

corresponding coordinates in the 3-D model and the image by computing the essential matrix. Unlike the stereo vision case, this method gives a solution without any scale ambiguity, if used with a 3-D model and an image, instead of two images. Moreover, the attitude estimation is very robust even in the case of small optical fields. This measurement is fed into a recursive motion estimator, which gives the current and the predicted values and derivatives of the translation and rotation components of the vehicle motion.

The 3-D coordinate systems of the model  $R_m$  and the sensor  $R_i$  are related by the composition of a rotation and a translation. The coordinates of a point in  $R_m$ ,  $[x_m \ y_m \ z_m]^T$ , and in  $R_i$ ,  $[x_i \ y_i \ z_i]^T$ , are related by:

$$[x_m \ y_m \ z_m]^T = R \left( [x_i \ y_i \ z_i]^T - \mathbf{T} \right)$$

The essential matrix introduced by Longuet-Higgins is defined by  $Q = RS$ , where  $S$  is the matrix representing the cross product by the translation vector  $\mathbf{T}$ . Using  $Q$ , we have the fundamental relation:

$$[x_m \ y_m \ z_m] Q [X_i \ Y_i \ 1]^T = 0$$

where  $X_i$  and  $Y_i$  are the angular coordinates of the image points, deduced from the pixel location  $(p_x, p_y)$  and the angular resolution in both directions  $(s_x, s_y)$ .

This equation, valid for  $P$  pairs of corresponding coordinates, can be rewritten as  $A\mathbf{b} = 0$ , with  $A$  of dimension  $P \times 9$  and  $\mathbf{b}$  of dimension 9 containing the coefficients of  $Q$  written in a single line. If the SVD of  $A$  is  $A = UDV^T$ , the optimal value of  $\mathbf{b}$  corresponds to the null space of  $A$  given by the direction associated with the smallest singular value of  $A$ . If the singular values are sorted in decreasing order, this direction will be given by the ninth column of  $V$ .

To obtain  $V$ , we actually compute the SVD of  $A^T A$  which is  $A^T A = VDV^T$ , as demonstrated in [1]. We can factorize the essential matrix into the product of  $R$  and  $S$  by using a SVD of  $Q$ . We obtain the rotation and translation component up to a sign ambiguity in  $\mathbf{T}$  and a symmetry ambiguity in  $R$ . We recover this ambiguity by checking the sign of the depth coordinate in both systems. We recover the real distance to the object by using the following relation:

$$\|\mathbf{T}\| = \frac{(X_i \mathbf{R}_3 - \mathbf{R}_1) [x_m \ y_m \ z_m]^T}{(X_i \mathbf{R}_3 - \mathbf{R}_1) \mathbf{T}}$$

We compute this value over all available points and average the measurements to obtain the final result. From the measurements of  $R$  and  $T$  over successive images, we compute the rotational and translational speed of the vehicle and give the predicted position and orientation of the sensor in the next image. A complete motion estimator would be obtained by the implementation of a vehicle state model and a noise measurement model into a Kalman filtering algorithm, which will not be dis-

cussed here. The predicted attitude of the sensor in the next frame is used to estimate the displacement of the image points between the two frames, which solves the point tracking problem.

## 5 Results

Experiments have been performed on sequences of both infrared images and synthetic images. The results on infrared images are presented from figures 1 through 5, showing the reconstruction of a building seen from a plane, and its recognition from a different viewpoint. We were not able to give a complete error characterization since we do not have a calibrated CAD-model of the building. The facet description of the building was computed from the 3-D point geometry of the scene in two steps: we first compute the 3-D location of the ground plane, assuming that most of the points we detected are on the ground. Next, we produce the geometry of the roofs using the information of linked points and complete the building facet description by assuming that the walls are vertical. The algorithms have been implemented in C on a Sparc station 10. The time needed for processing is about 1 minute for the reconstruction using the original sequence containing 50 frames each, of dimension 512 by 512. For the recognition phase, the time needed is 30s for a model of 50 points. Our future work will concentrate on using high level features such as segments and regions.

## 6 References

- [1] R. I. Hartley. *Calibration of cameras using the essential matrix*, in Proc. of the DARPA I. U. Workshop, pp 911-915, San Diego, CA, Jan 1992.
- [2] H.C. Longuet Higgins. *A computer algorithm for reconstructing a scene from two projections*, Nature, Vol. 293, pp 133-135, Sep 1981.
- [3] J.R. Bergen et al. *A three-frame algorithm for estimating two-component image motion*, IEEE PAMI, Vol 14, Sep 1992
- [4] T.J. Brodia and R. Chellappa. *Estimating the kinematics and structure of a rigid object from a sequence of monocular images*, IEEE Trans.PAMI., Vol 13, June 1991
- [5] B.K.P. Horn and B.G Schunck. *Determining optical flow*, Artificial Intelligence, vol 17, 1981.
- [6] D.P. Huttenlocher and S. Ullman. *Object recognition using alignment*, Proc. 1st ICCV, pp 102-111, London 1987.
- [7] D.H. Ballard. *Recognizing the Hough Transform to detect arbitrary shapes*, Patt. Recog. Vol. 13, pp 111-122, 1981
- [8] Y. Lamdan and H.J. Wolfson. *Geometric hashing: a general and efficient model-based recognition scheme*, in Proc. of the 2nd ICCV, 1988.
- [9] C. Tomasi and T. Kanade. *The factorization method for the recovery of shape and motion from image streams*, Image Understanding Workshop 1992.
- [10] S. Soatto and P. Perona et al. *Recursive motion and structure estimation with complete error characterization*, Computer Vision and Pattern Recognition, 1993
- [11] R. Haralick and L. Shapiro. *Computer and Robot Vision*, Vol 2.

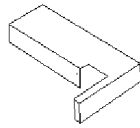


Figure 3. View of the reconstructed building.

$3 \times 3$  invertible matrix  $A$ . By writing that the sensor 3-D coordinate system is orthonormal at every frame, we can obtain linear metric constraints on the motion matrix. We compute the matrix  $A$  such as  $M = \tilde{M}A$  satisfies the metric constraints in the least sum-of-squares error sense. Once the matrices  $M$  and  $S$  have been computed, the F attitudes (position, orientation) of the sensor and the P 3-D positions of the points can be determined in the referential of the first image. The facet description of the object can be produced from its 3-D geometry by a model-based interpolation (polyhedral rigid objects) or by using an chain-coded description from the object edges. This algorithm is still under development and will not be described here.

### 3 The Recognition Phase

We address the problem of recognition of a previously reconstructed object. We suppose that an external system of navigation provide the approximate point-of-view of the sensor. However, this information is not accurate enough to predict the exact 2-D appearance of the object. The task of the system is to identify the object in the image, the matching of the image points of interest with the 3-D model points, and to give the exact 3-D attitude of the sensor relative to the object.

We find three general ways to consider the problem of matching 3-D model-based objects to 2-D image features. The alignment method [6] generates hypothetical transforms from features correspondences in the real and the predicted images. The correct transform, applied on the image features, will propagate the matching to the neighboring features and recursively to the whole image. To accomplish this we must reduce the image to a small set of distinguished high level features. This requires an “intelligent” interpretation of the image which is difficult to perform. Another is the Hough Transform [7] that assumes the model and the image are similar, up to a transform characterized by a set of parameters. The image features produce a set of weighted votes for different transforms. The final transform is the global maxima of the sum of the votes. A third way is to provide a mathematical formulation of the model which is invariant under geometric transforms and to recognize the model in the image, after reducing the image to the same formulation [8].

We use the predicted value of the sensor attitude to project the 3-D model into a set of 2-D points. Our recognition method suppose that approximating the deformation between the predicted 2-D appearance and the

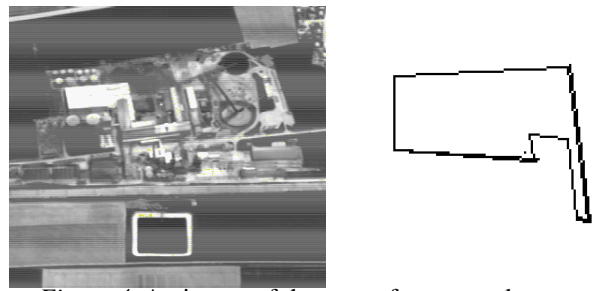


Figure 4. An image of the scene from an unknown viewpoint and the model predicted appearance.

real appearance of the object by a similarity is sufficient to find a large number of correspondences. We then refine the 3-D sensor attitude from this set of correspondences and solve the remaining ambiguities.

For recognition we use a geometric hashing formulation [8]. This method, based on the rigidity assumption, uses the fact that the coordinates of a 2-D object are invariant, up to a similarity, when expressed in a reference frame given by any pair of object points. We characterize the 2-D model in an orthonormal basis, which results from any “good” pair of model points. For each basis, the affine coordinates of the remaining points are calculated and stored in a two-dimensional hash table. Each entry of the hash table is a linked list, every node recording a pair of affine coordinates and the basis in which those coordinates were calculated. From the set of image points, we sort a list of 2-point bases that express the image-point coordinates. For each image base, we vote for each basis of the model that produced the same point affine coordinates. If a model base gets a sufficient number of votes, a match is declared successful between the image base and this model and, after a verification process, the global transform is computed over the set of matched points. The method is very sensitive to the number of bases which could be reduced by keeping the main segments or junctions of the image.

### 4 Self-Guiding

The problem of self-guiding can be decomposed into two major tasks. Tracking the object by matching the feature points from one image to another, in order to apply corrections to the orientation of the sensor (to keep the object visible and centered), and providing to the automated navigation system the estimation of the vehicle’s 3-D motion relative to the object, in order to command its future motion. The pose estimation of a 3-D model-based object in a single perspective image has already been studied with different approaches. These methods usually try to find the pose from the minimum number of correspondences. Therefore, they are sensitive to noise and complex to solve (non-linear optimization problem). Moreover, these methods are not dedicated to a tracking process.

We present a pose estimation method that determines the attitude of the sensor relative to the object from the

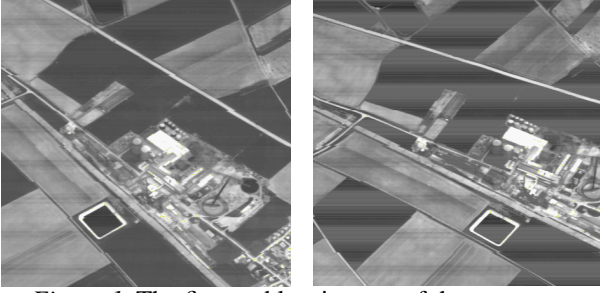


Figure 1. The first and last images of the sequence, showing the two extreme viewpoints.

can be significant and estimate the 3-D structure of an object viewed in the monocular sequence of images

There are two main approaches towards the problem of motion and structure from a monocular sequence of images. The optical flow approach computes the three-dimensional motion and structure parameters from the instantaneous brightness changes in the image [5]. The method is stable for a global estimation of the motion because it uses the textural information from the whole image, but the localization of specific points is not accurate. The second approach computes the motion and structure from a set of corresponding points. The rigidity assumption allows us to integrate (recursively or in batch mode) the information along the whole sequence and to compute the motion and structure, which is robust to the noise in the localization of the feature points [4]. The approach is accurate but has to solve the difficult feature matching problem

We present a feature-based reconstruction system that combines an intensity approach to compute a 2-D affine motion, followed by local matching to perform tracking, and a factorization of the measurement matrix in a paraperspective approximation to compute 3-D motion and structure. To perform the global motion estimation, we make use of the incremental approach developed in [3]. It assumes that the velocity components of the motion field on every pixel of the tracked region can be described by an affine transformation:

$$p_x(x, y, t) = a_x(t) + b_x(t)x + c_x(t)y$$

$$p_y(x, y, t) = a_y(t) + b_y(t)x + c_y(t)y$$

This approximation allows us to compute an image field that can combine the following motions: translation, scale, and rotation in the image plane. These transforms are the major components of the image motion due to displacement of the sensor. However, affine motion modeling does not take into account the variations of displacements of object features due to 3-D geometry. In our scenario, the sensor is relatively far from the object. Therefore the deviation of the point-of-view, which produces these displacements, is small. These small errors are corrected by the local matching. The coefficients of the affine transformation can be found by minimizing the brightness change constraint equation over the whole image, in a least-square sense.

$$\frac{d}{d(\text{coeff})} \left( \sum_{x,y} (I_t + p_x I_x + p_y I_y)^2 \right) = 0$$

The differentiation with respect to each of the parameters produces a linear system of six equations with six unknowns that can be easily solved. This estimation is extended to a general displacement by implementing two alignment procedures. The procedure is recursively applied to the first image and the warped second image, until the remaining affine transform is negligible. This alignment method is applied at all levels of a resolution pyramid of the images, starting at the lowest resolution. The algorithm produces the measurement of the best approximation of the motion field as an affine transform between the images. We apply the algorithm on the sequence and produce a predicted value for the next image pair. This is justified by the low frequency of the sensor attitude motion compared to the image frequency.

A complete 3-D model of the object contains the geometry and luminance of a set of facets. However, we limit the present work to the reconstruction of *geometry* only. To compute this CAD-like model, we detect and track interest points over the image sequence. We detect the interest points as the local extrema of a measure of similarity to a corner defined in [11]. Points in successive images can be matched using a combination of the global motion estimation and local matching procedure. The affine transform computed from the global motion estimation is applied to the points of the second image. A local matching solves the remaining ambiguities.

We perform 3-D reconstruction from the sequence of image points using the paraperspective model that approximates the perspective projection of a set of points as the projection onto a plane parallel to the image plane and passing through the object's center of mass. This approximation is justified in our case because we have a small optical field. A reconstruction algorithm has been developed in [9] by factorizing the  $[2F, P]$  measurement matrix (contains the image coordinates of  $P$  corresponding points during the sequence of  $F$  images) into a  $2F \times 3$  motion matrix and a  $3 \times P$  structure matrix. The factorization method extracts the best approximation of rank 3,  $W$ , from the noisy measurement matrix  $\hat{W}$  using the singular value decomposition. The SVD provides a possible decomposition of  $W$  into the product of two matrices  $\tilde{M}$  and  $\tilde{S}$ , defined up to any

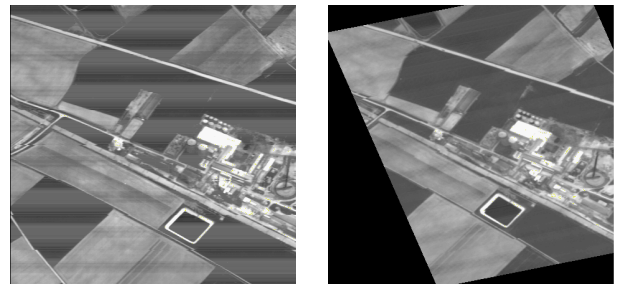


Figure 2. First original image (left) and last image warped to the same orientation as the first image.

# Learning, Recognition and Navigation from a Sequence of Infrared Images\*\*

Nicolas Milhaud\*

Société Anonyme de Télécommunications  
41 rue Cantagrel  
75013 Paris, France  
milhaud@satdod.dod.sat.fr

Gérard Medioni

Institute for Robotics and Intelligent Systems  
University of Southern California  
Los Angeles, CA 90089-0273  
medioni@iris.usc.edu

## Abstract

We address the problem where an autonomous system equipped with an infrared camera “learns” a designated rigid 3-D scene so that later it can recognize and guide itself relative to the reconstructed scene, starting from an approximately known viewpoint, to reach a given destination. This scenario is relevant to several domains, in particular military missions and robotic navigation. A goal of our system is that a real-time implementation be feasible on special hardware. In this paper, we describe the software version of such a system and show results on real infrared images.

## 1 Introduction

A major issue in autonomous vision systems is the recognition of objects from one or multiple images. This recognition can be performed automatically when the system has a sufficient a-priori knowledge of the object. This knowledge can be provided by several sources such as a human-made CAD model. An interesting scenario is to consider a system capable of learning an object structure from observations, so that it can recognize the object from a different attitude at a different time. We will impose the limitation that the point-of-view, although different, is approximately known, and that the objects are polyhedral, so that matched points correspond to the same physical 3-D point.

Once this recognition is done, a second useful task is to allow a self-guiding system to automatically evolve with a predefined trajectory relative to the model. The application is either to navigate in a rigid environment for moving automated systems, or to grasp a moving object for robots applications. For the first application, electronic devices provide a solution but are costly and not very accurate. For the second application, where the object is moving, image analysis is necessary.

The goal of this article is to present the design of a

system that attempts to solve the entire problem and works on the articulation of different existing methods and their contributions.

This article is in four parts, each describing a component of the system, showing the relation with previous work, and presenting results: We first describe the *learning phase*, where the system performs 3-D reconstruction of a designated object from a sequence obtained by a moving infrared sensor. It uses a feature extraction algorithm to reduce the object to a set of features, a tracking algorithm to perform a global analysis of the image motion, followed by a local matching of the image features and a reconstruction algorithm to recover the 3-D position of the features and the parameters of the camera motion.

Next we describe *recognition* of the reconstructed object from a different, approximately known, attitude. It involves a two-step 2-D to 2-D *correspondence* algorithm, that outputs the set of correspondences between the image features and the 3-D model features, and a correction of the estimated attitude of the input sensor. The third describes the *self-guiding* system that provides automatic tracking of the recognized object and a *recursive motion estimation* of the sensor attitude relative to the object. The fourth shows results on real infrared images.

## 2 The Learning Phase

We address here the problem of the 3-D structure estimation of a designated fixed object that is observed from a moving vehicle driven by a human operator. The automated system is able to track an object that is pointed to in the first image. After a short period of observation, during which the vehicle motion (manually controlled) produces a limited variation of the viewpoint, the system estimates both the object structure and the sensor motion. We suppose that the system does not have access to any external motion information. Therefore, the system has to solve the two following problems: automatically track a fixed object designated in one image from a moving sensor, to produce an image sequence of this object, when the inter-frame motion

\*Supported by Société Anonyme de Télécommunications.

\*\* This research was supported in part by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research under Contract No. F49620-90-C-0078 and/or Grant No F49620-93-1-0620.