Exploiting Image Descriptions for the Generation of Referring Expressions

Knut Hartmann*
Otto-von-Guericke-Universität Magdeburg
Institut für Informations und Kommunikationssysteme
P.O.Box 41 20, D-39016 Magdeburg, Germany
Email: hartmann@iik.cs.uni-magdeburg.de

Jochen Schöpp
software design & management GmbH & Co. KG
Thomas-Dehler-Str. 27
81737 München, Germany
Email: jochen.schoepp@sdm.de

1 Introduction

Intelligent multimedia representation systems [Bordegoni et al., 1996] require an explicit representation of the content of all multimodal expressions (in texts, graphics and animations) in order to select appropriate expressions in these different modes and to coordinate them.

In contrast to other projects like COMET [Feiner and McKeown, 1993] and WIP [André and Rist, 1994], computer graphics and animation in our approach are not generated on the fly in order to fulfill a communicative goal. In our project VisDok we select appropriate computer generated pictures and animations which suit a given communicative goal and coordinate them with generated texts. We decided us to use this approach because we feel that hand-made presentations have some advantages, as they are generated by a specialised designer, whose experience is hard to represent in formal rules\(^1\). Furthermore, we want to use computer generated graphics and animations for many different tasks.

This paper is organised as follows: in section 2 we describe how the content of a computer generated graphic is propositionally and analogous represented and propose an algorithm to compute a set of so-called characteristic objects. In

---

*the authors appear in alphabetic order

\(^1\)However, Thomas Rist describes strategies for an evaluation and partial refinement of computer generated graphics within technical documentation [Rist, 1996]
section 3 we give examples how the result of our algorithm can be applied to the generation of referring expressions and the creation of a graphical context stack. In section 4 we compare our algorithm with other proposed algorithms and interpret it in another framework. Section 5 and 6 discuss our results and propose further enhancements.

2 Description of the content of pictures

The content of a picture is described by encoding the visible objects in the KL-ONE [Brachman and Schmolze, 1985] like the knowledge representation formalism LOOM. This structure is referred to as the image description.

But how to represent the content of the picture in an efficient way? We decided to represent as few encoding information as possible, and to use LOOM's powerful inference mechanism. On the other hand, attributes of objects like their size, colour and the relative position of an object with respect to other objects can be gained by inference processes in other knowledge sources as the geometric model and the illumination model.

For reasons of simplicity, we enclose all objects in a picture with a separate bounding box. Those components of a complex object that are located on the surface of the "virtual" bounding box are especially marked. We represent this information within the is-located-on relation. Humans typically refer to some sides of the bounding box with lexemes like frontside, bottomside, topside etc. These lexems refer to directions within a system of coordinates, with 2 possible origins, either within the object itself or within the addressee of the generated presentation. Depending on the chosen origin of the coordinates, there are two possible interpretations, the intrinsic (with the origin within the object) and the deictic (where the origin lies within the addressee). In the presented work we favour the intrinsic interpretation.

Since we want the addressee of the multimodal presentation to identify objects and its intrinsic sides, we labeled the sides of the virtual bounding box with associated intrinsic sides. However, it is not feasible to assign each side of the bounding box an intrinsic side, as there is no unique relation of the sides of the virtual bounding box and an intrinsic side of the object. So we allow several sides of the virtual bounding box to be associated with an intrinsic side in the extreme case of a ball we label all virtual sides with the same intrinsic side.

Intrinsic sides provide us with a means to identify named sides of the objects, but how to refer to them? An intrinsic side of an object can be characterised by a combination of components unique to that side. When looking at a picture like the one in figure 1, humans can easily tell which sides of the presented object are visible and which sides are hidden by identifying exactly this characteristic combination of objects. Take, for instance, the front side of the toaster depicted in figure 1: This side can be identified unambiguously and hence be referred to
as "front" in the subsequent interactive discourse, because the user can identify control devices like the roast intensity selector or the elevating pushbutton. As another example, the top side of our toaster can be identified when recognising the bread slot or the mounted rolls support. We call a combination of objects that enable humans who are looking at a picture to identify an (intrinsic or deictic) side of the depicted object the characteristic objects of this (intrinsic or deictic) side.

![Diagram of a toaster with labelled components](image)

**Figure 1:** A complex object with some labelled components (Instructions for Use of the Toaster Siemens TT 621)

In the following, we assume that all objects are identifiable and distinguishable, that means that the colour of objects differs from their background, the illumination is perfect, etc. Given this assumption we can define a straightforward procedure to compute the characteristic objects. First, we give a formal criterion which the characteristic objects of a given (intrinsic or deictic) side have to fulfill.

Let \( \mathcal{O}_{s_i} \) be the set of objects located on the side \( i \). Let \( s \) be the side for which we want to compute the set of characteristic objects in relation to a set of other sides \( S \). In figure 2 we state a criterion that formally describes our characteristic object definition.

In our simple model we assume that one cannot distinguish instances of the same class, because we do not have explicit information about their colour and location. This is motivated by our strategy to represent only those information which are not represented otherwise, what prevents us from using this attributes.

---

2This set can be computed by the union of all objects on virtual sides of the bounding box which are associated with the same intrinsic side \( s_i \).
In case of location and colour, these information are implicitly represented in the geometric model and the illumination model and it is time-consuming to retrieve it. So we define the relation undistinguishably\((o_1, o_2)\) to be true iff the objects \(o_1\) and \(o_2\) belong to the same direct superconcept and false otherwise. We will show in section 3 how information about colour and location can be exploited using the results of our algorithm. Note, that if one can distinguish all objects \(o_1\) and \(o_2\), every object is a characteristic object for the side, on which it is located.

The basic idea of this definition is to state a criterion that ensures that the set \(C\) is a distinctive characterization under the equivalence relation \(\equiv\)undistinguishably. An algorithm that computes this set is shown in figure 3.

There can be no, one or several sets of characteristic objects for a given object \(o\), side \(s\) and \(S\). We can heighten our definition by adding a minimality condition.

Let’s look at a simple example. With \(s_i\) we denote the intrinsic sides of the object and with \(S\) the set of all the intrinsic sides \(s_1, \ldots, s_6\). With \(a_j, b_j, c_j, d_j,\) and \(e_j\) we denote objects which are located on these sides and which are instances of the concept \(A, B, C, D\) and \(E\).

If we apply the algorithm of figure 3 to the example given in figure 4, the set of characteristic objects of the side \(s_5\) is \(\{\{d_5\}, \{e_5\}, \{d_5, e_5\}\}\). That means that the addressee can identify the side \(s_5\) when recognising an instance of the concept \(D\) or an instance of the concept \(E\). In this case there exist two minimal sets. The set of characteristic objects of side \(s_6\) is \(\{\{a_6, b_6, c_6\}\}\), which means that the side \(s_6\) can be identified only when recognising an instance of the concept \(A\) or \(B\) and \(C\). The addressee has to identify a instance of all these concepts, as combinations of instances of two of these concepts can be found on the sides \(s_1, s_2, s_3\) and \(s_4\). The sides \(s_1, s_2, s_3\) and \(s_4\) cannot be identified by exploiting the knowledge which objects are located on these sides, as instances of the concept \(A\), \(B\) and \(C\) are located on side \(s_6\).

Using this model, we have a simple formalism to describe the visible objects within a picture, by describing the picture with an enumeration of the visible sides of the object’s bounding box. Together with the information, which parts
are located on which side of the bounding box (represented in the \textit{is-located-on-Relation}), we can reason about the visible objects and about the characteristic objects of the intrinsic sides. If we use the transitivity of the part-of relation, our algorithm becomes recursive. However, there is a problem with the transitivity of the part-of relation, as parts can be so small, that one cannot recognise them on the screen. This means that the part-of relation should only be transitive for visible objects.

3 Generation of Referring Expressions

Analyses of the interaction of text and pictures have shown that pictures provide a new context in the discourse by introducing visible objects as potential referents. Pictures frequently display a number of objects which are not mentioned in the associated text, as one rhetoric function of a picture is to provide the background in which a referring expression of the text can be interpreted [André and Rist, 1994]. The other objects depicted in the picture can serve as possible referents in the following discourse, so these objects become part of the potential focus [Grosz and Sidner, 1986].

Our algorithm simulates the recognition process of humans when learning the attributes of the intrinsic sides of a depicted object. How can we apply this knowl-

\[
\begin{align*}
\mathcal{C}(o, s, \mathcal{S}) & \quad \mathcal{O}_s := \{ p \mid \text{is-located-on}(p, s) \} \\
\mathcal{C} := \emptyset \\
\text{Candidates} := \text{PowerSet}(\mathcal{O}_s) \\
\text{while } (\text{Candidates} \neq \emptyset) \text{ do} \\
\quad \text{Candidate} := \text{member}(\text{Candidates}) \\
\quad \text{check} := \text{false} \\
\quad \text{for } s_i \text{ in } \mathcal{S} \text{ do} \\
\quad \quad \mathcal{O}_{s_i} := \{ p \mid \text{is-located-on}(p, s_i) \} \\
\quad \quad \text{if } (\text{Candidate} /\equiv \text{undistinguishably} \subseteq \mathcal{O}_{s_i} /\equiv \text{undistinguishably}) \\
\quad \quad \quad \text{then } \text{check} := \text{true} \\
\quad \quad \text{fi} \\
\quad \text{od} \\
\quad \text{if } (\text{check} = \text{false}) \text{ then } \mathcal{C} := \mathcal{C} \cup \text{Candidate} \text{ fi} \\
\text{Candidates} := \text{Candidates} \setminus \{ \text{Candidate} \} \\
\text{od} \\
\text{return } \mathcal{C}
\end{align*}
\]
Figure 4: A complex object example, column one denotes the intrinsic sides of the object, the second column displays the range of the is-located-on-relation and the third column depicts the result of our algorithm.

| side components | $c(a, s_i, S \setminus \{s_i\})$ |
|-----------------|----------------------------------|
| $s_1$           | $a_1, b_1$                       |
|                 | $\emptyset$                      |
|                 | $\{a_1, a_2, a_6\} \subseteq A$ |
| $s_2$           | $a_2, e_2$                       |
|                 | $\emptyset$                      |
|                 | $\{b_1, b_3, b_6\} \subseteq B$ |
| $s_3$           | $b_3, c_3$                       |
|                 | $\emptyset$                      |
|                 | $\{c_3, c_4, c_6\} \subseteq C$ |
| $s_4$           | $c_4$                            |
|                 | $\emptyset$                      |
|                 | $\{d_5\} \subseteq D$          |
| $s_5$           | $d_5, e_5$                       |
|                 | $\{\{d_5\}, \{e_5\}, \{d_5, e_5\}\}$ |
|                 | $\{e_5\} \subseteq E$          |
| $s_6$           | $a_6, b_6, c_6$                  |
|                 | $\{\{d_5\}, \{e_5\}, \{d_5, e_5\}\}$ |

dege in generating referring expressions and calculating the set of potential referents? Dale & Reiter assume that "a referring expression contains two kinds of information: navigation and discrimination. Each descriptor used in a referring expression plays one of these two roles. Navigational, or attention-directing information, is intended to bring the referent in the hearer’s focus of attention [while] discriminating information is intended to distinguish the intended referent from other objects in the hearer’s focus of attention" [Dale and Reiter, 1995]. In the following we show how the results of our algorithm can be used to provide navigational and discriminating information for the intended referent, especially for the components of a complex object.

Using the information gained by our characteristic object algorithm, one can decide whether navigational information has to be included in constructing the referring expression or not and from which object the intended referent has to be distinguished. Assume that the system has decided to present a picture like the one in figure 1 on which the frontside, the topside and the left side of a toaster are shown. The elevating push button and the roast intensity selector represent both a unary set of characteristic objects for the front side. Hence, one can refer unambiguously to these objects in an associated text, because the addressee can unambiguously distinguish these objects from all other objects on the frontside, the topside and the left side of the depicted toaster (the other objects in the hearer’s focus of attention) and no additional navigational or discriminating descriptions is needed.

Assume that an object is not included in a unary set of characteristic objects, but that it is included as an element in the characteristic object set. Figure 5 shows a clipping of an iron with two buttons on the top side. The example is taken from the doctoral thesis of Elisabeth André [André, 1995, page 80], a member of the WIP project. Both buttons together represent characteristic objects for the top side of the depicted electric iron. In this case, we have to provide further descriptors to enable the addressee to distinguish the spray button from the
other depicted button on the top side. Thus, we have to search for perceptually recognisable attributes of the spray button like its colour, shape or its relative location to the other button depicted. Generally, if the intended referent $r$ is an element of a no-unary set of characteristic objects $C$ and if $C'$ is the set difference of $C$ and $\{r\}$, we have to compute a set of discriminating attributes for the intended referent with respect to $C'$ using the colour and shape or the relative location with respect to the other objects in $C'$ by retrieving this information from the illumination model or the geometric model in order to enable the addressee of the presentation to identify the intended referent. As the colour and the shape of both buttons do not differ, we have to exploit the information about the relative location, which enables us to generate a sentence like "Press left button, which is the spray button". This establishes a coreferential connection between the referent of the "the spray button" nominal phrase and the left button on the top side, which can be exploited in the subsequent dialogue.  

For objects that are not elements of a characteristic object set navigational information (the side on which they are located) has to be included. In addition, we have to provide discriminatory descriptions for the intended referent that distinguishes the intended referent from all the other objects which are located on this side. This set of attributes can be computed by a traditional reference algorithm. In example 4, for instance, if the system wants in to refer to the component $a_1$ of side $s_1$, it would insert the name of the side $s_1$ as navigational information and the set of attributes which discriminates $a_1$ from $b_1$ as discriminative information.

\footnote{André and Rist proposed to augment the depicted graphic with an arrow in order to establish this coreference}
Furthermore, we propose to establish a new graphic context, which includes the members of the characteristic objects set of the visible sides (provided that the side of the object is visible as a whole) as actual referents and the other visible objects as potential referents.

4 Related work

In previous work algorithms to generate referring expressions were proposed [Dale and Reiter, 1995], [Horacek, 1996]. The main goal of these algorithms is to compute a referring expression for a given object, which enables the hearer to distinguish it from all other objects in the hearer’s focus of attention [Dale and Reiter, 1995]. The basic terms of Dale & Reiter’s framework are the referent and the contrast set, that are all the other objects in the hearer’s focus of attention. Each object is associated with a number of attribute-value pairs. The attributes include the type of the object and its perceptually accessible properties like colour, shape and size. The algorithm computes a subset of the referent’s attributes, with the property that there is no other object in the contrast set that possesses a combination of the referent’s attributes. Dale and Reiter proposed a number of algorithms that differ in their computational complexity. As the task to find the minimal set of attribute-value pairs with this property is NP-hard, a number of heuristics are used, which approximate the minimal set.

The algorithm proposed in this paper can be interpreted in Dale & Reiter’s framework. The aim of our algorithm is to compute all sets of attributes which enable the hearer to identify a side of an object (the intended referent in Dale & Reiter’s terms) of a given object in contrast to the set of other object sides (the contrast set in Dale & Reiter’s terms). In contrast to Dale & Reiter who compute at most one set of attributes, which distinguishes the referent from all other objects in the contrast set, our algorithm computes all the (minimal) attribute sets. The set of the attribute sets is called the characteristic objects of the intended side. Another contrasting feature is that Dale & Reiter use a number of single-valued attributes like type, size and colour, whereas our algorithm uses one, multiple-valued relation is-located-on.

5 Discussion

This work incorporates propositional and analogue representation as proposed by [Habel et al., 1995]. Within the VisDok-project we decided to combine the geometric information of the 3-D-Model and information about the colour, size and shape of the object which can be gained from the illumination model with

---

4The problem can be transformed into the problem to find the minimal size set cover, which is proven to be NP-hard [Garey and Johnson, 1979].
a propositional representation in a KL-ONE like knowledge base. As the blocks in the 3-D-Model are labeled with identifiers which correspond to instances in the knowledge base, both representation types can be combined. On the other hand, it is time-consuming to retrieve the perceptually recognisable attributes of objects like colour, shape and relative location, as they have to be gained by retrieving them within the 3-D-Model or the illumination model. The algorithm presented in this work factors out this kind of information employing the undistinguishably-relation and uses it only for constructing discriminating descriptors for few components of a side.

The computation of the characteristic object set enables us to reduce the contrast set in the algorithm of Dale & Reiter, that is the set of objects, for which we have to know the perceptually recognisable attributes of objects like colour, shape and relative location. On the other hand, the algorithm itself is far more expensive, because we calculate all minimal distinguishing descriptions using only the type attribute to speak in Dale & Reiter's terms. However, our algorithm enables us to minimise the time-consuming communication between separate processes in order to gain the perceptually recognisable attributes of objects like colour, shape and relative location while employing the type attribute, which is explicitly represented within the knowledge base.

6 Outlook

We plan to combine the approach presented in this work with the results of the Hyper-Renderer [Emhardt and Strothotte, 1992], which stores information about visible objects and their texture. These information are computed as a side effect of the rendering algorithm and can be used in our framework. Especially for complex objects, the is-located-on relation can be computed automatically and serves as the input data for our algorithm.

The result of this work can be exploited in an intelligent multimedia representation system like VisDok to select the content within the different media, to coordinate the media and to generate referring expressions. In contrast to other work like IDAS [Reiter et al., 1995] and KAMP [Appelt, 1985], the navigational part of the referring expression can exploit other sources then the part-whole relation or the spatial including relation.

7 Acknowledgement

The authors want to thank Brigitte Grote for all the discussions about the basic ideas presented in this paper and a careful reading of the first version of this paper. Thanks also to Oliver Brandt and Björn Höfling, who gave hints how our algorithm can be implemented.
References

[André and Rist, 1994] Elisabeth André and Thomas Rist. Referring to World Objects with Text and Pictures. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), Kyoto, Japan, 1994.

[André, 1995] Elisabeth André. Ein planbasierter Ansatz zur Generierung multimodaler Präsentationen. infix Verlag, 1995.

[Appelt, 1985] Douglas E. Appelt. Planning English Sentences. Cambridge University Press, Cambridge, UK, 1985.

[Bordegoni et al., 1996] M. Bordegoni, G. Faconti, T. Rist, S. Ruggieri, P. Trahanias, and M. Wilson. Intelligent multimedia presentation systems: A proposal for a reference model. In J.-P. Courtiat, M. Diaz, and P. Sénac, editors, Multimedia Modeling: Towards the Information Superhighway, pages 3–20. World Scientific, Singapore, 1996.

[Brachman and Schmolze, 1985] Ronald J. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. Cognitive Science, 9(2), 1985.

[Dale and Reiter, 1995] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19(2):233–263, 1995.

[Emhardt and Strothotte, 1992] Jürgen Emhardt and Thomas Strothotte. Hyper Rendering. In Proc. Graphics Interfaces ’92, pages 37–43, Vancouver, Canada, May 13–15 1992.

[Feiner and McKeown, 1993] Steve K. Feiner and Kathleen R. McKeown. Automating the generation of coordinated multimedia explanations. In M. T. Maybury, editor, Intelligent Multimedia Interfaces, pages 117–138. AAAI Press, Menlo Park, CA, 1993.

[Garey and Johnson, 1979] W. Garey and D. Johnson. Computers and Intractability: a Guide to the Theory of NP-Completeness. W. H. Freeman, San Francisco, 1979.

[Grosz and Sidner, 1986] Barbara J. Grosz and Candace L. Sidner. Attention, intentions and the structure of discourse. Computational Linguistics Journal, 12(3):175–204, 1986.

[Habel et al., 1995] Christopher Habel, Simone Pribbenow, and Geoffrey Simmons. Partonomies and depictions: A hybrid approach. In B. Chandrasekaran J. Glasgow, H. Narayanan, editor, Diagrammatic Reasoning: Computational and Cognitive Perspectives. AAAI/MIT Press, 1995.
[Horacek, 1996] Helmut Horacek. A new algorithm for generating referential descriptions. In Wolfgang Wahlster, editor, *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI’96)*, pages 577–581, Budapest, Hungary, August 11–19 1996. John Wiley & Sons LTD., Chichester, New York, Brisbane, Toronto, Singapore.

[Reiter et al., 1995] Ehud Reiter, Chris Mellish, and John Levine. Automatic generation of technical documentation. *Applied Artificial Intelligence*, 9, 1995.

[Rist, 1996] Thomas Rist. *Wissensbasierte Verfahren für den automatischen Entwurf von Gebrauchsgraphik in der technischen Dokumentation*. infix Verlag, 1996.