LS3D:
LEGO Search Combining Speech and Stereoscopic 3D

Pedro B. Pascoal, INESC-ID/Técnico Lisboa, Universidade de Lisboa, Lisboa, Portugal

Daniel Mendes, INESC-ID/Técnico Lisboa, Universidade de Lisboa, Lisboa, Portugal

Diogo Henriques, INESC-ID/Técnico Lisboa, Universidade de Lisboa, Lisboa, Portugal

Isabel Trancoso, INESC-ID/Técnico Lisboa, Universidade de Lisboa, Lisboa, Portugal

Alfredo Ferreira, INESC-ID/Técnico Lisboa, Universidade de Lisboa, Lisboa, Portugal

ABSTRACT

The number of available 3D digital objects has been increasing considerably. As such, searching in large collections has been subject of vast research. However, the main focus has been on algorithms and techniques for classification, indexing and retrieval. While some works have been done on query interfaces and results visualization, they do not explore natural interactions. The authors propose a speech interface for 3D object retrieval in immersive virtual environments. As a proof of concept, they developed the LS3D prototype, using the context of LEGO blocks to understand how people naturally describe such objects. Through a preliminary study, it was found that participants mainly resorted to verbal descriptions. Considering these descriptions and using a low cost visualization device, the authors developed their solution. They compared it with a commercial application through a user evaluation. Results suggest that LS3D can outperform its contestant, and ensures better performance and results perception than traditional approaches for 3D object retrieval.

Keywords: *3D Object Retrieval, Immersive Visualization, Multimodal Interaction, Speech Recognition, User Interfaces, Voice Search*

INTRODUCTION

The appearance of low-cost technologies that allow scanning of three-dimensional physical objects, such as the Microsoft Kinect[1], Asus Xition[2] or PrimeSense Sensor[3], along with the vulgarization of 3D modeling software, has resulted in a considerable increase of available 3D virtual objects. An implicit consequence of this growth is the increased complexity in searching for a specific 3D model desired by the user, which leads to a slow and tedious retrieval process.

When performing the retrieval of 3D objects and other types of multimedia objects, the intrinsic information contained in them, such as the corresponding files names, has proved insufficient, and new meta-data is often needed (Funkhouser et al., 2003; Smith and Chang, 1997).

DOI: 10.4018/IJCICG.2015070102

With the purpose of overcoming this challenge, several solutions for performing retrieval have been proposed. Some resort to textual annotations (Funkhouser et al., 2003; Smith and Chang, 1997), sketches (Liu et al., 2013; Santos et al. 2008), verbal descriptions (Lee et al. 2010; Wang et al. 2011), gestures (Holz and Wilson, 2011) or using an object as an example (Lavoué, 2011; Paquet and Rioux, 1997). However, all the proposed solutions have drawbacks. For instance, retrieval by example requires the user to have a similar object to use as an example, which may not always be available. The remaining solutions do not properly explore the descriptive power and potential of human interaction. Although some studies (Holz and Wilson, 2011; Kamvar and Beeferman, 2010; Lee and Kawahara, 2012) have followed this direction, albeit not applied in the context of the 3D object retrieval, these do not combine the multimodality of speech and gestures used in natural human interactions. Moreover, the current solutions for searching 3D objects visually present results in a grid of thumbnails (Funkhouser et al., 2003). This may be an inadequate representation, since it loses relevant 3D information.

In order to explore the descriptions of 3D objects, we conducted an experimental session with users in order to understand which verbal and gestural expressions are naturally used. Our aim was to understand if users resort more to speech, gestures or a combination of both. For this purpose, a scenario of building LEGO models was designed, in which one of the subjects had to request the necessary blocks from another subject, describing them as accurately as possible. Based on the results of this experiment we developed a system in which users can search for 3D objects using multimodal interactions. Both exploration and result analysis is performed through a 3D immersive environment, in order to provide a clear representation of the virtual LEGO blocks, as we can see in Figure 1. The use of LEGO blocks has already proven to be a good context to explore new concepts such as interactions (Mendes et al. 2011; Santos et al., 2008) or object tracking (Gupta et al. 2012; Miller et al. 2012). Although we use the LEGO blocks context, our solution can also be extended to different contexts.

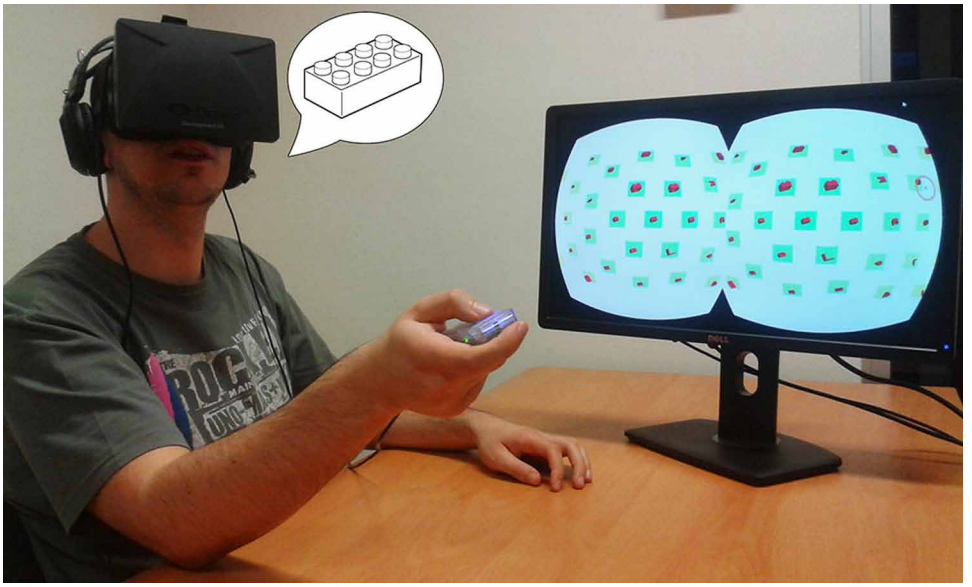
In the remaining of the paper we will discuss the state of the art for searching multimedia content, focusing on three-dimension content. Then we present the preliminary study and a discussion regarding its results. After the study it is presented our prototype, followed by an evaluation where we compare our prototype against a commercial application. Finally, we present our conclusions and point out some directions for future work.

RELATED WORK

With the increase of available digital objects of any type, retrieving specific information presents a challenge, and three-dimensional objects, such as any other type of multimedia content, are no exception. One of the traditional ways to perform retrieval consists of using textual queries. However, this method is not trivial, considering that objects do not usually contain sufficient intrinsic information. For instance, files name may not be related to the objects. 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. This concept has already been applied in image retrieval (Smith and Chang, 1997) and 3D object retrieval (Funkhouser et al., 2003). In the work of Funkhouser et al., synonyms of words taken from the texts are also used to of increase the information used to describe the 3D object and resolve vocabulary mismatch.

Despite all these solutions, the information to describe a 3D object is still insufficient, especially regarding its shape. Some of the proposed solutions use query-by-example to ease this process. The goal of search by example is to obtain similar objects in terms of visual aspects such as the

Figure 1. LS3D prototype



color (Paquet and Rioux, 1997) or shape (Lavoué, 2011). Moreover, this solution requires the user to already have an object identical to the one he is searching, which is usually not the case.

Retrieval by sketching addresses this problem, offering users the possibility of searching for similar shaped objects, for which they do not have a model to serve as an example (Pu et al. 2005). In the work of Santos et al. (2008), users can make sketches that match the dimensions of the desired objects which, in this case, are LEGO blocks. Funkhouser et al. also proposed a method of sketching several 2D views of the model. In a different approach, Liu et al. (2013) attempted to improve search by sketches taking into account the user profile, i.e. the habits of the users when drawing. This leads to improved results as users perform more searches.

Holz and Wilson (2011) followed a different approach in order to apply a method often used to describe physical objects. In their work, the authors focused on recognizing descriptions of three-dimensional objects through hand gestures and, with these descriptions, retrieve objects that are present in a database. This work consisted of capturing and interpreting gestures, exploring the spatial perception of the users. The shape and movement of the hands when the objects are described, are used to create a three-dimensional shape filling the voxels which user's hands crossed. The authors concluded that participants were able to keep the correct proportions relatively to the physical objects. Moreover, the authors observed that in areas with more details, gestures were performed more slowly.

Although this previous work explored mid-air gestures to describe objects, the presentation of results relied on the traditional approach of viewing the results, resorting to lists of thumbnails, like, Funkhouser et al. (2003). Nakazato and Huang (2001) presented 3D MARS, which demonstrates the benefits of using immersive environments for multimedia retrieval. Their work focused mainly on the presentation of query results for a content-based image retrieval system, using a CAVE like setup. Extending 3D MARS, Pascoal et al. 2012, showed that some challenges can be overcome when applying a similar approach for retrieving 3D objects. These results are distributed in the virtual space according to the similarity between them. Users can

then explore the results by navigating in the immersive environment, which is seen through a head-mounted display. Their system also allows a diversified set of different visualization and interaction devices, which was used to test multiple interaction paradigms for 3D object retrieval.

In the area of the information retrieval, we have recently witnessed a wide spread of search-by-voice on mobile devices. This new possibility has led to the preference of this search method over the traditional search by text (Kamvar and Beeferman, 2010). In most cases, when searching by voice, the speech is converted to text, which is then used as a search parameter (Lee et al., 2010; Wang et al., 2011). More recently, Lee and Kawahara (2012) performed a semantic analysis of the speech queries used to search for books, achieving a greater understanding of what the user desires to retrieve.

In short, the retrieval of multimedia content, in particular of three-dimensional objects, has been subject of previous research. Most solutions, although already started to explore more natural methods for describing objects, such as mid-air gestures, do not yet conveniently explore the potential of the interaction between humans and its descriptive power. In other areas, verbal descriptions are already being used for retrieving content. However, they have not yet been applied to 3D objects or complemented with other natural descriptive methods. When viewing retrieved results, some solutions provide immersive environments. Nevertheless, some of the results may appear overlapped if they are too similar. Traditional approaches do not overlap results, but their visualization based on thumbnails lacks an adequate 3D representation of objects.

PRELIMINARY STUDY

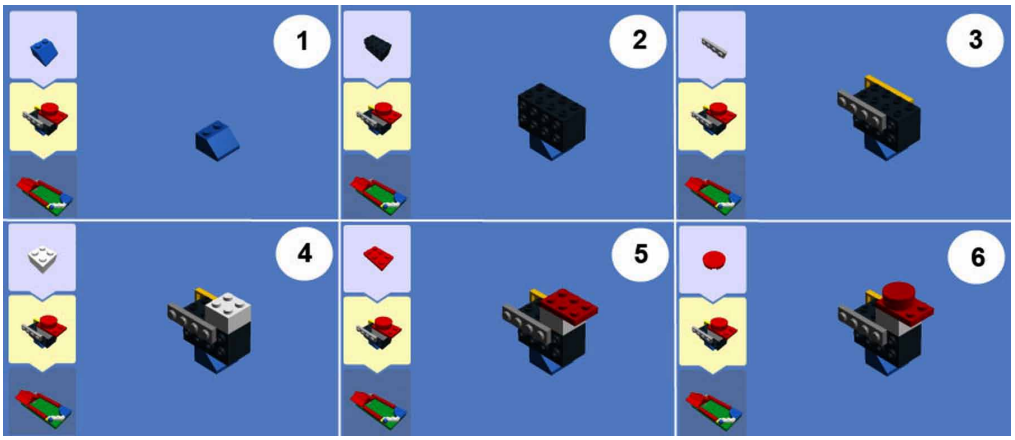
To understand which methods are most natural for people to describe three-dimensional objects, we conducted an experiment with ten pairs of participants. The scenario of building LEGO models was used, in which one participant had to request specific blocks from another, in order to build a model. Each participant performed two roles: once as a builder and once as supplier. After a preliminary introduction, step-by-step instructions to assemble a model were given to the builder, and the supplier was given a box of blocks, containing 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 experiment was performed with four different models, composed of different blocks, but with a similar geometric complexity. The instructions for each model included 20 steps. Figure 2 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 thought it was more suitable. The supplier had to search for the block and hand it to the builder. After completing their first model, the two subjects changed roles, repeating the process for a different model. After building the two models, they were asked to fill a short questionnaire.

The majority of the twenty participants (seven females) in our experiment were college students, whose ages ranged from 18 to 24 years old (55% of the participants), but older participants were also involved (all with less than 45 years). All participants owned a college degree. All of them were familiar with LEGO blocks. Except for four participants, they knew their partner prior to the experiment, a fact that originated very informal communications. All participants were native Portuguese speakers.

Results shown that eleven out of twenty participants did not use any gesture to describe a block. Other participants occasionally used gestures, but always only for the purpose of complementing the verbal description, without adding any significant information. For example,

Figure 2. Example of instructions provided to the builder, similar to those that are shipped with the original LEGO sets



to describe a corner piece, some participants used, the letter 'L' as a reference, making the corresponding gesture while referring the letter, as depicted in Figure 3. Another example of how users complemented their speech descriptions from time to time was by performing sketches in the air, in the shape of the desired block.

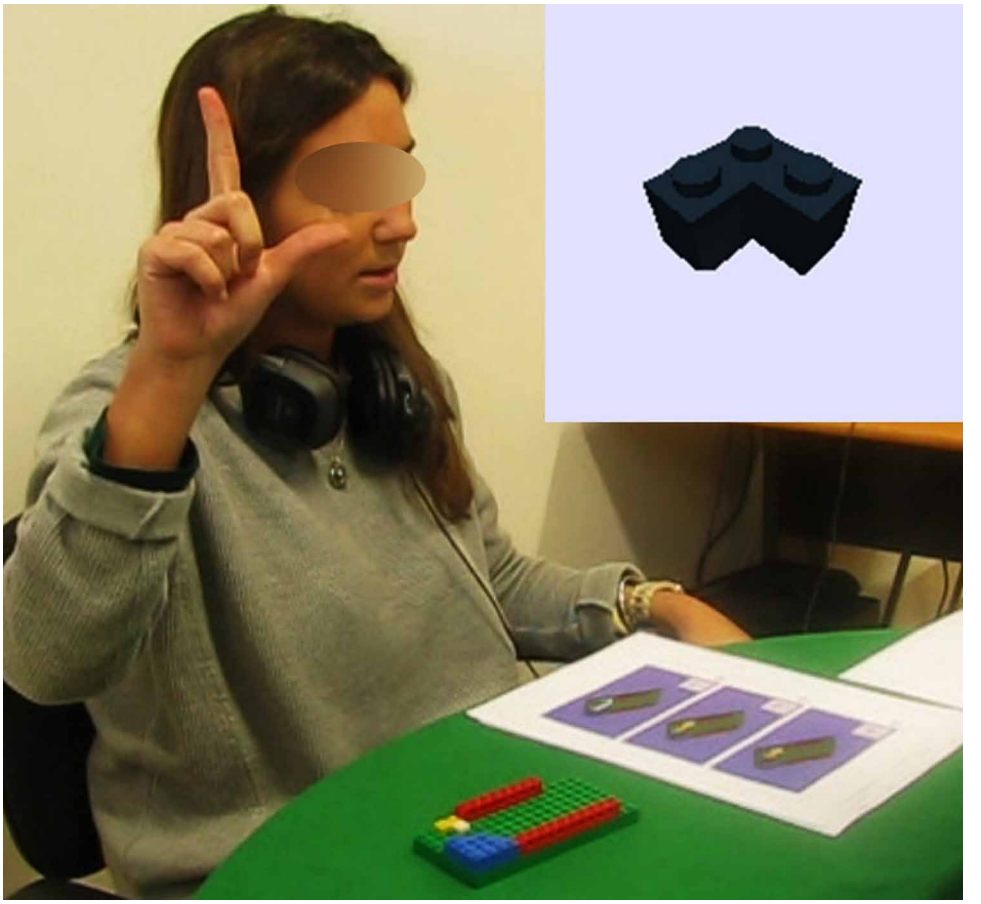
It was often observed that participants made several refinements to the descriptions. So, whereas most blocks took only a single step to be described, on average 1.4 descriptions were made for each step. On average, four steps per model had refinements to the original descriptions, but only two steps had refinements made because the supplier provided the wrong block. Even though half of these refinements were driven by the wrong block being returned, on several occasions the participants in the role of builder added more detail to the initial description by their own initiative.

The specification of each block usually began with its dimensions and color. Regarding dimensions, the unit used was the number of pins of the block, or the equivalent space for blocks without pins. Adjectives were used very often in order to avoid counting pins. When presented to a query such as "Now I want a long block, red, with lateral holes, and holes going from one side to the other.", the supplier just searched for the longest block with these characteristics, without counting the actual number of connectors or holes. 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.

Metaphors were also often used in descriptions. The most frequent ones were related to the already mentioned L-shaped blocks or roof-shaped, which took part in several models. The most unusual blocks were the ones that triggered the most creative descriptions, such as letters, teeth and the top of a trident.

Additionally, 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 easiness to describe objects. The classification was done using a Likert scale with four values (1-very difficult and 4-very easy). The participants showed a strong preference for using exclusively speech, compared to using gestures or a combination of modalities. Furthermore, the combination of gestures and speech was consistently preferred over using just gestures. More details on this experiment is provided by Henriques et al. 2014 (Interspeech).

Figure 3. Participant requests a corner block, using reference to the letter 'L'



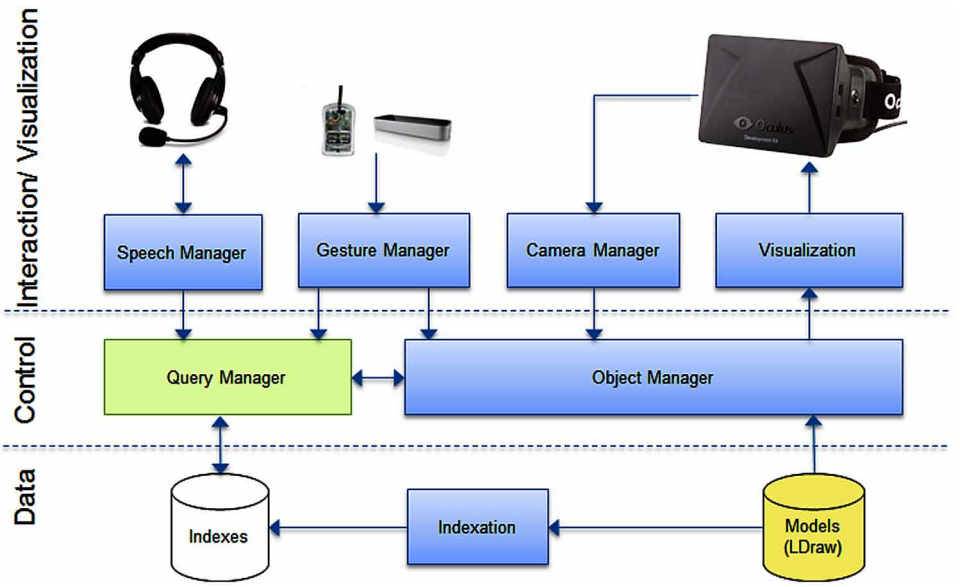
LS3D

After verifying that the participants clearly prefer describing the blocks through speech, we developed a prototype, LS3D, which provides a natural way of retrieving 3D objects. Our prototype used the context of LEGO as a toy-problem. To take advantage of speech descriptions, LS3D supports spoken queries, as well as it synthesizes text to speech to give feedback to the user. Also, we developed our prototype using an immersive environment, taking the advantage of recent devices of computer interaction and visualization.

Architecture

Our prototype was developed accordingly to the architecture presented in Figure 4. We devised three main layers with several modules: Interaction/Visualization, Control and Data. The layer Interaction/Visualization is where we implemented the Voice Interface (Speech Manager module). Here, the spoken queries of the user are recognized and audio feedback is synthesized to inform the user about what was understood by the system. We also implemented interactions with gestures (Gesture Manager module), allowing users to interact using the Leap Motion or

Figure 4. LS3D architecture



the Space Point Fusion devices. The Camera Manager controls the point of view of the user. Is also in this layer that the visualization of the objects as well the disposal of the retrieved results is done, in the Visualization module. For this purpose, we use OpenSG, an open source graph system that undertakes the conversion to OpenGL primitives and the rendering process.

The Control layer contains the Query Manager and Objects Manager modules. The Query Manager is where queries are created using users' input. The Objects Manager creates and displays objects returned as query results and allows the user to interact with them. Finally, the Data layer is where the information of the models is stored and indexed. The Indexing module uses both textual and shape descriptions. For the LEGO blocks' models we resorted to LDraw, an open-source standard and library that contains their geometric information and textual descriptions.

Indexing

Taking advantage of existing solutions for text matching, we applied the approach proposed by Gennaro et al. 2010, to 3D object retrieval. To extract the shape features we used the D2 descriptor that had shown promising results in previous works (Funkhouser et al., 2003). We extracted the shape features for each object, creating its own signature in a 128 dimensional feature vector. Based on the idea that similar objects will have similar view of the descriptors surrounding space (Gennaro et al., 2010), we selected a set of reference objects (RO) from the collection. The Euclidean distance of the signature of every object in the collection to each one of these RO is calculated. This distance is then used to sort all RO for each object. Assigning a unique term for each RO, we create a textual representation for an object repeating these terms more times the closer the respective RO is to the object.

We use 85 randomly selected RO from the collection to represent the 1810 objects, accordingly to the equation suggested by Amato and Savino (2008): $\#RO \geq 2\sqrt{\#X}$. To achieve an acceptable compromise between indexing and querying time and accurate results, we represent

each object with the 30 closest RO. This way, although comparing each object with 85 RO, we only use 30 in its representation, thus significantly reducing the size of the textual representations.

The generated representations in textual form of the objects, can be processed in a similar fashion as textual descriptions. For each object, its closest RO are associated in the shape index, giving a higher relevance to closer objects. For each RO, the objects of the collection that are closer to it are associated in the shape inverted index.

To retrieve objects in the collection, our system relies on the TF-IDF approach to calculate results relevance. We developed two types of queries: query-by-text and query-by-example.

The query-by-text follows a direct implementation of a standard search engine for textual documents. For each term in the query, the system will gather objects that contain the term from the corresponding inverted index, combining these results. Then, the similarity of each object to the query is calculated.

The query-by-example, is done selecting an example using both its textual description an shape information. Using both information we perform queries-by-text to the respective indexes. Since both approaches generate a similarity value contained in the interval [0, 1], they can be averaged, while keeping the final similarity value also on this interval. The main idea is that objects that appear similar to the query using both textual description and shape information will be more relevant.

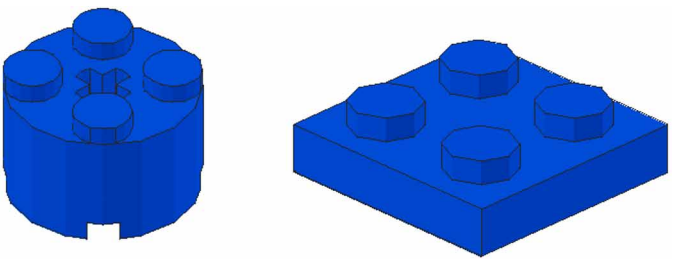
Query Specification

To query the system users can use spoken queries. For this purpose, the spoken interface was done integrating speech recognition (Meinedo et al. 2010) and synthesis (Paulo et al., 2008) modules. We built a grammar using GRXML[4], based on the most common expressions from the recordings of the preliminary study. The selected vocabulary included more than 500 different words, due to the inclusion of inflected forms of the words in the training data. This vocabulary covers 1810 different block shapes.

Our grammar allows users to describe a single block with more than one description. Users can describe blocks through their dimensions as using adjectives. For example considering the block in Figure 5 (left), which belongs to the class of rounds LEGO blocks, could be described with 14 different adjectives. The Figure 5 (right) which belongs to the class of plates LEGO blocks, can be described by 21 different adjectives.

In our system, every query starts with the keyword, “Acorda LEGO” (Wake up LEGO), to which the system verbally replies with a prompt “Sim”(Yes). 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). In our system, as every block can have 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

Figure 5. Examples of virtual LEGO blocks used



change its color in a new query, as in "Acorda LEGO. Pinta de encarnado" (Wake up LEGO. Paint it red). Given the very large amount of LEGO blocks, three additional strategies have been implemented to restrict the number of pieces shown. Users can verbally filter by a given characteristic (e.g. "Filtra por curva" (Filter by curved)), or exclude a type of blocks (e.g. "Exclui DUPLO" (Exclude DUPLO)) or get blocks similar to the selected one (e.g. "Dê-me semelhante a esta"(Give me similar to this one)).

Result Exploration

To view query results of the queries we implemented four modes of visualization, illustrated in Figure 6. We started by adapting the traditional grid approach to a 3D immersive environment, in order to be able to analyze its suitability. The Rectangular Grid shows results placed in vertical grid. The top left object in the grid is the higher ranked, while in the down right corner is placed the worst ranked one. The Square Grid is similar, but has equal height and width, being the higher ranked objects placed close to the center. Both these approaches appear in front of the user, within the three-dimensional space. We also implemented two approaches to better explore the three-dimensional space. With the Cylindrical and the Spherical approaches, we surround the user with objects, placing them on the surface of an invisible cylinder or sphere, respectively. In both these approaches, higher ranked objects are placed in front of the user, expanding from there.

In this stage navigation through results was done by freely moving user's head. To see beyond the user's field of view, or focus on a specific section of the results, users can place an open hand in mid-air in front of them and move it to rotate (in Cylindrical and Spherical approaches) or pan (in Rectangular Grid and Square Grid) results. When a detailed view of an object is desired, the user can point, with one finger, at the object. The pointed object then is brought closer to him and rotates, in order to be viewed in different angles. We used Leap Motion[5] to capture these mid-air gestures.

We evaluated these immersive gesture-based interactions against each other and a traditional 2D grid with thumbnails. The evaluation showed that immersive gesture-based interaction can compete with the traditional, which users are acquainted to. Participants agreed that immersive approaches are preferred over the traditional. Also participants mentioned the Cylindrical and Spherical as the preferred modes. More details on this experiment is provided by Henriques et al. 2014 (3DUI).

With this evaluation results in mind, we developed a new mode of visualization that combines these two preferred modes. The new mode, Barrel, shows results placed in a half of Barrel (Figure 7). This has the higher ranked objects placed in front of the user, expanding from there. Additionally, we added an ordered square to help understand the rank of the results, being green the higher ranked and red to last ranked, as show in Figure 8.

Figure 6. Modes for visualizing query results in an immersive environment: (a) Rectangular Grid; (b) Square Grid; (c) Cylindrical; (d) Spherical

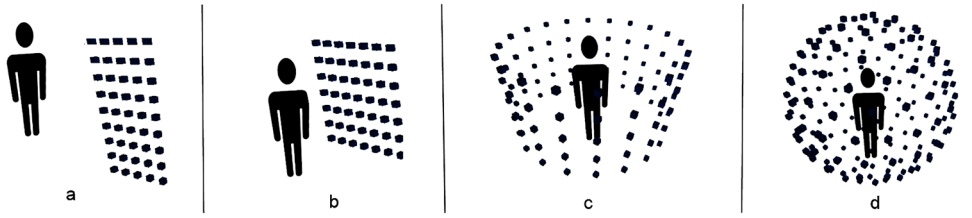
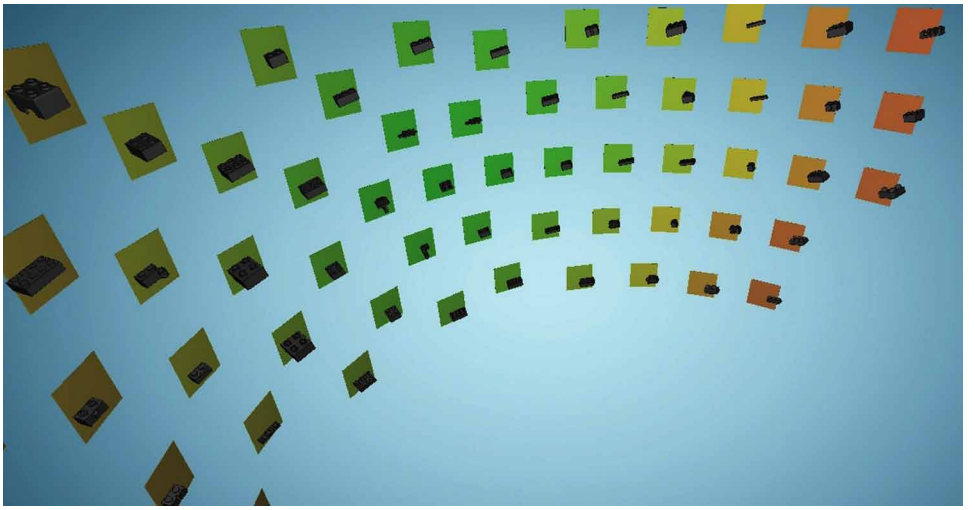


Figure 7. Mode of visualization Barrel



With the evaluation of the modes of visualization, we noticed that users often lost track of the volume where they could interact with gestures. Because of that, we explored different interactions using Space Point Fusion[6] device, which didn't present these restriction. With the gyroscope of this device, it is possible to point to a specific object (Figure 9) with a higher precision and without the restriction of the interaction volume of Leap Motion.

To select an object, users can press the left button of the Space Point Fusion to bring the object closer to the point of view of the user and rotate their hand to freely rotate the object, in order to be viewed in different angles, as illustrated in Figure 10. After the development of the interaction with Space-Point Fusion, we compared the two interactions through user tests. The test was executed by eight participants, 50% of them didn't have any experience with neither these devices, 25% had experienced with both devices and the remaining 25% just one of the devices. We verified that not only did users prefer to use Space Point Fusion and, they were faster performing the testes with it than with Leap Motion.

USER EVALUATION

To evaluate LS3D, we compared it against a commercial application, Lego Digital Designer (LDD). LDD uses a Windows Icons Menus and Pointing devices (WIMP) paradigm and the results are presented in a 2D traditional grid of thumbnails divided in categories (Figure 11).

PARTICIPANTS

The evaluation was carried out by twenty users (sixteen males and four females), whose ages ranged from 18 to 50 years old (with 70% between 24 and 29). All users were native Portuguese speakers, being most of them from Lisbon. Concerning previous experience, none of them knew the LDD application. They were, however, experienced with image search systems, using them on a daily basis. Only 30% of the users already used search engines for 3D objects. These engines represent objects by thumbnails. 60% of the users already tried systems that use spoken queries.

Figure 8. Disposal of objects by rank

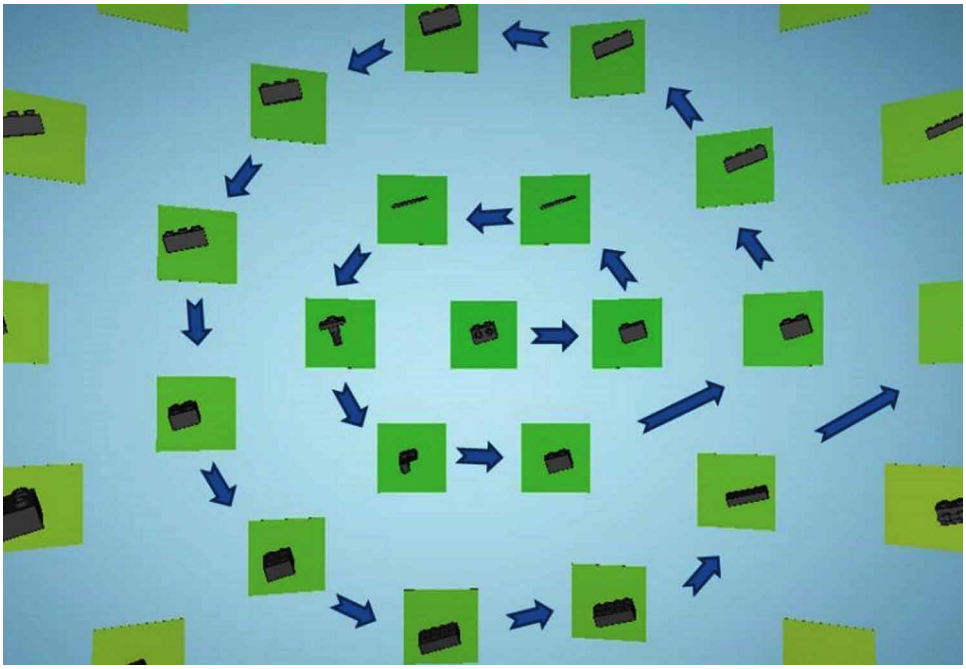


Figure 9. Selection of the objects using Space Point Fusion

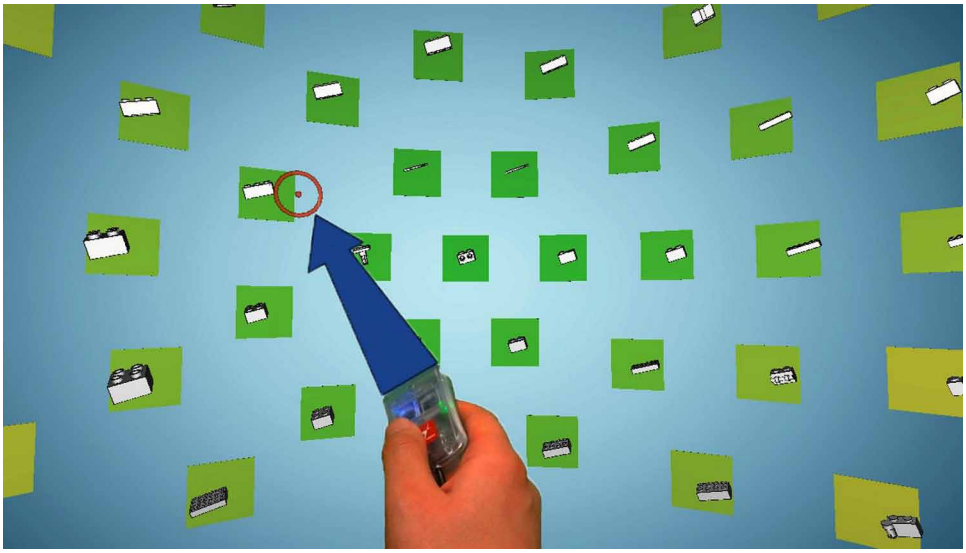
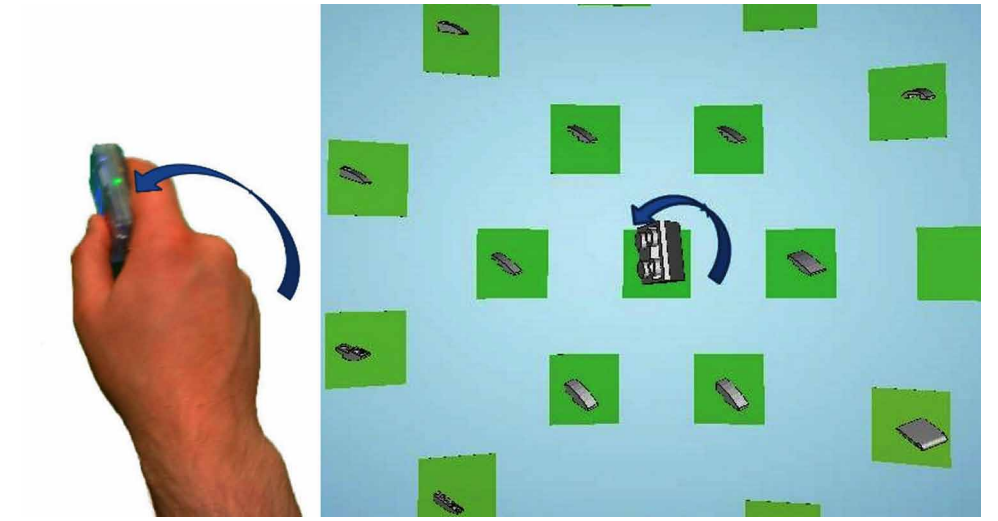


Figure 10. Rotation of the object according to rotation of Space Point Fusion



Regarding experience with stereoscopic devices, 60% of the users had already experienced and just 50% the users had previously tried Head Mounted Displays.

Setup

The tests were conducted in a controlled environment, without external influences. To run LDD, a computer with a standard screen with mouse was used, as shown in Figure 12 (left). LS3D was experienced using the SpacePoint Fusion for interactions, Oculus Rift for the visualization and a headset for voice interaction, as can be seen in Figure 12 (right).

Methodology

The user tests were structured in four stages: pre-test questionnaire to evaluate user profile and previous experience; briefing of the purpose of the tests; execution of the two tasks; and

Figure 11. Lego Digital Designer application

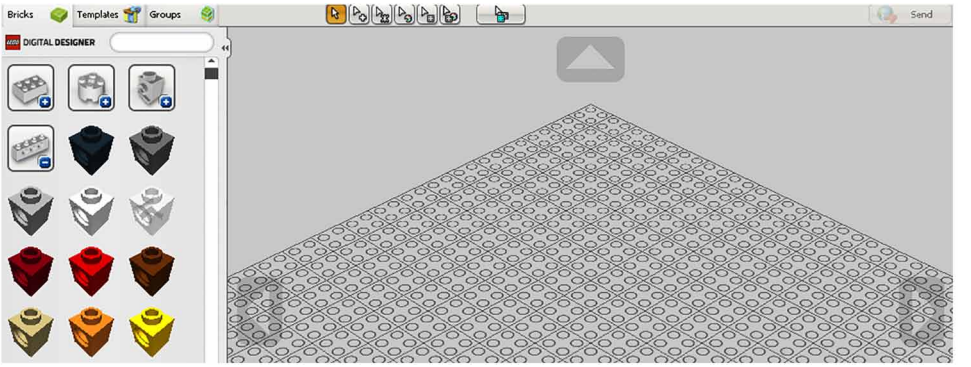


Figure 12. Participants in the evaluation session, using LDD (left) and LS3D (right)

questionnaires regarding the applications. To ensure even test distribution, the order of the two systems was selected randomly.

The experience with each system started with a brief adaptation time, in which users could experience its interface. After this, users were asked to search for eight blocks, which were shown to them using physical LEGO blocks (Figure 13). The order of the blocks was selected randomly, but different users would never do the same sequence, in order to ensure an even test distribution. After finishing searching for all blocks, users were asked to answer a small questionnaire about their experience regarding the evaluated systems.

RESULTS

We conducted three different perspectives on the analysis of the results from our user evaluation. In first place, we present a quantitative analysis taken from time taken for completing each search, as well as the number of wrong blocks selected. Then, a qualitative analysis based on the questionnaire answers. Finally, we discuss several observations captured over the test sessions.

Quantitative Analysis

To analyze results regarding time spent in each search, we used the Wilcoxon signed-ranks test, with which we concluded that statistically significant differences exist. Users did faster searches with LS3D than with LDD ($Z=-3.509$, $p=0.000$). In Figure 14, we can see the total time each user took to complete the eight searches, and the total of wrong blocks selected in both systems.

In the case of the number of wrong blocks selected, we also used the Wilcoxon signed-ranks test. We concluded that statistically significance differences exist. It was clear that LS3D solution was better than LDD ($Z=-6.697$, $p=0.000$), preventing more errors.

Qualitative Analysis

After completing the task in each system, users were asked to classify the search interface for both systems, using a 4 values Likert scale, regarding to: how fun it was to use; how easy it was to view the blocks; how easy it was to use. The Wilcoxon signed-ranks test was used to find statistically significant differences. Users strongly agreed that LS3D was the easiest to use ($Z=-2.441$, $p=0.015$). In our system, even with participants that did not used spoken queries managed to use it with success. Concerning the objects' visualization, users strongly agreed that LS3D

Figure 13. Set of eight LEGO blocks used

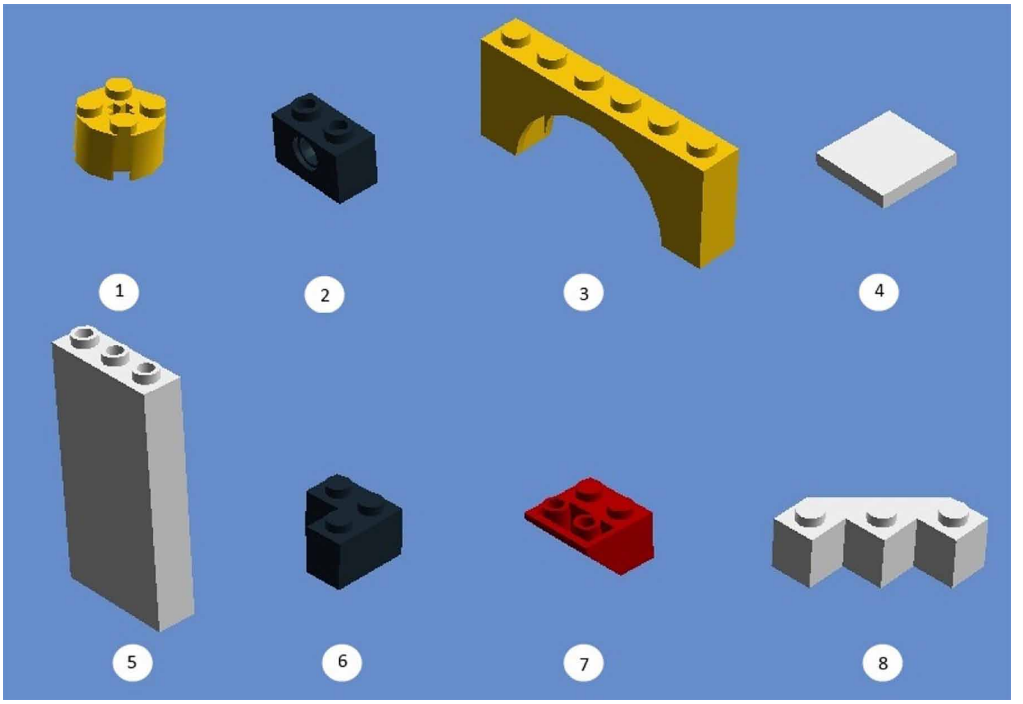
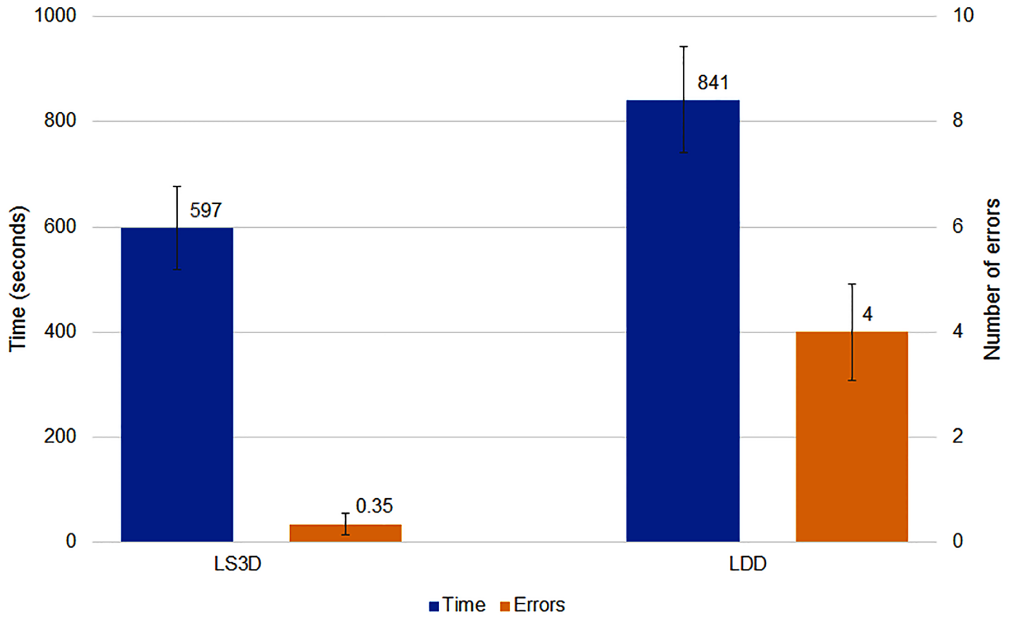


Figure 14. The total number of errors and time each user took from searching in both systems



was the easiest (Z=-2.780, p=0.005). Finally, they also strongly agreed that LDD was less fun than LS3D (Z=-3.626, p=0.000).

Observations

During the evaluations we noticed some relevant aspects from users' behavior as well as some comments made. Throughout the evaluation users mentioned that the division by categories of the LDD was helpful. However, this division failed in two of the selected blocks, where the blocks that represented the categories weren't similar at all to the desired object. This caused users to open all categories and browsing through them, which eased searching for the next blocks. Users that didn't started with one of these two blocks made comments reflecting some frustration while searching in the LDD, something that did not occurred in the LS3D, even on longer searches.

In LDD, users shown to have problems understanding the blocks depicted, which did not occur in LS3D. Comments from users suggested that our prototype gave to them a better perception of objects' size and details. This was particularly noticed in blocks without pins, where users often selected a block with a smaller size.

Furthermore, users commented that it was easy to use LS3D to search for objects by their dimension, despite the fact that some users mentioned their dimensions wrong. These mistakes where all noticed, and corrected when received the feedback from the system. We also noticed that users resorted more to metaphors and adjectives at their latter searches. When asked why this happened, they said that didn't know what words could be used at the beginning and at the end they already known the potential of our prototype's search.

We observed that users concluded that the refinements available in LS3D helped them to find the objects. Also, users mentioned that the exclusion refinement was the less needed, because the other two refinements, example and filter were enough. This was also noticed by the distribution of the refinements, where 55% of these were to filter, 36% examples and 9% exclusions.

CONCLUSION AND FUTURE WORK

With the easiness in creating 3D virtual content, the retrieval of specific objects within large libraries becomes an increasingly relevant challenge. One of these steps, is the query specification where users describe the object that they intend to retrieve. Existing solutions in the literature do not adequately explore natural ways for this specification.

In order to understand the more natural and simple methods to describe three-dimensional objects, we conducted a preliminary study, where we verified that participants preferred to use exclusively verbal descriptions. It was often observed that participants resorted to imaginative metaphors, comparing the most relevant characteristics of the blocks with other objects. Based on the knowledge attained, we developed our LS3D prototype, which uses spoken queries in an immersive environment, for retrieving virtual LEGO blocks.

We conducted an evaluation with twenty users, comparing our prototype with a commercial application of the LEGO Company. Results suggest that our prototype is able to compete with a traditional application that uses a 2D grid with thumbnails and, a more familiar interface to users. Through the results, we could conclude that our solution gave a better perception of the objects, being less susceptible to errors than the commercial application. Moreover, through users' comments and questionnaire analysis, we can conclude that most participants found it easy to use the search method of our solution, without extensive training.

As future work, we consider that would be interesting to invest in more sophisticated language models for multimedia content retrieval, as well as dialog management systems, sup-

porting error recovery strategies. Our experiment also showed that this setup can also be very suitable for entrainment studies.

In terms of showing query results, a possible improvement could be object clustering as available in LDD, but with more than one object describing each cluster. Also, our prototype gives an experience where the user is always in the same position. With the new version of Oculus Rift, coming up with a better resolution, and head tracking, we could bring better immersive environment without restricting users' position.

In our prototype, we used the LEGO blocks toy-problem as a proof-of concept. We believe that our solution can be applied to other scenarios, namely for automotive and construction industries.

ACKNOWLEDGMENT

The work described in this paper was partially supported by the Portuguese Foundation for Science and Technology (FCT) through the project TECTON-3D (PTDC/EEI-SII/3154/2012), doctoral grant SFRH/BD/91372/2012 and by national funds through FCT with reference UID/CEC/50021/2013.

REFERENCES

Amato, G., & Savino, P. (2008). Approximate Similarity Search in Metric Spaces Using Inverted Files. *Proceedings of the 3rd International Conference on Scalable Information Systems (pp. 28:1–28:10)*, ICST, Brussels, Belgium. Retrieved from http://dl.acm.org/citation.cfm?id=1459693.1459731

Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., & Jacobs, D. (2003). A search engine for 3D models. *ACM Transactions on Graphics*, *22*(1), 83–105. doi:10.1145/588272.588279

Gennaro, C., Amato, G., Bolettieri, P., & Savino, P. (2010). An Approach to Content-Based Image Retrieval Based on the Lucene Search Engine Library. *Proceedings of the European Conference on Digital Libraries*. doi:10.1007/978-3-642-15464-5_8

Gupta, A., Fox, D., Curless, B., & Cohen, M. (2012). DuploTrack: A Real-time System for Authoring and Guiding Duplo Block Assembly. *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (pp. 389–402). New York, NY, USA: ACM. http://doi.org/ doi:10.1145/2380116.2380167

Henriques, D., Mendes, D., Pascoal, P., Trancoso, I., & Ferreira, A. (2014). Poster: Evaluation of immersive visualization techniques for 3D object retrieval. Proceedings of the 2014 IEEE Symposium on 3D User Interfaces (3DUI) (pp. 145–146). http://doi.org/10.1109/3DUI.2014.6798862

Henriques, D., Trancoso, I., Mendes, D., & Ferreira, A. (2014). Verbal description of LEGO blocks. *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*.

Holz, C., & Wilson, A. (2011). Data miming: inferring spatial object descriptions from human gesture. *Proc. of the 2011 annual conference on Human factors in computing systems* (pp. 811–820). ACM. doi:10.1145/1978942.1979060

Kamvar, M., & Beeferman, D. (2010). Say What? Why users choose to speak their web queries. In Interspeech.

Lavoué, G. (2011). Bag of Words and Local Spectral Descriptor for 3D Partial Shape Retrieval. *Proceedings of the Eurographics Workshop on 3D Object Retrieval 3DOR*, Aire-la-Ville, Switzerland (pp. 41–48). Switzerland: Eurographics Association. doi:10.2312/3DOR/3DOR11/041-048

Lee, C., & Kawahara, T. (2012). Hybrid vector space model for flexible voice search. Proceedings of the 2012 Asia-Pacific Association Annual Summit and Conference on Signal Information Processing (APSIPA ASC) (pp. 1–4).

Lee, C., Rudnicky, A., & Lee, G. G. (2010). Let's Buy Books: Finding eBooks using voice search. *Proceedings of the Spoken Language Technology Workshop (SLT)* (pp. 85–90). IEEE.doi:10.1109/SLT.2010.5700827

Liu, Y.-J., Luo, X., Joneja, A., Ma, C.-X., Fu, X.-L., & Song, D. (2013). User-Adaptive Sketch-Based 3-D CAD Model Retrieval. *IEEE Transactions on* Automation Science and Engineering, *10*(3), 783–795. doi:10.1109/TASE.2012.2228481

Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I., & Neto, J. (2010). The L2F broadcast news speech recognition system. *Proc. of FAlA* (pp. 93–96).

Mendes, D., Lopes, P., & Ferreira, A. (2011). Hands-on Interactive Tabletop LEGO Application. *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology* (pp. 19:1–19:8). New York, NY, USA: ACM. doi:10.1145/2071423.2071447

Miller, A., White, B., Charbonneau, E., Kanzler, Z., & LaViola, J. J. (2012). Interactive 3D Model Acquisition and Tracking of Building Block Structures. *IEEE Transactions on* Visualization and Computer Graphics, *18*(4), 651–659. doi:10.1109/TVCG.2012.48 PMID:22402693

Nakazato, M., & Huang, T. S. (2001). 3D MARS: Immersive Virtual Reality for Content-Based Image Retrieval. *Proceedings of 2001 IEEE International Conference on Multimedia and Expo (ICME2001)*. doi:10.1109/ICME.2001.1237651

Paquet, E., & Rioux, M. (1997). Nefertiti: a query by content software for three-dimensional models databases management. Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling (p. 345). Washington, DC, USA: IEEE Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=523428.825366 doi:10.1109/IM.1997.603886

Pascoal, P., Ferreira, A., & Jorge, J. (2012). In M. Spagnuolo, M. Bronstein, A. Bronstein, & A. Ferreira (Eds.), *Towards an Immersive Interface for 3D Object Retrieval* (pp. 51–54). Cagliari, Italy: Eurographics Association. Doi:10.2312/3DOR/3DOR12/051-054

Pascoal, P. B., Ferreira, A., & Jorge, J. (2012). Im-O-Ret: Immersive object retrieval. Proceedings of the Virtual Reality Short Papers and Posters (VRW), 2012 (pp. 121–122). IEEE. doi:10.1109/VR.2012.6180912

Paulo, S., Oliveira, L. C., Mendes, C., Figueira, L., Cassaca, R., Viana, C., & Moniz, H. (2008). DIXI --- A Generic Text-to-Speech System for European Portuguese. *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language* (pp. 91–100). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-540-85980-2_10

Pu, J., Lou, K., & Ramani, K. (2005). A 2D Sketch-Based User Interface for 3D CAD Model Retrieval. *Computer-Aided Design and Applications*, *2*(6), 717–725. doi:10.1080/16864360.2005.10738335

Santos, T., Ferreira, A., Dias, F., & Fonseca, M. J. (2008). Using Sketches and Retrieval to Create LEGO Models. *Proceedings of the Fifth Eurographics Conference on Sketch-Based Interfaces and Modeling,* Aire-la-Ville, Switzerland, Switzerland (pp. 89–96). Eurographics Association. doi:10.2312/SBM/SBM08/089-096

Smith, J. R., & Chang, S. F. (1997). Visually searching the web for content. *MultiMedia*, IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=621578

Wang, R., Paris, S., & Popović, J. (2011). 6D Hands: Markerless Hand-tracking for Computer Aided Design. *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 549–558). New York, NY, USA: ACM. doi:10.1145/2047196.2047269

ENDNOTES

1 Microsoft Kinect, http://www.xbox.com/en-US/kinect
2 ASUS Xtion, http://www.asus.com/Multimedia/XtionPRO/
3 PrimeSenses Senor, http://www.primesense.com/solutions/sensor/
4 Speech Recognition Grammar Specification, http://www.w3.org/TR/speech-grammar/
5 Leap Motion, https://www.leapmotion.com/
6 PNI Sensor Corporation, the SpacePoint Fusion, http://www.pnicorp.com/products/spacepoint-gaming

Pedro B. Pascoal is a PhD student at Instituto Superior Técnico since 2013. He received his MSc (2011) degree in Information Systems and Computer Engineering from Technical University of Lisbon (Instituto Superior Técnico), during which time he also worked in the Visualization and Intelligent Multimodal Interfaces Group (VIMMI) at INESC-ID for the 3DORuS project. He is also a researcher at Microsoft Language and Development Center (MLDC), Lisbon, Portugal. His main research interests are multimedia information retrieval, information visualization and multimodal interaction.

Daniel Mendes is a PhD student at Instituto Superior Técnico since 2013. He received his MSc (2011) and BSc (2008) degrees in Information Systems and Computer Science from Instituto Superior Técnico / Univertsity of Lisbon. He is also a researcher in the Visualization and Intelligent Multimodal Interfaces Group (VIMMI) at INESC-ID Lisbon. His main interest areas are virtual reality, multimodal interfaces, 3D user interfaces and touch / gesture-based interactions.

Diogo Henriques received his MSc (2014) degree in Information Systems and Computer Science from Instituto Superior Técnico / Univertsity of Lisbon. His dissertation focused on multimodal interfaces for 3D object retrieval.

Isabel Trancoso received the Licenciado, Mestre, Doutor, and Agregado degrees in electrical and computer engineering from Instituto Superior Técnico, University of Lisbon, Portugal, in 1979, 1984, 1987, and 2002, respectively. She has been a Lecturer at this University since 1979, having coordinated the EEC course for 6 years. She is currently a Full Professor, Chair of the EEC Department where she teaches speech processing courses. She is also a Senior Researcher at INESC-ID Lisbon, having launched the speech processing group, now restructured as L2F, in 1990. Her first research topic was medium-to-low bit rate speech coding. In October 1984–June 1985, she worked on this topic at AT&T Bell Laboratories, Murray Hill, New Jersey. Her current scope is much broader, encompassing many areas in spoken language processing, with a special emphasis on the Portuguese language. She was a member of the ISCA (International Speech Communication Association) Board (1993–1998), the IEEE Speech Technical Committee and the Permanent Council for the Organization of the International Conferences on Spoken Language Processing. She was elected Editor in Chief of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2003–2005), Member-at-Large of the IEEE Signal Processing Society Board of Governors (2006–2008), Vice-President of ISCA (2005–2007) and President of ISCA (2007–2011). She chaired the Organizing Committee of the INTERSPEECH'05 Conference that took place in September 2005, in Lisbon. She launched and coordinates the Dual Degree Ph.D. Program of the Carnegie Mellon Portugal Partnership in Language Technologies. She chaired the IEEE James L. Flanagan Speech and Audio Processing Award Committee, and is a member of the IEEE Fellows Committee, of the IEEE Publication Services and Products Board, and of the Editorial Board of the IEEE Signal Processing Magazine. She chairs the Distinguished Lecturer Selection Committee of ISCA and is part of the ISCA Advisory Committee. She is Vice-President of the European Language Resources Association. She received the IEEE Signal Processing Society Meritorious Service Award in 2009, was elevated to IEEE Fellow in 2011, and to ISCA Fellow in 2014.

Alfredo Ferreira is an Assistant Professor at the Instituto Superior Técnico (IST), Technical University of Lisbon. He received his Ph.D. (2009), MSc (2005) and BS (2002) degrees in Information Systems and Computer Science from IST/TU Lisbon. He is also a researcher of the Visualization and Intelligent Multimodal Interfaces Group at INESC-ID since 2002. He works on 3D object analysis classification and retrieval, virtual and augmented reality, natural user interfaces, immersive environments and augmented reality. He participated in the SmartSketches, Eurotooling21 and MAXIMUS projects, funded by EC. His involvement in these projects, focused on researching multi-user, large-scale and multi-display environments, and sketch-based interfaces. At the national level he participated in the DecorAR project and is now responsible for the 3DORuS project, which focuses in 3D object retrieval. He is also involved in the nationally funded MIVis and CEDAR projects, researching user interfaces for procedural modelling and design review.