Exploiting Visual Salience for the Generation of Referring Expressions

John Kelleher
Media Lab Europe,
Sugar House Lane,
Dublin 8, Ireland.
john.kelleher@medialabeurope.org

Josef van Genabith
National Centre for Language Technology,
School of Computing,
Dublin City University,
Dublin 9, Ireland.

Abstract

In this paper we present a novel approach to generating referring expressions (GRE) that is tailored to a model of the visual context the user is attending to. The approach integrates a new computational model of visual salience in simulated 3-D environments with Dale and Reiter's (1995) Incremental Algorithm. The advantage of our GRE framework are: (1) the context set used by the GRE algorithm is dynamically computed by the visual saliency algorithm as a user navigates through a simulation; (2) the integration of visual salience into the generation process means that in some instances underspecified but sufficiently detailed descriptions of the target object are generated that are shorter than those generated by GRE algorithms which focus purely on adjectival and type attributes; (3) the integration of visual saliency into the generation process means that our GRE algorithm will in some instances succeed in generating a description of the target object in situations where GRE algorithms which focus purely on adjectival and type attributes fail.

Introduction

The focus of the Linguistic Interaction with Virtual Environments (LIVE) project is to develop a natural language framework to underpin natural language virtual reality (NLVR) systems. An NLVR system is a computer system that allows a user to interact with simulated 3-D environments through a natural language interface. The central tenet of this work is that the interpretation and generation of natural language (NL) in 3-D simulated environments should be based on a model of the user's knowledge of the environment. In the context of an NLVR system, one of the user's primary information sources is the visual context supplied by the 3-D simulation. In order to model the flow of information to the user from the visual context, we have developed and implemented a visual saliency algorithm that works in real-time and across different simulated environments. This paper describes the visual saliency algorithm and illustrates how it is used to underpin the LIVE generation of referring expressions (GRE) algorithm, which is tailored to ground the GRE process in a model of the visual context the user is attending to.

First, we discuss the distributational properties of visual perception. Second, we review work on computationally modelling visual salience. Third, we present the LIVE visual saliency algorithm. Fourth, we review previous work on generating referring expressions. Fifth, we show how the LIVE visual saliency algorithm can be integrated with Dale and Reiter's (1995) GRE algorithm enabling the generation of underspecified yet sufficiently detailed references. Sixth, we present preliminary testing of the new algorithm. Finally, we conclude and outline future work.

Perception and Attention

Although visual perception seems effortless, "psychophysical experiments show that the brain is severely limited in the amount of visual information it can process at any moment in time" (Reynolds 2001, pg. 1). In effect, there is more information perceived than can be processed.

The human faculty of attention is the "selective aspect of processing" (Kosslyn 1994, pg. 84). Attention regulates the processing of perceived visual stimuli by selecting a region within the visual buffer for detailed processing. Our knowledge of the human attention process is not complete, "but it appears to consist of a set of mechanisms that exhibit different, sometimes opposing effects" (Hewett 2001, pg. 9). For example, (Landragin, Bellalem, & Romary 2001) lists: visual familiarity, intentionality, an object's physical characteristics, and the structure of the scene. This multiplicity makes the modelling of visual perception difficult.

A priori, one of the major functions of visual attention is object identification. With this in mind, an important factor when considering modelling visual attention is the difference between foveal and peripheral vision. The fovea is a shallow pit in the retina which is located directly opposite the pupil, consisting of cones and is the site of highest visual acuity, the ability to recognise detail. It "drops 50 percent when an object is located only 1° from the centre of the fovea and an additional 35 percent when it is 8° from the centre" (Forgus & Melamed 1976, pg. 228). Identifying an object requires the use of foveal vision, occurring when a person looks directly at the object, causing the image of the object falling on the retina to be centred on the fovea. The dependence of object identification on foveal vision implies a relationship between foveal vision and attention. Moreover, the gradation of visual acuity is congruent with the

gradation of attention theory. This theory posits that "attention is greatest at a single point, and drops off gradually from that point" (Kosslyn 1994, pg. 90).

Following this, the more central a location is with respect to the centre of an eye fixation the higher the location's salience. Indeed, the most common computational mechanism for modelling visual attention is a filtering of visual data by removing portions of the input located outside a spatial focus of attention (Hewett 2001).

Computationally Modelling Visual Salience Previous Work

Many computational models of vision use spatial attention as a visual filtering mechanism; most have been developed for robot navigation (for a review see (Hewett 2001)).

However, there are two reasons why the models of vision created for robotic systems are not suitable for NLVR systems. First, nearly all of these systems have a connectionist or neural net architecture. This form of system requires training. As a result, these models are restricted to the domains described by or sufficiently similar to the training set given to the system. A system that requires retraining when shifting from one visual domain to another is not suitable as a model of rendered environments which may change drastically from program to program or within the one application. Second, the major difficulties facing robotic vision (pattern recognition, distance detection, and the binding problem (Renault et al. 1990)) do not impact on NLVR systems because the visual scene is already modeled.

There have been several models of visual perception developed that use 3-D graphics techniques. These models can be classified based on the graphics techniques they use: ray casting and false colouring. (Tu & Terzopoulos 1994a; 1994b) implemented a realistic virtual marine world inhabited by autonomous artificial fish. The model used ray casting to determine if an object met the visibility conditions. Ray casting can be functionally described as drawing an invisible line from one point in a 3-D simulation in a certain direction, and then reporting back all the 3-D object meshes this line intersected and the coordinates of these intersections. It is widely used in offline rendering of graphics; however, it is computationally expensive and for this reason is not used in real-time rendering.

Another graphics based approach to modelling vision was proposed in (Noser *et al.* 1995). This model was used as a navigation system for animated characters. The vision module comprised a modified version of the world being fed into the system's graphics engine and scanning the resulting image. In brief, each object in the world is assigned a unique colour or "vision-id" (Noser *et al.* 1995, pg. 149). This colour differs from the normal colours used to render the object in the world; hence the term false colouring. An object's false colour is only used when rendering the object in the visibility image off-screen, and does not affect the renderings of the object seen by the user, which may be multicoloured and fully textured. Then, at a specified time interval, a model of the character's view of the world, using the false colours, is rendered. Once this rendering is finished, the viewport[1] is copied into a 2-D array along with the z-buffer[2] values. By scanning the array and extracting the pixel colour information, a list of the objects currently visible to the actor can be obtained. (Noser *et al.* 1995) used this vision model as part of a navigation system for animated characters. Another navigation behavioral system that used false colouring synthetic vision was proposed by (Kuffner & Latombe 1999). (Peter & O'Sullivan 2002) also used a false-colouring approach to modelling vision. They integrated their vision model as part of a goal driven memory and attention model which directed the gaze of autonomous virtual humans.

The LIVE Visual Saliency Algorithm

The basic assumption underpinning the LIVE visual saliency algorithm is that an object's prominence in a scene is dependent on both its centrality within the scene and its size. The algorithm is based on the false colouring approach introduced in the previous section. Each object is assigned a unique ID. In the current implementation, the ID number given to an object is simply 1 + the number of elements in the world when the object is created. A colour table is initialised to represent a one-to-one mapping between object IDs and colours. Each frame is rendered twice: firstly using the objects' normal colours, textures and normal shading. This is the version that the user sees. The second rendering is off-screen. This rendering uses the unique false colours for each object and flat shading. The size of the second rendering does not need to match the first. Indeed, scaling the image down increases the speed of the algorithm as it reduces the number of pixels that are scanned. In the LIVE system the false colour rendering is 200 x 150 pixels, a size that yields sufficient detail. After each frame is rendered, a bitmap image of the false colour rendering is created. The bitmap image is then scanned and the visual salience information extracted.

To model the size and centrality of the objects in the scene, the LIVE system assigns a weighting to each pixel using Equation 1. In this equation, P equals the distance between the pixel being weighted and the centre of the image, and M equals the maximum distance between the centre of the image and the point on the border of the image furthest from the centre; i.e., in a rectangular or square image, M is equal to the distance between the centre of the image and one of the corners of the image.

$$Weighting = 1 - \left(\frac{P}{M+1}\right) \quad (1)$$

This equation normalises the pixel weightings between 0 and 1. The closer a pixel is to the centre of the image, the higher its weighting. After weighting the pixels, the LIVE system scans the image and, for each object in the scene, sums the weightings of all pixels that are coloured using that

[1] A viewport is the rectangular area of the display window. It can be conceptualised as a window onto the 3-D simulation.

[2] The z-buffer stores for each pixel in the viewport the depth value of the object rendered at that pixel

object's unique colour. This algorithm ascribes larger objects a higher saliency than smaller objects since they cover more pixels and objects which are more central to the view will be rated higher than objects at the periphery of the scene as the pixels the former cover will have a higher weighting. This simple algorithm results in a list of the currently visible objects, each with an associated saliency rating.

It is important to note that the scanning process in the LIVE visual salience algorithm differs from those in the previous false colour synthetic vision models (Noser *et al.* 1995; Kuffner & Latombe 1999; Peter & O'Sullivan 2002). The previous false colouring algorithms simply recorded whether the object had been rendered or not. The LIVE algorithm records whether an object has been rendered and ascribes each object a relative prominence within the scene. What is more, the LIVE algorithm naturally accounts for the effects of (partial) object occlusion. It is this difference that allows the LIVE system to rank the objects based on their visual salience. We do not claim that this algorithm accommodates all the perceptual factors that affect visual salience. However, it does define a reasonable model of visual saliency that operates fast enough for real-time systems.

GRE Previous Work

GRE is an essential component of natural language generation. GRE focuses on the semantic questions involving the factual content of the description, and does not concern itself with the linguistic realisation of the description. There have been many GRE algorithms proposed (Appelt 1985; Dale 1992; Dale & Reiter 1995; Krahmer & Theune 2002; van Deemter 2002). Most of these algorithms deal with the same problem definition: given a single target object, for which a description is to be generated, and a set of distractor objects, from which the target object is to be distinguished, determine which set of properties is needed to single out the target object from the distracters. On the basis of these properties a distinguishing description of the target object can be generated; i.e., a distinguishing description is a description of the target object that excludes all the elements of the distractor set.

The current state of the art for GRE is the Incremental Algorithm (Dale & Reiter 1995), with most later algorithms extending this. The Incremental Algorithm "sequentially iterates through a (task-dependent) list of attributes, adding an attribute to the description being constructed if it rules out any distractors that have not already been ruled out, and terminating when a distinguishing description has been constructed" (Dale & Reiter 1995, pg. 247). If the end of the list of attributes is reached before a distinguishing description has been generated the algorithm fails. It should be noted that in the Incremental Algorithm the target object's type is always included in the generated description even if it has no distinguishing value.

The ouput of the Incremental Algorithm is, to a large extent, determined by the context set it uses. However, Dale and Reiter do not define how this set should be constructed, they only write: "[w]e define the context set to be the set of entities that the hearer is currently assumed to be attending to" (Dale & Reiter 1995, pg. 236). *A priori*, there is a domain of discourse D, the total set of entities that can be referred to. However, always using D as the context set can result in the Incremental Algorithm generating longer descriptions than are necessary: depending on the linguistic and or perceptual context a reduced description may suffice and may be more natural to the discourse.

Theune (2000) Chapter 4[3] discusses whether restricting the context set to a proper subset of D containing those entities of D that have been referred to before would enable the Incremental Algorithm to generate reduced anaphoric descriptions. She concludes that restricting the context set in this way has an unwanted consequence: "the descriptions of all domain entities will be made relative to this restricted set" (2000, pg. 106); as a result, the descriptions generated for new entities entering the discourse by the algorithm are not sufficiently detailed to distinguish the target object from the other entities in the domain that are not in the context set. Theune's solution is to structure D by marking certain entities as more linguistically prominent than others. This is achieved using a framework for modelling linguistic salience that is a synthesis of the hierarchical focusing constraints of Hajicová (1993) and the constraints of Centering Theory (Grosz, Joshi, & Weinstein 1995). Essentially, in Theune's extension to the Incremental Algorithm, "an entity that is being newly introduced into the discourse must be distinguished from all other entities in the domain, whereas an entity that has been previously mentioned can have a reduced description"(2000, pg. 106). The underlying idea of Theune's extension is to modify the definition of a distinguishing description as:

> "A definite description 'the N' is a suitable description of an object d in a state s iff d is the most salient object with the property expressed by N in state s." (2000, pg. 101)

Theune's structuring of D focuses on linguistic salience and enables the Incremental Algorithm to generate reduced *anaphoric* references. Similar to (Theune 2000) we use a saliency measure to restructure D. However, our saliency model is based on visual rather than linguistic salience. Consequently, our modified algorithm can generate *undetermined*[4] *exophoric*[5] references.

Integrating Visual Salience and GRE

It has been shown in psycholinguistic experiments that subjects can easily resolve ambiguous or underdetermined references: "In order to identity the intended referent under these circumstances [where there is more than one entity in the discourse domain whose properties fulfil the linguistic description of the referent], subjects rely on perceptual salience" (Duwe & Strohner 1997). We have developed a version of the Incremental Algorithm that uses the LIVE visual saliency algorithm and exploits subjects's abilities to

[3] See also (Krahmer & Theune 2002).

[4] An underdetermined or underspecified reference is a reference that breaks the singularity constraint: i.e., there is more than one candidate referent.

[5] An exphoric reference denotes an entity in the spatio-temporal surroundings that is new to the discourse (Byron 2003).

H1	Green	1.0000
H2	Red	0.8646
H3	Yellow	0.0235
H4	Red	0.0111
H5	Brown	0.0149

Table 1: The LIVE system's analysis of Figure 2 listing the object ID's color attribute's and visual saliency score's.

resolve underdetermined references given a visual context. This enables us to generate underdetermined yet sufficiently detailed descriptions. We integrate our visual salience algorithm with Dale and Reiter's Incremental Algorithm in two steps.

Firstly, the output of the LIVE visual saliency algorithm is used to create the context set used by the GRE algorithm. For each scene rendered, the LIVE visual saliency algorithm produces a list of the objects in the scene each with a visual saliency score and a set of attributes. This excludes all the objects in the world that are not currently visible and improves the ability of the algorithm to generate relevant references. A further advantage is that the objects in the context set can be ordered using their visual saliency scores.

Secondly, similar to (Theune 2000), we modify the definition of a distinguishing description to exploit the visual saliency scores associated with each object in the context set. However, our definition of a distinguishing description, unlike (Theune 2000), requires that the target object should not only be the most salient object fitting the generated description in the scene but its salience should exceed the salience of the elements in the distractor set that fulfil the description by a predefined confidence interval; i.e., *a description is distinguishing if it excludes all the distractors that have a visual salience greater than the target object's salience minus a predefined confidence interval*. The motivation for this definition is that a small difference in visual salience is not normally sufficient to resolve underspecified references. Based on preliminary testing, reported in the next section, we have set the LIVE system's confidence interval to 0.6. Of course, this interval can be adjusted to model a stricter or looser interpretation.

When using the LIVE system, the user specifies the target object by left-clicking on it using the mouse. Once the target object has been specified the system uses the algorithm listed in Figure 1 to generate a description of the object.

Figure 2 illustrates a scene from the LIVE system and Table 1 lists the LIVE system's visual saliency ranking of the objects in the scene. Note that in this scene all the objects have the same absolute height, width and depth values It is the effect of perceiving the scene from a particular point of view as the user navigates through the simulation that is captured by the visual saliency algorithm.

Given this context, assuming the user selects H2 as the target object, the system would begin trying to generate a description using the target type attribute: **house**. None of the objects in the scene are excluded by this description. Moreover, the target object's visual saliency score does not ex-

Input: The context set containing the target object and the distractor set for the current scene, as computed by the LIVE GRE visual saliency algorithm. Each element in the context set has a visual salience scores ascribed to it.

Output: A description of the target object that excludes all the elements of the distractor set that have a visual salience score greater than the target object's visual salience minus a predefined confidence interval.

1. Sequentially iterate through the list of preferred attributes <type, colour, tall, short, wide, narrow, deep, shallow>.
2. For each attribute create a set containing the objects that have that attribute plus all the previously accepted attributes.
3. If the number of elements in this set is less than the number of elements in the set created using the previously accepted attributes add the current attribute to the list of accepted attributes.
4. Terminate when the visual salience scores associated with the target object exceeds the visual salience scores associated with the other objects in the newly created set by the predefined confidence interval.
5. Always include the target object's type in the set of accepted attributes.

Figure 1: The LIVE GRE Algorithm

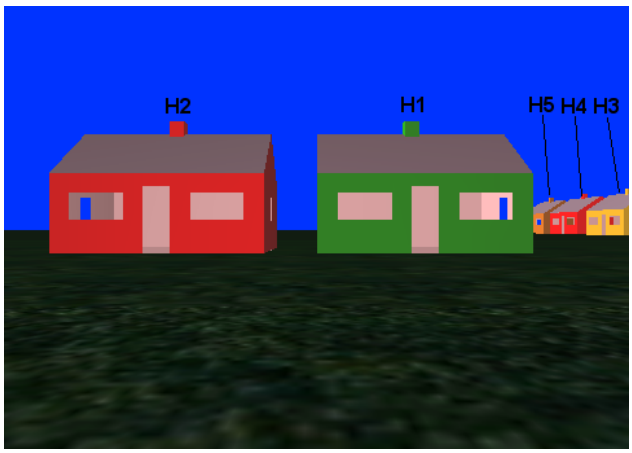

Figure 2: The Visual Context

ceed all the scores associated with the elements in this set by the predefined confidence interval. Indeed, H1's visual saliency score exceeds that of the target object. As the first attribute in the preferred attribute list does not result in a distinguishing description the system tries to generate a description by using the first and second attributes in the list: type and colour. This results in the set of objects fulfilling the description **red house**. Only two objects in the scene fulfil this description: the target object H2 and one of the distractor objects H4. Furthermore, H2's visual saliency rating exceeds H4's by more than the predefined confidence interval (0.6). Consequently, the description **red house** is deemed to be a distinguishing description and the generation algorithm terminates. It should be noted that in this instance Dale and Reiter's (1995) algorithm would fail to generate a description as it would not have been able to distinguish between H2 and H4 using only adjectival and type attributes.

Testing

We have carried out preliminary testing of our algorithm. This involved generating images of visual scenes from the LIVE system and recording the visual saliency values ascribed to each object in each scene by the system. After this, for each image a caption (of the form "*the [adj] N*") intended to denote one of the objects in the scene was decided on. Some of the object descriptions were underspecified, some were not. The underspecified descriptions were designed to examine what would be a good confidence interval; i.e., the difference between the target object's visual saliency rating and that of the next most salient object not excluded by linguistic description was varied. Ten subjects took part in the testing. Each subject was given a randomly selected set of images and asked to either mark which object in the scene was described by the caption or to tick a box if they felt the description was ambiguous. Table 2 lists the results. Column 1 lists the interval between the most salient object in the scene matching the description in the caption and the next most salient object matching the description. Column 2 lists the number of instances where the subject selected the most salient object, as rated by the system, as the referent for the description. There was no case where a subject selected an object other than the most salient object as the referent. Column 3 lists the number of times subjects found an image-caption pairing ambiguous.

Although the sample size of the test was relatively small, the general trend in the results does support the hypothesis that subjects use visual salience to resolve underspecified references. It is also evident from the results that there is a relatively sharp drop off in subject's ability to resolve underspecifed references once the salience interval drops below 0.6. Based on this we have set the confidence interval for the LIVE system to 0.6.

Salience Interval	Correct	Ambiguous
int > 0.9	9	1
$0.9 <$ int > 0.8	8	2
$0.8 <$ int > 0.7	8	2
$0.7 <$ int > 0.6	8	2
$0.6 <$ int > 0.5	4	6
$0.5 <$ int > 0.4	1	9
$0.4 <$ int > 0.3	2	8
$0.3 <$ int > 0.2	2	8
$0.2 <$ int > 0.1	1	9
$0.1 <$ int $>= 0.0$	0	10

Table 2: Breakdown of testing results.

Conclusions

The LIVE GRE framework integrates a visual saliency algorithm with an adaptation of Dale and Reiter's (1995) incremental GRE algorithm. The visual saliency model is suitable for real-time 3-D simulations and is a novel application and extension of the false colouring graphics technique. The LIVE GRE algorithm integrates visual saliency into the generation process by modifying the definition of a distinguishing description. The advantages of the LIVE GRE framework are: (1) our modification does not make the Incremental Algorithm more complex than Dale and Reiter's (1995) version: it still has a polynomial complexity and its theoretical run time, $n_d n_i$, depends solely on the number of distractors n_d and the number of iterations n_i (i.e., the number of properties realised in the description); (2) for each scene, the LIVE visual saliency algorithm dynamically computes the context set for the generation algorithm; as a result the context set is updated as the user moves through the world; (3) the LIVE GRE algorithm's utilisation of visual saliency information means that it can generate underspecifed, but sufficiently detailed, exophoric descriptions of the target object; consequently, in some instances the LIVE GRE algorithm will succeed in generating a description in contexts where GRE algorithms which focus purely on adjectival and type attributes fail.

Future Work

The focus of this paper has been on integrating visual salience with the Incremental Algorithm. As a result, our algorithm exhibits some of the limitations of Dale and Reiter's (1995) algorithm. Like (Dale & Reiter 1995) we only discuss the generation of descriptions of the form "*the [Adj] N*". In future work, we plan to integrate our GRE algorithm with the multimodal discourse framework developed in (Kelleher 2003). This would involve integrating a model of linguistic salience, similar to Theune's (2000) framework, with our model of visual salience. Such an integration would allow us to generate reduced anaphoric and exphoric references in multimodal environments. We also intend to extend the algorithm to generate relational descriptions, in particular those involving locative expressions. Locative expressions are often used to intend on objects in a visual domain. This could be done using the computational model of projective prepositions developed in (Kelleher & van Genabith 2004).

References

Appelt, D. 1985. Planning English referring expressions. *Artificial Intelligence* 26(1):1–33.

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

Dale, R., and Reiter, E. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Dale, R. 1992. *Generating Referring Expressions: Building Descriptions in a Domain of Objects and Processes*. MIT Press.

Duwe, I., and Strohner, H. 1997. Towards a cognitive model of linguistic reference. Report: 97/1 - Situierte Kunstlicher Kommunikatoren 97/1, Univeristat Bielefeld.

Forgus, R., and Melamed, L. 1976. *Perception A Cognitive Stage Approach*. McGraw-Hill.

Grosz, B.; Joshi, A.; and Weinstein, W. 1995. Centering: A framework for modelling local coherence of discourse. *Computational Linguistics* 21(2):203–255.

Hajicová, E. 1993. Issues of sentence structure and discourse patterns. In *Theoretical and Computational Linguistics*, volume 2.

Hewett, M. 2001. *Computational Perceptual Attention*. Ph.D. Dissertation, University of Texas, Texas.

Kelleher, J., and van Genabith, J. 2004. Forthcoming: A computational model of the referential semantics of projective prepositions. In Saint-Dizier, P., ed., *Computational Linguistics: Dimensions of the Syntax and Semantics of Prepositions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kelleher, J. 2003. *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. Ph.D. Dissertation, Dublin City University.

Kosslyn, S. 1994. *Image and Brain*. The MIT Press.

Krahmer, E., and Theune, M. 2002. Efficient context-sensitive generation of referring expressions. In van Deemter, K., and Kibble, R., eds., *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. Standford: CLSI Publications.

Kuffner, J., and Latombe, J. 1999. Fast synthetic vision, memory, and learning models for virtual humans. In *Proceedings of Computer Animation Conference (CA-99)*, 118–127. Geneva, Switzerland: IEEE Computer Society.

Landragin, F.; Bellalem, N.; and Romary, L. 2001. Visual salience and perceptual grouping in multimodal interactivity. In *Proceeding of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD)*.

Noser, H.; Renault, O.; Thalmann, D.; and Magnenat-Thalmann, N. 1995. Navigation for digital actors based on synthetic vision, memory and learning. *Computer Graphics* 19(1):7–9.

Peter, C., and O'Sullivan, C. 2002. A memory model for autonomous vitual humans. In *Proceedings of Europraphics Irish Chapter Workshop (EGIreland-02)*, 21–26.

Reynolds, J. 2001. Visual salience, competition, neuronal responce synchrony and selective attention. In *Sloan/Swartz Centers for Theoretical Neurobiology Annual Summer meeting*. The Swartz Foundation.

Theune, M. 2000. *From data to speech: language generation in context*. Ph.d., Eindhoven University of Technology.

Tu, X., and Terzopoulos, D. 1994a. Artificial fishes: Physics, locomotion, perception, behaviour. In *Proceedings of ACM SIGGRAPH*, 43–50.

Tu, X., and Terzopoulos, D. 1994b. Perceptual modelling for behavioural animation of fishes. In *Proceedings of the Second Pacific Conference on Computer Graphics and Applications*, 185–200.

van Deemter, K. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics* 28(1):37–52.