3D object recognition through a process based on semantics and consideration of the context
Jean-Jacques Ponciano, Frank Boochs, Alain Trémeau

To cite this version:
Jean-Jacques Ponciano, Frank Boochs, Alain Trémeau. 3D object recognition through a process based on semantics and consideration of the context. Photogrammetrie, Laserscanning, Optische 3D-Messtechnik, Beiträge der Oldenburger 3D-Tage 2020, 2020. hal-02651917

HAL Id: hal-02651917
https://hal.science/hal-02651917
Submitted on 29 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
3D object recognition through a process based on semantics and consideration of the context

Jean-Jacques PONCIANO, Frank BOOCHS, Alain TRÉMEAU

Abstract

In the domain of computer vision, object recognition aims at detecting and classifying objects in data sets. Model-driven approaches are typically constrained through their focus on either a specific type of data, a context (indoor, outdoor) or a set of objects. Machine learning-based approaches are more flexible but also constrained as they need annotated data sets to train the learning process. That leads to problems when this data is not available through the specialty of the application field, like archaeology, for example. In order to overcome such constraints, we present a fully semantic-guided approach. The role of semantics is to express all relevant knowledge of the representation of the objects inside the data sets and of the algorithms which address this representation. In addition, the approach contains a learning stage since it adapts the processing according to the diversity of the objects and data characteristics. The semantic is expressed via an ontological model and uses standard web technology like SPARQL queries, providing great flexibility. The ontological model describes the object, the data, and the algorithms. It allows the selection and execution of algorithms adapted to the data and objects dynamically. Similarly, processing results are dynamically classified and allow for enriching the ontological model using SPARQL construct queries. The semantic formulated through SPARQL also acts as a bridge between the knowledge contained within the ontological model and the processing branch, which executes algorithms. It provides the capability to adapt the sequence of algorithms to an individual state of the processing chain and makes the solution robust and flexible. The application of this approach to data acquired by low-cost technology for the recognition of small objects shows its flexibility and its capability of adaptation to the recognition of different objects in different contexts.

1 Introduction

Object recognition requires to detect the representation in the data of the objects and geometries of the digitized scene. The detection of objects and geometries depends on their characteristics (e.g., size, shape) on the one hand, and the characteristics of the data on the other hand (e.g., density, noise, occlusion, roughness). The characteristics of the data depend on the acquisition process (e.g., technology and methodology used) that generates them. The characteristics of the digitized objects (e.g., material, reflectance, roughness, size), the context of the scene (e.g., urban outdoor, indoor building, ruin excavation), and various other factors external to the acquisition process (e.g., ambient light, light intensity, weather conditions, movement of the measuring instrument or digitized objects) that influence the acquisition process. Small variations in one of these factors can strongly influence the characteristics of the data. These variations in characteristics generate divergences between the expected
and obtained characteristics. Therefore, research work conducted in Reference (Ponciano JJ, 2019.d) has searched for a 3D object recognition able to adapt itself to the diversity of data and their context.

1.1 Related work

The two main categories of approaches for 3D object recognition in the literature are the Model-driven approaches and the Data-driven approaches.

The Model-driven approaches such as (VOSSELMAN & DIJKMAN, 2001; BORRMANN et al., 2011; PONCIANO et al., 2016; ANAGNOSTOPOULOS et al., 2016; SCHNABEL et al., 2007) mainly consider part of the information on the scene context and the objects contained in the data. These approaches use information about the geometry and shape of objects to build corresponding models. These approaches examine the data to identify sections that are similar to the models constructed. However, these approaches use only part of the information on the acquisition context and do not exploit information on acquisition methodologies or acquisition technologies. These approaches are, therefore, not able to compensate for weaknesses in data quality. The results of these approaches depend directly on the quality of the data and are specific to the object searched.

The category of data-driven is largely led by Machine Learning approaches that provide efficient results. The approaches based on Machine Learning, such as (DAI et al., 2016; DONG et al., 2014; GIRSHICK, 2015; HE et al., 2016). provide better flexibility than model-driven approaches. They substitute the lack of understanding of these factors through the identification of reliable patterns. These patterns have to be learned in a training stage. That is why "reliable" Machine Learning requires a vast amount of data. However, in some application domains, such as cultural heritage, annotated data are not available or are not sufficient. Thus, the lack of annotated data makes Machine Learning irrelevant to detect objects.

Moreover, Machine Learning approaches work as long as the data considered is representative of the content to be understood. The more variation in object or appearance occurs, the more data is required for training. Machine learning approaches remain unflexible and are unable to detect the object or the geometry for which they are not trained. Furthermore, just a small change of the acquisition process or external factors that influence the data can result in a change for which the Machine learning approaches are not trained.

1.2 Proposed solution

To overcome the limits of approaches from literature, we present a fully semantic-guided approach that attempts to understand the origin of data characteristics by understanding the influences of the characteristics of the acquisition process and the different factors influencing it. This understanding requires explicit knowledge of the domains of data, scene, and data processing, as well as knowledge of the influence between these domains. The knowledge about these domains and their influences allows for establishing an efficient reasoning base.

Therefore, the presented approach uses this explicit knowledge to guide data processing. Computer Vision algorithms from various libraries such as PCL (RUSU & COUSINS, 2011) or OpenCV, are used to process the data. Similarly, Semantic Web technologies such as SPARQL (HOMMEAUX & SEABORNE, 2008) and OWL2 (GRAU et al., 2008) allow for knowledge management. The approach presented combines the knowledge paradigm and the
data processing paradigm through both a technical and a conceptual bridge. This bridge allows knowledge to fully drive the data processing process step by step, through a continuous exchange between the two paradigms. This continuous exchange allows the object and geometry detection process to be adapted according to the newly acquired knowledge at each step of the data processing.

This solution has yet shown its capability to be applied to application contexts such as archaeological context (PONCIANO et al., 2019.b), which make difficult the 3D object recognition for other approaches. Its capability of object description adaptation according to a context and its performance compared to other approaches has been demonstrated in (PONCIANO et al., 2019.c) for the detection of rooms in an indoor point cloud. However, its efficiency in these two application cases has been shown on data acquired by a high technology of acquisition, which is the laser scanner and on big object recognition that are the floor, walls, and rooms. Therefore, we present the efficiency of this solution on data acquired by a low-cost technology and a recognition of smaller objects by considering the internal context of the scene. The use case of this article is presented in the next section.

1.3 Use case

The use case used in this paper aims to recognize specific objects in different indoor scenes for monitoring. In this use case, the objects sought are furniture such as chairs and tables, whose shape and color differ from one scene to another. Thus, the geometric and physical characteristics of objects are diverse. The scenes chosen for this use case are digitized by (LAI et al., 2011) using Microsoft’s Kinect that produces 3D point clouds. Each point cloud is composed of more than 1 million points. The use of Microsoft’s Kinect produces low-quality of the data set with mostly irregular surfaces.

Moreover, the acquisition process of these point clouds produces many missing parts due to occlusions and sparse acquisition. These characteristics increase the difficulty of object detection. The main challenge, in this use case, is to detect objects even if their characteristics and shapes change considerably (e.g., round table, square table, office chair, wooden chair).

2 Methodology

The proposed approach is fully guided by semantics. That is why its main component is an ontological model contained into a knowledge base and representing the semantic knowledge. Experts provide this ontological model of knowledge, which is further explained in section 2.1. The second main component is the set of algorithms that allows the processing of unstructured data. SPARQL queries play the role of a bridge between the semantic knowledge and the algorithms. Indeed, these queries allow for executing algorithms according to the semantic knowledge and enriching the knowledge with the semantic representation of algorithm execution results. Results of algorithm execution integrated into the knowledge base are the base for logical reasoning, allowing for classifying segments and objects. These two steps of algorithms execution through SPARQL queries and logical reasoning constitute the main steps of the detection process, presented in section 2.2. The detection process is followed by a knowledge-based self-learning, presented in section 2.3, which improves the data understanding. Figure 1 illustrates the components of the methodology and their interactions.
2.1 Knowledge modeling

The ontological model aims at representing knowledge of objects representation inside the data sets and of the algorithms which address this representation. Therefore, the knowledge modeling concerns the data domain, the scene domain (i.e., description of objects linked to their context), the data processing domain (i.e., description of algorithms with their parameters and requirements), and the relationships between these three domains. The knowledge about the factors that influence the data characteristics is described using rules of inference and SPARQL queries. Such rules and queries allow the system to deduce the different possible representations of objects or to delimit the search field of this type of object. This enrichment of the knowledge base aims at providing a knowledge description more adapted to the processed use case. Then, the knowledge base is used to guide the process of detection. The more precise the description of objects is, the better their detection in the data is. It is, therefore, essential to enrich the knowledge base with accurate object representations adapted to the context of the data. It is essential to provide an optimal knowledge model that gives a maximum of information for guiding the detection process.

2.2 Detection process

The detection process is a cycle composed of a data processing step and a classification step. The data processing step consists in selecting, configuring, and executing algorithms. If an algorithm detects a characteristic of an object (e.g., size, shape, orientation), then it is selected for the use case considered. Moreover, algorithms that meet the prerequisites of a selected algorithm are also selected (e.g., normal estimation algorithm for a prerequisite of estimated normal or denoising algorithm for a prerequisite of denoised data) as it is explained in Reference (PONCIANO et al., 2017). The selected algorithms are configured on the basis of available knowledge about their parameters and data characteristics. Algorithms, whose parameters have been configured and prerequisites have been met, are then executed through SPARQL queries. The SPARQL query that executes an algorithm retrieves a semantic representation of the algorithm results to add it to the knowledge base and classify it automatically. Such a detection process based on the parameters and prerequisites configuration allows for providing the capability to adapt the sequence of algorithms to an individual state of the processing chain.

After having considered the context of a scene to adapt the object description (c.f. different walls representation according to archeological scene (PONCIANO et al., 2019 .b), outdoor
scene (PONCIANO et al., 2019.a) or indoor scene (PONCIANO et al., 2019.c)), the process has been optimized by considering the internal context of the scene. The internal context of the scene corresponds to the description of topological links between the objects of the scene. Such consideration of the internal context aims at applying a hierarchical strategy of detection. This strategy depends thus, on the topological links between the objects and their size. The topological relationship provides information about the location of an object, according to others. This information allows for reducing the area of object search, and the size determines the priority of object detection (i.e., the biggest objects obtain the highest priority). For example, a can is defined as "being on a table," and a table is defined as "being on the floor". A can is smaller than a table and a table smaller than the floor; the detection process begins by detecting the floor, then the table, and finally, the can. According to the description of knowledge of the use case, the detection process begins with the floor detection. Then it detects objects on the floor such as walls, tables, chairs. Among these objects on the floor, the walls are the biggest objects. Therefore, they are detected in priority. Finally, it detects objects relatively link to other objects such as cans, bowls, cereal boxes, and cups located on a table for the use case.

**Figure 2:** Illustration of topological hierarchy between objects.

### 2.2.1 Dynamic Segmentation

According to the hierarchical strategy, the detection process begins by detecting floors. Floors are defined as having a single orientation (i.e., horizontal) and as segments (i.e., data portion), so it is necessary to use segmentation algorithms to divide the data into segments. Moreover, floors being defined by their orientation specified by a normal, the algorithm of Normal Region Growing is selected as relevant for detecting floor segments. If a characteristic of color has defined a floor, then a color region growing would have been selected. However, the normal region growing algorithm requires filtered normal. That is why the normal filtering algorithm (that filters the data according to the orientation of the objects sought) is selected. This filtering algorithm requiring a point cloud with estimated normal, the algorithm of normal estimation is also selected. Then, information about data characteristics allows for configuring the normal estimation, which satisfies the constraint of execution. The results of its execution enrich the knowledge base, allowing for the algorithm of normal filtering to satisfy, in turn, the execution constraints. The normal filtering algorithm is configured according to the orientation of the searched objects. It produces new data composed only of the portions having an orientation corresponding to the searched objects. It allows, thus, reducing the search field of objects.
Finally, thanks to the execution results of normal filtering, the normal region growing is configured and executed on the use case studied. Figure 3 shows step by step the floor segmentation process through the consecutive execution of algorithms.

Similarly, various algorithms for estimating characteristics such as dimensions (height, width, length), volume, area, or topological relationships (e.g., parallel, perpendicular, in contact) are applied to the segments. The results of such algorithms are integrated into the description of the segments contained in the knowledge base. The knowledge about the segments is then analyzed in the classification process to identify objects in the data.

2.2.2 Dynamic Classification

The ontological model is formalized through OWL2, which allows classifying a segment by the application of automatic logical reasoning. The classification of segments into objects is based on the semantic description of the objects, which is composed of geometric characteristics, object-specific characteristics, and topological links. A segment is thus, classified as an object if it satisfies the constraints defined for this object.

After each algorithm execution and the adding of its results into the knowledge base, the reasoning is applied to classify all elements that can be identified by the adding of information resulting from the algorithm execution. In the use case, a floor is defined as a horizontal plane with a surface area greater than 1 square meter. After the execution of the normal region growing, the horizontal segments resulting from this algorithm are classified as horizontally oriented planes. After the execution of the feature extraction algorithms, each of these segments is characterized. The surface and volume characteristics are automatically calculated from the segment dimensions. Thus, among the segments identified as horizontal planes, those with an area greater than 1 square meter are classified as floors.

2.3 Learning stage

The application of the detection process provides the first detection results. These detection results depend strongly on the knowledge modeling. Therefore, objects that do not respect the object modeling are not classified and thus, not identified. Therefore, a knowledge-based self-learning has been developed to fill in this lack produced by the approach. This self-learning aims at using the result of the first detection process to adapt the knowledge modeling to the specific data context and to allow better detection. This learning stage analyzes all characteristics of objects in a category to make a hypothesis corresponding to hypothetical
object modeling. The hypotheses are then tested to assess their validity. The process of hypothesis design and assessment is further explained in Reference (Ponciano et al., 2019, p. 16). The new object descriptions are finally used to apply a new detection process. The repetition of the detection process followed by a learning stage is applied until no improvement can be drawn from the learning stage. Such a learning stage can, for example, allow for identifying a can on the floor, whereas a can was first modeled as on a table.

### 3 Results

The use case used is composed of four indoor scenes representing a workroom and three lounging rooms. Figure 4 presents for each of these scenes, its original part of the point cloud, and the visual results of its understanding. In these scenes, three main categories of objects are described in the knowledge base. The first category of objects are objects defining a room as floors (in green in Figure 4) and walls (in blue in Figure 4). The second category of objects are desk furniture as seats (in bright green in Figure 4), chairs (in dark green in Figure 4), tables (in orange in Figure 4), computer tower (in pink in Figure 4), and boxes (in bright purple in Figure 4). The last category of objects is kitchen elements, which are on tables, like cereal boxes (in blue-purple in Figure 4), cups (in purple in Figure 4), cans (in brown in Figure 4), and bowls (in grey in Figure 4). Not classified points, which do not belong to noise, appears in black in Figure 4, and also in red in Figure 4 (b).

Visually, two sources of unclassified points can be observed. Figure 4 (b) highlights the two different sources of not classified points. The first source resulting in black points corresponds to segments that the approach has not classified due to a lack of knowledge description of these objects. For example, on the right of Figure 4 (b), a backpack appears as a black shape. The no-understanding of this segment is due to the absence of a semantic description of a backpack. The second source of classification lack that is highlighted in red corresponds to ambiguous segments. These segments could be classified into two object types and are therefore not classified neither in both types to avoid inconsistency of the knowledge base, neither in one type to avoid arbitrary choice that conducts to wrong classification. These red segments can also be not classified because they do not fit with the semantic description. For example, a table tray is described semantically as a horizontal plane; however, the table has round edges. Therefore, the process detects the table by considering only the plane surface of the tray and not the round edges. These red segments can be considered as errors because they should be classified into an object type. Thus, these results illustrate that the approach produces "edge effects" in the data understanding process. Errors produced by this approach have been quantified and presented in Table 1.
Figure 4: Results on the use case: the same color represents each object classified as belonging to the same type, elements colored in black and red are unclassified elements.
Table 1 provides three pieces of information about the results. Firstly, it provides the percentage of classified objects according to objects described in the ontological model and present in the scene. Secondly, it gives the distribution of points between points representing noise, classified, and unclassified points. Finally, it shows an estimation of success and fails percentage according to the proportion of classified and unclassified points, respectively, without considering points belonging to noise. The percentage of classified objects shows that all considered objects (in a sense, semantically described in the knowledge base) have been detected in all scenes. The percentage of noisy points in each scene highlights variations of noise among the different scenes. Scene 4 is particularly noisy, with around 15\% of noisy points. This noise can impact the result of the approach since the approach obtains its higher fail percentage in this scene. However, despite the noise, the approach succeeds in detecting all objects of this scene. Also, the approach has an accuracy of 94\% of points successfully classified. Moreover, the approach shows good robustness to the noise around 5\% by obtaining an accuracy of around 99\% for the understanding of scene 2 and 3. Finally, the approach obtains an average accuracy of 97.66\% for the understanding of this application case, and thus, a lack of 2.23\% of data understanding.

### Table 1: Results of objects detection in scenes of the use case

<table>
<thead>
<tr>
<th>Studied criteria</th>
<th>Scene 1</th>
<th>Scene 2</th>
<th>Scene 3</th>
<th>Scene 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of detected object</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Percentage of noise</td>
<td>0.57%</td>
<td>5.43%</td>
<td>4.90%</td>
<td>14.98%</td>
<td>6.47%</td>
</tr>
<tr>
<td>Percentage of classified points</td>
<td>97.12%</td>
<td>93.50%</td>
<td>94.91%</td>
<td>80.16%</td>
<td>91.42%</td>
</tr>
<tr>
<td>Percentage of not classified points</td>
<td>2.31%</td>
<td>1.07%</td>
<td>0.19%</td>
<td>4.86%</td>
<td>2.11%</td>
</tr>
<tr>
<td>Success (without considering noise)</td>
<td>97.67%</td>
<td>98.87%</td>
<td>99.8%</td>
<td>94.3%</td>
<td>97.66%</td>
</tr>
<tr>
<td>Fail (without considering noise)</td>
<td>2.33%</td>
<td>1.13%</td>
<td>0.2%</td>
<td>5.7%</td>
<td>2.34%</td>
</tr>
</tbody>
</table>

### 4 Conclusion

This paper has presented an approach fully guided by semantics.

This approach uses information from (i) acquisition technology (i.e., Microsoft’s Kinect) to adapt the choice of algorithms and their configuration to noised data, (ii) application context (i.e., indoor) to adapt object description, and (iii) the internal scene context through topological relationship between objects for a hierarchical strategy of detection. This context-related
information is inferred from the ontological model containing semantic modeling of objects, data, data processing, and their relation. Thanks to SPARQL queries, the ontological model guides the choice and the execution of algorithms. SPARQL queries also allow for enriching the ontological model with the semantic representation of the executed algorithm results. Reasoning on these enrichments classifies segments and objects automatically. This detection process is improved by a learning stage that adapts the knowledge modeling to better fit with the representation of objects in the data. The use case presented in this paper has highlighted an edge effect limit of the approach for some objects. However, the quantity of error remains low, although it is estimated by aggregating edge effects and points belonging to categories of objects that are not semantically described. It has also allowed for quantifying the impact of noise on the approach efficiency. Finally, this use case has shown robustness face to noise with a full detection of all searched objects and a success average of classified points of 97.66% for data with a noise average of 6.47%. In addition to its robustness against noise, the approach has shown its flexibility and its adaptation capability to another application context (i.e., application for building monitoring with a low-cost technology), but also its adaptation capability to smaller and more complex objects than rooms.

**Literature**


3D object recognition through a process based on semantics and consideration of the context


Ponciano, J-J., Boochs, F., & Tremeau, A. (2019, February) (a). Identification and classification of objects in 3D point clouds based on a semantic concept. 3D-Tage. https://hal.archives-ouvertes.fr/hal-02014831


Ponciano, Jean-Jacques  
Hochschule Mainz, Institut für Raumbezogene Informations- und Messtechnik  
jean-jacques.ponciano@hs-mainz.de

Boochs, Frank  
Hochschule Mainz, Institut für Raumbezogene Informations- und Messtechnik  
frank.boochs@hs-mainz.de

Trémeau, Alain  
Jean Monnet Universität, Hubert Curien Labor  
alain.tremeau@univ-st-etienne.fr