IAPR Workshop on CV - Special Hardware and Industrial Applications OCT.12-14, 1988, Tokyo

A VISION SYSTEM FOR THE INTERPRETATION OF 3D COMPLEX MOVING SCENES

Daniele D. Giusto, Sebastiano B. Serpico, and Gianni Vernazza

Dept. of Biophysical and Electronic Engineering
University of Genoa

Via all'Opera Pia 11A, I-16145 Genoa, Italy

ABSTRACT

In this paper we propose a vision system aimed to interpret 3D scenes with moving objects which may give rise to partial or total occlusions. The system employs the knowledge-based approach to exploit both information extracted by the optical flow analysis and by a-priori knowledge about object models and perspective laws. In particular, the different information is used to apply a hypothesis generation-and-test paradigm. Optical flow is obtained by establishing correspondences among segmented regions in successive frames, by analyzing the feature space (average grey-level, area, shape). At the same time, the recognition process is performed on single frames, to provide rough interpretations of the scenes. At this point, the two above kinds of results (optical flow and rough interpretation) are appropriately merged to solve ambiguous situations. All the information disposable is exploited to generate an interpretation hypothesis of the moving scene, that is tested according to rigid motion hypothesis and perspective laws. The above system has been tested on inner scenes acquired in our lab, considering real objects of simple shapes. Preliminary results are promising and confirm the interest of our approach.

INTRODUCTION

Recognition of 3D moving scenes is one of the most challenging problems in the computer vision and image understanding field. Visual analysis of motion generally involves two successive steps. The first step consists in the examination of the variations that have occurred in an image due to relative shiftings between the scene and the viewpoint. This procedure makes it possible to determine the so-called "optical flow" (the projection, on the image plane, of the vectors that indicate the shiftings of objects in a 3-D scene). Starting from this information, the motion parameters are then extracted (second step), also utilizing various kinds of constraints and a-priori knowledge.

Different systems for 3D scene interpretation have been proposed in literature, usually based on TV camera images and/or range data. The reconstruction of a 3D scene from 2D information is in general difficult and not always solvable. This problem has been faced by using direct or derived (by stereo analysis) range information, by using perspective laws, shape from shading, etc.; usually, some assumptions are made to consider simplified worlds.

Purpose of this paper is to present a system for moving scene interpretation. To determine the optical flow we use a technique based on the search for significant matching points in the various frames, by employing virtual points (i.e., a set of features of the main regions into which an original image has been appropriately segmented). Matching probabilities are provided by an iterative algorithms that has proved reliable and rather fast. The interpretation process is performed by a knowledge-based system obtained by using the basic architecture of a previous system developed for medical image understanding [1]. Numerical processing, and content of knowledge bases have been changed to take into account the peculiarity of the considered application.

SEGMENTATION AND OPTICAL FLOW CALCULATION

The low-level processing module is based on the following main steps:

- identification of significant regions inside an image (segmentation);
- extraction of feature vectors for each region;
- matching of the regions in the feature space.

In order to facilitate the segmentation process, a filtering procedure (edge-preserving smoothing) is adopted, which reduces the noise averaging only in almost uniform areas [2]. The segmentation procedure is of the region-growing type and is based on the analysis of local histograms of the differential values to determine an adaptive segmentation threshold. Through this analysis, the various pixels are processed and elementary regions (zones whose grey levels are similar) are detected [2]. Usually, the number of elementary regions in a frame is very large, but most of such regions are not very significant, since they are due to illumination disuniformity, to a too noise-sensitive segmentation threshold, etc..

The significant regions are a subgroup of the initial elementary regions, and are described by specific parameters. Consequently, in order to perform the merging step, a preliminary feature extraction process has to be activated for each elementary region. The merging module examines, basically, the area of each region and, if a region is smaller than a threshold, the module tries to merge it into a neighbouring one. After the area check, a region can be merged into the most similar neighbouring one, on the basis of some parameters, such as: average grey level along the adjacency and length of adjacency.

The parameter extraction module provides a feature vector for each region. Even a very small number of parameters can be sufficient to describe a region to a good accuracy, without considerably increasing the computation time.

The parameters can be of the geometric type (centroid, area, shape factor), intensity type (average grey level, contrast level), texture type, and colorimetric type (different wavelength response). The selected features should be of the "invariant type", i.e., their values should be independent of the region position or rotation. Many features have this peculiarity; however, for the sake of simplicity, the considered features at this stage are: area, contrast level, and simplified shape factor. The adopted shape factor is the elongatedness of a region, defined as the ratio between the moments associated with the principal axes of inertia of the region, which provide information also about its lengthening direction. This shape factor has been adopted since it does not depend on the perimeter, which is heavily modified by the filtering and segmentation steps.

The motion algorithm allows a comparison between two successive frames, according to the following sequence: |3|

- extraction of suitable matching points;
- calculation of the matching probability for each couple of points, and decision on the most likely matchings.

The main problem is the choice of appropriate matching points, which usually correspond to the points that exhibit sharp variations in their bright intensity (typically, contour points or contour lines). Instead, in this paper, motion analysis is made using virtual points (i.e., a set of features of the main regions into which an original image has been appropriately segmented). The basic reason for choosing virtual points lies in the consideration that one single pixel can be easily modified within a frame sequence; instead, it is more difficult that a virtual point, achieved as a "synthetic description" of a region clearly defined in the original image, should disappear or be heavily modified in a frame sequence. Consequently, the matching phase may turn out to be more advantageous if performed in the feature space.

The matching method adopted makes it possible to calculate the matching probabilities of each couple of regions in two successive frames by using an iterative algorithm. For each region belonging to the first frame, a vector is defined that contains the probabilities of matching each region of the second frame, taking also into account that a region might disappear, and then no matching with another region would be possible. Initialization of match probabilities is performed, taking into account the available information about regions of both frames (feature vectors). Match probabilities, for a region, are iteratively updated analyzing also the movements of neighbouring regions. The final result is represented by probabilities generally very close both to 1 for the right matches, and to 0 for the wrong matches.

SCENE INTERPRETATION

A knowledge-based system has been designed to perform the interpretation of the scene on the basis of a priori knowledge, perspective constraints, and motion anlysis.

The basic control strategy is represented by a hypothesis generation-and-test cycle which allows the best hypothesis to be chosen and to be subsequently accepted or discarded by using all disposable sources of knowledge.

The system architecture is based on an Interpreter of Production Rule, a domain knowledge base, a global database (GDB), and various routines.

The domain knowledge base contains production rules (representing procedural knowledge) and a net of frames (representing descriptive knowledge); production rules are hierarchically structured according to a task/subtask organization formalized by a net of frames, too. The global database contains the input data and the progressive results of the system processing, still in the form of frame networks.

The rule interpreter, on the basis of the current problem status stored inside the GDB, activates a rule which specifies how to apply the system procedures to the data in the GDB, according to the domain knowledge. In this way, the current problem status changes and the rule interpreter starts again. The cycle is repeated untill a termination condition is reached.

At present, we consider object of simple shapes: cubes, parallelograms, spheres, cones, and cylinders. Their models are represented inside the domain knowledge as frame networks containing information about their external surfaces, and about their 'qualitative views' (e.g., in simplified condition, a cube has 3 qualitative views: 1-sqaure, 2-rectangles, and 3-parallelograms views). Also objects (instances of models) are described, by giving specific object properties (e.g. edge-lenght for an instance of cube).

The interpretation process is performed on a symbolic representation of the input data, that is, by using a frame network in which elementary regions are described. In particular, the intrinsic features used for motion analysis are considered, in addition to adjacency to other regions, geometrical shape (square, circle, etc.) if it can be defined, correspondence with regions of previous and subsequent frames (according to motion analysis results).

Starting from models and symbolic representation of a frame in the sequence, production rules make the system operate as follows. First of all an hypothesis of background regions location is made. To this end, some heuristic rules are used which focus on regions extending all over one image dimension, or however on very large regions, motionless or very slowly moving, with concavities complementary to some model shape. These are not necessary conditions, but are hints to identify background regions.

The remaining regions are subdivided into groups of regions adjacent to one another. Such image subparts are analyzed separately, since their recognition hypotheses do not interact (at least in a single frame).

A list of all possible side shapes for the considered models (circle, rectangle, square, parallelogram, ellypse, mixed arcs-lines figures) is computed, indicating to which qualitative view they relate and, consequently, to which model they could belong.

Starting from the most discriminant one (i.e. the one relating to the least number of models) all shapes in the above list are searched for among the elementary regions shapes.

When a region is found corresponding to a selected shape, all possible qualitative views it could belong to are considered. Among them, a first selection is performed by testing adjacency information. If the region is adjacent (on a whole edge) to other regions, those hypotheses are privileged, if any, that include similar adjacency constraints in the related model qualitative views; the other hypotheses are suspended.

The second hypothesis selection is based on some perspective constraints linked to the qualitative views. Some of them depend on properties of models, while other can be applied after hypothesizing also the instances of models.

As an example, for the 2-rectangle qualitative view of cubes, the edge of contact of the two rectangles must be the longest edge for both of them (largest size); moreover, the other size of the largest rectangle must be greater than the square root of sin 45^ by the largest size.

Occlusions make interpretation more complex, so the system tries to start with nonoccluded parts of qualitative-views. Anyway, occlusions add some constraints on relative object position (occluded objects are more distant than occluding ones).

Notwithstanding all these attempts, some undecidable situations may remain (e.g. a circle that may be the view of a cone or a sphere or a cylinder). Then the system tries to scan the frame sequence to find new information sufficient to solve the situation. Obviously, not always a solution can be found.

The system could be increased by adding some heuristic rules derived from 'shape from shading' studies. A remarkable modification to obtain a more powerful system would be the use of a range sensor; it should be sufficient to obtain range data regarding few points selected by the system to solve ambiguities in the interpretation.

Some specific rules (with the related routines) manage the hypothesis tree, and all related information (current, active, and suspended global scene hypotheses; currently considered elements: image, group of regions, region-shape and elementary region for selecting views, selected view, model, instance, etc.).

RESULTS

The 3 images shown in Fig.1 have been considered for system testing. They were acquired by a TV camera, sampled, and stored as 512x512x8 bit images. After filtering (5 iterations), segmentation provided about 400 regions in all of the 3 cases; most of them were merged because of very small area or very low contrast on the whole border. At the end of this process, 9 or 10 regions were obtained.

The motion analysis provided the results in Fig.2a,b: some problems arose due to the different overlapping situations in the image of Fig.1c with respect to the one in Fig.1b.

Starting from the first image, the recognition system recognized the cube and the parallelepiped without any problem. On the contrary, it could not choose about sphere, cone, and cylinder hypotheses generated from the 2 circles. By analyzing the third image, a cylinder was found that allowed to solve one of the above ambiguities. The other one could not be solved. The recognition of the cylinder (starting from its lateral surface) allowed the view parts of both cube and cylinder to be reconstructed, so also motion problems could be solved. Final results are displayed in Fig.2c.

The system run on a μVAX for numerical processing (C and FORTRAN languages) and on a TI Explorer for symbolic processing (in Common-LISP). It took about 1 hour for low-level processing, 10 seconds for motion analysis, and 5 minutes for interpretation, globally for the 3 considered images.

DISCUSSION

Our system is still under development, in fact, while the numerical image processing (including also the motion analysis) has already been widely tested, the interpretation phase has been applied only to the presented example.

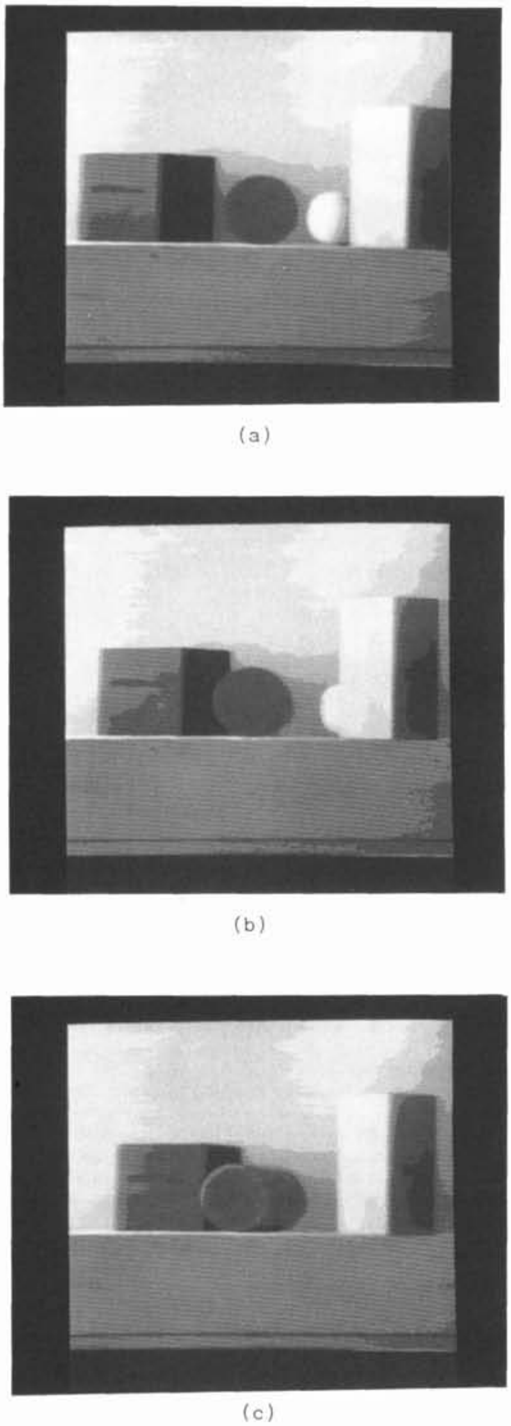
Anyway, in our opinion, the system we have described is an interesting framework in which other knowledge and procedures can be easily added to obtain a powerful system. Moreover, the proposed strategy (hypothesis generation-and-test) allows to make a restricted use of perspective laws (only as much as required for hypothesis validation, and driven by heuristic rules) so avoiding an heavy computational load. The merging of 3D interpretation and motion analysis is another important feature of our system.

REFERENCES

[1] G.L.Vernazza, S.B.Serpico, and S.G.Dellepiane, A knowledge-based system for biomedical image processing and recognition, IEEE Trans. on Circuits and Systems, vol. CAS 34, No 11, Nov. 1987, pp. 1399-1416.

[2] M.Nagao, and T.Matsuyama, A structural analysis of complex aerial photographs, Plenum Press, New York, 1980.

[3] D.D.Giusto, and G.Vernazza, Optical flow calculation from feature space analysis through an automatic segmentation process, Signal Processing, Dec. 1988 (in press).

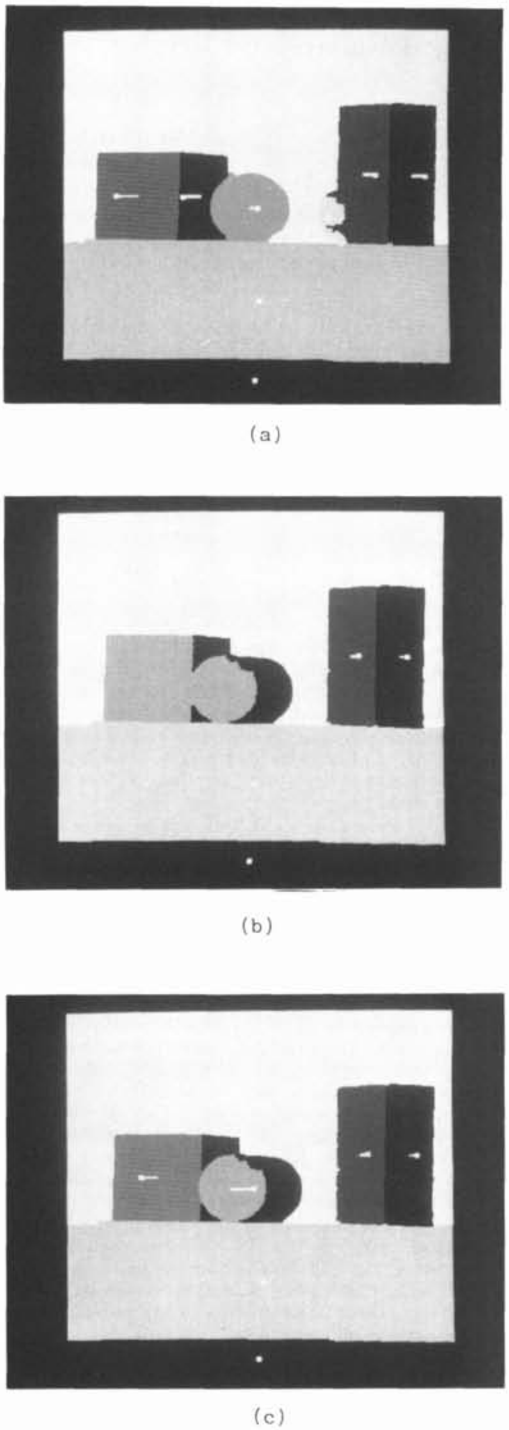


(a)

(b)

(c)

Fig.1. Original images acquired by TV camera. Some changing in occlusions, and region appearances, and disappeareances can be observed.



(a)

(b)

(c)

Fig.2. Results of motion analysis considering the images in Fig.1a and Fig.1b (a); and the images in Fig.1b and Fig.1c (b). In (c) unsolved situations of (b) are recovered by means of recognition results. Shifting vectors are indicates by white lines (small squares refer to the positions of the centroids in the previous frame).