Recognition of 3-D objects on complex backgrounds using model based vision and range images*

E. Natonek and C. Baur

Vision Lab
Swiss Federal Institute of Technology
EPFL, IMT-DMT
1015 Lausanne, Switzerland

phone: +41 21 693-3826
fax: +41 21 693-3866
natonek@imtsg5.epfl.ch

ABSTRACT

One of the active research fields in computer vision is the recognition of complex 3-D objects. The task of object recognition is tightly bound to background understanding or suppression. Current literature describes the top down approaches as promising but not complete and the bottom-up approaches as not robust.

This paper describes a model based vision system in which a commercial 3-D computer graphics system has been used for object modeling and visual clue generation. Given the computer generated model image (i.e. color, depth,..) a conventional CCD camera image and the corresponding scanned 3-D dense range map of the real scene, the object can be located in it.

This paper will focus on how this is done using newly developed segmentation algorithms extracting "focus features" from range images (depth map) of the scene. Our system uses image pyramid of resolution and prediction-verification process. First we generate a hypothesized in a low resolution description , giving rough clues for the object boundaries, position and orientation. These regions of interest are then used as the field of comparison with higher resolution models. Such an iterative process is repeated until a given threshold of similarity is reached.

Next an intensity image of the model in the scene is created using the available a priori knowledge; 3-D object description, photometric attributes and the information extracted from the range image. Direct correlation is then performed between the model and the "focus feature" of the scene.

Illustrative examples of object recognition in simple and complex scenes are presented.

KEY WORDS: *Correlation, 3-D, model based, range images, ray tracing, global features, view independent*

1. INTRODUCTION

The research in computer vision has dramatically increased over the last thirty years. The main applications areas include medical diagnosis, target detection, character recognition and remote sensing. Only recently the computer vision researcher had growing interest in robotics and more precisely the automation of manufacturing processes and quality control. The reason for the slow progress in commercial machine vision system for robotics is the high complexity of the visual interpretation of the various manufacturing tasks. The four main industrial demands in machine vision for the automatic assembly task are *low cost*, *high speed*, *accuracy* and *flexibility*.

In order to be robust and effective in industrial environments, the vision system should recognize multiple parts types and should also determine their position and orientation. The perfect robot vision should be able to extract and locate the salients features of the parts in order to extract view invariant 3-D features for assembly and handling operations and be able to verify the success of these operations. Most industrial parts-recognition systems are *model-based.* This recognition scheme involves matching of an input image with a set of predefined 3-D models of parts. The task of 3-D object recognition is tightly bound to background understanding or suppression. So far in the literature the top down approaches are promising but not complete and the bottom-up approaches are not robust.

*This work is supported by the FNSRS 5003-34336-SPP-IF

0-8186-6250-6/94 $3.00 © 1994 IEEE

This paper describes a model based vision system in which a commercial 3-D computer graphics system is used for object modeling. Given the computer generated model image (i.e. color, depth,..) a conventional CCD camera image and the corresponding scanned 3-D dense range map of the real scene, the object can be located in it.

First a brief description of general 2-D, 2½D and 3-D model based system will be exposed, followed by the presentation of our 3-D model based vision system using range images to extract object clues. Finally illustrative examples of object recognition in simple and complex scenes are presented.

2. MODEL BASED OBJECT RECOGNITION

A model based vision system can be divided in to the training phase and the classification phase. As illustrated by figure 2.1 the *feature extraction* and *object modeling* are the main components of the *training* phase which is traditionnaly a bottom-up processes. The scene feature extraction and matching process are the central aspects of the *classification.*

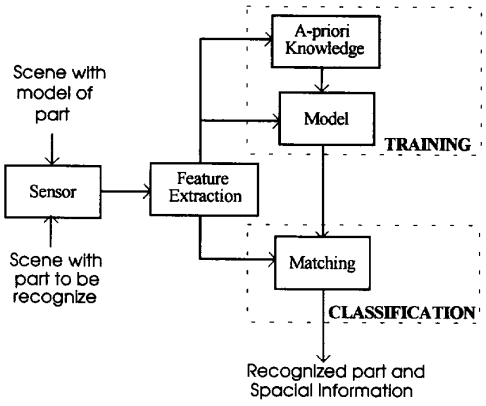

Figure 2.1: Model based recognition system

Before looking at the general goal of these three segments as specified by [Chin], lets have a look at the typical constraints in industrial environments. Typically the number of parts are usually small (1-50) and their geometrical aspects are exactly specified with known tolerances, and possibly, their CAD-CAM database is available. Usually the parts have distinctive geometrical features (holes, edges, corners) and the scenes with multiple parts can have different possible configurations with touching or overlapping parts.

Models:

The central issue in model generation is twofold. First we have to resolve the problem of extracting features from an image which best describes the physical properties and their spatial relation in the scene. Secondly we have to point out what constitutes an adequate representation for these features. In other words, how we should combine these exacted features to create a model able to characterize all parts present in the image. So far several research groups have been trying to resolve this problem using 2-D, 2½D and 3-D models.

2-D spatial descriptions, are represented in "image space" and have the advantage of being easily constructed from a set of viewpoint. Each 3-D part is modeled by a set of one or more distinct views, or by extracting these features directly from a CAD-CAM database. We can distinguish *global features*, *structural features* and *relational graph methods*. We will not examine each method in detail but references can be found in [Chin]. We should point out, that the weakness of 2-D models lies in the non trivial construction of 3-D object representation starting from 2-D views.

2½D models have attributes of 2-D and 3-D representations and are defined in "surface space", depending on surfaces properties instead of boundaries. Even if these representations are more accurate than 2-D they still are viewpoint dependent and they still deal with 3-D objects in terms of 2D features. For examples see [Chin].

Finally the 3-D models are represented in "object space" and allow the complete representation of the objects and are viewpoint independent. Their geometrical attributes and physical volume can be derived directly from CAD-CAM databases. 3-D models have the disadvantage of needing complex 2-D to 3-D correspondence procedures. Current 3-D system seldom have surface reflectance or physical surface proprieties as in [Bur].

Features:

The problem of selecting the appropriate feature is tightly bound to the model space representation (2-D, 2½D or 3-D). Thus image features for 2-D space would be edges, corners, lines as opposed to the distribution of surface-orientation normal for 3-D space representation [Horn]. In order to have object descriptions that are less sensitive to Gaussian noise, the spatial relation and the object features are combined. Usually the feature extraction is task dependent. So if the task is the identification of solid objects in the robotics

environment, the salient feature should be the geometric description of the part. But if the robotics task is to find a red box, then the feature could be the color.

***Matching*:**

The matching process is the task of finding a set of salient features, in a given image, that matches the model's features. Matching 2-D models usually consist in locating parts with a few key features, using global feature or graph-matching techniques. Most of this matching are fast, but are too sensitive to noise and image distortion. For the matching of 2½D models, the comparison of planar sets and curved surfaces patches are being made. These is achieved by direct comparison of features derived from these patches or by finding the best fitting region described in the model. Matching 3-D models requires the most computational time. Some of the model based systems use direct correlation between the bit map of the model and the image bit map [Bau]. It is difficult to use a real-time object recognition system since the transformation from 2-D to 3-D is computationaly expensive.

3. 3-D MODEL BASED VISION SYSTEM

The Model Based Vision System (MBVS) attempts to bridge some of the gaps between the image analysis and image synthesis communities. To perform the interpretation of a scene, this system first constructs and modifies an internal 3-D model of the world. These models are computer generated images, thus creating what is referred to the *virtual world.* The initial 3-D surface models are composed of the information the system knows about the world. In the remaining of this discussion we shall use the term *a priori* knowledge to describe this aspect.

Model creation :

In the generation of models we need to take into account the most outstanding properties of the real objects. Models used by MBVS are twofold. The first part reflects the *geometry* of the object, scale one to one. The second part defines what we call the *attributes* of the object. MBVS takes into account several attributes. Among them we have those which describe the way the object interacts with light: the *photometric attributes.* We also consider parameters such as degrees of freedom, or levels of complexity, to improve the scene interpretation process. In this paper we will focus on the geometric and photometric attributes.

To generate the *geometric* database, 3 different techniques have been used. Firstly we use the conventional technique of defining objects with the keyboard. Generally considered as tedious. The second and so far the more interesting technique uses 3-D geometric databases of different commercial CAD packages. EXPLORE, the Thomson Digital Image developed software used by MBVS is able to read different CAD-CAM standards. The advantage of this solution, especially in the case of automated assembly, is that

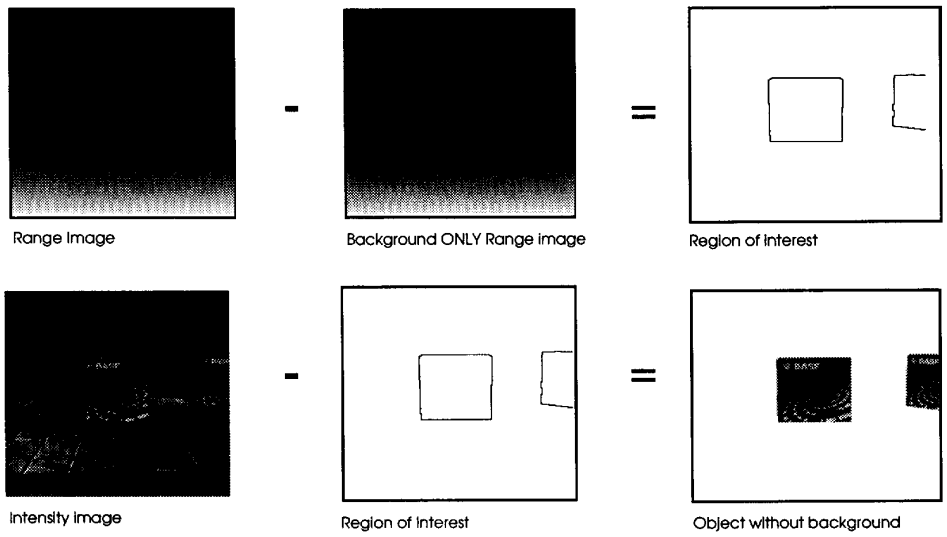

Figure 3.1: Background suppression using range images and a priori knowledge

we can take full advantage of the work already done and to rely on a precise and complete database of objects. The third technique relies on the use of 3-D scanners which are used more and more often when the outer surface dimensions of an object have to be determined. The scanner resolution can vary from one dealer to another.

In the synthetic world we can use several different techniques to generate the interaction of light sources with matter. For example, using a method like *radiosity* we try to create a perfectly diffuse world. In our system we use more classical render techniques such as *scan line algorithms* or *ray tracing*. More details on the these different algorithms can be found, for example, in [Tha].

Feature extraction:

As mentioned before, the problem of selecting the appropriate feature is closely linked to the model space representation (2-D, 2½D or 3-D). Our model-based system has a 3-D *multi-feature representation* because it uses a combination of 2½D and 3-D space features. Actually it uses range images to extracts the boundaries of the object it's size and possible orientation. The system first extract edges in the range image using a conventional gradient operator. A multidirectional tracking algorithm is used to find smoothly curved edges. This algorithm is executed several times until all edges are connected forming surfaces. Using the "a-priori" knowledge and the *virtual range* images we can easily suppress unwanted backgrounds as shown by figure 3.1.

To do so first we have to, thanks to the a-priori knowledge, build a *virtual range* image of the table without any objects. One of the strenghs of range images is the absence of textures and shadows. This virtual representation of the evironment is used to isolate the unknown objects and to build several clues as to their position and size.

With such a technique, complex background can be easily extracted from the intensity image using the range image information and the *a-priori* knowledge. This leads us to a robust and fast *region of interest* extraction technique, using the range image as a *focus feature*.

For each region, a set of global features is computed including surface type (planar, cylinder, cone, etc.), compactness, occlusion and mean as well as standard deviation.

Matching:

The recognition mechanism involves several steps. First the system analyses the *a-priori* knowledge and extracts the relevant feature to build an internal 3-D model. Our system uses a pyramid of resolution and prediction-verification processes. To optimize the object recognition scheme, it first forms a set of hypotheses about the objects present in the scene and then proceeds by trying to

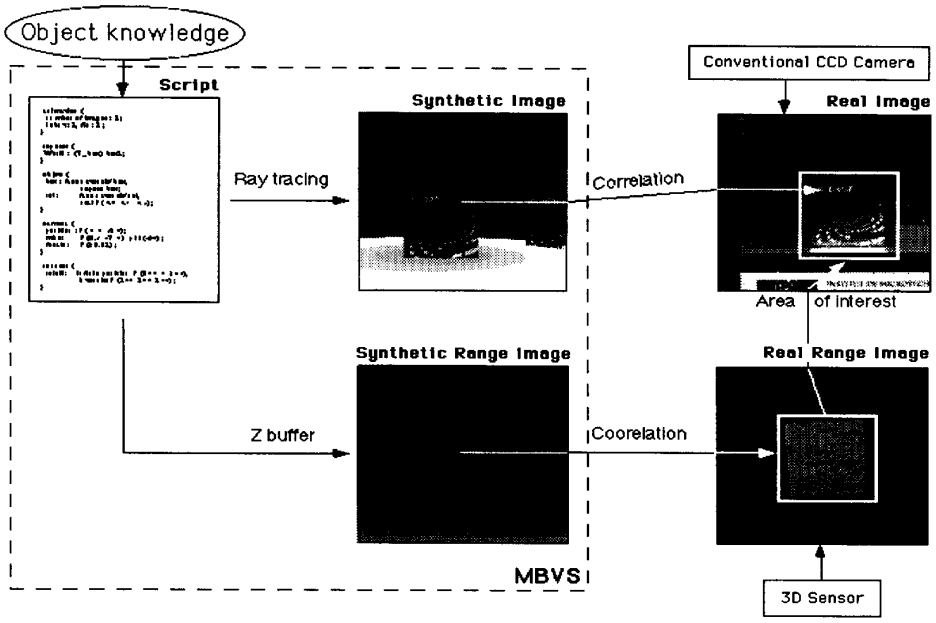

Figure 3.2: Global schema of 3-D MBVS using range images for object size and pose estimation

confirm/reject them. If any part of the object hypothesis is missing, the system uses the object model to predict the shape, location and orientation of that missing part. Next analyzing the range image we verify the main feature and we determine roughly the region occupied by the object, its possible dimension and it orientation. By doing so we find what we call regions of interest. These are the regions of the image where something is happening. These regions will speed up the search process by excluding non interesting zones of the image. This step is crucial because all 3-D model based systems based on correlation are rotation and size dependent. Thus a position and orientation estimation is essential for the object-model generation. Finally, the complete object is rendered and classical correlation is achieved between the virtual model and the real scene. This is illustrated by figure 3.2

The geometric database does not always need to be entirely complete. What is necessary is that it be task adapted. If we know what type of information will be required by vision system we can set the highest resolution necessary to perform the task. The upper limit being fixed, the adaptation of lower levels can be done in several ways. It can be done *a priori*, on-line or off-line. By *a priori* we mean that the detail limit (resolution) is fixed during the modeling. We decide that certain given details are not relevant for a search; this means, for example, that either the cost to take them into account or to handle them during the rendering is too high.

Adaptation can also be done on-line either with dedicated techniques which simplify the geometric information or with techniques used during the rendering. The adaptation can also be done off line using the hierarchical representation defined during the modeling. We can then use conventional simplification of the trees encoding the hierarchy.

After being modeled, an object is manipulated by a script which is a file interpreted by a dedicated program that allows the user to describe the topology of a scene. We then compose a scene by instancing these objects at various locations. Objects are then manipulated by the referential attached to them within a common reference frame.

4. 3-D OBJECT RECOGNITION

In this section we will comment on the recognition of two objects in a complex backgroud. Starting from analyzing the a-priori knowledge and the exact database of both objects, the system will first build several hypotheses. Figure 4.1 Shows in

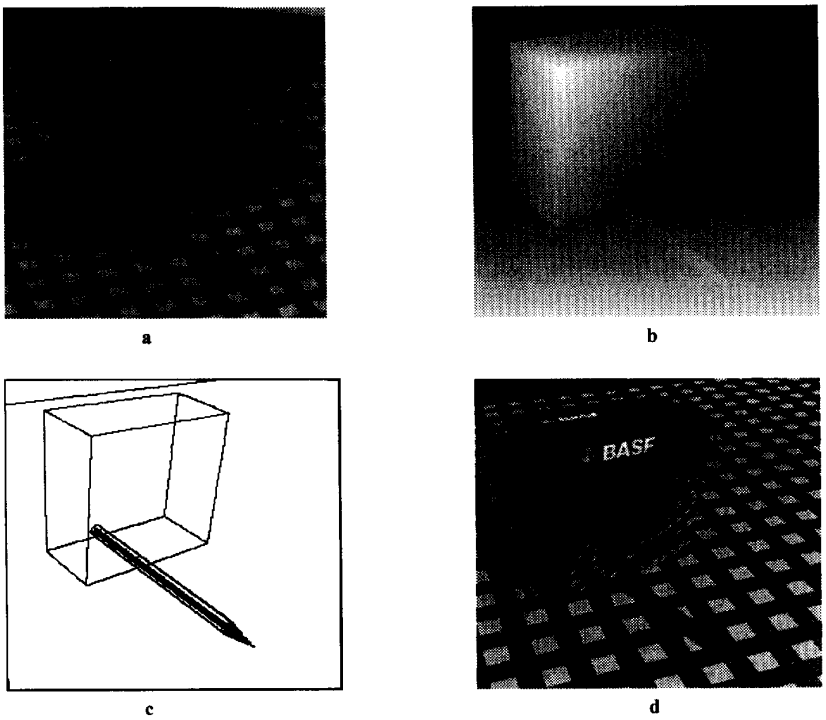

Figure 4.1: Example of 3-D object recognition using MBVS and range images

a) the conventional CCD camera image, in b) the dense range map of a) and c) shows the geometrical identification of the searched objects and finally image d) shows the reconstructed virtual model representing the real objects.

To achieve such results, MBVS first builds a pyramid of resolution and prediction-verification processes. To optimize the object recognition scheme at the range image stage, the system first forms a set of hypotheses about the objects present in the scene and then proceeds by trying to confirm/reject them. As described above newly developed range images segmentation techniques will extract edges and surfaces for feature verifications. Thus by using the a priori knowledge and the depth map it will suppress the background and extracting the size and exact positioning of both objects. This is shown by figure 4.1 c). At this step both objects are recognized in a purely geometrical way. For examples see [Nat].

Next this information is used to construct a more realistic 3-D model using the geometrical database and the photometric attributes available. Figure 4.1 d) shows the *virtual image*. Finally, to perform the last recognition, direct correlation is performed between the virtual image (Figure 4.1 c) and the real CCD camera image (Figure 4.1 a). Several techniques described in [Bau] are used to speed up the correlation between the virtual and real images. One of the methods, is to choose a salient correlation feature to in order to assure a useful recognition based on the correlation results. In our example the logo BASF™ is a salient correlation feature that will ensure the correct object recognition.

5. CONCLUSION

In this paper we have described a model based vision system in which a commercial 3-D computer graphics system has been used for object modeling and visual clue generation. The use of 3-D models is essential for the recognition of randomly oriented and positioned objects in robotics workspaces. If a computer generated model image (i.e. color, depth,..) is given and the corresponding scanned 3-D dense range map of the real scene is available, then the object is located in the scene. We have shown than the a-priori knowledge combined with range images is a powerful tool for background elimination without interfering with the rest of the image quality. Thus with the availability of object databases and with the progress in 3-D scanners, the technique showed, is very promising for 3-D object recognition, and could be used as a machine vision system for industrial assembly.

So far our system uses virtual range images. Several test done so far on real range images, from commercial 3-D scanners, are promising [Nat]. The direction followed by our 3-D model based vision system is a multi-sensor and multi focus-features vision system. So far only 8-bit intensity and range images where used by our system, the actually extensions are color [Vua] and movement.

6. ACKNOWLEDGMENTS

The authors would like to thank all the members of the *Vision Lab* of the department of Microengineering (IMT) of the Swiss Federal Institute of Technology. We also thank B. Romanowicz for his help with the English version of this text.

7. REFERENCES

[Bau] Baur C. :"Un système de vision fondé sur les modeèles utilisant l'imagerie de synthèse et la corrélation normalisée: MBVS", These N 998, Lausanne EPFL 1992

[Bur] Burckhardt C.W., Baur C. Natonek E. : "MBVS: The Fusion of Image Synthesis and Computer Vision", *Processing of 22nd ISIR*, p 3.53-3.64, October 1991

[Chin] Chin t. Roland and Dyer C.R. ," Model Based Recognition in Robot vision", *Selected Paper on Model-Based Vision* , pp 334-375, 1986

[Hor] Horn B.K.P., "Robot Vision", MIT Press, McGraw-Hill Book Company

[Nat] Natonek E. and Baur C.,"Model based 3-D object recognition using intensity and range images: *submitted SPIE OE Aerospace Sensing,* Orlando 1994

[Tha] Thalmann D. and Magnenat-Thalmann N.: "Image Synthesis: Theory and Practice". Springer-Verlag, 1987

[Vua] Vuagnat D. and Baur C.: "Depth by focus et traitement de couleur". Projet de Diplôme, EPFL 1993.