# Model-based 3D object recognition by a hybrid hypothesis generation and verification approach

Philippe Gingins, Heinz Hugli

Institut de Microtechnique, University of Neuchatel
Tivoli 28, CH-2003 Neuchâtel, Switzerland

gingins@imt.unine.ch, hugli@imt.unine.ch

## ABSTRACT

We present a model-based 3D object recognition architecture that combines pose estimation derived from range images and hypothesis verification derived from intensity images. The architecture takes advantage of the geometrical nature of range images for generating a number of hypothetical poses of objects. Pose and object models are then used to reconstruct a synthetic view of the scene to be compared to the real intensity image for verification. According to the architecture a system has been implemented and successful experiments have been performed with boxes of different shapes and textures. Recognition with our approach is precise and robust. In particular verification can detect false poses resulting from wrong groupings. In addition, the system provides the interesting features to recognise the true pose of shape-symmetrical objects and also to recognises objects that are ambiguous from their sole shape.

**Keywords**: 3D vision, range images, rendering based vision, knowledge based vision.

## 1. INTRODUCTION

In object recognition the basic task for a machine is to recognise objects from the real world and to locate them as precisely as possible. By model-based recognition, we understand that explicit knowledge is available to the system regarding the objects to be recognised. Model-based recognition is adequate for applications where an exact knowledge of the different types of objects is possible. For example, quality control where the recognised object has to be compared with some reference, or in assembly where a robot has to grasp and manipulate known objects. In applications like robot manipulation in hostile environments, the knowledge extends to the complete environment and is named a virtual world.

Because object recognition by 3D vision is usually a difficult task, it is important that it takes best advantage of the available knowledge. Information of sensed data and information of model data must be brought into interaction in a purposive and efficient way. To do so we consider in this paper an approach based on the hypothesis generation and verification scheme[10]. Adequate methods must be selected for hypothesis generation and verification.

In virtual worlds, where all objects are modeled and the knowledge about the environment is complete, it is possible to use image rendering to generate any view of the environment[3]. Comparing such a rendered view with the current image provided by a camera constitues therefore a promising verification method we decide to integrate as one component of our approach.

Regarding the hypothesis generation in the frame of 3D vision, it requires an estimate of the object pose in 3D. With intensity images, as provided by video cameras, only 2D spatial location are given and the derivation of 3D information is rather complex. Range images however provide explicit 3D spatial information which is more suitable to be used for 3D pose estimation[1]. In the context of this paper, we therefore choose to build hypothesis generation on range data.

This approach defines a promising hybrid hypthesis generation and verification architecture combining information from range and intensity image. Some principles were presented earlier[9]. It is now analysed both in principle and in practice. Section 2 briefly presents the different existing methods for model-based 3D vision and some of their limitations.


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

Section 3 describes the proposed architecture and its advantages. Section 4 presents the experimental system we implemented. Then section 5 presents experimental results and discusses them.

## 2. EXISTING SOLUTIONS FOR MODEL BASED 3D VISION

Without doing a complete typology of existing model based 3D vision systems, we consider the major division between them which is the kind of real world information they use: range information or light intensity information.

### 2.1. Range image vision system

Some sensors, such as laser scanners, yield range images. For each pixel of the image, the range to the visible surface of the objects in the scene is known. Therefore, spatial location is determined for a great number of points on this surface.

It is then possible to recognise and locate some known objects' shapes. Many proposed systems are limited to some class of shapes, for example polyhedrons. By first locating flat surface patches or other features in the image, they try to match these to some models[1,5,6]. Free-form recognition is an active field of research[4,7].

Of course, such systems cannot distinguish between objects with the same or similar shape, neither between symmetric positions of the same object (for example: is some box upside-down or not?). Furthermore, they may generate different hypotheses from one single image. Selecting the correct ones may be rather difficult, resulting in added complexity to the system.

In conclusion, these systems are able to generate object identity and pose hypotheses, but are unable to use texture information and need a way to validate their hypotheses.

### 2.2. Intensity image vision system

Normal video cameras provide light intensity information in a flat, 2D image, without any explicit range information. One major and difficult problem is then to reconstruct this range information, or at least to determine at what distance are lying the visible objects.

#### 2.2.1. Bottom-up

A few general methods permit recovery of spatial information from intensity image in a bottom-up way: stereo vision reconstructs some of it by combining two viewpoints. Other possible means are motion, texture or shading, but they need some special conditions and usually result in sparse spatial data.

If spatial information is correctly reconstructed, the problem is then very similar to range-based vision.

#### 2.2.2. Top-down

Making assumptions about the objects seen may enable some spatial reconstruction, for example by analysing edges, especially relations between them.

Another top-down method is to use rendering. If we want to check if a given object is present in some image in a known orientation and if we have a complete shape and texture model of it, it is possible to render a synthetic view of it, and to use correlation to detect it in the image[3]. This method is very selective, as it uses all the texture information. Thus a hypothesis validated this way is very reliable. But it is impossible to try all the possible poses for all the objects. As efficient rendering software and hardware is becoming more popular, this method may be quite simple and efficient.

## 3. MIXING BOTH KIND OF SYSTEM

### 3.1. Architecture

To overcome the limitations of any single system, we brought two very different systems together. The first one, the *hypothesis generation system*, is a range image vision system, using segmentation and simple model matching. The second one, the *hypothesis verification system*, is an intensity image system, using photo-realistic rendering and fast image correlation. This architecture is represented in the figure 1.

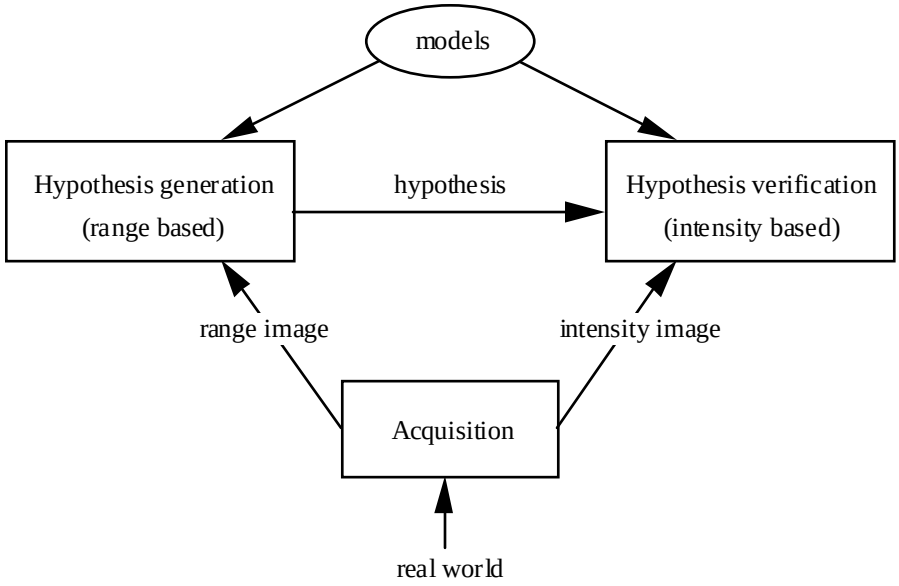



*Fig. 1: system architecture*

First, the generation system generates a few hypotheses about the objects in the scene, using shape data only. Therefore we call them *shape hypotheses*. Then the verification system uses them as a basis for his work, using full texture data. Each initial hypothesis may lead to a few different secondary hypotheses, when some objects are ambiguous in shape, or when the object is symmetrical in shape but not in texture. These are then checked by comparing a rendered image of the model with the real intensity image.

### 3.2. Advantages

Combining a range based vision system and an intensity vision system brings many advantages. We see mainly three of them.

#### 3.2.1. Robustness

As the two systems are very different in their principles, false hypotheses asserted by the first one are easily rejected by the second one. Alternatively, the second system receives only quite meaningful hypothesis and does not have to explore all possibilities.

As it uses full shape and texture information, the system is very reliable.


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

### 3.2.2. Robustness brings speed

A consequence is that the two subsystems may be quite simple, as false hypotheses generated by one are rejected by the other. Simpler systems result in better speed. For example, we use for range image segmentation a direct method that involves only local operations. It is a fast method, but it may confuse a smooth surface with a flat surface and often underestimates the area of the flat surface. These imperfections may create wrong hypotheses, that are normally rejected by the verification system. Therefore we do not need a more efficient but slower method.

### 3.2.3. Differentiating objects ambiguous in shape or in texture

Even the most perfect range vision system could not differentiate two objects having the same shape but different textures, as are for example two floppy boxes from two different brands. Nor could it determine if such a box is in normal position or upside-down.

On the other hand, similarly textured objects with different shapes may be very difficult to differentiate with an intensity image vision system.

A hybrid system easily handles all those difficult situations.

## 4. HYBRID EXPERIMENTAL RECOGNITION SYSTEM

### 4.1. System architecture

Our experimental system depicted in figure 2 is composed of three main subsystems: acquisition, range image vision and intensity image vision. To test our proposed architecture, we first aimed at an easy case: recognising simple boxes (rectangle parallelepiped). Hence we could easily implement our test system with a very interesting case, since boxes commonly exhibit shape similarity and symmetry. They are also very common.

### 4.2. Acquisition

We use a range image acquisition system built around a commercial ABW 320 LCD stripes projector and a standard, low-cost black & white camera. The software employed was originally developed by Marjan Trobina at the ETHZ. This system generates both range and intensity images, such as the ones in figures 4 and 5. Noise is quite important in the range image (more important than with most of the laser scanners).

### 4.3. Range image analysis

Box recognition and pose estimation from range images are achieved through three processes, leading from the range image to a set of shape hypotheses. Each hypothesis contains the class of the recognised object (only boxes so far), the estimation of the size of the box (object parameters) and its pose, i.e. the coordinates of its center and three normalized vectors giving the orientation. All these data are grouped into a box hypothesis, stored in a file.

#### 4.3.1. Flat surfaces segmentation

First the range image is segmented in smooth surfaces by using an algorithm detecting local discontinuities at different orders[8]. We call segments each smooth and connected set of pixels. In the next steps, they are always considered flat. A plane is then fitted on each segment. We call it the *plane of the segment* and its normal *the normal of the segment*. The latter is chosen so that it points towards the camera.

As it may happen that a face of the real object is detected as two or more segments, specially when partially occluded, coplanar segments are merged into a single segment.

4.3.2. Grouping of orthogonal segments

Then we search for all sets of three pairwise orthogonal segments, in the idea that they may be three visible faces of the same box.

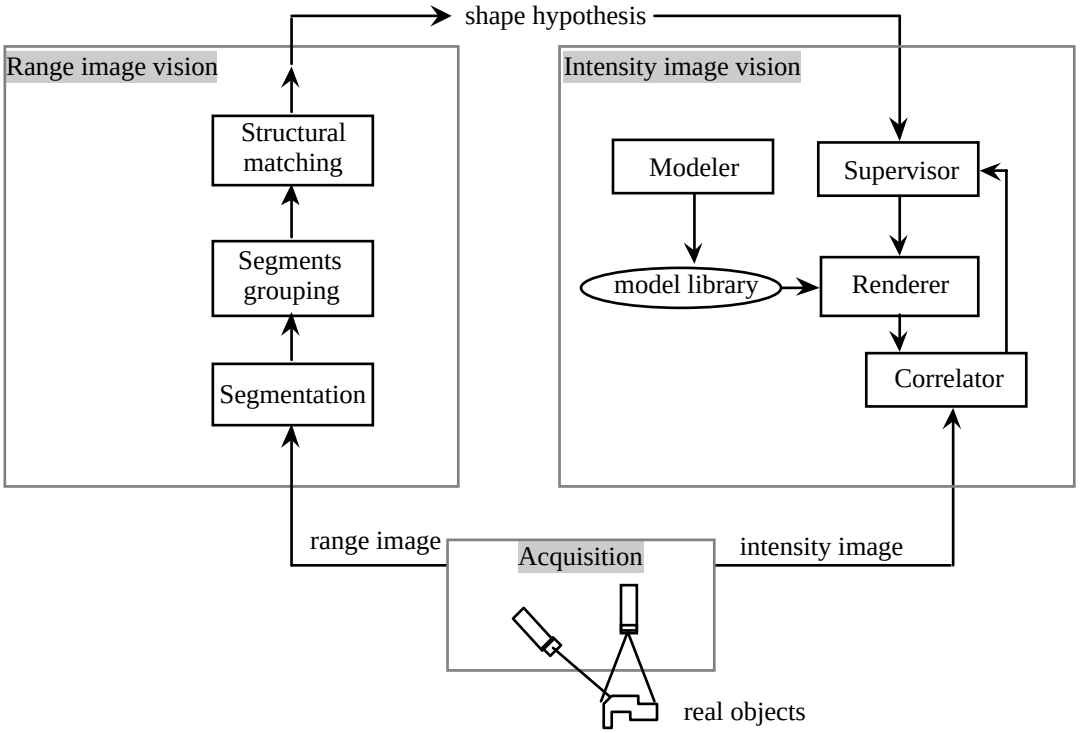



*Fig. 2: System description*

4.3.3. Structural matching

For each one of these triples of segments, we test if they may be three faces of the same box. In the affirmative, the result of this process will be some estimations of the size of the box, and a pose estimation, i.e. its spatial position and orientation.

As a box is convex, the three segments of the triple should form a convex set of points. We apply this constraint by testing if all the points of each segment are below the planes of the two others. We reject all triples that do not satisfy this constraint.

From this point, we do the hypothesis that the three segments are part of the same box. Therefore the orientation of the segments gives the orientation of the box (i.e. the normals to each segment are parallel to the axis of the box).

The intersection of the planes of the segments is also one vertex of the box. The situation is depicted onto figure 4, where *S* is this vertex. Three of its edges lie along the intersections of each couples of these planes. $E_3$ is one of these edges.


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

To estimate the minimal size of the box, we compute the smallest box containing all points of the selected segments. In figure 4, $l_{min}$ is the minimal length of $E_3$.

The segmentation process usually provides segments smaller than the real faces of the box. Hence we need to estimate this erosion for a better estimation of the dimensions of the box. We use for this purpose the minimal distance from the plane of each segment to the two other segments. We get this way a better estimate called $l_{best}$ on figure 4, where $d_1$ is the corresponding estimate of the erosion. By using $d_2$, the distance from the plane of the top segment to the remaining segment, we get a third estimate, $l_{large}$.

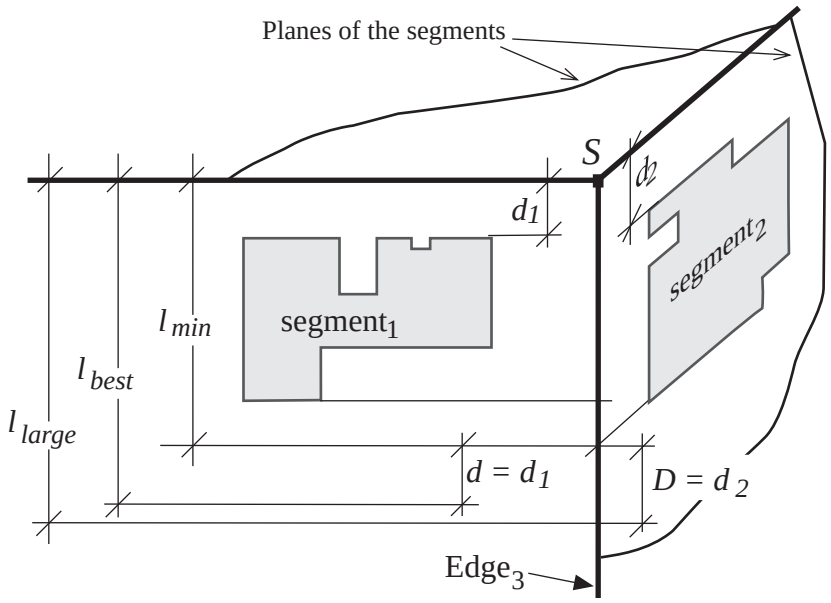



*Fig. 3: Estimation of the length of one edge of the box*

Once all these triples are treated, all segments used in any hypothesis are removed of the set of the segments and a similar process is performed on the remainder, searching for boxes with only two visible faces. In other terms, we search for couples of orthogonal segments, with all their points below the plane of the other, we compute the minimal box, and we use the distance from the segment to the single known edge as an estimation of the erosion.

In this simple and experimental system, we do not search for boxes with a single face visible.

4.4. Rendering based vision

The basic principle of this system is for a given hypothesis to render a synthetic image of the supposed object in the expected position, and to validate the hypothesis when the correlation factor of this image with the actual image is high. For each model, a complete texture and shape knowledge is assumed. For speeding up the process only part of the image may be rendered and used. The availability of specialised hardware for rendering and correlation makes this principle quite practical.

In our case, each shape hypothesis read from the range system corresponds to a few hypotheses for the intensity system. Indeed different models may match the sizes given in the hypothesis and, as boxes are symmetrical, a few different orientations of the box result in the same apparent shape (exactly 4 for a box with three different dimensions, 8 when 2 dimensions are the same and 24 for a cube). For common manufactured boxes, the printing is not symmetrical and thus these different positions are distinguishable.


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

The architecture of the intensity based vision system is depicted in figure 3. The main process is called the supervisor. For each shape hypothesis, it generates all corresponding hypotheses. For each one, it sends to the renderer the geometry of the scene (position of the camera, of the light projector) and the name of model to use and its position.

The renderer retrieves the model from the model library and generates a synthetic view of the scene using ray-tracing. For speeding up the process, only a part of the image is rendered and used, chosen among the discriminating spots, as brand labels.

Then a correlator checks if the rendered image may be found in the real image. If not, the whole process is repeated for the other hypotheses until satisfaction. When satisfied, a complete image is rendered for operator control.

### 4.5. Hardware and software

The range image system is running on a Sun Sparc 10 workstation, with custom software.

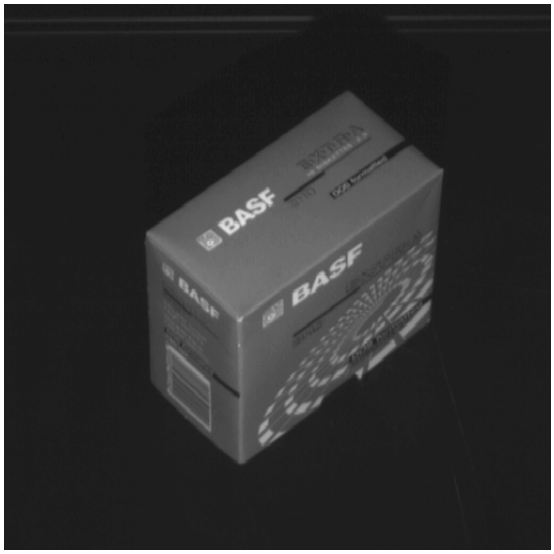
The intensity image system is composed by a Silicon Graphics workstation and by a Matrox board in a PC machine providing fast correlation. A commercial rendering software (TDI Explore) is used for object modelling, model storing and rendering.

The system is thus composed of several machines. The two sub-systems are located in two remote cities and the communication between them is implemented by a file directory, exported on the internet by NFS. This directory is used as a blackboard, each system writing or reading items as files on it.

## 5. RESULTS

Practically the system works as expected. We used some floppy boxes of different brands as test objects. Models for these were created and stored in the model library. We put one or two of them in front of the acquisition camera to test if the system was able to recognise and locate them. The range system always located them as long as the range image quality was good and at least two faces were clearly visible. The intensity image system could then determine precisely the object (i.e. distinguishing between different brands) and could also determine its correct orientation.

Figures 4 to 10 present typical cases, with three or two visible faces and partial occlusion in figure 10.



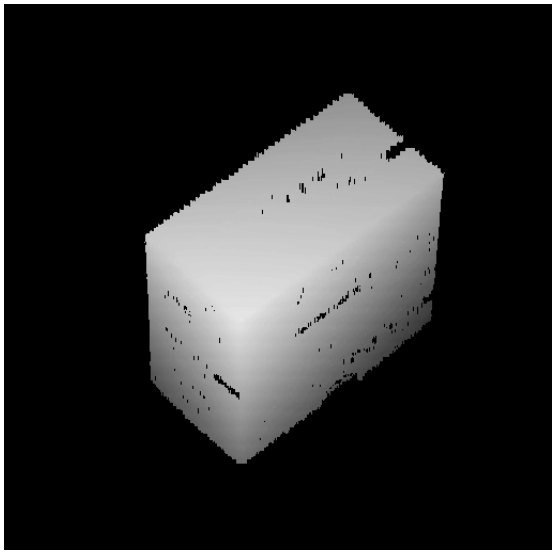



*Fig. 4 and 5: intensity and range image of a floppy box.*

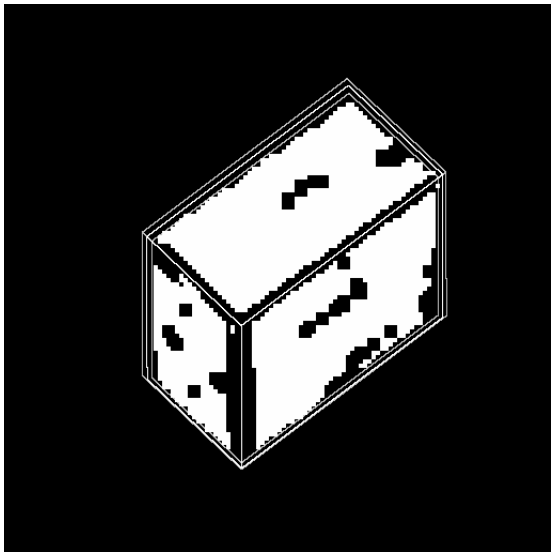


*Fig. 6: Result of segmentation, together with the wire frame of shape hypothesis. The three frames correspond to the three different size estimations.*


*Fig. 7: Rendered view of the final hypothesis.*


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

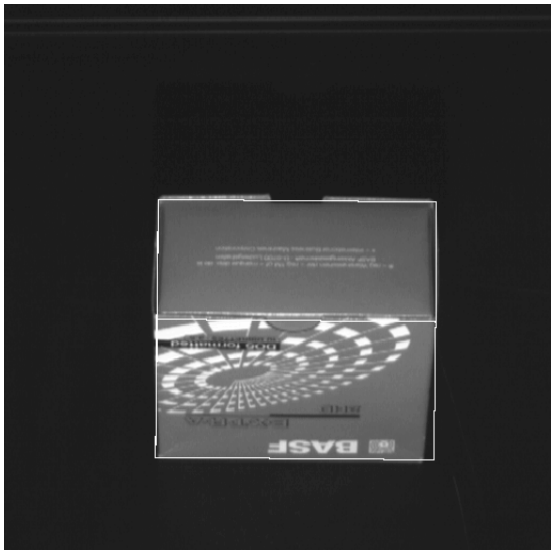


*Fig. 8: Real intensity image together with the wire frame of the shape hypothesis (best estimation only).*

*Fig. 9: Rendered view of the final interpretation corresponding to figure 8.*

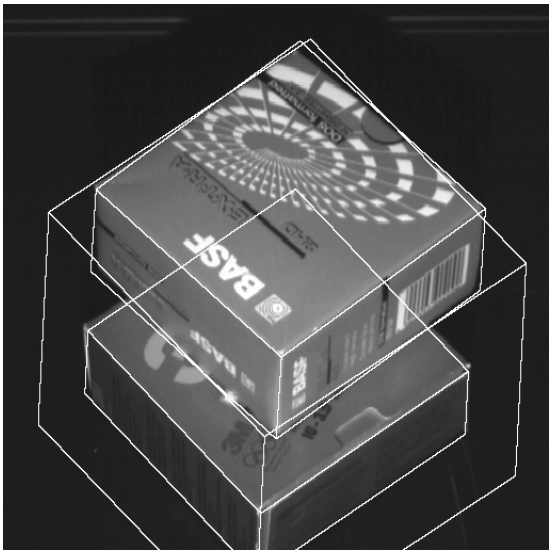


*Fig. 10: Intensity image with wire frames of box hypotheses. A false hypothesis is made due to a wrong grouping.*

5.1. Precision and robustness

In the case of a box with 3 visible faces, as in figures 4 to 7, the entirely visible vertex and the 3 visible edges are located precisely, with a typical error of less than 2 mm. When 2 faces are visible, as in figure 8 to 9, the entirely visible edge is then located with the same tolerance. The size estimation error is typically 5 mm.


Philippe Gingins & Heinz Hügli, "Model-based 3D object recognition by a hybrid hypothesis generation and verification approach", Intelligent Robots and Computer Vision XIII: Algorithms and COmputer Vision, SPIE 2353, Boston, Oct. 1994

When more than one box are present in the scene, the results are still good even in case of moderate occlusion, as in figure 10. Due to the very simple method, some completely erroneous hypotheses, due to wrong grouping, are sometimes generated by the range image vision system, as the biggest box frame in figure 8. These are easily rejected by the second system.

The second system finds the correct model and its correct orientation. Figure 7 and 9 show the images produced by the rendering system for figure 4 and 8 respectively. Note that the box is in normal position on figure 4 and upside-down on figure 8. Typical normalized correlation scores are higher than 75%, thus final decision is easy.

### 5.2. Quickness

The quickness of the solution is not clearly shown out by our experimental system, as all the elements were implemented without any attempt to be efficient. A complete processing, not taking in account the acquisition time, takes 3 to 5 minutes, most of it for the rendering. The efficiency may be improved in several ways:

- The segmentation takes approximately 20 seconds. As the method used involves only local operations, it may easily speeded up by using a specialised hardware for image processing.
- Exploring systematically all the grouping of the segments for building box hypotheses takes from 2 seconds to 15 seconds. It is a cubic complexity method in the number of segments, but testing for orthogonality is very simple and computing box extents may be very fast if we precompute the convex hulls of the segments.
- Ray tracing rendering is also slow, up to 4 minutes. This method was used only for practical reasons. Using fast rendering methods, we may expect to do it in a fraction of a second, at the cost of a reduced quality.

Correlation is done in a fraction of a second, using specialised hardware.

### 5.3. Limitations

As the two subsystems form a chain, the weakest one limits the whole system. For example, our current experimental system is limited to boxes recognition only due to the limited aims of the hypothesis generation system. Nevertheless it would be still possible to use each subsystem for itself, with a reduced reliability.

The verification system may face a real problem of hypotheses number explosion. Each shape hypothesis generated by the first system may be consistent with many models. Then each model may have many different symmetrical orientations (for example, 24 for a cube). Thus it may be necessary to have some indexing or pre-filtering scheme.

Range data acquisition may very difficult, as for shiny or transparent objects.

## 6. CONCLUSIONS

We presented a model-based 3D object recognition architecture that combines two vision systems in a *hypothesis generation then verification* architecture: first a range image vision system generates object hypotheses from shape information and then an intensity image based vision system using full texture information verifies these hypotheses.

Experiments with the implemented architecture were successful and showed the feasibility of this hybrid approach and robustness and precision of the recognition. We showed how floppy boxes are recognised. In these experiments, all objects are recognised correcly. False hypotheses are rejected. The size estimation error of the objects is in the order of 5 %. We also demonstrated a further advantage of the combined range and intensity vision by recognising correctly objects which are ambiguous either in shape or in texture.

The hybrid model-based 3D object recognition architecture presented here describes a general approach for recognition. It is not limited to boxes and can be extended beyond the performed experiments. It can be generalised to the recognition of arbitrary objects and is most suitable in the context of virtual worlds.

## 7. ACKNOWLEDGEMENTS

The presented work is sponsored by the *Swiss National Fund for Scientific Research* under project number 5003-34336 and is a collaboration with Charles Baur and Emerico Natonek of the *Swiss Federal Institute of Technology in Lausanne (EPFL)*.

## 8. REFERENCES

1. H.P. Amann & H. Hugli, "An algorithm for the inexact matching of high level 3D polyhedric representations", *Selected papers on model-based vision*, Ed. Hatem Nasr, SPIE Milestone Series, Vol. MS 72, pp. 313-321, SPIE, Bellingham, 1993.

2. D. Ballard & C. Brown, "Computer vision", Prentice Hall Inc. 1982

3. C. Baur, *Un système de vision fondé sur les modèles utilisant l'imagerie de synthèse et la corrélation normalisée (MBVS)*, Ph.D. Thesis No 998, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, 1992.

4. P. Besl & N. MkKay, "A method for registration of 3-D shapes", PAMI, vol 14(2), pp 239-256, 1992.

5. T.J. Fan, G. Medioni, R. Neviata, "Recognizing 3-D Objects Using Surface Descriptions", PAMI, Vol 11(11), pp 1140-1157, November 1989.

6. O. Faugeras & M. Hebert, "The representation, recognition and locating of 3d objects", Intl. J. Robotics Res., Vol 5(3), pp 27-52, 1986.

7. J. Feldmar & N. Ayache, "Rigid and Affine Registration of Smooth Surfaces using Differential Properties", in *Computer Vision - ECCV'94*, J.-O. Eklundh ed., vol II, pp 397-406, Springer-Verlag, Berlin, 1994.

8. G. Maître & H. Hügli., "Range image segmentation by controlled-continuity spline approximation for parallel computation", in *Curves and Surfaces in Computer Vision and Graphics II*, Proc. SPIE Vol 1610, pp 238-249, SPIE, Boston, 1991.

9. E. Natonek & C. Baur, "Model based 3-D Object Recognition using Intensity and Range images", *Proc SPIE Intl. Symposium on aerospace sensing*, vol 2234-01, SPIE, Orlando, 1994.