3D semantic map computation based on depth map and video image

Włodzimierz Kasprzak[1] and Maciej Stefańczyk[2]

[1] Industrial Research Institute for Automation and Measurements
Al. Jerozolimskie 202, 02-486 Warszawa, Poland
wkasprzak@piap.pl, WWW: http://www.piap.pl/
[2] Institute of Control and Computation Engineering, Warsaw University of Technology, ul. Nowowiejska 15-19, 00-665 Warsaw, Poland
W.Kasprzak@elka.pw.edu.pl, WWW: http://ia.pw.edu.pl/

Abstract. A model-based object recognition in video and depth images is proposed for the purpose of semantic map creation in mobile robotics. Three types of objects are modeled: a human silhouette, a chair/table and corridor walls. A bi-driven hypothesis generation and verification strategy is outlined. The object model includes a hierarchic semantic nets, combined with a graph of constraints and a Bayesian network for hypothesis generation and evaluation. For the purpose of model-to-image matching we define an incomplete constraint satisfaction problem and solve it. Our CSP-search allows partial assignment solutions and uses a stochastic inference to provide judgments of such solutions. The verification of hypotheses is due to a top-down occlusion propagation process, that explains why some object parts are hidden or occluded.

Keywords: Bayesian net, constraint satisfaction problem, depth map, object recognition, semantic map

1 Introduction

Three general paradigms for object classification and recognition in images are most often distinguished: the stochastic Bayesian approach [1], the neuro-computational and biological approach [2] and the rule-based approach [3]. Although of different nature these approaches share the concept of *rationality*, as the recognition and understanding processes in all paradigms need to satisfy some appropriate optimization criteria.

In model-based image analysis fundamental problems are: model representation language, a control and evaluation of partial model-to-data matches. Here we shall follow an object-oriented framework build around semantic networks, and we shall integrate it with another three general-purpose tools: (1) a modified search for constraint satisfaction problems ([3]), applied as control of partial model-to-data matches (hypothesis generation), (2) the Bayesian approach to statistical inference [1] (applied for evaluation of hypotheses), and (3) rules of

an attributed structure grammar [4] (applied for hypothesis verification). All these abstract tools parts are of dominating declarative nature and there exists well-known machine learning approaches for them, e.g. inductive inference for the concept learning, and ML- or MAP-estimation for the learning of Bayesian net probability distributions [5].

The object recognition system is applied for labeling of 3D environment maps in mobile robotics, i.e. the creation of 3D semantic maps [6]. At first a 3D environment map need to be reconstructed from measurements that combine laser scans (if a scan line laser is used) and video images. At first the individual scan lines are integrated into a cloud of 3D points (e.g. the ICP (iterated closest point) algorithm [7]. Next, the point set is approximated by triangle faces (e.g. Delaunay triangulation) [8] or by fitting superquadrics surfaces patches [9]. The map texturing steps follow - an addition (or stretching) of the video image content onto the 3D surfaces [10]. The final map segmentation step is to approximate the triangle net by larger planar of curved surface patches - using incremental growth [8] or point elimination-based algorithms [11].

In section 2, the application scenario is outlined and the object recognition approach is introduced too. The main 3 recognition steps are explained in sections 3, 4 and 5. Two implementation results in section 6 complete the paper.

2 3D semantic map

We are interested to recognize solid objects, like chair, table and wall, and a human silhouette in the neighborhood of a service mobile robot. The person can eventually sit in a chair or stay in front of a wall or behind the table or chair, i.e. the human is fully or only partially visible.

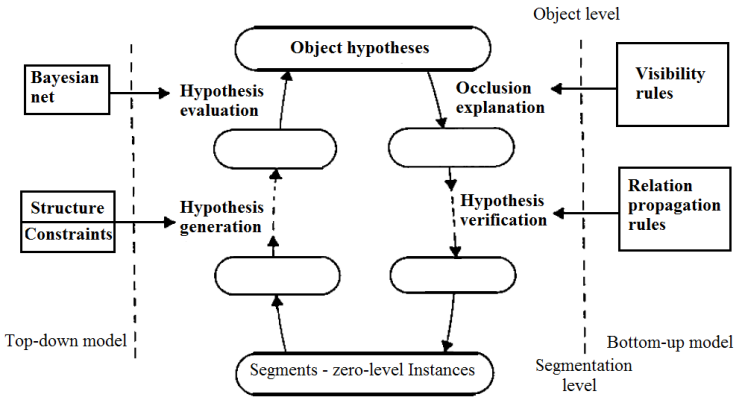

Fig. 1. The 3D object recognition approach

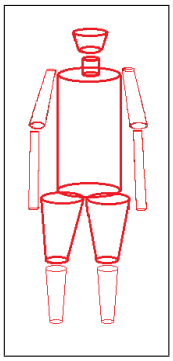

Fig. 2. A human silhouette (based on [12])

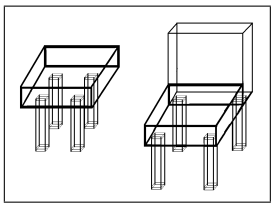

Fig. 3. A chair and table model model

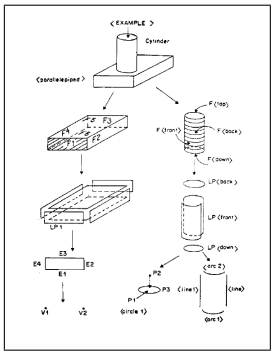

Fig. 4. The decomposition of a cylinder and parallelepiped shell onto faces and edge loops

2.1 The object recognition approach

The 3D object recognition approach consists of following processes (Fig. 1): hypothesis generation (model-to-data matching), hypothesis evaluation (stochastic inference), object visibility test, hypothesis verification (occlusion propagation).

To accomplish the overall task we have to define four models: the hierarchic structure of concepts and graphs of constraints per concept, a *Bayesian network* for quality judgement of an *instance* or *modified concept*, the mutual object occlusion relations, and the occlusion propagation rules.

The model-to-data matching is seen as a specific constraint satisfaction problem [3], that needs to be satisfied only partially. The judgement (score) of instance is estimated by a stochastic inference in the Bayesian net, linked to given concept.

The hypothesis verification process is a bottom-up explanation of possible mutual object occlusions and self-occlusions. The hidden parts are added as evidence and a Bayesian inference for given instance is repeated with additional evidence variables set to synthesized parts.

3 Object hypotheses generation

3.1 The 3D object model

Common to semantic networks is the explicit structuring of domain knowledge along two hierarchies: the decomposition (vertical) hierarchy and the specialization (horizontal) hierarchy of concepts. Starting from the pixel level the vertical hierarchy expresses increasingly abstract representation levels ("part" or "concrete" links). Simple elements are combined into more complex one, being parts of objects and scenes. Specialization links ("spec") represent inheritance relations between elements at the same abstraction level.

Every node (called "concept") represents some object category and it contains a parameter vector (called "attributes"), where every parameter is evaluated by some *term*, and every concept defines a set of constraints, evaluated by *predicates*, among its parts and related concepts.

The generic object types, required for 3D map labeling, take the form of wireframe models (Fig. 2, 3). There are default dimensions of object parts provided - this especially allows to constrain the human object hypotheses.

3.2 Partial CSP

A discrete *Constraint Satisfaction Problem* is defined in terms of states, actions and the goal test. A state set S, where a particular *state*, $\mathbf{s} = (d_1, d_2, ..., d_n)$, is defined by assignments to its variables, $X = x_1, x_2, ..., x_n$, where each $x_i, (i = 1, ..., n)$, can take values from a domain D_i. The *actions*, $a \in A$, mean transitions between states: $a_k : s_i \rightarrow s_j$. The *goal test* checks a set of constraints, $C(X)$, which induces allowed combinations of assignment values for subsets of state variables. A *solution state* is every state that satisfies the goal test. In particular, in our problem: the variables in X correspond to parts of some model concept, the values in domain D represent the current data entities (instances) and an action is assigning a value to some variable in given state. The variables and the set of constraints, $C(X)$, can be represented as a graph, $G(X, C(X))$ where nodes X represent variables and arcs $C(X)$ represent constraints between particular variables. For example, typical constraints for edges are: A = *edges are connected*; B = *edges are parallel*; D = *edges are of similar length.*

A modified CSP search is proposed that allows partial solutions (some variables may have no assigned value). While starting from an empty assignment the goal is to match (assign) eligible image segments (values) with model entities (variables). We introduced two *modifications* to the basic CSP search. The first modification is due to the definition of a *Bayesian network* for every problem. The subfunction *Score* calculates probability value of a partial solution, that consists of eligible assignments to variables. This score is due to a stochastic inference process performed in a dedicated Bayesian net, created for current CSP problem.

The basic algorithm for CSP is a depth-first tree search with a backtracking step, performed when the path is not consistent with given constraints. The second modification of a typical CSP is that now partial paths can be potential solutions. The backtrack step is performed now when currently selected (extended) path does not satisfy the constraints of given problem or its score is lower than the score of predecessor path. In our view this is not a general failure but a situation where the previous state corresponds to a partial solution. The current path is stored as a possible partial solution only if it has higher score than the previous best one.

4 Hypothesis evaluation

4.1 Bayesian network

This is a simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions. Syntax of a BN: 1) a set of nodes, one per variable; 2) a directed, acyclic graph (link means that "direct influence") - incoming links of given node represent a conditional distribution for this node given its parents, $P(X_i|Parents(X_i))$. In the simplest discrete case, conditional distribution is represented as a conditional probability table (CPT), giving the distribution over X_i for each combination of parent values.

For an concept of a semantic net model, presented in previous section, the structure of a corresponding Bayesian model is automatically generated (Fig. 5).

A Bayesian network will here represent stochastic dependencies between the solid type "parallel", intermediate-level "views" and "faces", and low-level "edges" (that corresponds to image segments). The "face" concepts consist of 4 edges. The constraints in CSP model now correspond to additional evidence variables (nodes) in the Bayesian net. There are evidence nodes that represent constraints between faces (fA, fB) and constraints between squares (A, B, D). The rank, to which a particular constraint is satisfied, can be measured after its "parents" (the "edge" variables) have been assigned to image segments.

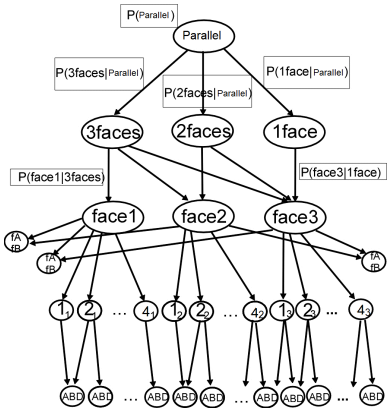

Fig. 5. The Bayesian net for a "parallelepiped" solid

4.2 Score by stochastic inference

The score of a partial solution (assignment in terms of CSP), in which some variables X_i have already been assigned to image segments l_k but not all of them, is obtained due to stochastic inference in Bayesian net. For example the computation of posterior probability of a "cube" instance (that is a *cause* in terms of BN) given its parts (that are *evidences* in BN). For example, if segments are assigned to X_0 and X_1 then one need to compute the probability: $P(cube|X_0 = l_1, X_1 = l_2)$.

This leads to a summation of pdf over all domain values for remaining (non-evidence) variables, $X_2, ..., X_l$. Thus, scores of partial matches or a complete match, between image segments and model entities, are naturally obtained by the same evaluation method.

5 Hypothesis verification

As a result of the hypothesis generation process many competitive object instances exist. In general, to find a best consistent subset one needs a search procedure. In our test implementation we make a simplifying assumption that at most one instance per object type can exist. This allows us to make a systematic check of all the subsets.

The top-down verification process for a selected subset of hypotheses consists of two steps: the initial generation of occlusion relations (between objects) and the propagation of occlusion relations from an upper level L to a lower level $L-1$ (Fig. 6). There are three types of relations used: "potentially hidden", "partially hidden", "hidden". A set of generic propagation rules is used. In general, when an instance is "potentially hidden" then its visibility case has to be resolved at the lower level, with regard to its parts. As a result of such check the relation will be canceled, kept or replaced by "partially hidden" or "hidden". The last two labels induce additional evidence (support) for a hypothesis, as they explain why a given part has not been matched with image data.

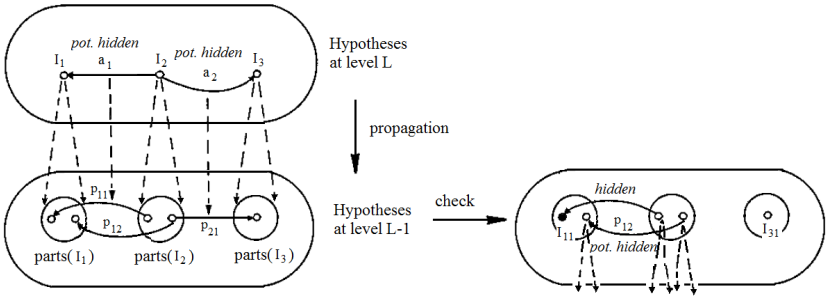

Fig. 6. Illustration of the occlusion propagation and visibility check. The relations a_1 and a_2, specified at level L as "potentially hidden", are resolved among the parts of instances I_1, I_2 and I_3. Assume, the visibility check turned some relations into a "hidden" status (terminate instance I_{11}), some others were rejected (terminal instance I_{31} and remaining relations were propagated to level $L-2$.

6 Results

The first example demonstrates the color image analysis in the absence of a depth map. The number of possible objects is limited to at most one per modeled type (Fig. 7 - 12). An important step is the color-based human skin detection (the human region detection is based on [12]). Also image regions of small size are eliminated, whereas large regions are eventually split into convex parts. In order to detect faces of a solid, sufficiently strong line segments are set in correspondence (geometrical proximity) with the post-processed regions. If sufficient "face evidence" is available a solid and object hypotheses are generated. Eventually

multiple hypotheses of given type are still allowed at this stage. Finally, subsets of hypotheses are eventually verified by considering the visibility relations.

Fig. 7. Example of a room scene

Fig. 8. Region-based image segmentation

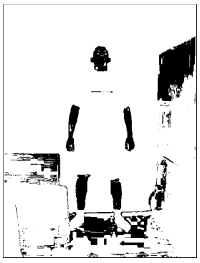

Fig. 9. Skin color filtering

Fig. 10. After morphological filtering

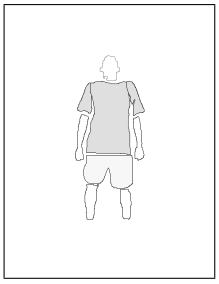

Fig. 11. Regions related to human instance

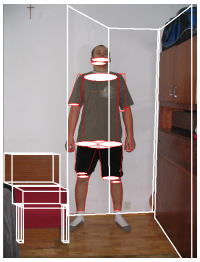

Fig. 12. Detected 3 object instances.

The second example demonstrates surface generation if corresponding video image and depth map are both available (Fig. 13 - 18). The depth map and surface patches follow the work [13]. The laser scanner SICK LMS 200 acted as the acquisition device. Here, the 3D surface patches are approximated by planar surfaces. Accordingly, strong ling segments in the video image are defined at places with large discontinuity of depth information. Now face hypotheses are generated from four line segments "enclosing" a planar surface. The next processing steps are the same as in the first example.

7 Summary

A model-based object recognition in video and depth images waqs proposed for the purpose of 2D and 3D map labeling in mobile robotics. Four types of objects were modeled: corridor walls, a human silhouette, a chair and a table. A bi-driven hypothesis generation and verification strategy was defined. The object model is expressed by a hierarchic semantic net, where each concept

Fig. 13. An empty corridor with no obstacles

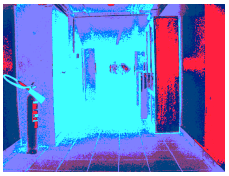

Fig. 14. Image regions

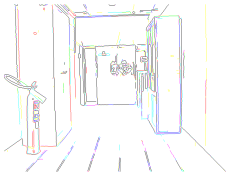

Fig. 15. Edge image

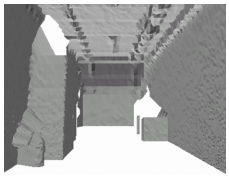

Fig. 16. 3D surface patches in depth image

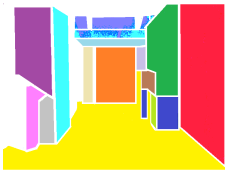

Fig. 17. 3D surfaces

Fig. 18. Wall detection

is also characterized by a graph of constraints (for matching) and a Bayesian network (for instance evaluation). The verification stage is modeled as a top-down propagation of occlusion relations, repeated for every subset of hypotheses. A successful verification means, that the evidence part set of given instance is extended by hidden parts, increasing the score of such an instance.

References

1. A.K. Jain, R.P.W. Duin and J. Mao: Statistical Pattern Recognition: A Review, IEEE Tran. on Pattern Analysis and Machine Intelligence, 22, 4-37 (2000), No. 1
2. D. Marr, Vision: A computational investigation into the human representation and processing of visual information. New Freeman, New York, 1982
3. S. Russel and P. Norvig, Artificial Intelligence. A modern approach. Prentice Hall, second edition, 2002
4. W. Kasprzak, A Linguistic Approach to 3-D Object Recognition, Computers & Graphics, 11, 427-443 (1987), No. 4 Pergamon Journals, London, UK
5. R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification and Scene Analysis. 2nd edition. J. Wiley, New York, 2001
6. H. Surmann, A. Nuchter, and J. Hertzberg: An autonomous mobile robot with a 3d laser range finder for 3d exploration and digitalization of indoor environments. Journal Robotics and Autonomous Systems, 45, 181–189 (2003)
7. Z. Zhang: Iterative Point Matching for Registration of Free-Form Curves. International Journal of Computer Vision, 13, 119–152 (1994)
8. O. Faugeras, M. Hebert, P. Mussi, and J. D. Boissonnat: Polyhedral approximation of 3-D objects without holes. In Computer Vision, Graphics, and Image Processing, 25, 169–183 (1984)
9. A. Jaklic, A. Leonardis, F. Solina: Segmentation and Recovery of Superquadrics. Computational imaging and vision, 20 (2000), Kluwer, Dordrecht
10. P. Dias, V. Sequeira, F. Vaz, J. G. M. Gonalves: Registration and Fusion of Intensity and Range Data for 3D Modeling of Real World Scenes. Proc. 4th International Conference on 3-D Digital Imaging and Modeling, 418-425 (2003)

11. M. Soucy, D. Laurendeau: Multiresolution surface modeling based on hierarchical triangulation. In Computer Vision and Image Understanding, 63, 1–14 (1996)
12. A. Wyszomierski: Detekcja osb w obrazach otoczenia autonomicznego pojazdu. B.Sc. diploma, ICCE WUT, Warsaw, 2007.
13. K. Przedniczek: Building and simplification of three-dimensional maps of an environment using 3D laser scanner *in Polish*). B.Sc. diploma, ICCE WUT, Warsaw, 2008.