Verbal description of LEGO blocks

Diogo Henriques, Isabel Trancoso, Daniel Mendes, Alfredo Ferreira

INESC-ID, Lisbon, Portugal
IST, Universidade de Lisboa, Portugal

diogo.henriques@tecnico.ulisboa.pt, isabel.trancoso@inesc-id.pt
danielmendes@tecnico.ulisboa.pt, alfredo.ferreira@inesc-id.pt

Abstract

Query specification for 3D object retrieval still relies on traditional interaction paradigms. The goal of our study was to identify the most natural methods to describe 3D objects, focusing on verbal and gestural expressions. Our case study uses LEGO® blocks. We started by collecting a corpus involving ten pairs of subjects, in which one participant requests blocks for building a model from another participant. This small corpus suggests that users prefer to describe 3D objects verbally, rarely resorting to gestures, and using them only as complement. The paper describes this corpus, addressing the challenges that such verbal descriptions create for a speech understanding system, namely the long complex verbal descriptions, involving dimensions, shapes, colors, metaphors, and diminutives. The latter connote small size, endearment or insignificance, and are only very common in informal language. In this corpus, they occurred in one out of seven requests. This experiment was the first step of the development of a prototype for searching LEGO® blocks combining speech and stereoscopic 3D. Although the verbal interaction in the first version is limited to relatively simple queries, its combination with immersive visualization allows the user to explore query results in a dataset with virtual blocks.

Index Terms: multimodal corpus, 3D objects, voice search

1. Introduction

The number of three-dimensional objects in digital format has significantly increased in recent years, mostly motivated by the appearance of low-cost technologies that allow scanning of physical objects. This increased complexity led to a slow and tedious retrieval process, triggering the need for adequate 3D object retrieval systems and raising the challenge of query specification. Several approaches have been proposed to address this challenge, resorting to different modalities for performing the search: text, sketches, verbal descriptions, gestures, or using an object as an example, something that is not always available. Despite significant progress, these approaches are still very far from properly exploring the descriptive power and potential of human interaction.

Our study aims to identify the most natural methods to describe 3D objects, focusing on verbal and gestural expressions. As a case study, we adopted LEGO® blocks, since we believe such controlled context comprises many challenges faced in other domains. The first step in our study was the collection of a small multimodal corpus involving twenty participants, in which one participant requests blocks for building a model from another participant. This small corpus suggests that users in this context prefer to describe 3D objects verbally, rarely resorting to gestures, and using them only as complement. The paper describes this corpus, addressing the challenges that such verbal descriptions create for a speech understanding system, namely the long complex verbal descriptions, involving dimensions, shapes, colors, metaphors, and diminutives.

The first practical application of this corpus was in the development of a multimodal prototype for searching LEGO® blocks, combining speech and stereoscopic 3D. Although the verbal interaction in the first version is limited to rather simple queries, its combination with immersive visualization allows the user to explore query results in a dataset with virtual blocks, avoiding the traditional grid of thumbnails that typically loses relevant 3D information about the objects.

After a necessarily brief review of the state of the art for the retrieval of 3D objects in Section 2, this paper will describe the collection and main characteristics of the multimodal corpus (Section 3). This analysis is followed by the presentation of our prototype in Section 4. The last section summarises the main findings and future directions for this work.

2. Related Work

With the increase of any content type, retrieving specific information will always be a challenge, and three-dimensional objects, or other multimedia content, are no exception. One of the traditional ways to perform this retrieval consists of using textual queries. However, this method is not trivial, since the objects do not usually contain sufficient intrinsic information for a description of it, for instance the file name may not be related to the objects [1, 2]. Generically, search engines often use text associated with objects such as captions, references to the objects or even when it comes to contents scattered over the Internet, the links of the objects or file names. The use of synonyms is also possible. However, the information to describe 3D objects is still insufficient, especially regarding their shape.

Retrieval by example may ease the process of recovery by shape, searching by similar objects in terms of visual aspects such as color [3] or shape [4] of the object. For our LEGO® search scenario, however, this solution is not possible.

Retrieval by sketch [5, 6] offers users the possibility of searching for similar shaped objects, for which they do not have an example. In fact, allowing users to make sketches that match the dimensions of the desired objects was the solution proposed for our scenario in [7].

Retrieval by gestures captures and interprets gestures of the user, exploring the human spatial perception. In [8], the shape and movement of the hands when describin objects were used

to create 3D sketches. The authors concluded that participants were able to keep the correct proportions relatively to physical objects and in areas with most significant details when the gestures were performed more slowly.

Voice search has recently become a very widespread alternative to the traditional search by text, particularly on mobile devices. But what began as a convenient hands-free option for these devices is now also making its way onto our desktops. Despite the enormous progress in this area spurred by major companies, searching 3D objects by voice is still a relatively unexplored topic.

All the above mentioned types of retrieval attempt to explore more natural methods for the description of objects, but they do not yet conveniently explore the potential of the interaction between humans and their descriptive power.

The visualization of the results of the 3D search is also an object of research. Traditional approaches based on thumbnails lack an adequate 3D perception of the objects. The benefits of using immersive environments in the context of multimedia retrieval have been shown in [9], which used a CAVE like setup for presenting the query results for a content-base image retrieval system. Extending this approach for retrieving 3D objects, Pascoal et al. [10] showed that some of the main issues when presenting 3D object retrieval results can be overcome. These results are distributed in the virtual space accordingly to their similarity. Users can then explore the results by navigating in the immersive environment, through a head-mounted display. Their system also allows a diversified set of different visualization and interaction devices, used to test multiple interaction paradigms for 3D object retrieval.

3. Multimodal Corpus

Our small corpus involved ten pairs of subjects, in which one participant requested blocks for building a model from another participant. Each subject participated in two tasks: once as a builder and once as a supplier. After a preliminary introduction to the task with an example of the step-by-step instructions for building a model, the two participants were ready to start the first task. The builder was given the instructions to assemble a new model, and the supplier was given a large box of blocks, much more than those needed to complete the model. A small barrier between the two participants prevented the builder from seeing the box and the supplier from seeing the instructions, but they could see each other's faces and hand gestures.

The experimental setup was done with four different models, involving different blocks, but a similar geometric complexity. The instructions for each model included on average 20 steps. Figure 1 illustrates some of the steps of one of the models. For each step, the builder had to request the corresponding block from the supplier, describing it as he/she thought it was more suitable. The supplier searched for the block and handed it to the builder. After completing their first model, the two subjects changed roles, repeating the process for another model. After building the two models, they were asked to fill a short questionnaire.

The majority of the participants were university students, with ages ranging from 18 to 24 (60%). All other participants had a university degree, with ages ranging from 25-34 (35%), or 35-45 (5%). 7 of the 20 participants were female. All participants were familiar with LEGO® blocks. Except for two speaker pairs, participants knew each other prior to the experiment, a fact that originated a very informal interchange, and was reflected in the use of jargon and jokes. All participants were native Portuguese speakers.

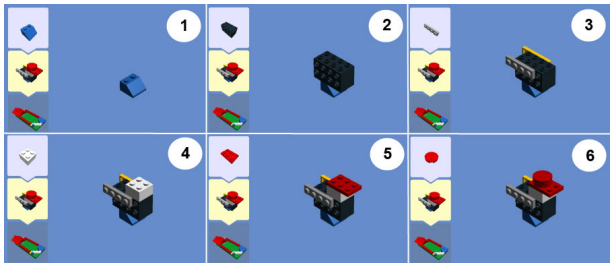

Figure 1: *Example of instructions provided to the builder, similar to those that follow the original LEGO® models.*

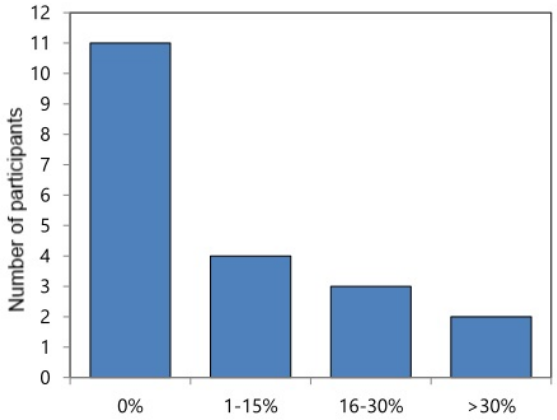

Figure 2: *Percentage of steps in which participants described blocks using gestures.*

The description of each block took on average 7.7s, whereas searching for each block took twice that long (15.3s), since the available blocks were presented in a box, and therefore were not arranged by any color, shape or dimension. The fact that the total time (5m:33s, on average) is less than the sum of the times for description (2m:22s) and search (4m:39s) was due to the fact that suppliers often started searching for the block before the builder finished its description. This added noise to the recordings, a fact that unfortunately had not been anticipated.

3.1. Main characteristics

The use of gestures to describe the blocks was one of the major issues in this study. As shown in the graphic in Figure 2, 11 out of 20 participants did not use any gesture in any of the steps of the session, being mostly of the vocal type. One participant was clearly of the gestural type (11 gestures). Occasionally, participants used gestures, altogether in roughly 10% of the blocks, but always for the purpose of complementing the verbal description, not adding any significant information. For example, to describe a corner piece, some participants used as a reference the letter *L*, making the corresponding gesture while referring to the letter (Figure 3, left). Height is one of the block characteristics that was most often complemented by gestures, as illustrated in Figure 3, right. Sometimes, the participants performed sketches in the air, with the shape of the desired block. A frequent word accompanying gestures was *assim* (like this), as in for instance, *Uma de três por um inclinada, assim.* (one

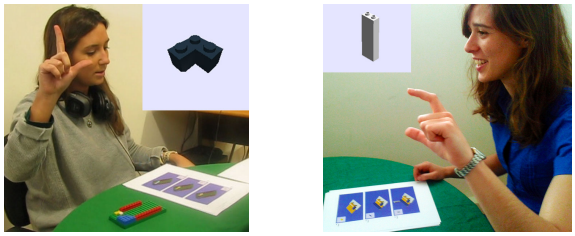

Figure 3: *Gestures denoting L-shape (left) and height (right).*

[block] of three by one inclined, like this.)

The word *agora* (now) was very often used to start a new query, followed by *quero* (I want) , *preciso* (I need) or *dá-me* (give me), although these two forms were also frequent as query starters. Alternatively, queries could start without any verbal form, just the block description itself. E.g. *Peça branca, fininha, dois por oito.* (White piece, quite thin, two by eight.)

The last example illustrates how even in more straight-to-the-point queries, among unfamiliar speakers, the use of diminutives (with the suffix *-inha* in this case) can be quite frequent. Diminutives connote small size, endearment or insignificance, and are only very common in informal language. In this corpus, they occurred in one out of seven requests.

The specification of each block descriptions usually began with its dimensions and color. Regarding the dimensions, the unit used was the number of pins of the block, or the equivalent space in blocks without pins. Adjectives were very often used to avoid counting pins, taking advantage of the limited number of blocks of a given color in the box. When confronted by a query as for instance: *Agora quero uma peça comprida, vermelha, com os buracos laterais, com buracos de um lado ao outro.* (Now I want a long block, red, with lateral holes, with holes going from one side to the other.), the supplier just searched for the longest block with these characteristics, without the need to count the number of connectors or holes. Saying things such as *around 12 connectors* would be another way to avoid the counting task in long blocks. Likewise, the height of each block, when differing from *normal* was either referred to as *thin*, or *high*. Blocks with slopes were often referred to as ramps or roof-shaped, followed by the dimensions at the base and top.

Participants in the role of builders often made several refinements to their initial descriptions. On average, each participant refined the description of 4 of his/her blocks. Even though most refinements were caused by suppliers returning the wrong block (on average, this occurred in two blocks for each participant), builders often added more detail to the initial description on their own initiative.

Metaphors were often used in descriptions. The most frequent ones were related to the already mentioned roof-shaped or L-shaped blocks, which were part of many models. The most unusual blocks were the ones that triggered the use of more creative descriptions, as expected. Blocks such as the one shown in Figure 4a were described as the prow of a boat, the shape of the letter *A*, or a Space Invader, whereas the one in Figure 4b triggered metaphors such as shark teeth or the top of a trident.

Confirmations were very frequent. Words like *Okay*, *sim* (yes), *exacto* (exact), *correcto* (right), *isso* (that's it) or equivalent expressions such as *É esta* (that's the one) were often used for this purpose. Acknowledgements were also present, but not that frequent.

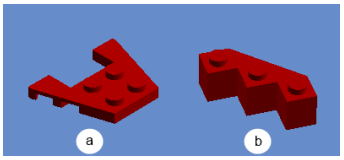

Figure 4: *Examples of shapes that triggered metaphors.*

It is interesting to notice how participants typically improve their query strategy as the session progresses. Some participants, however, are remarkably consistent in their style, for instance, always starting their query with a color specification.

The corpus was recorded both in audio and in video. The audio recordings were made using a lapel microphone at 16kHz, 16-bits per sample. The corpus was divided into training and testing subsets. Table 1 describes the main statistics of each subset. Comments by the speakers that were not related to the task were not included in the transcriptions. Unlike our other dialogue corpora, such as the Map Task corpus [11], we only recorded and transcribed audio-visual data for the participants in the role of builders. The participants acting as suppliers can only be heard in the background.

Table 1: *Corpus statistics.*

	Train	Test	Total
No. speakers	16	4	20
No. queries	306	66	372
No. queries w/ gestures	26	11	37
No. tokens	6088	1015	7103
No. different words	503	167	542
No. filled pauses	119	21	140
No. of diminutives	130	15	145
No. of confirmations	156	32	188
No. of metaphors	46	11	57

3.2. Survey results

At the end of the session, each participant was asked to fill a questionnaire about the use of the modalities - speech, gesture, and the combination of both - in terms of ease in describing objects. The classification was done in a Likert scale with four values, where 1 corresponds to very difficult, and 4 to very easy. The results are presented in Table 2. The analysis of the questionnaire responses was made using the Wilcoxon test. The participants showed a strong preference for using exclusively speech, compared to using gestures ($Z = -4.018$, $p = 0.000$) or a combination of modalities ($Z = -3.502$, $p = 0.000$). Furthermore, the combination of gestures and speech was consistently preferred over using just gestures ($Z = -3456$, $p = 0.001$).

Table 2: *Survey results in terms of ease in describing blocks for three modalities (median, inter-quartile amplitude).*

Description	Classification
Verbal	4(1)
Gestural	1(1)
Combined	2(1)

4. The LS3D Prototype

This section is devoted to a necessarily brief description of the LS3D prototype (LEGO® Search combining Speech and Stereoscopic 3D). The first version of this prototype was aimed at investigating different immersive visualization strategies which could be combined with a verbal interaction. For this purpose, the spoken interface was kept as simple as possible, integrating in-house speech recognition [12] and synthesis [13] modules. Given the naturally poor performance of the speech recognition module in the LEGO® domain, using the most generic existing language model (the one trained for broadcast news), the simplest option was to build a very limited grammar (GRXML[1], based on the most common expressions of our training data. The selected vocabulary included more than 500 different words, due to the inclusion of inflected forms of the words in the training data (e.g. *inclinado/inclinada*). This vocabulary covers 1810 different block shapes.

In this simple version, the semantic model was manually implemented. The main reason was the lack of adequate semantic resources for Portuguese, and the fact that the block descriptions are in English. The example *6143.dat - brick 2 x 2 round* illustrates the LEGO® database description of bricks to which the class of (14) adjectives denoting round shapes may be applied. The example *3022.dat - plate 2 x 2* , on the other hand, can fit the database description of plates to which the class of (21) adjectives denothing thinnness may be applied. In fact, the part of the XML grammar file that deals with shapes is naturally more complex than the one dealing with dimensions or colors.

Every query starts with a keyword (Acorda LEGO®), to which the system verbally replies with a prompt (Yes). In this very simple interaction, errors can only be addressed by a new query, and confirmations are not done verbally. An example of a query is *Acorda LEGO. Quero uma peça fina, 2 por 2* (Wake up LEGO. I want a thin block, 2 by 2). This example also illustrates another difference between this type of limited interaction, and the one that was observed during the corpus collection: here, as every block can virtually exist in a multitude of colors, users typically concentrate in describing the shape and dimensions in the first query. Once the desired block is selected, they can change its color from the default one in a new query, as in *Acorda LEGO. Pinta de encarnado.* (Wake up LEGO. Paint it red.). Given the very large universe of LEGO® blocks two additional strategies have been implemented to restrict the number of pieces shown. Users can verbally filter by a given characteristic (e.g. filter by curved), or exclude a type of blocks (e.g exclude DUPLO®).

Given the focus on visualization of this prototype version, we started by adapting the traditional grid approach with 2D thumbnails to a 3D immersive environment, to be able to analyse its suitability. To create the fully immersive visualisation of the environment, a set of Oculus Rift is used, and to enable a gesture-based interaction, we use the Leap Motion device as illustrated in Figure 5.

Several visualization modes have been tested in order to better explore the 3D space. In the cylindrical mode which is currently integrated, we surround the user with objects, placing them on the surface of an invisible cylinder. The higher ranked objects are placed in front of the user, expanding from there. Navigation through results is made by freely moving the user's head. To see beyond the users field of vision, or focus on a specific section of the results, a user can place an open hand in mid-air in front of him/her and move it to rotate left and right, or pan up and down the results. When a detailed view of an object is desired, the user must point with one finger at the object, which is then brought closer, and rotates in order to be viewed from different angles.

[1]Speech Recognition Grammar Specification, http://www.w3.org/TR/speech-grammar/

Figure 5: *Using the LS3D prototype.*

Despite the very simple verbal interaction, users prefer to follow the examples of the spoken queries that are initially given to illustrate how the prototype works. Since they do not know the speech understanding capabilities of the system, they focus on the novelty of the 3D glasses, and opt for not risking complex verbal queries. This is for instance illustrated by the absence of diminutives in these first tests, although the grammar does take them into account. This shows how a user can entrain to the system.

5. Conclusions

The focus of this paper is on the multimodal corpus that has been collected to investigate the most natural ways in which humans describe 3D objects, in this case, LEGO® blocks. The experiment showed the relevance of verbal descriptions vs. gestural ones, which occurred only as a complement.

The experiment served as a basis for building the first version of a prototype combining verbal interaction with immersive visualization, although we cannot directly compare the task of searching for blocks in a limited number of pieces, in which the supplier presents a single block as a query result, with the task of searching for blocks in the universe of LEGO® blocks, in which the system typically presents a large number of blocks as the result of a first query.

Nevertheless, the experimental setup is very rich for studying models of human communicating patterns, namely in what concerns the many different ways of combining modalities. Future work will be devoted to research much more complex dialog management systems, supporting error recovery strategies, confirmations, etc. It would be also interesting to investigate how to make use of multilingual semantic ontologies to automatize the association of verbal descriptions to database descriptions. Our on going experiments with the prototype also show that this setup can also be very suitable for entrainment studies.

6. Acknowledgements

The authors would like to thank their colleagues Carlos Mendes, José David Lopes, Hugo Meinedo and Alberto Abad for many helpful suggestions. This work was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under grants SFRH/BD/91372/2012, PTDC/EIA-CCO/122542/2010 and PEst-OE/EEI/LA0021/2013, and also by EU-IST FP7 project SpeDial under contract 611396.

7. References

[1] J. Smith and S. Chang, "Visually searching the web for content," *MultiMedia, IEEE*, 1997.

[2] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs, "A search engine for 3D models," *ACM Transactions on Graphics*, vol. 22, no. 1, pp. 83–105, Jan. 2003.

[3] E. Paquet and M. Rioux, "Nefertiti: a query by content system for three-dimensional model and image databases management," *Image and Vision Computing*, vol. 17, pp. 157–166, 1999.

[4] H. Laga, T. Schreck, A. Ferreira, A. Godil, I. Pratikakis, and R. Veltkamp, "Bag of words and local spectral descriptor for 3d partial shape retrieval," 2011.

[5] J. Pu, K. Lou, and K. Ramani, "A 2 d sketch-based user interface for 3 d cad model retrieval," *Computer-Aided Design and Applications*, vol. 2, no. 6, pp. 717–725, 2005.

[6] Y.-J. Liu, X. Luo, A. Joneja, C.-X. Ma, X.-L. Fu, and D. Song, "User-Adaptive Sketch-Based 3-D CAD Model Retrieval," *IEEE Transactions on Automation Science and Engineering*, pp. 1–13, 2013.

[7] T. Santos, A. Ferreira, F. Dias, and M. J. Fonseca, "Using sketches and retrieval to create lego models," *Proceedings of the Fifth Eurographics conference on Sketch-Based Interfaces and Modeling*.

[8] C. Holz and A. Wilson, "Data miming: inferring spatial object descriptions from human gesture," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 811–820, 2011.

[9] M. Nakazato and T. Huang, "3D MARS: Immersive virtual reality for content-based image retrieval," *ICME*, pp. 44–47, 2001.

[10] P. Pascoal, A. Ferreira, and J. Jorge, "Im-o-ret: Immersive object retrieval," in *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, 2012, pp. 121–122.

[11] I. Trancoso, C. Viana, I. Duarte, and G. Matos, "Corpus de diálogo coral," in *Proc. PROPOR'98 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, Porto Alegre, Brazil, Nov. 1998.

[12] H. Meinedo, A. Abad, T. Pellegrini, J. Neto, and I. Trancoso, "The L2F broadcast news speech recognition system," 2010, fALA 2010.

[13] S. Paulo, L. C. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, and H. Moniz, "Dixi - a generic text-to-speech system for european portuguese," 2008, pROPOR 2008 - 8th International Conference on Computational Processing of the Portuguese Language.