Planning Referential Acts for Animated Presentation Agents

Elisabeth André, Thomas Rist
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
Email: {andre,rist}@dfki.uni-sb.de

Abstract

Computer-based presentation systems enable the realization of effective and dynamic presentation styles that incorporate multiple media. In particular, they allow for the emulation of conversational styles known from personal human-human communication. In this paper, we argue that life-like characters are an effective means of encoding references to world objects in a presentation. We present a two-phase approach which first generates high-level referential acts and then transforms them into fine-grained animation sequences.

1 Introduction

A number of researchers have developed algorithms in order to discriminate referents from alternatives via linguistic means (cf. (RD92)). When moving from language discourse to a multimedia discourse, referring expressions may be composed of several constituents in different media. Each constituent conveys some discriminating attributes which in sum allow for a proper identification of the referent. However, to ensure that a composed referring expression is intelligible, the system has to establish cohesive links between the single parts (cf. (AR94)).

In this paper, we argue that life-like characters are particularly suitable for accomplishing referring tasks. For example, a life-like character can:

- draw the viewer's attention to graphical object representations by pointing with body parts, and additional devices such as a pointing stick.
- make use of facial displays and head movements as an additional means of disambiguating discourse references,
- effectively establish cross-references between presentation parts which are conveyed by different media possibly being displayed in different windows;
- enable new forms of deixis by personalizing the system as a situated presenter.

For illustration, let's have a look at two example presentations taken from the PPP system (Personalized Plan-Based Presenter, (RAM97)). In Fig. 1, a pointing gesture is combined with a graphical annotation technique using a kind of magnifying glass.

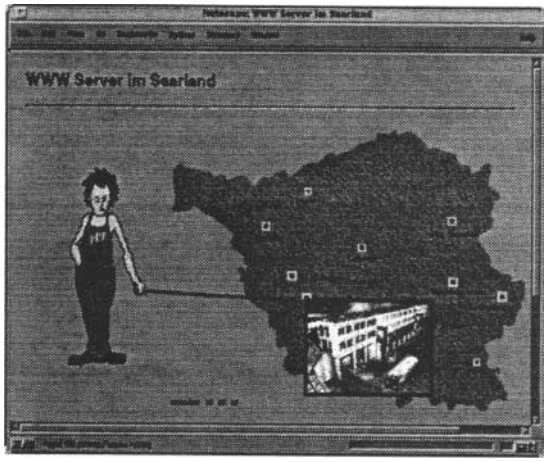

Figure 1: Referring to Objects Using a Magnifying Glass

The Persona provides an overview of interesting sites in the Saarland county by uttering their names and pointing to their location on a map. In addition, Persona annotates the map with a picture of each site before the user's eyes. The advantage of this method over static annotations is that the system can influence the temporal order in which the user processes an illustration. Furthermore, space problems are avoided since the illustration of the corresponding building disappears again after it has been

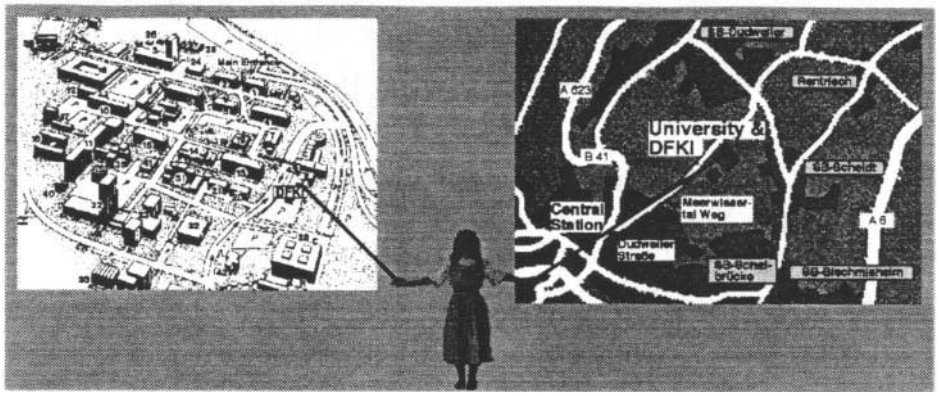

Figure 2: Establishing Cross-Media References

described. The example also demonstrates how facial displays and head movements help to restrict the visual focus. By having the Persona look into the direction of the target object, the user's attention is directed to the target object.

Whereas in the last example, the pointing act of the Persona referred to a single graphical object, the scenario in Fig. 2 illustrates how cross-media links can be effectively built up between several illustrations. In this example, the Persona informs the user where the DFKI building is located. It utters: "DFKI is located in Saarbrücken" and uses two pointing sticks to refer to two graphical depictions of DFKI on maps with different granularity.

As shown above, life-like characters facilitate the disambiguation of referring expressions. On the other hand, a number of additional dependencies have to be handled since a referring act involves not only the coordination of document parts in different media, but also the coordination of locomotion, gestures and facial displays. To accomplish these tasks, we have chosen a two-phase approach which involves the following steps:

(1) the creation of a script that specifies the temporal behavior of the constituents of a referential act, such as speaking and pointing

(2) the context-sensitive conversion of these constituents into animation sequences

2 Representation of the Multimedia Discourse

A few researchers have already addressed referring acts executed by life-like characters in a virtual 3D environment (cf. (CPB+94; LVTC97)). In this case, the character may refer to virtual objects in the same way as a human will do in a real environment with direct access to the objects. A different situation occurs when a character interacts with objects via their presentations as in the example scenarios above. Here, we have to explicitly distinguish between domain objects and document objects. First, there may be more than one representative for one and the same world object in a presentation. For example, in Fig. 2, DFKI is represented by a schematic drawing and a colored polygon. Furthermore, it makes a difference whether a system refers to features of an object in the domain or in the presentation since these features may conflict with each other. To enable references to objects in a presentation, we have to explicitly represent how the system has encoded information. For instance, to generate a cross-media reference as in Fig. 2, the system has to know which images are encodings for DFKI. Inspired by (Mac86), we use a relation tuple of the form:

(Encodes carrier info context-space)

to specify the semantic relationship between a presentation means, and the information the means is to convey in a certain context space (cf. (AR94)). In our approach, the third argument refers to the context space to which the encoding relation corresponds to and not to a graphical language as in the original Mackinlay approach. This enables us to use one and the same presentation means differently in different context spaces. For example, the zoom inset in Fig. 1 is used as a graphical encoding of the DFKI building in the current context space, but may serve in another context as a representative building of a certain architectual style. In addition, we not only specify encoding relations between individual objects, but also specify encoding relations on a generic level (e.g., that the property of being a red polygon on a map encodes the property of being

a built-up area in the world).

Furthermore, we have to explicitly represent the Persona's current state since it influences both the contents and the form of a referring expression. For instance, the applicability of deictic spatial expressions, such as "on my left", depends on the Persona's current position.

3 Highlevel Planning of Referential Acts

Following a speech-act theoretic perspective, we consider referring as a goal-directed activity (cf. (AK87)). The goal underlying a referring expression is to make the user activate appropriate mental representations in the sense of picking them out of a set of representations which are already available or which have to be built up (e.g., by localizing an object in a user's visual field). To plan referential acts which accomplish such goals, we build upon our previous work on multimedia presentation design (cf. (AR96)). The main idea behind this approach was to formalize action sequences for designing presentation scripts as operators of a planning system. Starting from a complex communicative goal, the planner tries to find a presentation strategy which matches this goal and generates a refinement-style plan in the form of a directed acyclic graph (DAG). This plan reflects not only the rhetorical structure, but also the temporal behavior of a presentation by means of qualitative and metric constraints. Qualitative constraints are represented in an "Allen-style" fashion (cf. (All83)) which allows for the specification of thirteen temporal relationships between two named intervals, e.g. *(Speak1 (During) Point2)*. Quantitative constraints appear as metric (in)equalities, e.g. *(5 ≤ Duration Point2)*. While the top of the presentation plan is a more or less complex presentation goal (e.g., instructing the user in switching on a device), the lowest level is formed by elementary production (e.g., to create an illustration or to encode a referring expression) and presentation acts (e.g., to display an illustration, to utter a verbal reference or to point to an object).

If the presentation planner decides that a reference to an object should be made, it selects a strategy for activating a mental representation of this object. These strategies incorporate knowledge concerning:

- *the attributes to be selected for referent disambiguation*
  To discriminate objects from alternatives, the system may refer not only to features of an object in a scene, but also to features of the presentation model, their interpretation and to the position of objects within a presentation, see also (Waz92).
- *the determination of an appropriate media combination*
  To discriminate an object against its alternatives through visual attributes, such as shape or surface, or its location, illustrations are used. Pointing gestures are planned to disambiguate or simplify a referring expression or to establish a coreferential relationship to other document parts.
- *the temporal coordination of the constituents of a referential act*
  If a referrring expression is composed of several constituents of different media, they have to be synchronized in an appropriate manner. For instance, a pointing gesture should be executed while the corresponding verbal part of the referring expression is uttered.

After the planning process is completed, the system builds up a schedule for the presentation which specifies the temporal behavior of all production and presentation acts. To accomplish this task, the system first builds up a temporal constraint network by collecting all temporal constraints on and between the actions. Some of these constraints are given by the applied plan operators. Others result from linearization constraints of the natural-language generator.

For illustration, let's assume the presentation planner has built up the following speech and pointing acts:

```
A1: (S-Speak Persona User (type pushto
      modus (def imp tense pres number sg)))
A2: (S-Speak Persona User
      (theagent (type individual
                  thediscourserole
                  (type discourserole value hearer)
                  modus
                  (def the ref pro number sg))))
A3: (S-Speak Persona User
      (theobject (type taskobject
          thetaskobject
          (type namedobject
           thename S-4
           theclass
           (type class value on-off-switch)))))
A4: (S-Speak Persona User
      (thegoal (type dest
                 thedest (type destloc value right))))
A5: (S-Point Persona
      User image-on-off-switch-1 window-3)
```

At this time decisions concerning word orderings are not yet made. The only temporal constraints

which have been set up by the planner are: *(A5 (During) A3)*. That is the Persona has to point to an object while the object's name and type is uttered verbally.

The act specifications A1 to A4 are forwarded to the natural-language generation component where grammatical encoding, linearization and inflection takes place. This component generates: "Push the on/off switch to the right". That is, during text generation we get the following additional constraints: *(A1 (meets) A3), (A3 (meets) A4)*.[1]

After collecting all constraints, the system determines the transitive closure over all qualitative constraints and computes numeric ranges over interval endpoints and their difference. Finally, a schedule is built up by resolving all disjunctions and computing a total temporal order (see (AR96)). Among other things, disjunctions may result from different correct word orderings, such as "Press the on/off switch now." versus "Now, press the on/off switch." In this case, the temporal constraint network would contain the following constraints: *(Or (S-Speak-Now (Meets) S-Speak-Press) (S-Speak-Switch (Meets) S-Speak-Now)), (S-Speak-Press (Meets) S-Speak-Switch), (S-Point (During) S-Speak-Switch)*. For these constraints, the system would build up the following schedules:

Schedule 1
1: **Start** S-Speak-Now
2: **Start** S-Speak-Press, **End** S-Speak-Now
3: **Start** S-Speak-Switch, **End** S-Speak-Press
4: **Start** S-Point
5: **End** S-Point
6: **End** S-Speak-Switch

Schedule 2
1: **Start** S-Speak-Press
2: **Start** S-Speak-Switch, **End** S-Speak-Press
3: **Start** S-Point
4: **End** S-Point
5: **Start** S-Speak-Now, **End** S-Speak-Switch
6: **End** S-Now

Since it is usually difficult to anticipate at design time the exact durations of speech acts, the system just builds up a partial schedule which reflects the ordering of the acts. This schedule is refined at presentation display time by adding new metric constraints concerning the duration of speech acts to the temporal constraint network.

4 Context-sensitive Refinement of Referential Acts

The presentation scripts generated by the presentation planner are forwarded to the Persona Server which converts them into fine-grained animations. Since the basic actions the Persona has to perform depend on its current state, complex dependencies have to be considered when creating of animation sequences. To choose among different start positions and courses of pointing gestures (see Fig. 3), we consider the following criteria:

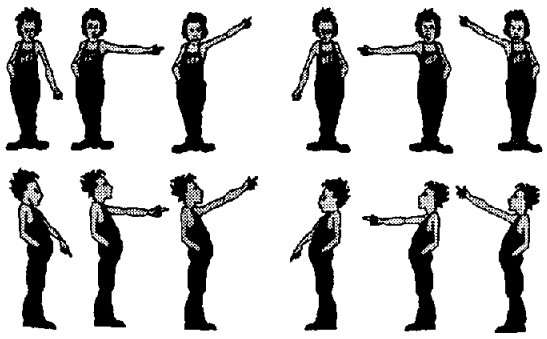

Figure 3: Different Pointing Gestures

- *the position of the Persona relative to the target object;*
  If the Persona is too far away from the target object, it has to walk to it or use a tele-scope pointing stick. In case the target object is located behind the Persona, the Persona has to turn around. To determine the direction of the pointing gesture, the system considers the orientation of the vector from the Persona to the target object. For example, if the target object is located on the right of the Persona's right foot, the Persona has to point down and to the right.

- *the set of adjacent objects and the size of the target object;*
  To avoid ambiguities and occlusions, the Persona may have to use a pointing stick. On the other hand, it may point to isolated and large objects just with a hand.

- *the current screen layout;*
  If there are regions which must not be occluded by the Persona, the Persona might not be able to move closer to the target object and may have to use a pointing stick instead.

- *the expected length of a verbal explanation that accompanies the pointing gesture;*
  If the Persona intends to provide a longer verbal explanation, it should move to the target object and turn to the user (as in the upper row in Fig. 3). In case the verbal explanation is very short, the Persona should remain stationary if possible.

[1] Note that we don't get any temporal constraints for A2 since it is not realized on the surface level.

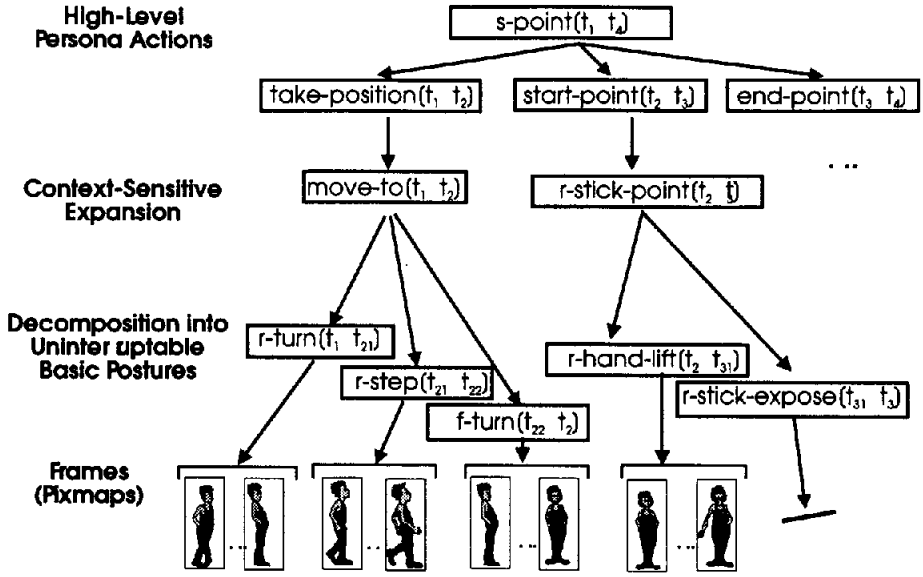

Figure 4: Context-Sensitive Decomposition of a Pointing Gesture

- *the remaining overall presentation time.* While the default strategy is to move the Persona towards the target object, time shortage will make the Persona use a pointing stick instead.

To support the definition of Persona actions, we have defined a declarative specification language and implemented a multi-pass compiler that enables the automated generation of finite-state automata from these declarations. These fine-state automata in turn are translated into efficient machine code (cf. (RAM97)).

Fig. 4 shows a context-sensitive decomposition of a pointing act delivered by the presentation planner into an animation sequence. Since in our case the object the Persona has to point to is too far away, the Persona first has to perform an navigation act before the pointing gesture may start. We associate with each action a time interval in which the action takes place. For example, the act *take-position* has to be executed during $(t_1\ t_2)$. The same applies to the *move-to* act, the specialization of *take-position*. The intervals associated with the subactions of *move-to* are subintervals of $(t_1\ t_2)$ and form a sequence. That is the Persona first has to turn to the right during $(t_1\ t_{21})$, then take some steps during $(t_{21}\ t_{22})$ and finally turn to the front during $(t_{22}\ t_2)$. Note that the exact length of all time intervals can only be determined at runtime.

5 Conclusion

In this paper, we have argued that the use of life-like characters in the interface can essentially increase the effectiveness of referrring expressions. We have presented an approach for the automated planning of referring expressions which may involve different media and dedicated body movements of the character. While content selection and media choice are performed in a proactive planning phase, the transformation of referential acts into fine-grained animation sequences is done reactively taking into account the current situation of the character at presentation runtime.

The approach presented here provides a good starting point for further extensions. Possible directions include:

- *Extending the repertoire of pointing gestures* Currently, the Persona only supports punctual pointing with a hand or a stick. In the future, we will investigate additional pointing gestures, such as encircling and underlining, by exploiting the results from the XTRA project (cf. (Rei92)).

- *Spatial deixis*
  The applicability of spatial prepositions, such as "on the left", depends on the orientation of the space which is either given by the intrinsic organization of the reference object or the location of the observer (see e.g. (Wun85)). While we assumed in our previous work on the semantics of spatial prepositions that the user's location coincides with the presenter's location (cf. (Waz92)), we now have to distinguish whether an object is localized from the user's point of view or the Persona's point of view as the situated presenter.

- *Referring to moving target objects*
  A still unsolved problem results from the dynamic nature of online presentations. Since image attributes may change at any time, the visual focus has to be updated continuously which may be very time-consuming. For instance, the Persona is currently not able to point to moving objects in an animation sequence since there is simply not enough time to determine an object's coordinates at presentation time.

- *Empirical evaluation of the Persona's pointing gestures*
  We have argued that the use a life-like character enables the realization of more effective referring expressions. To empirically validate this hypothesis, we are currently embarking on a study of the user's reference resolution processes with and without the Persona.

Acknowledgments

This work has been supported by the BMBF under the grants ITW 9400 7 and 9701 0. We would like to thank Jochen Müller for his work on the Persona server and the overall system integration.

References

D. Appelt and A. Kronfeld. A computational model of referring. In *Proc. of the 10 th IJCAI*, pages 640–647, Milan, Italy, 1987.

J. F. Allen. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(11):832–843, 1983.

E. André and T. Rist. Referring to World Objects with Text and Pictures. In *Proc. of the 15 th COLING*, volume 1, pages 530–534, Kyoto, Japan, 1994.

E. André and T. Rist. Coping with temporal constraints in multimedia presentation planning. In *Proc. of AAAI-96*, volume 1, pages 142–147, Portland, Oregon, 1996.

J. Cassell, C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression,gesture and spoken intonation for multiple conversational agents. In *Proc. of Siggraph'94*, Orlando, 1994.

J. Lester, J.L. Voerman, S.G. Towns, and C.B. Callaway. Cosmo: A life-like animated pedagogical agent with deictic believability. In *Proc. of the IJCAI-97 Workshop on Animated Interface Agents: Making them Intelligent*, Nagoya, 1997.

J. Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Transactions on Graphics*, 5(2):110–141, April 1986.

T. Rist, E. André, and J. Müller. Adding Animated Presentation Agents to the Interface. In *Proceedings of the 1997 International Conference on Intelligent User Interfaces*, pages 79–86, Orlando, Florida, 1997.

E. Reiter and R. Dale. A Fast Algorithm for the Generation of Referring Expressions. In *Proc. of the 14 th COLING*, volume 1, pages 232–238, Nantes, France, 1992.

N. Reithinger. The Performance of an Incremental Generation Component for Multi-Modal Dialog Contributions. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation: Proceedings of the 6th International Workshop on Natural Language Generation*, pages 263–276. Springer, Berlin, Heidelberg, 1992.

P. Wazinski. Generating Spatial Descriptions for Cross-Modal References. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 56–63, Trento, Italy, 1992.

D. Wunderlich. Raumkonzepte. Zur Semantik der lokalen Präpositionen. In T.T. Ballmer and R. Posener, editors, *Nach-Chomskysche Linguistik*, pages 340–351. de Gruyter, Berlin, New York, 1985.