a type of object which has sides; hence "the side" is most likely the side of the carton, and we can again use co-indexation across structural descriptions for visual ECs to establish a relation.

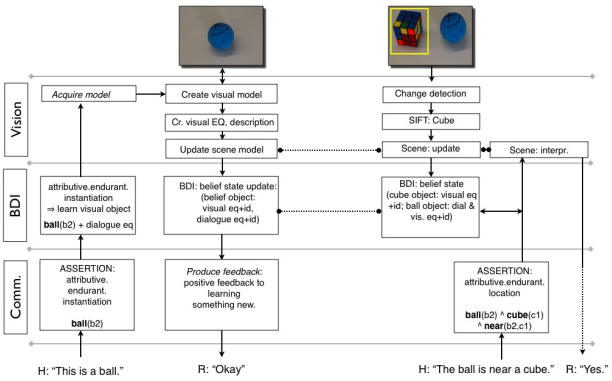

Fig. 4. Changing scenes

Figure 4 illustrates how we can use ECs to deal with changing scenes. After the visual system has learned a model for a new object ("ball"), we reorganize the scene, placing the ball next to a (known) cube. In this new scene, the original position of the ball has changed. We store the new scene in a ordered scene *history*. Now we say "The ball is near a cube", asserting the position of a (known) object. To interpret this assertion, we first need to resolve the new sighting of a ball object as an instance of the visual EC which corresponds to the discourse EC for "ball". Provided we can do so, we can then evaluate that the asserted spatial relation holds [Kelleher and Kruijff, 2005], and respond.

Finally, Figure 5 illustrates how we deal with references to previous scenes. As previously discussed, scene changes are inherent to dynamic environments but provide a problem for current approaches to vision/language-fusion (§2). To address this problem, we again use the ECs to relate different sightings of the same object, and provide history over the scenes in which these different sightings occurred

When the human asks "Where was the ball at first?", the communication subsystem analyzes the utterance and determines its meaning as expressing a question about the location of an object. The meaning also indicates that the location is temporally circumscribed, and that we are after the location of the ball. BDI mediation uses the dialogue EC for "ball" to determine for which visual EC we should retrieve a scene in which we had a sighting. It then queries the scene history for a past scene – specifically, for the first scene as indicated by the temporal modifier "at first".

Based on the retrieved scene, BDI mediation establishes the dialogue ECs for the visual objects. This information is then provided to the communication subsystem, to-

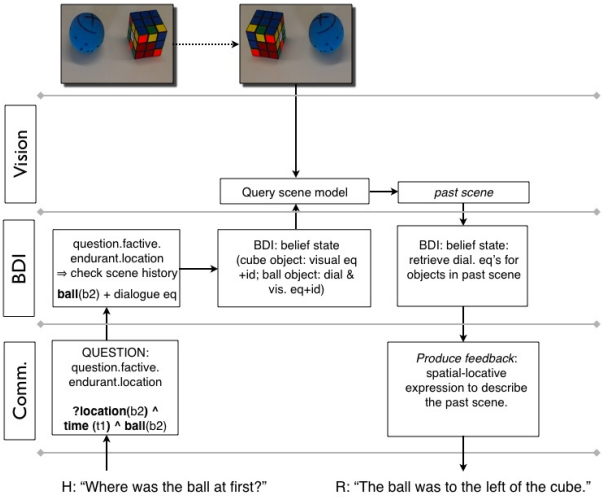

Fig. 5. Previous scenes

gether with the scene, so that we can generate a spatial-locative expression to describe the past scene [Kelleher and Kruijff, 2005]: "The ball was to the left of the cube."

6 Conclusions

We have presented an approach to reference resolution for situated human-robot dialog. The dynamics of the environment and the noise inherent in robot perception makes this task genuinely difficult. The framework distinguishes between intra- and inter-modal fusion. Intra-modal fusion establishes whether different percepts from a single modality concern one and the same object. We introduced the notion of an equivalence class to represent the set of percepts from a single modality that relate to the same object. Inter-modal fusion relates the use of object references across different modalities, e.g. the resolution of an exophoric linguistic reference against an object in the robots perceptual field. Within the framework, inter-modal fusion results in the binding of equivalence classes from different modalities. A key element of the inter-modal fusion process is the use of ontology-based mediation to provide a mapping between conceptual systems to establish whether we can relate percepts from different modalities. One of the main advantages of this framework is that it provides a mechanism for dealing with the temporal dimension of situated reference. In contrast with previous approaches, the inter-modal binding process does not restrict linguistic reference to the current perceptual state. Rather, due to the fact that equivalence classes retain all the prior percepts relating to an object, a linguistic reference can leverage any of the prior perceptual states of the object.

References

[Alexandersson and Becker, 2003] Alexandersson, J. and Becker, T. (2003). The formal foundations underlying overlay. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages pp. 22–36, Tilbugy, The Netherlands.

[Allen et al., 2000] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2000). An architecture for a generic dialogue shell. *Journal of Natural Language Engineering*, 6(3):1–16.

[Asher and Lascarides, 2003] Asher, N. and Lascarides, A. (2003). *Logics Of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom.

[Baldridge and Kruijff, 2002] Baldridge, J. and Kruijff, G.-J. M. (2002). Coupling CCG and hybrid logic dependency semantics. In *Proceedings of ACL 2002*, Philadelphia, Pennsylvania.

[Baldridge and Kruijff, 2003] Baldridge, J. and Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammar. In *Proceedings of EACL 2003*, Budapest, Hungary.

[Bos et al., 2003] Bos, J., Klein, E., and Oka, T. (2003). Meaningful conversation with a mobile robot. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Li nguistics (EACL'03)*, Budapest, Hungary.

[Byron, 2003] Byron, D. (2003). Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*.

[Cheyer and Martin, 2001] Cheyer, A. and Martin, D. (2001). The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1):143–148.

[Coradeschi and Saffiotti, 2003] Coradeschi, S. and Saffiotti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96.

[Gurevych et al., 2003] Gurevych, I., Porzel, R., Slinko, E., Pfleger, N., Alexandersson, J., and Merten, S. (2003). Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT-NAACL Workshop on Text Meaning*, pages pp. 14–21, Edmonton, Canada.

[Kelleher and Kruijff, 2005] Kelleher, J. D. and Kruijff, G.-J. M. (2005). A context-dependent model of proximity in physically situated environments. In *Proceedings of the ACL-SIGSEM workshop The Linguistic Dimension of Prepositions*, Colchester, England.

[Kruijff et al., 2006a] Kruijff, G., Kelleher, J., Berginc, G., and Leonardis, A. ((Under Review) 2006a). Structural descriptions in human-assisted robot visual learning. In *Human Robot Interaction*, Salt Lake City, Utah.

[Kruijff, 2005] Kruijff, G.-J. M. (2005). Contextually appropriate utterance planning for CCG. In *Proc. of the 9th European Workshop on Natural Language Generation*, Abderdeen, Scotland.

[Kruijff et al., 2006b] Kruijff, G.-J. M., Kelleher, J., Berginc, G., and Leonardis, A. (2006b). Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*, Salt Lake City, UT.

[Loutfi et al., 2005] Loutfi, A., Coradeschi, S., and Saffiotti, A. (2005). Maintaining coherent perceptual information using anchoring. In *Proc. of the 19th IJCAI Conf.*, Edinburgh, UK. Online at http://www.aass.oru.se/~asaffio/.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. In *Int. Jnl. Computer Vision*, pages 91–110.

[Poesio, 1994] Poesio, M. (1994). *Discourse Interpretation and the Scope of Operators*. Ph.d. dissertation, University of Rochester.

[Wache et al., 2001] Wache, H., Vögele, T., Visser, U., and G. Schuster, H. S., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *Proceedings of IJCAI 2001 Workshop "Ontologies and Information S haring"*, Seattle WA.