Integration of Visual Perception in Dialogue Understanding for Virtual Humans in Multi-Party interaction.

David Traum and Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292
{traum,morency}@ict.usc.edu

ABSTRACT

While the dialogue functions of speech in two-party dialogue have been extensively studied, there has been much less work on either multi-party communication, multimodal communication, and especially vision in a multi-party face-to-face setting. In this paper we report on one such effort to apply state of the art real-time visual processing to enhance a dialogue model of multi-party communication. We are concerned with situations in which there are at least three parties involved in conversation (at least one of whom is a human participant and at least one of whom is a virtual human). We focus on the visual behaviors of head orientation, head nods and head shakes, and examine how these behaviors influence several aspects of a multi-layer dialogue model, including addressee identification, turn-taking, referent identification, social affiliation, grounding, and question answering. We describe the extensions to the dialogue model and the implemented techniques for recognizing these behaviors and their impact on the dialogue models in real time, in realistic conversational settings from people participating in dialogue with virtual humans. We present several case studies (with accompanying videos) of dialogue fragments of the virtual agents with and without the recognition of these behaviors. Future work involves detailed studies on both the context recognition rates for this task as well as overall subjective impact on user satisfaction and dialogue efficiency.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Languages, Theory, Human Factors

Keywords

Dialogue Understanding, Visual Perception

1. INTRODUCTION

Face-to-face dialogue is a multi-modal process, in which participants use multiple channels to communicate and coordinate their communication and other activities. Probably the most important channel is speech/audio, which is used for most of the propositional content and main illocutionary force. However, the visual channel is also very important, especially for coordinating group communication, which is why people generally prefer face-to-face meetings over voice conference calls, sometimes even when extensive travel is required.

While the dialogue functions of speech in two-party dialogue have been extensively studied there has been much less work on either multi-party communication, multimodal communication, and especially vision in a multi-party face-to-face setting. Such studies are important primarily for better understanding the communication process as a whole, but are also needed for constructing Virtual Humans who can use these processes to communicate naturally with people and with each other in such a way that the communication is comprehensible by people.

In this paper we report on one such effort to apply state of the art real-time visual processing to enhance a dialogue model of multi-party communication. We are concerned with situations in which there is at least one human participating with at least one virtual human, and in which there are at least three parties involved in the conversation. There are a number of non-verbal behaviors that may influence the meaning and coordination of multi-party dialogue. In this initial effort, we focus on recognition of three behaviors: head orientation, head nods, and head shakes, and the various meanings that such behaviors can have in different dialogue contexts.

The following section describes related work on dialogue modelling, visual perception, and integration with embodied conversational agents (ECAs). Section 3 presents the dialogue model used as a basis to integrate visual information. Section 4 highlights important ways in which visual information influences the dialogue and methods for recognizing behaviors using vision. Section 5 shows how we adapted the dialogue model to incorporate the visual features. Section 6 presents a proof of concept implementation of our adapted dialogue model in a multi-party scenario. We conclude in Section 7, with our research program for extending this work.

2. RELATED WORK

There has been a substantial, interdisciplinary body of work on modelling dialogue structure and the information needed by participants to understand and engage in dialogue. Conversation analysts have looked at aspects of the fine interactive structure, such as how people coordinate who is to speak when [29]. Turn-taking has also been studied by social psychologists, e.g. [15]. Speech act theorists [3, 30] has been important at looking how people "do things with words". Linguists have also noted that dialogue is multi-functional, comprising not just a single speech act function, but different kinds of action [33, 2]. Allwood describes expressive, evocative and feedback functions of dialogue utterances. Sinclair and Coulthard have a series of hierarchical "ranks" of action, including

acts, moves, exchanges, and transactions. Clark and Shaefer[12] model the *Grounding* process of reaching mutual understanding as collaborative contributions, composed of proposal and acceptance phases. Ginzburg models the structure of dialogue context as a "gameboard", including a structure of *questions under discussion* (QUD), that is used for interpreting short answers [16].

The dialogue modelling work has been used and extended in dialogue systems that can use this model to participate in conversation. Grosz and Sidner [17] model the structure of discourse (including dyadic task-oriented dialogue) as having hierarchical focus spaces, which mirror the intentional structure of the task, and can be used to resolve underspecified referring expressions. Novick [27] Traum and Hinkelman [43] and Bunt [9] have developed multi-level dialogue act schemes that include control functions such as turn-taking and grounding as well traditional speech act functions. The information state approach to dialogue modelling [38], models aspects of dialogue context as an*information state*, with *dialogue acts* that represent the meaning of conversational behaviors in context, and act as functions to *update* the information state.

There has been considerable work on gestures with embodied conversational agents. Bickmore and Cassell developed an embodied conversational agent (ECA) that exhibited many gestural capabilities to accompany its spoken conversation and could interpret spoken utterances from human users [4]. Sidner *et al.* have investigated how people interact with a humanoid robot [31]. Nakano *et al.* analyzed eye gaze and head nods in computer–human conversation and found that their subjects were aware of the lack of conversational feedback from the ECA [26]. Numerous other projects (e.g. [39, 10]) explore aspects of gestural behavior in human-ECA interactions. Physically embodied ECAs—for example, ARMAR II [13, 14] and Leo [6]—incorporate the ability to perform articulated body tracking and recognize human gestures.

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. Stiefelhagen developed several systems for tracking face pose in meeting rooms and has shown that face pose is very useful for predicting turn-taking [35]. Takemae *et al.* also examined face pose in conversation and showed that if tracked accurately, face pose is useful in creating a video summary of a meeting [36]. Siracusa *et al.* developed a system that uses head pose tracking to interpret who was talking to who in conversational setting [34]. The position and orientation of the head can be used to estimate head gaze which is a good estimate of a person's attention.

A few research groups have begun to look at use of vision in multiparty dialogue. Matsusaka et al [22] constructed a robot named ROBITA, who could carry on conversations in Japanese with two people about baseball. Vision was used for face identification, face direction, and posture detection. Face direction information was used for turn-taking: the next speaker was the participant that the current speaker is looking at when finishing a turn. More recently, Bohus and Horvitz created a system that involves a virtual agent who can engage in open world dialogue in domains such as quiz games and receptionist tasks [5]. Their system recognizes faces and tracks locations, as well as tracking head orientation. Head orientation is used to decide who the addressee is, as well as for turn-taking (a release-to-system action is recognized if a human speaker ends the turn while addressing the system). Neither of these systems considers the problem of vision used to support multiple artificial agents engaged with humans.

3. MULTI-PARTY DIALOGUE MODEL

The dialogue model we are using is that of [39, 41]. Here conversational meaning is modeled using the information state approach, and partitioning the information into *layers*, each of which consists of a set of components of the information state, a set of dialogue moves, and update functions linking the moves and components. Each layer represents a distinct dialogue function, similar in many respects to dialogue act schemes such as [33, 27, 20], though without the organization into *ranks* of [33] or the strict requirements of *dimensions* of [8]. The layered model of [39, 41] is summarized in figure 1 and described briefly below.

One of the basic aspects of the information state is a set of *participants* who can be involved in interaction. The participant model includes basic information such as the name of the participant, the participant's focus, behaviors performed by the participant, as well as relational information to the agent who is doing the modeling, such as the participant's contact and attention status and trust. The *contact* layer manages whether this participant is available to engage in communication, using actions of making contact and breaking contact. The *attention* layer manages the focus of the participant.

Multiple *conversation*s can be active at a time, and each one has it's own internal structure. Conversation structure includes

- a list of participants in the conversation (who are assumed to understand the grounded contributions), divided into active participants who perform speaker and/or addressee roles in utterances of the conversation, and overhearers (who don't).
- modality of the conversation (face-to-face, radio, etc)
- the turn-holder (a participant, or none)
- the initiative-holder (a participant or none)
- the purpose of the conversation (e.g. to negotiate a task), if any
- a history of utterances that are part of the conversation
- a history of recently mentioned concepts
- a structure of questions under discussions
- a grounding structure, consisting of a bounded stack of common ground units [42]

The *turn-taking* layer manages who has the turn for each conversation (either one of the participants, *none*, or *conflict*). The *initiative* layer manages which of the participants (or *none*) has the initiative and is expected to move the conversation forward by initiating new subtopics and performing forward-looking acts. The *grounding* level manages how material is added to the common ground of material understood by the participants.

Finally the Core Speech Acts [43] include the forward and backward functions [1]. These have effects on aspects of the conversation, such as recency of mention (used for interpreting definite references such as pronouns), questions under discussion (QUD), on the negotiation state, and on the social state including establishing and relieving obligations to act and commitments to propositions. Core speech acts that have not been grounded are represented in the contents of grounding units but not yet in the main part of the information state.

Each of the dialogue acts described above has at least an *actor*, who performs the dialogue act, a *addressee* that is the main intended audience, and a dialogue act type. Most will also have other components, such as semantic content, that provide enough information for the update functions.

Layer	Info state components	Dialogue Acts
contact	participant contact	make-contact, break-contact
attention	focus	show, request, accept
conversation	conversation, topic conversation participants	start-conversation, end-conversation, confirm-start, deny-start, identify-topic, join, leave
turn-taking	conversation turn	keep-turn, hold-turn, release-turn, assign-turn
initiative	conversation initiative	take-initiative release-initiative
grounding	conversation CGUs	initiate, continue, repair, acknowledge, request-repair, cancel
Core	Social state (obligations, commitments, trust), QUD, Negotiation, CGU contents	**Forward:** assert, info-req, order, request, thank, greeting, closing, express, check, suggest, promise, offer, apology, encourage, accuse, intro-topic, avoid
		Backward: accept, reject, address, answer, divert, counterpropose, hold, check, clarify-parameter, redirect

Figure 1: Summary of Information state and dialogue acts

The information state represents a snapshot of an agent's model of an ongoing interaction with other participants. When a participant speaks (or otherwise performs communicative action), this will be interpreted as performing a (possibly empty) set of dialogue acts, and the information state will be updated with the effects of this communication.

4. VISUAL PERCEPTION AND MULTI-PARTY INTERACTIONS

There are a number of ways in which people use visual information to impact the aspects of the dialogue model presented in the previous section. In this section we describe these aspects, as well as the model of the actions recognized by the visual channel, and finally a message API for communication between the visual recognizer and the dialogue manager.

4.1 Aspects of the model influenced by visual recognition

Many aspects of dialogue understanding can be influenced by visual information about the speaker's behavior. We start by examining the role of a few behaviors in several phenomena, namely addressee recognition, turn-taking, grounding, focus of attention, inter-personal relationships, and feedback actions. Each of these are briefly described below.

As mentioned above, all dialogue acts have an addressee as one of their parameters. There are several aspects of information used to recognize an addressee, including explicit call by name, context of who the previous addressee and speaker are, as well as expectation of who is most likely to be addressed with such an act (or set of acts). Visually recognized information such as the orientation of the body, head, and eyes of the participants can also play a role [19]. All things being equal, one expects the speaker to gaze toward the addressee.

Turn-taking is also greatly influenced by head and eye gaze of the speaker and addressee [21, 7], especially at moments of silence or at utterance completion points[15]. At this point a speaker will often look away to keep the turn, or look at a next speaker to assign the turn to that participant (or more weekly, invite that participant to take up a next turn).

Grounding can be achieved by using physical behaviors as well as verbal behavior. Gaze directed to an object of discussion can signal understanding [11], as can mirroring gestures and head nods. Just as with speech, other behaviors can also be used to display grounding such as contextually relevant responses of various sorts, such as those described below. Lack of understanding can be signaled e.g., with facial expressions such as furrowed brow, looking away or staring without moving.

Gaze and pointing can be used to display or signal focus of attention, and thus can be used for reference resolution.

Pointing, orientation and physical proximity [18] can be used to signal inter-personal relationship status, such as affiliation.

Head nods and head shakes are commonly used to answer yes-no questions, accept or reject offers, or express agreement or disagreement.

4.2 Visual Perception

Human interactions with an embodied conversational agent are often prolonged so the tracking algorithm needs to be robust enough to not drift over time. The visual perception module was built with the following requirements in mind:

- Automatic initialization
- User independence
- Robustness to different environment (lighting, moving background, etc.)
- Sufficient sensitivity to recognize natural (subtle) gestures
- Real-time processing
- Stability over a long period of time

Based on the previous discussion, we identified three important visual features:

GAZE To correctly model gaze, we define several states: *look-away* and *look-at-X* (one for each participant X). The look-at-X states mean that the human participant is currently facing one of the other participants (participant X), virtual or human. In the general case, look-at-X can also be used for deictic gestures where X is an object of interest. The look-away state means that the human participant is currently not facing any of the known participants (or salient object in the scene). To correctly estimate the current state, we first estimate the head position and orientation of the human participant and then project the head gaze vector in the 3D world to determine if it intersects a virtual agent.

In our visual perception module, head position and orientation is estimated using the Adaptive View-Based Appearance Model [24]. In this framework, key frames are acquired online during tracking and used later to bound the drift. When the head pose trajectory crosses itself, the view-based model can track objects undergoing

large motion for long periods of time with bounded drift. This approach is able to track subtle movements of the head for a long periods of time. When compared with an inertial sensor (*Inertia Cube*^{2}),the head pose estimation has a rotational RMS error smaller than the 3° accuracy of the inertial sensor[24]. The position and orientation of the head can be used to estimate head gaze which is a good estimate of the person's attention. When compared with eye gaze, head gaze is more accurate when dealing with low resolution images and can be estimated over a larger range than eye gaze [23].

HEAD NODS To correctly recognize head nods, the visual perception module will analyze the head motion during the last 1.2 seconds (average length of a head nod) and decide if a head nod is happening or not.

We compute likelihood measurements of head gestures (head nods and head shakes) using the computed head velocities as input features to a multi-class head gesture recognizer. In our perception module trained a multi-class Support Vector Machine (SVM) with two different classes: head nods and head shakes [25]. The head pose tracker outputs a head rotation velocity vector at each time step. We transform the velocity signal into a frequency-based feature by applying a windowed Fast-Fourier Transform (FFT) to each dimension of the velocity independently using a 32-sample, 1.2-second window. The multi-class SVM was trained on this input using an RBF kernel.

HEAD SHAKES While head shakes do not happen as often as head nods, they do have a strong correlation with negative feedback and should not be ignored.. This visual feature is created in a similar fashion to the head nods features. A separate pattern recognizer was trained for head shakes.

4.3 Communication API

The vision system sends messages to the dialogue manager with the results of visual recognition. The messages have the form:

```
vrVision facing gazee timestamp
vrVision gesture type timestamp
```

The first message tells the dialogue manager that the person being observed has shifted gaze to *gazee*. The second says that the person has performed one of the gestures head-nod or head-shake. Timestamps are in milliseconds.

5. UPDATED DIALOGUE RECOGNITION ALGORITHMS

In this section we describe how the vision information from the previous section is used to recognize aspects of dialogue described in Section 3. There are at least three ways that visual information can be used to update the model:

Behavior as utterance Visually recognized actions can be interpreted autonomously as performing one or more dialogue acts, similar to the way speech is interpreted. This is the most natural approach for recognizing conventionalized gestures that have proposition content or relational content that refers to previous (spoken or non-verbal) actions rather than simultaneous speech. Head-nods and head-shakes are good examples of these, as described below. Bows or a "thumbs up" as evaluation would be other examples.

Behavior setting context Visual information can be used as context to influence the interpretation of speech. A good example is gaze, in which the specific head or eye movements may happen much earlier than accompanying speech which may occur while someone is still looking at someone or something.

multi-modal fusion Vision and speech recognition can be used together as input channels to a multi-modal fusion recognition process. A good example would be iconic gestures, where one must interpret hand motions relative to the words that are being said at the same time in order to fully recognize the intended concepts.

We currently use each of the first two techniques but defer the last to future work.

5.1 Non-verbal "Utterances"

When head-nods and head-shakes are received, the dialogue manager treats these just as if the person being observed had said "yes" or "no" respectively. Specific meaning in terms of dialogue acts depends on the rest of the context.

HEAD NODS As mentioned above, head-nods are treated just as if the observed participant had said "yes". In this case, a set of rules will interpret potentially multiple dialogue acts. As with a verbal backchannel, performing a head-nod while someone is speaking does not take a turn. If there is some content that is ungrounded, needing only an acknowledgment, the nod will be seen as this kind of grounding act. If there is a yes-no question under discussion in a conversation that the nodder is a participant of, this will serve as an answer, both committing the nodder to the positive proposition, as well as resolving the question from QUD. Likewise, if there is a proposal on the table, the nodder is seen as accepting this. Nodding also acts as an indication of attention on the conversation.

HEAD SHAKES Head shakes are interpreted as if the observed participant had said "no". This functions the same way as a head nod in terms of grounding and attention, but opposite in polarity with respect to answering questions and addressing proposals.

5.2 Contextual Interpretation of Nonverbal Behaviors

The participant model includes a gaze field, which is updated continuously when new *facing* messages are received, which refer to a new gazee. This information is in turn used as context for a number of other updates, as described below.

ADDRESSEE RECOGNITION The previous addressee recognition algorithm used a set of information in a simple decision tree, as described in [37]. While this algorithm worked very well for the Mission Rehearsal Exercise [45], it suffered from two problems for more general application, as pointed out in [28]: lack of use of gaze information, and inadequate handling of group addressing. The former is now corrected with the following revised algorithm in Figure 2, in which step 2 is new.

TURN-TAKING The turn-taking model uses two sets of rules to recognize acts: proposal rules that look at some features of an utterance to decide that such an act might have happened, and selection rules that arbitrate in case multiple acts are proposed. We use the gaze information as follows.

If a speaker is looking away at the end of an utterance, then a hold-turn act is proposed. If speaker is looking at someone then the an assign-turn to the gazee is proposed. Other rules for proposing an assign-turn act include asking a question or making a repair or request-repair grounding move (assigning the turn to the

1. **If** utterance specifies addressee (e.g., a vocative or utterance of just a name when not expecting a short answer or clarification of type person)
   then Addressee = specified addressee
2. **else if** speaker facing someone
   then Addressee = faced participant
3. **else if** speaker of current utterance is the same as the speaker of the immediately previous utterance
   then Addressee = previous addressee
4. **else if** previous speaker is different from current speaker
   then Addressee = previous speaker
5. **else if** unique other conversational participant
   then Addressee = participant
6. **else** Addressee unknown

Figure 2: Addressee Identification Algorithm

addressee). Performing a counter-proposal triggers a rule to propose a release-turn action, as does an utterance by a speaker who does not have the initiative.

The selection rules currently prefer a hold-turn over both of the other acts, and an assign-turn over a release-turn. Thus the gaze will override the other end of turn indicators. The turn will be assigned at the end of turn gazee unless the addressee is someone else and a question is asked.

6. APPLICATION & TESTING

We have implemented the above theory and tested in the context of the Stability And Support Operations: Extended Negotiation (SASO-EN) domain [44].

6.1 SASO-EN

Our current test scenario is an expansion of that used in [40]. This scenario involves a negotiation about the possible re-location of a medical clinic in an Iraqi village. As well as the virtual Doctor Perez and a human trainee playing the role of a US Army Captain, there is a local village elder, al-Hassan, who is involved. The doctor's main objective is to treat patients. The elder's main objective is to support his village. The captain's main objective is to move the clinic out of the marketplace, which is considered an unsafe area. Figure 3 shows the doctor and elder in the midst of a negotiation, from the perspective of the trainee. There are three main issues under discussion, corresponding to different options for and plans to accomplish the location of the clinic:
– whether to move the clinic near to the US Base (the captain's preferred option, unsuitable for the elder)
–whether to keep the clinic in the marketplace (the preferred option of both the elder and the doctor, though initially with negative utility, unsuitable for the captain)
–whether to move the clinic to an old hospital location in the center of the village (no one's preferred option because of the large amount of work needed to make it viable, but with potential for positive utility).

For this scenario, the trainee can look either at the Doctor, the Elder, or away from both.

6.2 SASO-EN Visual Perception

Figure 3: SASO-EN Negotiation in the Cafe: Dr Perez (left) looking at Elder al-Hassan

The visual perception module described in Section 4.2 was customized for the SASO-EN scenario in three different aspects: definition of *gazee* regions, IR-support and optimized messaging protocol. Since the SASO-EN scenario is designed for one human participant and two virtual human, two regions were defined for the gaze estimation, one for the elder and one for the doctor. Also, the SASO-EN setup is in a dark environment, so we update the visual perception module to work with infra-red camera, under low-illumination. New camera calibration was necessary. Finally, we created an optimized network protocol between the visual perception module and the dialogue manager where only messages are sent when the visual state changes (instead of using a set frame rate). These customizations enabled multimodal interaction in the SASO-EN scenario.

6.3 Proof of Concept Validations

While we have not yet had a chance to do a full evaluation of the impact that the inclusion of visual dialogue act recognition has on a user's negotiation experience, we have done preliminary testing on several case studies that show improvement in the expected behavior. We illustrate some of these here and provide videos showing example interaction with and without vision.

ADDRESSEE RECOGNITION Using the previous algorithm of [37], there are some points where an addressee can not be determined. Of particular concern is at the beginning of a conversation, where little context is available for guidance. If you come up to the two agents and say "Hello" without vision, the agents will not know who is being referred to. In this case, a likely response is for the agents to clarify the addressee.

With vision, if you are looking at one of the participants, a simple "hello" will be interpreted as being addressed to the gazee, as shown in Figure 4.

QUESTION ANSWERING AND GROUNDING This example in Figure 5 shows that a question (in this case a clarification question) can be answered without taking a speaking turn. The doctor's response is thus very quick.

TURN-TAKING Without vision or an explicit verbal turn-taking act by a speaker who has the initiative, the dialogue manager assumes

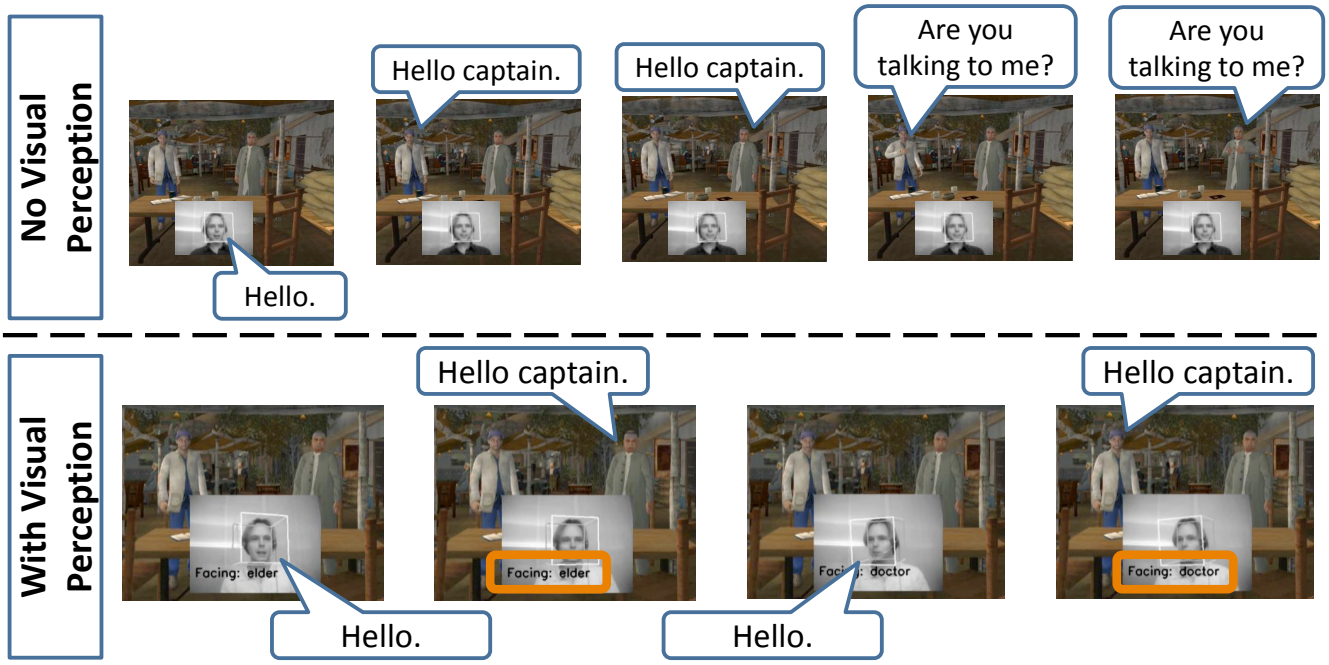

Figure 4: Illustration of dialogue using head orientation for addressee recognition

that the speaker continues to hold the turn, and waits for turn to timeout before continuing, as in the upper part of Figure 6. With vision, gaze at end of turn is used to recognize an assign-turn. No delay is needed, as shown in the lower part of Figure 6.

7. CONCLUSIONS AND FUTURE DIRECTIONS

We have presented an integration of several non-verbal behaviors, recognized using the visual signal, into a multi-layer dialogue model. The model works for multi-party conversation including small groups of human and artificial agents. We have implemented the model and tested in the SASO-EN negotiation domain. This model is, as far as we are aware, the most comprehensive implemented system involving visual recognition to support multi-party dialogue: the models support multiple virtual agents, they involve head gestures with multifunctional meaning as opposed to just head tracking. Head orientation is used to influence addressee detection and turn-taking, as in [22, 5], but unlike those systems our approach also uses verbal information and context information, so that e.g. a participant can ask a second participant something about at a third participant while looking at the third person, without being interpreted as addressing and giving the turn to the third person.

There are several directions in which we want to take this work. We describe them briefly here.

First, we would like to formally evaluate the impact of the vision recognition on human subjects. We have so far done only tests of specific cases as described in the previous section and tested whole dialogues with about 5 users. These tests show that the addition of vision has not broken the system and that people are able to negotiate, but are not sufficient to show whether there is any measurable difference in metrics such as user satisfaction, task efficiency, or even dialogue act recognition accuracy.

Second, we would like to expand the set of visual inputs used. We have started to experiment with pointing gestures and other proxemic cues. We would also like to recognize facial expressions and body posture shifts.Third, we would like to improve the dialogue recognition algorithms. Given sufficient data it may be possible to learn appropriate conditions and or weights for the various factors rather than use a strict preference ordering on signals.

Fourth, we would like to look at multi-modal fusion of input and look at the detailed timing between spoken word and visual movements by both speaker and addressee. Finally, we would like to use the information in the dialogue model to improve recognition of non-verbal behaviors, in a manner similar to [32].

Acknowledgments

We would like to thank the other members of the Virtual Human project at ICT for providing a framework in which to embed this research. The effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

8. REFERENCES

[1] J. Allen and M. Core. Draft of DAMSL: dialog act markup in several layers. available at: http://www.cs.rochester.edu/research/trains/annotation, Draft, 1997.

[2] J. Allwood. *Linguistic Communication as Action and Cooperation*. PhD thesis, Göteborg University, Department of Linguistics, 1976.

[3] J. A. Austin. *How to Do Things with Words*. Harvard University Press, 1962.

[4] T. Bickmore and J. Cassell. *J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.

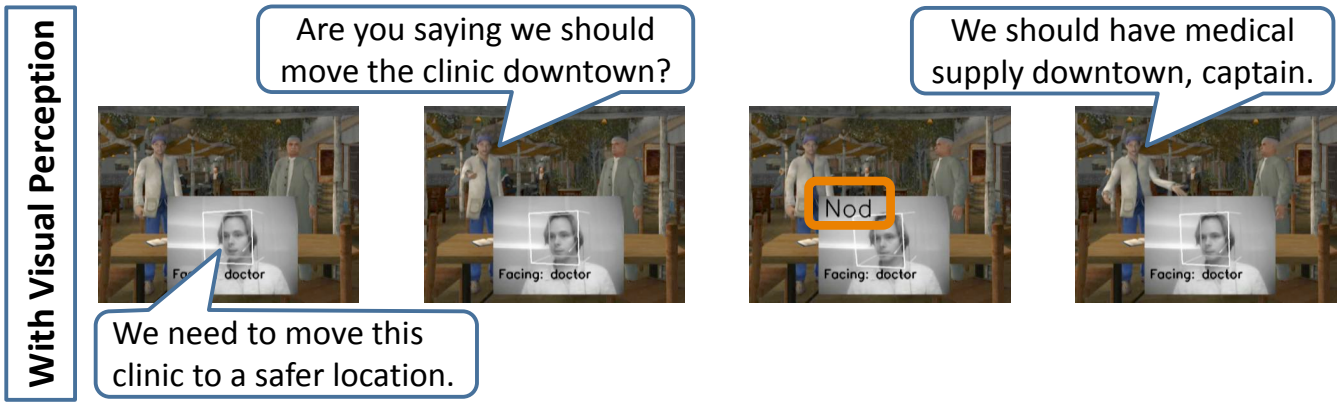

Figure 5: Illustration of dialogue using head nod to answer a clarification question

[5] D. Bohus and E. Horvitz. Dialog in the open world: Platform and applications. In *Proceedings of ICMI-MLMI 2009*, Boston, MA, October 2009.

[6] Breazeal, Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. In *The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, pages 1028–1035. ACM Press, July 2004.

[7] D. N. Brian, B. Hansen, and K. Ward. Coordinating turn-taking with gaze. In *in Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP-96*, pages 1888–1891, 1996.

[8] H. Bunt and J. Girard. Designing an open, multidimensional dialogue act taxonomy. In *Proceedings of DIALOR'05, the 9th Semdial Workshop on the Semantics and Pragmatics of Dialogue*, Nancy, France, 2005.

[9] H. C. Bunt. Dynamic interpretation and dialogue theory. In M. M. Taylor, F. Néel, , and D. G. Bouwhuis, editors, *The Structure of Multimodal Dialogue, Volume 2*. John Benjamins, Amsterdam, 1999.

[10] D. Carolis, Pelachaud, Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *Proceedings of IJCAI*, Seattle, September 2001.

[11] J. Cassell, Y. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 114–123, Toulouse, France, July 2001. Association for Computational Linguistics.

[12] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[13] Dillman, Becher, and P. Steinhaus. ARMAR II – a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, 1(1):143–155, 2004.

[14] Dillman, Ehrenmann, Steinhaus, Rogalla, and R. Zoellner. Human friendly programming of humanoid robots–the German Collaborative Research Center. In *The Third IARP Intenational Workshop on Humanoid and Human-Friendly Robotics*, Tsukuba Research Centre, Japan, December 2002.

[15] S. Duncan, Jr. and G. Niederehe. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–47, 1974.

[16] J. Ginzburg. Interrogatives: Questions, facts and dialogue. In S. Lappin, editor, *The Handbook of Contemporary Semantic Theory*. Blackwell, Oxford, 1996.

[17] B. J. Grosz and C. L. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[18] E. T. Hall. Proxemics. *Current Anthropology*, 9(2/3):83–108, apr 1968.

[19] N. Jovanovic and R. op den Akker. Towards automatic addressee identification in multi-party dialogues. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA, USA, 2004.

[20] S. Keizer and H. Bunt. Multidimensional dialogue management. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 37–45, Sydney, Australia, July 2006. Association for Computational Linguistics.

[21] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psyghologica*, 26:22–63, 1967.

[22] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multiperson conversation via multi-modal interface – a robot who communicates with multi-user. In *In Proceedings of Eurospeech*, pages 1723–1726, 1999.

[23] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environment. In *Proceedings of the Int. Conference on Automatic Face and Gesture Recognition*, pages 375–380, 2002.

[24] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 803–810, 2003.

[25] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the International Conference on Multi-modal Interfaces*, October 2005.

[26] Nakano, Reinstein, Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.

[27] D. Novick. *Control of Mixed-Initiative Discourse Through Meta-Locutionary Acts: A Computational Model*. PhD thesis, University of Oregon, 1988. also available as U. Oregon Computer and Information Science Tech Report CIS-TR-88-18.

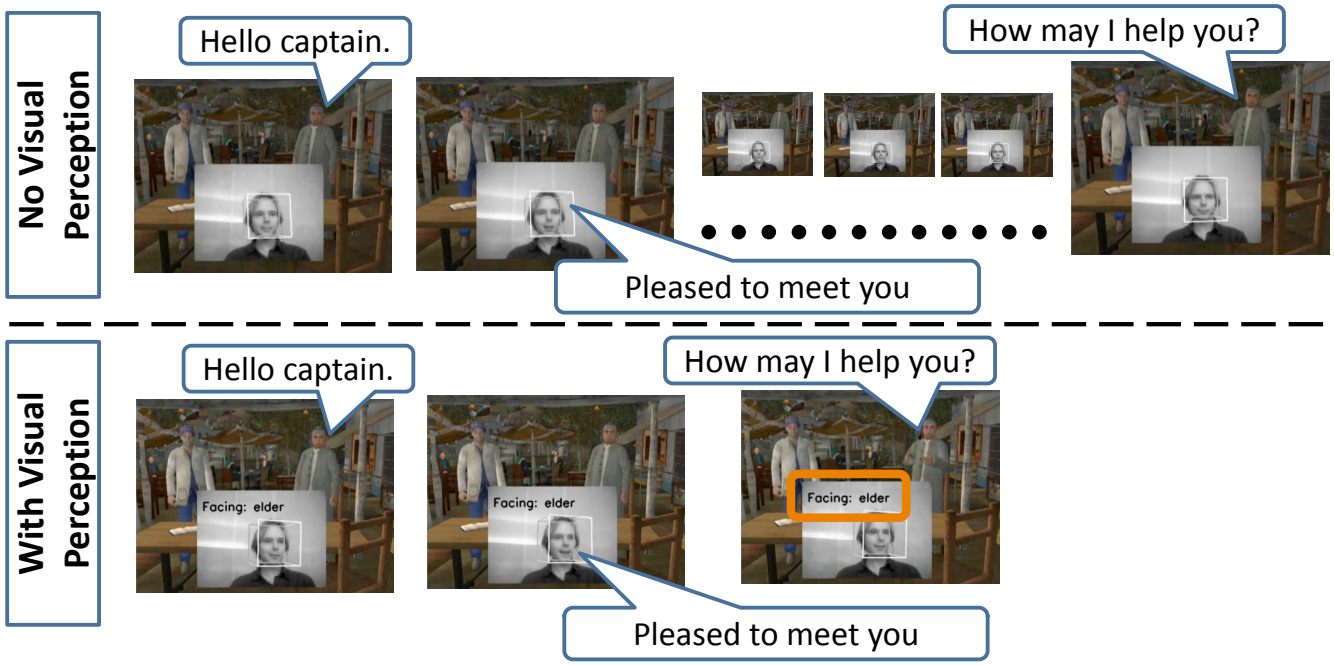

Figure 6: Illustration of dialogue using head orientation for turn-taking

[28] R. op den Akker and D. Traum. A comparison of addressee detection methods for multiparty conversations. In *Proceedings of DIAHOLMIA'09, the 13th Semdial Workshop on the Semantics and Pragmatics of Dialogue*, pages 99–106, Stockholm, Sweden, 2009.

[29] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.

[30] J. R. Searle. *Speech Acts*. Cambridge University Press, New York, 1969.

[31] C. Sidner, C. Lee, C.D.Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164, August 2005.

[32] C. Sidner, C. Lee, L.-P. Morency, and C. Forlines. The effect of head-nod recognition in human-robot conversation. In *HRI 2006*, March 2006.

[33] J. M. Sinclair and R. M. Coulthard. *Towards an analysis of Discourse: The English used by teachers and pupils.* Oxford University Press, 1975.

[34] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell. Haptics and biometrics: A multimodal approach for determining speaker location and focus. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, November 2003.

[35] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of International Conference on Multimodal Interfaces*, 2002.

[36] Y. Takemae, K. Otsuka, and N. Mukaua. Impact of video editing based on participants' gaze in multiparty conversation. In *Extended Abstract of CHI'04*, April 2004.

[37] D. Traum. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, LNAI 2922, pages 201–211. Springer Verlag, 2004.

[38] D. Traum and S. Larsson. The information state approach to dialogue management. In J. van Kuppevelt and R. Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. Kluwer, 2003.

[39] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of AAMAS '02*, pages 766–773, 2002.

[40] D. Traum, J. Rickel, S. Marsella, and J. Gratch. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of AAMAS 2003: Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 441–448, July 2003.

[41] D. Traum, W. Swartout, J. Gratch, and S. Marsella. A virtual human dialogue model for non-team interaction. In L. Dybkjaer and W. Minker, editors, *Recent Trends in Discourse and Dialogue*. Springer, 2008.

[42] D. R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Department of Computer Science, University of Rochester, 1994. Also available as TR 545, Department of Computer Science, University of Rochester.

[43] D. R. Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599, 1992. Special Issue on Non-literal language.

[44] D. R. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In H. Prendinger, J. C. Lester, and M. Ishizuka, editors, *IVA*, volume 5208 of *Lecture Notes in Computer Science*, pages 117–130. Springer, 2008.

[45] D. R. Traum, S. Robinson, and J. Stephan. Evaluation of multi-party virtual reality dialogue interaction. In *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1699–1702, 2004.