This paper is a preprint version of an article published by Springer-Verlag. The original publication is available at http://link.springer.com/chapter/10.1007/978-3-642-04617-9_68.

Formalizing Joint Attention in Cooperative Interaction with a Virtual Human

Nadine Pfeiffer-Leßmann and Ipke Wachsmuth

Artificial Intelligence Group, Faculty of Technology
Bielefeld University
{nlessman,ipke}@techfak.uni-bielefeld.de

Abstract. Crucial for action coordination of cooperating agents, joint attention concerns the alignment of attention to a target as a consequence of attending to each other's attentional states. We describe a formal model which specifies the conditions and cognitive processes leading to the establishment of joint attention. This model provides a theoretical framework for cooperative interaction with a virtual human and is specified in an extended belief-desire-intention modal logic. keywordscooperative agents, attention, alignment, BDI, modal logic

1 Introduction

A foundational skill in human social interaction, joint attention is receiving increased interest in human-agent interaction. Attention has been characterized as an increased awareness [1] and intentionally directed perception [2] and is judged to be crucial for goal-directed behavior. Joint attention can be defined as simultaneously allocating attention to a target as a consequence of attending to each other's attentional states [3]. In contrast to joint perception (the state in which interactants are just perceiving the same object without further constraints concerning their mental states), the intentional aspect of joint attention has been stressed, in that interlocutors have to deliberatively focus on the same target while being mutually aware of sharing their focus of attention [2] [4].

The computational modeling of joint attention mechanisms or prerequisites thereof, such as perceptional attention focus, convincing gaze behavior, gaze following skills, has been addressed in cognitive robotics, e.g. [3] [5], and research on virtual humans and embodied conversational agents, e.g. [6] [7]. However, aspects of intentionality and explicit representation of the other's mental state are not accounted for in these approaches altogether.

In this paper, we address the cognitive challenges of joint attention in action coordination of cooperating agents [8]. According to Pickering and Garrod [9] successful communication is based on joint processes, called alignment, which realize action coordination between interlocutors without an explicit exchange of information states. In previous work we have argued [10] that one central condition of such alignment processes consists of joint attention and that activation of a dynamic working memory and a partner model are crucial constituents.

We investigate joint attention in a cooperative interaction scenario with the virtual human Max, where the human interlocutor meets the agent face-to-face in 3D virtual reality. The human's body and gaze are picked up by infrared cameras and an eye-tracker [11]; e.g., Max can follow the human's gaze. The agent's mental state is modeled in the BDI (Belief-Desire-Intention) paradigm. In order to establish joint attention, the interlocutors need to be aware of each other's current epistemic activities. The human interlocutor's focus of attention is inferred from her overt behaviors, and focused objects are activated as salient in the agent's dynamic working memory; for detail cf. [10].

In this paper we describe a formal model which specifies the conditions and cognitive processes leading to the establishment of joint attention. This model provides a theoretical framework for a cooperative interaction scenario with the virtual human Max and the CASEC cognitive architecture introduced in [10]. In Section 2, we firstly introduce the use of activation values in modal logic and derive a definition of attention in Section 3. In Section 4, a formal definition of joint attention with regard to the required mental states is presented. In Section 5, we formally examine the action chain and skills involved bringing about the mental states requisite for joint attention. Section 6 presents a conclusion.

2 Formal Specification

To establish joint attention an agent must employ coordination mechanisms of understanding and directing the intentions underlying the interlocutor's attentional behavior, cf. [10]: The agent needs to (r1) track the attentional behavior of the other by gaze monitoring and (r2) derive candidate objects the interlocutor may be focusing on. Further, the agent has to (r3) infer whether attentional direction cues of the interlocutor are uttered intentionally. The agent has to (r4) react instantly, as simultaneity is crucial in joint attention and in response should (r5) use an adequate overt behavior which can be observed by its interlocutor.

Important in our approach is a dynamic working memory, which is inspired by Oberauer [12] who conceptualizes working memory in three successive levels characterized by increased accessibility for cognitive processes: (1) The activated part of long-term memory pre-selecting information over brief periods of time; (2) the region of direct access keeping a limited number of representational "chunks" available for ongoing cognitive processes; (3) the focus of attention holding the particular chunk selected for the immediate cognitive operation to be applied.

2.1 Beliefs

Our CASEC architecture (Cognitive Architecture for a Situated Embodied Cooperator) [10] adopts the BDI paradigm of rational agents [13] applying modal logic as a specification language, but additionally integrates a dynamic working memory. The formalism used to specify goals and beliefs builds on the possible worlds approach. We use a (doxastic) modal logic KD45 for modeling beliefs. In accordance with [13], we use the three modal connectives BEL, GOAL, and INTEND as atomic modalities.

Definition 1. *Any first-order formula is a state formula. If φ_1 and φ_2 are state formulae then also $\neg\varphi_1$ and $\varphi_1 \vee \varphi_2$. If φ is a formula then $BEL(\varphi)$, $GOAL(\varphi)$, and $INTEND(\varphi)$ are state formulae [13]. If i is an agent, then $(BEL_i\, \varphi)$ is an abbreviation denoting that agent i believes formula φ [14].*

Hereafter "formula" is to mean "state formula". To account for the dynamics of agent i's beliefs, we extend the formalism to include activation values (for further motivation cf. Section 3).

Definition 2. *If $(BEL_i\ \varphi)$ is a formula, then also $(BEL_i\ \varphi\ a), a \in \Re^+$ is a formula. $Acti(BEL_i\ \varphi\ a) = a$ returns the formula's current activation value a.*

Also terms are extended to contain an activation value:

Definition 3. *For a given formula φ with n terms, let $t\_set(\varphi)$ denote the set of terms of φ, $t\_set(\varphi) := \{e_i | e_i$ term of $\varphi, i = 1, ..., n\}$. Each term e_i with term value $\|e_i\|$ is augmented by an activation value a. Therefore we define $\widehat{e}$ to consist of: $\widehat{e} := (\|e\|, a), a \in \Re^+$. The function $Acti(\widehat{e}) = a$ returns the term's current activation value.*

Activiation values influence the beliefs' accessibility for mental operations. They are calculated by an ACT-R-like function for modelling recency effects and decay. Additionally, automatic activation impulses of different origins with own decay rates are included to model the overall saliency of a belief.

The activation value of a formula φ consists of the average of the contained terms' activations, $\#e_i$ denoting the number of terms (i=1,...,n):

$$Acti(\varphi) = \sum \frac{Acti(\widehat{e_i})}{\#e_i}, e_i \in t\_set(\varphi) \tag{1}$$

The set of current beliefs is defined as follows:

Definition 4. *Let $Beliefs_i$ denote the entire set of agent i's beliefs. Then we define ${}^{cur}Bels_i := \{b_k | b_k \in Beliefs_i \wedge Acti(b_k) > \theta BEL_{acti}, k = 1, ..., n\}$*

θBEL_{acti} represents a threshold which is dynamically tuned so that only a limited number of items reside in the set of ${}^{cur}Bels_i$ modeling the region of direct access of Oberauer's working memory model (see Section 1). Figure 1 illustrates the extension of the classical set of beliefs to a dynamic model including activation values. Activation values can be seen as adding an additional dimension which allows for filtering mechanisms. Thus we model "increased awareness" by use of activation values for aligning a candidate set of mental operations to the current context as well as to the interaction partner.

In addition to the modal connectives introduced above, we follow [15] in adding HAPPENS and DONE to the atomic modalities. If α is an action then $(HAPPENS\, \alpha)$ states that action α will happen next and $(DONE\, \alpha)$ means that action α has happened. These basic temporal operators are augmented by the operator ";" responsible for describing event sequences e.g. $(\alpha; \beta)$ denotes that first action α and then action β is executed. Additionally, <> denotes the modal operator *possibly* and [ ] the modal operator *always* [14].

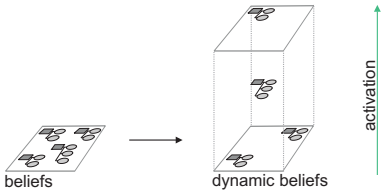

Fig. 1. Extending beliefs to dynamic beliefs with activation values

2.2 Goals - Intentions - Plans

Like [16] we see intentions as not reducible to beliefs and goals but as primitive modal connectives. However, they can be decomposed as follows (the modal operators PLAN, COMMIT are not formally introduced here).

Definition 5. *An intention is decomposed into the respective goal, the adopted plan and the commitment to use this plan as a means to achieve the goal:*

$$(INTEND_i\ \varphi) ::= (GOAL_i\ \varphi) \wedge (PLAN_i\ \varphi) \\ \wedge (COMMIT_i((GOAL_i\ \varphi), (PLAN_i\ \varphi)))$$

Whereas commitment is not directly relevant for the focus of attention, the parameters of the goal and the plan formulae directly apply to it. To cover the object related aspects of the formulae the function *t_set* (see Def. 3) is applied.

Definition 6. *The termsets of the modal connectives dissolve to the termset of the respective formula involved:* $t\_set(GOAL_i\ \varphi) := t\_set(\varphi)$, $t\_set(PLAN_i\ \varphi) := t\_set(\varphi)$. *The termset of an intention derives from the termset of the current goal:* $t\_set(INTEND_i\ \varphi) := t\_set(CurrentGoal_i(INTEND_i\ \varphi))$.

The *CurrentGoal* is the highest activated goal of the set of goals associated with the current intention. This set of goals consists of the intention's goal specification and the subgoals invoked in executing the adopted plan.

3 Defining the Focus of Attention

Like beliefs, also intentions and plans are qualified by activation values. We use activation values as a measure for saliency, i.e. an object with a higher activation value is more salient than one with a lower one. Whenever an object gets in the agent's gaze focus or is subject to internal processing, activation values are increased. That is, the set of ${}^{cur}Bels_i$ models *the region of direct access* proposed by Oberauer [12]. Depending on the processing step a new derived belief, a chosen intention, or an executed action of a plan corresponds to the focus of attention. We define the current belief and intention by use of activation values. The current plan step corresponds to the action of the currently adopted plan, an acyclic graph of nested goals covering the actions to be executed next:

Definition 7.

$$^{cur}BEL_i := \{b_x | b_x \in {}^{cur}Bels_i \wedge \forall b \in {}^{cur}Bels_i \wedge b \neq b_x : Acti(b_x) > Acti(b)\}$$
$$^{cur}INT_i := \{n_x | n_x \in Ints_i \ \wedge \ \forall n \in Ints_i \wedge \ n \neq n_x : Acti(n_x) > Acti(n)\}$$

$$^{cur}PLAN\_STEP_i := \{action_x | (COMMIT_i(GOAL_i\ \varphi), {}^{cur}PLAN_i\ \varphi) \wedge \ action_x \in Acy\_graph({}^{cur}PLAN_i\ \varphi)) \ \wedge \ (HAPPENS_i\ action_x)\}$$

As these processes of perception and cognition run concurrently, we conjoin all three aspects in our concept of focus of attention.

Definition 8.

$$ATT_i := \{t\_set({}^{cur}BEL_i) \cup t\_set({}^{cur}INT_i) \cup t\_set({}^{cur}PLAN\_STEP_i)\}$$

The focus of attention is part of dynamic working memory and is modulated by the changing beliefs and intentional states of the agent. Figure 2 illustrates the classical BDI model extended by the incorporation of activation values.

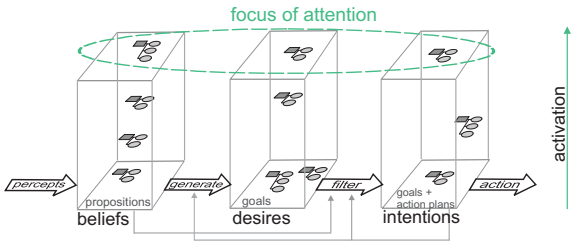

Fig. 2. BDI and Focus of Attention

4 A Definition of Joint Attention

In accordance with [2] we conceive of joint attention as an intentional process. Meeting the requirements of Sec. 2, we describe the mental state required for an agent i to believe in joint attention while focusing conjointly with its interlocutor j on a certain target ϑ (see Figure 3 for illustration and explanation next page).

Definition 9. $(BEL_i(JOINT\_ATT\ i\ j\ \vartheta))$ *iff*

1. *(being aware of other)* $BEL_i(ATT_j\ \vartheta) \wedge BEL_i(INTEND_j(ATT_j\ \vartheta))$
2. *(ascribing goal)* $BEL_i(GOAL_j(ATT_i\ \vartheta \wedge ATT_j\ \vartheta))$
3. *(adopting goal)* $GOAL_i(ATT_i\ \vartheta \ \wedge ATT_j\ \vartheta)$
4. *(feedback)* $BEL_i(BEL_j(ATT_i\ j))$
5. *(focus state)* $HAPPENS(< T(\vartheta) >_i \ \wedge < P(< T(\vartheta) >_j) >_i)$

(1) (being aware of other) By representing the explicit belief about the interlocutor's focus of attention $BEL_i(ATT_j\,\vartheta)$ the agent meets requirement (r1). To meet (r3) the agent additionally needs to infer whether the interlocutor intentionally draws its focus of attention on the object, $BEL_i(INTEND_j(ATT_j\,\vartheta))$.
(2) (ascribing goal) Agent i must believe that agent j has the goal that both agents draw their attention focus to the target $BEL_i(GOAL_j(ATT_i\vartheta \wedge ATT_j\vartheta))$. This belief can be evoked by an *initiate-act* of agent j e.g. by gaze-alternation.
(3) (adopting goal) The agent then needs to adopt the interlocutor's goal $GOAL_i(ATT_i\,\vartheta \wedge ATT_j\,\vartheta)$. To meet requirements (r4) and (r5), the agent as a recipient needs to employ an observable *respond-act*.
(4) (feedback) But for mutual belief, an additional *respond-act* is required from the initiator j so that agent i comes to believe $BEL_i(BEL_j(ATT_i\,j))$.
(5) (focus state) When agent i draws its focus of attention on the target $(< T(\vartheta) >_i)$ while perceiving that its interlocutor also focuses on the target $(< P(< T(\vartheta) >_j) >_i)$, then from agent i's perspective a joint attention state has been established. For definition of T (*test-if*) and P (*perceive-that*) see Sec. 5.

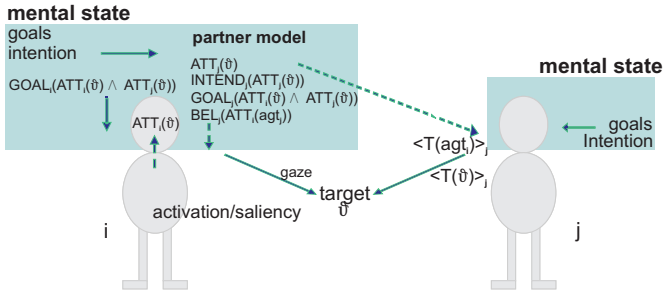

Fig. 3. Joint attention from agent i's perspective

5 Grounding Modal Connectives in a Logic of Action

After defining the mental state required for joint attention, we need to specify the epistemic actions that lead to the respective beliefs and goals. To this end, a logic of action is required. Like [14] we adopt standard *propositional dynamic logic*. In this logic, epistemic actions of perceptual kind are applicable to all formulae (propositions and actions) but do not allow direct perception of mental states. However an agent can perceive overt actions of its interlocutor as well as propositions of objects. We adopt the definition of two epistemic actions of [14]:

- ***Perceive-that***: Action of perceiving some ϑ in the external world.
  $< P(\vartheta) >_j \varphi$ (always φ is true after agent j has perceived ϑ)
- ***Test-if***: (precursor of *Perceive-that*) Test-if actions observable and testable from other agents as they are expected to have an observable counterpart.

By default we assume that, whenever an agent perceives something, it believes what it has perceived: $[P(\vartheta)]_i \rightarrow (BEL_i \vartheta)$. As time constraints and coordination are crucial in joint attention, a representation of time is needed.

Definition 10. *For α being an action expression, $Begin(\alpha) :=$ time t_{begin} at which execution of α starts, $End(\alpha) :=$ time t_{end} at which execution of α ends. The duration resolves to $Dur(\alpha) := End(\alpha) - Begin(\alpha)$. We write $(\alpha)_{j[t_{begin}, t_{end}]}$ to describe the points in time of agent j's action α beginning and ending.*

Test-action While an agent's *test-if* actions are observable [14], additional information is required to resolve the target. We use the dynamic working memory as a source of background information marking relevant objects and a partner model to account for the interlocutor's perspective. The candidate set of target objects are the objects in the interlocutor's line of gaze. Incorporating activation values in the reference resolution allows a fast and easy adjustment of the candidate set (meeting requirement (r2)). If the agent perceives the interlocutor's behavior as a *test-action* and is able to resolve a candidate object, the agent infers that the interlocutor's focus of attention must reside on that object.

$$< P(< T(\varphi) >_j) >_i \rightarrow \ (BEL_i(ATT_j\ \varphi)) \tag{2}$$

Beliefs about the interlocutor's focus of attention are updated dynamically, leading to new beliefs or increasing a belief's activation respectively. If the interlocutor focuses several times on an object (or for a long duration) the agent interprets this as the attention focus being *intentionally* drawn upon the target (cp. [10]):

$$< P(< T(\varphi) >_j) >_i; < P(< T(\varphi) >_j) >_i \rightarrow (BEL_i(INTEND_j(ATT_j\ \varphi))) \tag{3}$$

Initiate-actions One way to perform an *initiate-act* consists of gaze alternation. An object has to be the focus of attention for several times with additional short glances addressing the interlocutor inbetween (triadic relation).

$$\begin{aligned} &< P(< T(\varphi) >_j) >_i \ ; < P(< T(i) >_j) >_i \ ; < P(< T(\varphi) >_j) >_i \ ; \\ &< P(< T(i) >_j) >_i \rightarrow (BEL_i(GOAL_j(ATT_i\ \varphi)) \wedge (ATT_j\ \varphi)) \end{aligned} \tag{4}$$

Respond-actions Respond-actions play a crucial role to backup the actions the agents have sought to perform. They can consist of *smiling at*, *focussing on*, and *nodding to* the interlocutor. The *respond-actions* can be applied to establish mutual understanding between the interlocutors. E.g. after agent i performed a respond-act, it checks whether agent j has noticed its response:

$$\begin{aligned} &(DONE(< T(j) >_i)_{[t_{end}]} \ \wedge \ (HAPPENS < P(< T(i) >_j) >_i)_{[t_{begin}]} \wedge \\ &(Dur(< T(j) >_i \geq 2s)) \wedge ((t_{begin} - t_{end}) \leq 5s) \rightarrow BEL_i(BEL_j(ATT_i\ j)) \end{aligned} \tag{5}$$

(Heuristics: Empirical research, not quoted here, has shown that the recipient's response to an agent initiating joint attention needs to take place in a 5s time frame, with a signal duration of more than 2s.)

6 Conclusion

We presented work on equipping a cooperative agent with capabilities of joint attention. To this end, a formal definition of joint attention has been introduced. The required initiate- and respond-acts have been specified and grounded in a logic of action. The formalizations provide a precise means as to which requirements need to be met and which inferences need to be drawn to establish joint attention by aligning the mental states of cooperating agents. Implemented in the CASEC cognitive architecture [10] for our virtual human Max, they form the basis for the study of joint attention in a cooperative interaction scenario.

Acknowledgments. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center SFB 673. This paper is a preprint version of an article published by Springer-Verlag. The original publication is available at http://link.springer.com/chapter/10.1007/978-3-642-04617-9_68

References

1. Brinck, I.: The objects of attention. In: Proc. of ESPP2003, Torino. (2003) 1–4
2. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences **28** (2005) 675–691
3. Deak, G.O., Fasel, I., Movellan, J.: The emergence of shared attention: Using robots to test developmental theories. In: Proc. of the First Intl. Workshop on Epigenetic Robotics, Lund University Cognitive Studies, 85. (2001) 95–104
4. Hobson, R.P.: What puts the jointness into joint attention? In Eilan, N., Hoerl, C., McCormack, T., Roessler, J., eds.: Joint attention: communication and other minds. Oxford University Press (2005) 185–204
5. Doniec, M., Sun, G., Scassellati, B.: Active learning of joint attention. In: IEEE/RSJ International Conference on Humanoid Robotics. (2006)
6. Kim, Y., Hill, R.W., Traum, D.R.: Controlling the focus of perceptual attention in embodied conversational agents. In: Proceedings AAMAS. (2005) 1097–1098
7. Gu, E., Badler, N.I.: Visual attention and eye gaze during multiparty conversations with distractions. LNCS Intelligent Virtual Agents **4133** (2006) 193–204
8. Kaplan, F., Hafner, V.: The challenges of joint attention. Interaction Studies **7(2)** (2006) 135–169
9. Pickering, M.J., Garrod, S.: Alignment as the basis for successful communication. Research on Language and Computation **4**(2) (2006) 203–228
10. Pfeiffer-Lessmann, N., Wachsmuth, I.: Toward alignment with a virtual human - achieving joint attention. In: LNCS KI 2008, Springer (2008) 292–299
11. Pfeiffer, T.: Towards gaze interaction in immersive virtual reality. In: Virtuelle und Erweiterte Realität - Fünfter VR/AR Workshop, Shaker Verlag (2008) 81–92
12. Oberauer, K.: Access to information in working memory: Exploring the focus of attention. J. of Exp. Psych.: Learning, Memory, and Cognition **28** (2002) 411–421
13. Rao, A., Georgeff, M.: Modeling rational behavior within a BDI-architecture. Proc. Intl. Conf. on Principles of Knowledge Repr. and Planning (1991) 473–484

14. Lorini, E., Tummolini, L., Herzig, A.: Establishing mutual beliefs by joint attention: towards a formal model of public events. In: Proc of CogSci05. (2005)
15. Cohen, P., Levesque, H.: Intention is choice with commitment. Artificial Intelligence **42** (1990) 213–261
16. Bratman, M.: Intention, Plans, and Practical Reason. Havard Univ Press (1987)