# Adapting Descriptions of People to the Point of View of a Moving Observer

**Gonzalo Méndez** and **Raquel Hervás** and **Pablo Gervás**
**Ricardo de la Rosa** and **Daniel Ruiz**
Facultad de Informática - Instituto de Tecnología del Conocimiento
Universidad Complutense de Madrid (Spain)
{gmendez,raquelhb,pgervas}@fdi.ucm.es, {ricarosa,danruiz}@ucm.es

## Abstract

This paper addresses the task of generating descriptions of people for an observer that is moving within a scene. As the observer moves, the descriptions of the people around him also change. A referring expression generation algorithm adapted to this task needs to continuously monitor the changes in the field of view of the observer, his relative position to the people being described, and the relative position of these people to any landmarks around them, and to take these changes into account in the referring expressions generated. This task presents two advantages: many of the mechanisms already available for static contexts may be applied with small adaptations, and it introduces the concept of changing conditions into the task of referring expression generation. In this paper we describe the design of an algorithm that takes these aspects into account in order to create descriptions of people within a 3D virtual environment. The evaluation of this algorithm has shown that, by changing the descriptions in real time according to the observers point of view, they are able to identify the described person quickly and effectively.

## 1 Introduction

The task of Referring Expression Generation (REG) has traditionally been considered in static contexts, where neither the properties of the objects being described nor their relation to the observer change over time. This is a good starting point to address the problem because it includes the elements that are involved in more complex situations. The case where the observer is moving along a static context is a slight departure from the basic static case, with two significant advantages: many of the mechanisms available for static contexts may be applied with small adaptations, and it introduces the concept of changing conditions into the task of referring expression generation. For this reason, it is a worthwhile problem to explore.

A challenge when trying to address this problem is the need to continuously gather data on the relevant conditions – the field of view of the observer, his relative position to the people being described, and the relative position of these people to any landmarks around them in terms of how they appear in the field of view of the observer.

Gathering these data in a real life context may be very difficult, but if the situation is modeled in a 3D environment that represents the chosen scene, with a camera following the observer in first person mode, the compilation of all these data becomes a feasible task, and the generation of descriptions in real time becomes possible.

We have studied different proposals to solve similar problems and have developed a meta-algorithm based on the work depicted in (Méndez et al., 2017), where the authors studied the behavior of classic REG algorithms applied to this problem (section 3). Then, we have built a 3D scene and have populated it with people in order to test this meta-algorithm when the observer can move around the scene (section 4). The results of this evaluation have shown that the descriptions can be improved in order for the observers to find the target person more easily, so we have extended the previous algorithm to include additional information to the descriptions (section 5). We have subsequently evaluated the new algorithm using the same scenes (section 6) and the results show that the observers are able to find the target person faster and with a much higher hit rate than before.

## 2 Related Work

A Referring Expression (RE) is a description created with the intention of distinguishing a certain element (i.e. *referent*) from a number of other elements (i.e. *distractors*). It must identify the referent unambiguously, effectively ruling out all the distractors. Therefore, any expression that meets these criteria can be called a referring expression. However, not all of them can be considered equally good: they may be too long or too short, they may not contain enough information or they may have too many unhelpful details that hinder the listener.

The field of Referring Expression Generation (REG) has been widely explored for several decades (see (Krahmer and van Deemter, 2012) for an extensive survey), and there have been many studies for generating appropriate REs in different contexts. However, most of these solutions have approached the problem considering static contexts where neither the objects being described nor the point of view of the observer change over time.

In (Méndez et al., 2017) the authors assume that people and objects are described in different ways, since attributes such as size, shape or color, used to describe objects, are not so suitable for describing people. In order to identify what features are relevant for individuals when they have to describe other people, they conducted a number of surveys with human evaluators. These studies provided two important insights. The first one was that distance (from the viewer and to landmarks) influences the identification of referents, and REs that include information about nearby objects or people appeared to be easier to understand. The second insight obtained from the study was a list of preferred attributes when describing people in crowded environments. Based on these results, they proposed as future work the creation of a meta-algorithm that, depending on the particular circumstances pertaining to a given scene, selected a particular referring expression generation algorithm out of a set of the classic solutions studied.

Additionally, in recent years, computational approaches to REG have increasingly explored the task of adapting to dynamic contexts. The generation of appropriate referring expressions in the context of interactive dialogues is one of the problems that has received a lot of attention. Stoia et al. (2006) presented a REG system in dialogues that takes into account the current field of view of the speakers, how distant they are from the target, and the dialogue history. Similarly, Fang et al. (2014) describe two approaches to REG in situated dialog with artificial agents, both of which generate multiple small expressions that lead to the target object with the goal of minimizing the collaborative effort between the human and the agent. Janarthanam and Lemon (2009) explored a method for automatically adapting referring expressions to the lexical knowledge of users. Gatt et al. (2011) proposed a new model for interactive REG which incorporated both property preferences and priming effects and obtained good results in comparison with human experimental data. Garoufi and Koller (2014) presented a model of effective reference generation in situational contexts which distinguishes speaker helpfulness in a certain situation with the aim of modelling helpful speaker behaviour. Baltaretu et al. (2017) describe an approach that discusses the use of moving landmarks to generate route directions and how the listeners evaluate these instructions. The results show that listeners understand these instructions without much effort, but speakers tend to use stable landmarks more often. Unlike these approaches, which take advantage of situational dialogue and interaction with the user, the work described in this paper does not assume that the interaction with the user is possible or desirable, that is, we cope with the dynamics of the environment and try to provide the users with the best possible description, rather than requiring their collaboration to generate it.

Considering the physical context when generating REs, there are some works which have explored the REG problem in the context of 3D environments. The GIVE challenges (Byron et al., 2009; Koller et al., 2010; Striegnitz et al., 2011) focused on the generation of instructions in a virtual 3D environment to help a user solve a treasure-hunt task. One interesting aspect of using a virtual environment was that spatial and relational expressions played a bigger role than in other NLG tasks, and the necessary information to create the descriptions was already present in the environment. Garoufi et al. (2015) present an interesting work which has used the GIVE environment to study how a generation system that uses listener gaze to provide rapid feedback improves the generation of REs in comparison with two systems that do not consider the listener's gaze.

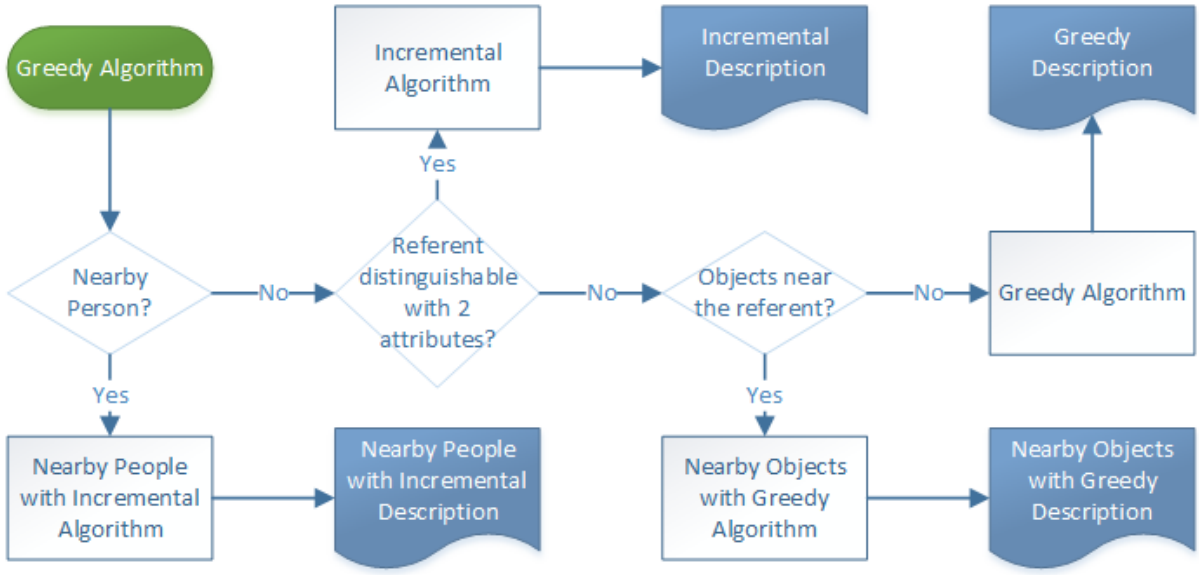



Figure 1: Design of the meta-algorithm

## 3 Design of a Meta-Algorithm for Character Descriptions

Based on the results and conclusions described in (Méndez et al., 2017), we decided to design and implement a meta-algorithm that, based on classic REG solutions, could dynamically decide which of them was more suitable to describe a given situation.

The classic algorithms considered were: *Incremental* (Reiter and Dale, 1992; Dale and Reiter, 1995), *Greedy* (Dale, 1989, 1992), *Nearby People with Incremental* – extend the description with relation to the nearest person, using the Incremental algorithm to describe that person –, and *Nearby Objects with Greedy* – extend the description with relation to nearest object, using the Greedy algorithm to describe the referent.

Taking into account the results obtained in the empirical evaluation of these algorithms in (Méndez et al., 2017), the meta-algorithm works as follows (see Figure 1 for a graphical description of the process).

First, the meta-algorithm tries to create a Nearby People with Incremental description. In order to do that, the meta-algorithm uses the Greedy algorithm to determine if there is a nearby person that is very easily identifiable (can be described by using only two attributes). If there is, the meta-algorithm returns the Incremental description of the referent plus the Greedy description of the nearby person.

If there is no other character near the target that is sufficiently distinguishable, the meta-algorithm goes on to find out if the referent stands out in the scene (can be referred to by using exactly two attributes). If this is the case, the meta-algorithm generates an Incremental description for the referent.

If the referent does not stand out, the meta-algorithm then tries to use the Nearby Objects with Greedy Algorithm. We use the Greedy Algorithm here only to describe the referent. Because of the low number and variation of objects in the scenes, we have considered the name of the object to be descriptive enough.

If there are no distinguishable objects near the referent, the meta-algorithm finishes by generating the description of the referent using the Greedy Algorithm.

The evaluation of this meta-algorithm with 38 users (15 women and 23 men) and 9 different scenes showed a total hit rate of 95% (324 correct answers out of 342). Even though in the evaluations used to design the meta-algorithm the users had shown a slight preference for the descriptions generated by the Nearby People with Incremental Algorithm, in this last evaluation the results were a little better when the descriptions were generated by the Nearby Objects with Greedy Algorithm, since all the users found the right target when this algorithm was used. In addition, after the evaluation some of the users reported that some of the mistakes they had made had to do with the difficulty to remember long descriptions.

| Algorithm | Description | Hits | Scene |
|---|---|---|---|
| Greedy | The girl in the blue shirt standing, leaning on a table | 96% | 10 |
| Nearby Objects | The boy in the red rolled up sleeves shirt near the window | 89% | 7 |
| Nearby People | The girl in the blue sweater with black hair. She is near, next to the girl in the yellow tank top | 89% | 3 |
| Nearby People | The girl in the green tank top who is standing up. She is near, next to the girl standing pointing at something | 85% | 6 |
| Nearby Objects | The boy in the black shirt sitting near the window | 85% | 1 |
| Nearby Objects | The boy in the blue t-shirt with black hair sitting near the window | 67% | 2 |
| Greedy | The girl in the red shirt with redhead hair | 63% | 5 |
| Nearby Objects | The boy in the green shirt with redhead hair sitting near the column | 48% | 9 |
| Incremental | The boy in the red t-shirt with spike blond hair who is sitting down. He is far | 48% | 4 |
| Incremental | The boy in the blue rolled up sleeves shirt with spike redhead hair. He is near | 44% | 8 |

Table 1: Results of the meta-algorithm evaluation

## 4 Perspective-Based Evaluation of the Meta-Algorithm

One of the difficulties when generating descriptions in changing environments is to decide when and how to change the referring expression we use to describe an element's situation, even more if we take into account that not everybody refers to an element in the same manner. In order to generate this kind of descriptions, the first step we took was to test the behavior of the meta-algorithm described in section 3 in dynamic conditions, in order to check the suitability of the generated descriptions as the user's viewpoint changed.

A survey was carried out in order to study how the changes in the user's point of view affected the perceived accuracy of the descriptions generated by algorithms thought to work in static conditions. The survey was completed by 27 people (45% of women and 55% of men), with ages from 20 to 45 years old. The users were shown ten scenes in a 3D virtual environment, together with the description of the target character they had to identify in each of them, generated by the meta-algorithm.

The description was presented to the users as a written message on the top part of the screen, and it was kept there until the users clicked on what they considered to be the target character. They were not told whether they could see the target character or not, and they could move around the environment in order to find the described person. The users had to click on it once they thought they had found it, but the provided description did not change as they moved.

All the scenes were reproductions of pictures taken in our canteen (so they all represent real situations), and all of them included more than 30 characters, both male and female, most of them between 18 and 25 years old, with varied characteristics, and typical actions included people speaking, drinking or working together, either standing up or sitting down. The scenes and characters were selected so that they put to test some difficult situations, such as characters that were initially out of sight, other characters that might get out of sight as the users moved around the scene, or some others that were difficult to see from a long distance and that looked similar to other characters close to them.

Table 1 shows the description generated for each scene of the evaluation, along with the algorithm selected by the meta-algorithm to generate it and the percentage of users that found the described character. Most correct clicks were achieved when the descriptions were generated by the nearby objects or people algorithms to describe the target; some of these descriptions made reference to the posture of the target to describe it.

In contrast, the incremental algorithm has got low hit rates in the two scenes where it was selected by the meta-algorithm to generate the descriptions. The reason behind it is that this algo-

rithm is selected when there are no salient objects or characters than can be used by other algorithms, and the incremental algorithm does not provide enough discriminating power when there are too many characters that look like the target one. This, in turn, has more to do with the difficulty to describe the characters in these scenes than with the algorithm itself, as it has been selected by the meta-algorithm precisely because it works better than the rest in these situations.

An advantage of the incremental algorithm is the inclusion of the distance between the user position and the target character as a descriptor. However, as the user moves around the scene, the meta-algorithm fails in updating this reference, thus making the description invalid. This points to the need of changing the description as the user moves, to keep it aligned with the user perspective of the scene, which will be included in the algorithm described in the next section.

Many users got some scenes wrong when they had to find easy to identify persons, because there was a character that looked like them in the users' field of view at the start. A way to fix this is to specify in the description if the target is in the user's field of view or not, and if it is near or far. Regarding the distance, the users were sometimes confused by the indication of the target being far, when they considered that it was not that far. Thus, a finer distinction of the distance to the target may also improve the quality of the descriptions.

## 5 Implementation of a Perspective-Based Algorithm to Describe People

With the results of the previous survey, and using the graphical engine Unity 3D, an extension to the meta-algorithm described in section 3 has been developed to generate descriptions of characters in real time that change according to the user's position within the environment.

The developed algorithm was implemented in a game where the user had to find the character that was being described, for which he could move around the environment and the provided description changed accordingly. The content of the description is based on the character's physical appearance, which does not change, its position within the environment, and its situation with respect to other relevant characters and objects that are present in the environment. Therefore, a complete description consists in the composition of two different parts:

- *attribute-based description*, which refers to the static characteristics of an individual and its environment, and they cannot change during the simulation. This part of the description is generated using the meta-algorithm;
- *perspective-based description*, which has to be generated in real time according to the situation of the user relative to the situation of the described character.

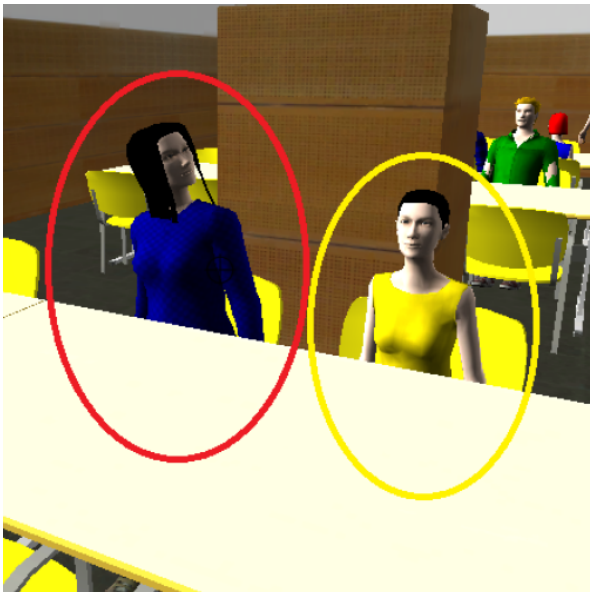
The perspective-based description of a character is composed of sub-descriptions, which are generated according to the data that is obtained from the scene. There are three possible types of sub-descriptions:

- *description of reference points*: this description contains information related to the reference points scattered all over the scene, such as the end of a corridor or a corner;
- *description of the visibility*: it contains the information about the visibility of the target from the user's point of view, such as the fact that the described person is behind a column or another person, or even behind the user;
- *distance between the described person and the user*: it contains a textual explanation of the distance between the described person and the user: near, a little far (i.e. medium distance) and far. This is a source of mismatch with the meta-algorithm, as it only considered that targets might be either near or far.

These sub-descriptions are generated and updated in real time, and they are shown to the users as soon as the conditions used to generate the current description change (e.g. the described character starts being visible, or the user gets close to the target character), according to the following rules:

1. First the algorithm checks the distance between the user and the reference points previously placed into the scene. If the distance from the user to one of these points is greater than a predefined constant (that depends on the dimensions of the scene), the generated description is updated with information about the proximity of the described character to that point (e.g. *the described person is at the end of the corridor*, if the user is far from the end of the corridor).

(a) Zoom of the initial situation of scene 3. The target person is a girl that is not near the observer, hidden behind a column. The provided initial description is *The girl in the blue sweater with black hair, next to the girl in the yellow tank top. The described person is behind a column. She is not far*



(b) Final situation of scene 3. The observer has moved closer to the target person and is looking at her from the other side. The provided description is *The girl in the blue sweater with black hair, next to the girl in the yellow tank top. She is near. You can see the described person*

Figure 2: Sample scene used in the evaluation. A red circle has been drawn around the described person

2. Then, the algorithm checks if the target character is in the user's point of view (i.e. approximately in front of the user). If not, the generated description must contain the positional references of the described person to the user: it indicates whether the described person is to the left, right or behind the user.

3. If the described person is within the user's field of view, the algorithm checks the absolute distance between the target and the user and indicates the user whether he is near, not far or far from the described person.

Finally, the description that is shown to the user has to be composed. First, if the description provided by the meta-algorithm contains information about the distance from the user to the target character, it is removed, as the new algorithm may treat this information differently. Then, by combining sub-descriptions, and using the previous rules, the perspective-based descriptions are generated (e.g. *The boy in the black shirt sitting near the window. The described person is behind another person. He is not far*).

## 6 Evaluation of the Perspective-Based Algorithm

A second evaluation was carried out six months after the first one, using the same conditions and instructions as in the first one. The main objective of this evaluation was to test the improvements added to the meta-algorithm by comparing the obtained results with those of the first survey. Therefore, the people that had to be found by the observer, and the scenes used for it, were the same as in the previous one. This way, a reliable comparison could be made between both versions of the meta-algorithm in order to test their effectiveness. A sample scene used for this evaluation can be seen in Figure 2. The number of people that completed the survey was twenty seven, the same as in the first survey. 85% of them were between 20 and 30 years old, and the remaining 15% were between 30 and 40. 63% of the participants were male, and the remaining 37% were female. Five of the evaluators had also completed the first survey, but after six months they assured they did not notice the scenes and characters were the same as in the first one.

Table 2 shows the results obtained in this evaluation. The column corresponding to the description only shows the initial descriptions of the tar-

| Algorithm | Initial Description | Hits | Scene |
|---|---|---|---|
| Greedy | The girl in the blue shirt standing, leaning on a table. You can see the described person. She is a little far | 96% | 10 |
| Nearby Objects | The boy in the red rolled up sleeves shirt near the window. The described person is far from you | 92% | 7 |
| Nearby People | The girl in the blue sweater with black hair, next to the the girl in the yellow tank top. The described person is behind a column. She is a little far | 92% | 3 |
| Nearby People | The girl in the green tank top who is standing up. The described person is a little far from you | 92% | 6 |
| Nearby Objects | The boy in the black shirt sitting near the window. The described person is behind another person. He is a little far. | 89% | 1 |
| Nearby Objects | The boy in the blue t-shirt with black hair sitting near the window. The described person is far from you | 89% | 2 |
| Nearby Objects | The boy in the red t-shirt near the column. He is at the back of the canteen | 89% | 4 |
| Nearby Objects | The boy in the green shirt sitting near the column. The described person is far from you | 85% | 9 |
| Incremental | The boy in the blue rolled up sleeves shirt with spike redhead hair. The described person is a little far from you | 85% | 8 |
| Greedy | The girl in the red shirt with redhead hair. The described person is far from you | 78% | 5 |

Table 2: Results of the perspective-based meta-algorithm evaluation

get characters, so that they can be compared with the ones in Table 1. An example of the initial and final descriptions for scene 3 in shown in Figure 2.

All the scenes have obtained an increased hit rate, except for the first one, which scores the same as in the first evaluation (96%). The first five scenes in the first evaluation still occupy the same positions in the second one, but with higher hit rates, as we have mentioned. So does the sixth one, but with a much higher hit rate than before. The last four ones have also experienced improvements in their hit rates, with slight variations in their relative positions.

Some remarkable differences can be found between the descriptions in Tables 1 and 2. In scenes 6 and 8, the target is not described as being *near* any more, but *a little far*. This is due to the finer distinction that the new algorithm makes for describing distances. In addition, in scene 4, the algorithm used to generate the description is different in both evaluation. This is caused by the inclusion of a landmark in the scene (i.e. *the back of the canteen*) which causes the meta-algorithm to change the algorithm selected to generate the description.

On average, the new perspective-based meta-algorithm has a hit rate of 88% $(240/270)$. Comparing it to the previous algorithm that got 71% $(194/270)$, we can see an improvement in the algorithm's capabilities to adapt the descriptions to different points of view.

A lot of factors have influenced the overall improvement in the results. For example, the participants of the second survey who had also completed the first survey provided us with some feedback about the improvements they had perceived. One of their comments was that they felt more confident looking for the target character if they knew at the beginning where to start looking for it. This confirms that having the algorithm detail the distance of the observer to the target character and specifying if he/she was in the field of view of the observer has provided better indications for the users to find the described person.

The change of the description in real time has helped the observers in a more realistic way, mimicking how a real person would be providing the description. Again, the users' feedback shows that they get lost less frequently when the algorithm offers them clues about where the target person is.

In both evaluations, we measured the time it took the users to click on the person they though

| Scene | First Eval | Second Eval |
|---|---|---|
| 10 | 0.82 | 0.90 |
| 7 | 2.99 | 4.30 |
| 3 | 3.01 | 3.50 |
| 6 | 2.12 | 4.40 |
| 1 | 3.79 | 5.00 |
| 2 | 4.88 | 5.30 |
| 4 | 3.50 | 2.90 |
| 9 | 1.66 | 0.48 |
| 8 | 0.70 | 0.30 |
| 5 | 7.40 | 4.70 |

Table 3: Average response times (in seconds) for each scene

that was being described. Table 3 shows the average response times for each scene, sorted descendingly according to the hit rates obtained in the second evaluation. At first sight, it seems that the results obtained with the new version of the algorithm are slightly worse than those of the first version. However, a careful analysis of the collected data has shown that this is due to the increased hit rate of the second evaluation. In the first evaluation, the users who clicked on the wrong target answered much faster than the ones who tried to find the right one. Comparing the ones who took the right choice, the average response times are slightly better using the second version of the algorithm, although the difference is not significant and may be even due to the users ability to play in first person games.

Even though this evaluation is not statistically significant, provided that there were only 10 scenes and 27 evaluators, the improvement obtained using the perspective-based algorithm was quite consistent across all the scenes, so we can conclude that adding information regarding the location and visibility of the target character, along with updates in the descriptions when the user's point of view changes, allow the users to better find the person that is being described, very much in line with some the previous work described in section 2.

## 7 Discussion

The current work has focused on describing people in dynamic contexts in which the observer can move around the environment, while the rest of the people are static.

The first question that arises is whether the described approach only works for people or if it is possible to generalize it to describe other entities. As far as we can tell, the way in which people are described differs from the way in which other entities are. Previous results presented earlier in this paper suggest that, when describing people, the attributes and order in which they are used differ from those used to describe objects. The algorithms we have used to describe people are not specifically tailored for the situations and environments in which the experiments have been run, so it is certainly possible to adapt them to describe other entities. It is not the case, however, of the meta-algorithm and the perspective-based meta-algorithm, as their design is based on experimental results drawn exclusively from descriptions of people, so further study is required in order to figure out how the adaptation to describe objects or other entities might be carried out.

The second question that arises is whether the proposed approach should have been used to describe objects instead of people, as the environment is static, or whether we should have been immersed in a more realistic, dynamic environment where the rest of the characters could also move. The answer to the first part of the question is that our main interest was in describing people, as much less research work seems to have been carried out in this area. This links with the second part of the question, for which the answer is that describing characters that can move around the environment is a much more complicated problem, since they can change their position, posture, the way they dress or, more important for some of the algorithms we have used, they can become or stop being a reference element in the description (e.g. *Nearby People with Incremental*), which introduces a high degree of complexity in the descriptions and requires the algorithms to monitor many more variables when deciding which elements to include in the descriptions. This work provides a first approach to deal with more dynamic environments where not only the observer's point of view is to be considered, but also other elements that move around the environment.

There are other limitations to the current approach, such as the lack of references to groups of people (or even objects) doing something (e.g. *the boy in the red t-shirt sitting near the girls playing Scrabble*), which becomes even more complicated in dynamic contexts where groups may form and

break, and which also leads us to evaluate under what conditions should we consider that some people are forming a group or not.

Some of the limitations of the proposed algorithm have not been studied yet, such as the results it would produce in environments where there is little variation in the aspect of the characters being described (e.g. all the characters are wearing a uniform). Another limitation is the fact that we have tested the algorithm in scenes where the number of relevant objects that may be included in the description is small, so just using the name of the objects in the descriptions was enough, but additional decisions on how to use object descriptions may be necessary if this situation changes.

## 8 Conclusions and Future Work

Throughout this work we have seen that the problem we have addressed – describing people when the observers can change their point of view – poses challenging issues that have been satisfactorily solved, although there is still space for improvement. He have shown that, by using the techniques that have been used traditionally to describe static situations, we can generate acceptable descriptions when we shift to more dynamic environments, closer to real life situations, in which the observer can move to get a better perspective of the person being described. In contrast with the works described in section 2, which assume that the users can interact and collaborate to let the system generate small bits of the description that takes them progressively closer to the target, we do not take that for granted, so we always provide users with a full description of the target from the current point of view, which is updated as the users move across the scene.

In addition, we have put forward that, if we take into account certain aspects that change as the observer moves, the descriptions we generate can be more accurate and can help the observer identify the target person more easily. The aspects we have taken into account in this work have been: the distance between the observer and the target subject; the visibility that the observer has of the described person; and the relative position among the observer, the referent and significant landmarks that can help locate the objective more easily.

The proposed solution to generate descriptions is based on the use of crisp values to determine thresholds in order to generate linguistic labels to refer, for example, to distances. However, this specific aspect can benefit from the use of fuzzy logic to generate descriptions of spatial relationships. Although we have not been able to find an approach of this kind in the reviewed literature, there have been some efforts to solve similar problems in the fields of image analysis and computer vision, as described by Bloch and Ralescu (2003), who have subsequently developed several methods to describe spatial relations between objects (Hudelot et al., 2008). Other authors have tackled the problem of automatic scene descriptions in image analysis using fuzzy rule-based systems (Keller and Wang, 2000) and fuzzy sets (Matsakis et al., 2001), through the use of histograms of angles and forces and a dictionary of labels.

In addition, the generation of descriptions in more realistic environments, where the elements of the scene can move and change, is another problem that still needs to be tackled and that poses even more challenging issues to solve.

## Acknowledgements

The work presented in this paper has been partially funded by the projects IDiLyCo: Digital Inclusion, Language and Communication, Grant. No. TIN2015-66655-R (MINECO/FEDER) and InVITAR-IA: Infraestructuras para la Visibilización, Integración y Transferencia de Aplicaciones y Resultados de Inteligencia Artificial, UCM Grant. No. FEI-EU-17-23.

## References

Adriana Baltaretu, Emiel Krahmer, and Alfons Maes. 2017. Landmarks on the move: Producing and understanding references to moving landmarks. *Spatial Cognition & Computation*, 17(3):199–221.

I. Bloch and A. Ralescu. 2003. Directional relative position between objects in image processing: a comparison between fuzzy approaches. *Pattern Recognition*, 36(7):1563—-1582.

Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation (Special session on Generation Challenges)*.

Robert Dale. 1989. Cooking up referring expressions. In *Proc. of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, University of British Columbia,Canada.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, MA, USA.

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

Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2014. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1544–1550. AAAI Press.

Konstantina Garoufi and Alexander Koller. 2014. Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience*, 29(8):986–1001.

Konstantina Garoufi, Maria Staudte, Alexander Koller, and Matthew Crocker. 2015. Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*.

Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Attribute preference and priming in reference production: Experimental evidence and computational modeling. In *Proc. of the 33rd Annual Meeting of the Cognitive Science Society (CogSci'11)*, Austin, TX. Cognitive Science Society.

C. Hudelot, J. Atif, and I. Bloch. 2008. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929—1951.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions in spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 74–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. M. Keller and X. Wang. 2000. A fuzzy rule-based approach to scene description involving spatial relationships. *Comp. Vision and Image Understanding*, 80(1):21–41.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In *Proceedings of the Sixth International Natural Language Generation Conference (Special session on Generation Challenges)*, Dublin.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

P. Matsakis, J.M. Keller, L. Wendling, J. Marjamaa, and O. Sjahputera. 2001. Linguistic description of relative positions in images. *IEEE Transactions on Systems, Man and Cybernetics*, 31(4):573–88.

Gonzalo Méndez, Raquel Hervás, Susana Bautista, Adrián Rabadán, and Teresa Rodríguez-Ferreira. 2017. Exploring the behavior of classic reg algorithms in the description of characters in 3d images. In *International Natural Language Generation Conference (INLG2017)*, Santiago de Compostela, Spain.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 232–238, Nantes, France.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariet Theune. 2011. Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation (Special session on Generation Challenges)*, Nancy.