RECONSTRUCTION OF STRUCTURE AND TEXTURE OF PLANAR ENVIRONMENTS BY DYNAMIC VISION TECHNIQUES

M. Cossi, G.M. Cortelazzo, R. Frezza

D.E.I., University of Padova
via Gradenigo 6/a, 35131 Padova, Italy
Tel. +39 49 8277825; fax: +39 49 8277826
e-mail: frezza@dei.unipd.it

ABSTRACT

This work is concerned with the estimate of structure and texture of buildings from a video sequence. The goal includes the recovery of metric information. The results could be conceivably used for many purposes ranging from photogrammetric applications to CAD models that could be applied, for example, for virtual visits of sites of artistic and historical significance.

We present an original algorithm to estimate both structure and texture of environments composed by planes like the interiors of most buildings. From a video sequence of a decorated wall the algorithm computes a plane that approximates the wall (structure estimation) and composes a mosaic of the single images to reproduce the decoration (texture estimation). The data are organized so that it is possible to observe the wall from an arbitrary point of view.

1 INTRODUCTION

Possible applications of the work presented in this paper range from photogrammetry to electronic publishing. The diffusion of electronic publishing products demands, in particular, the development of tools for easy authoring of virtual visits of remote sites. The work presented in this paper is a significant first step in this sense. The goal of the research project is the development of a tool that would allow automatic authoring of virtual visits from a simple sequence of images taken by an operator wondering through the site with a videocamera.

The approach proposed here is based on the consideration that, if the 3-D structure of the building and the position of the videocamera at each time instant were known, it would be possible to project the image texture data back on the walls, ceilings and floors creating a data structure that allowing for the computation of virtual images from any point of view. The presented algorithm can estimate the structure of multiplanar environments, from a video sequence, register the single images together laying texture on structure data and create a high resolution 3-D "mosaic".

The problem of estimating multiplanar environments from a video sequence has been studied by many researchers see, for example, [5, 2, 7, 8]. In most of the previous work the estimates are the planes fitting, in a least square sense, the 3-D position of a number of point features. One original aspect of our approach is that we assume an a-priori *probabilistic* model of the distribution of the point features. Most features do not actually belong to the wall, but are either slightly in front of the wall (paintings, hanging coats, etc.) or behind the wall (dents, window frames, etc.). These perturbations can be modeled in the a-priori distribution of the point features. We estimate the unknown parameters of the probabilistic distribution by maximum-likelyhood techniques, in order to obtain an unbiased estimate of the wall. Using standard photogrammetric techniques to compute the "ground-truth" we verify the precision of the estimate and compare it with standard least square techniques.

2 STRUCTURE ESTIMATION

Many algorithms have been proposed for feature based motion and structure estimation from a monocular sequence of images, see [6] for a survey and a unifying perspective on the most succesful algorithms presented by different researchers. For the application presented in this paper, we used a scheme specifically derived for features that are distributed on planes originally proposed by J. Weng, T. S. Huang and N. Ahuja [10, 11] and then, slightly modified, by O. Faugeras [2]. The output of the algorithm are the parameters of motion that describe the trajectory of the videocamera with respect to a reference frame fixed with its initial position and the 3-D position of the features points $P_k(x_k, y_k, z_k)$, $k = 1, \ldots, N$ always w.r.t. the same frame.

The next step in the interpretation of the images consists in a higher level analysis in which feature points are grouped together when belonging to the same plane. Associating more points to single objects allows for better estimates of their position. Clearly, this kind of processing needs a set of a-priori models describing those objects in the scene that one wants to identify and locate. H. Maître [5], for example, assumes that the feature points may belong to planes or quadratic surfaces such as

$$z = Ax + By + C$$

$$z = Ax^2 + By^2 + Cxy + Dx + Ey + F$$

He estimates with a least squares algorithms the parameters $\{A, B, C, D, E, F\}$ in order to locally describe the structure of the environment.

Here, we propose a probabilistic model of the distribution of the features. Further details may be found in [1]. We assume, in particular, that the features appearing in one image may:

i) belong to a wall;

ii) be close to a wall, but not actually belong to it (they may, for example, belong to objects hanged to the wall);

iii) be far from the wall (they may, for example, belong to another wall).

For each of these three classes one may introduce a probability density describing the position of the feature points relative to the wall. Let the wall be described by the plane of equation

$$ax + by + cz - d = 0 \qquad (1)$$

The distance r_k of a point of coordinates $P_k = (x_k, y_k, z_k)$ from the plane is given by

$$r_k = x_k + by_k + cz_k - d.$$

Given parameters $[a, b, c, d]$ of plane (1), the features belonging to class (i) can be described by the probability density

$$f_1(r|[a, b, c, d]) = \delta(r),$$

those belonging to class (ii) by

$$f_2(r|[a, b, c, d,]) = \frac{1}{\lambda}\exp(r/\lambda),$$

finally, since those belonging to (iii) can be anywhere, they can be described by a uniform probability density

$$f_3(r|[a, b, c, d]) = 1/r', \quad -r' \leq r \leq 0.$$

Choosing $r' > d$ one can also model dents in the wall.

The model must also take into account that the estimates of the position of the feature points P_k are noisy. If r_t is the true distance from the wall of a feature point, we assume that its measured distance r is distributed normally according to

$$f(r|r_t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{r - r_t}{\sigma}\right)^2\right].$$

We obtain, therefore, the following probabilistic model for the distribution of the feature points conditioned to the parameters of the plane (1)

$$\begin{aligned} f(r|[a, b, c, d]) \quad = \quad & k_1 f_{1,n}(r|[a, b, c, d]) + \\ & + k_2 f_{2,n}(r, [a, b, c, d]) + \\ & + k_3 f_{3,n}(r|[a, b, c, d]), \end{aligned} \qquad (2)$$

with $k_1 + k_2 + k_3 = 1$ and where, since the density of the sum of two independent r.v.'s is the convolution of the densities, the observations of the features belonging to class (i) are distributed according to

$$f_{1,n}(r|[a, b, c, d]) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{r}{\sigma}\right)^2\right],$$

those to class (ii) according to

$$f_{2,n}(r|[a,b,c,d]) = \frac{1}{\lambda}\exp\frac{\sigma^2}{2\lambda^2}\exp r\lambda Q(\frac{r}{\sigma}+\frac{\sigma}{\lambda}),$$

where Q is the complement of the Gaussian

$$Q(y) = \int_y^{+\infty}\frac{1}{\sqrt{2\pi}}\exp-\frac{y^2}{2}dy = 1-\Phi(y),$$

and those of class (iii) according to

$$f_{3,n}(r|[a,b,c,d]) = \frac{\Phi(\frac{r}{\sigma})-\Phi(\frac{r+r'}{\sigma})}{r'}.$$

The unknown parameters of this probabilistic model, including those describing the plane, are $[a,b,c,d]$, λ, σ and r'. We estimate them by maximizing the likelyhood (2) of the observations.

The estimates, denoted as $\phi_{M.L.}$, are smoothed and refined by a Kalman filter integrating them in time. Let T and R be the translation and the rotation of the videocamera from one time instant t to the next $t+1$. The relationship between the coordinates P of one feature point w.r.t a reference frame fixed with the camera at the two consequent time instants is

$$P_{t+1} = R(P_t - T).$$

If at time t the wall is described by equation $n_t^T P_t = d_t$, at the next time instant, w.r.t. the camera frame, it will be described by

$$n_t^T R^T P_{t+1} + n_t^T T = d_t \rightarrow n_{t+1}^T P_{t+1} = d_{t+1}$$

where $n_{t+1} = Rn_t$ and $d_{t+1} = d_t - N_t^T T$. It is, then, possible to apply a K.F. to the dynamic model

$$\begin{aligned}\begin{bmatrix} d(t+1) \\ n(t+1)\end{bmatrix} &= \begin{bmatrix} 1 & -\hat{T} \\ 0 & \hat{R}\end{bmatrix}\begin{bmatrix} d(t) \\ n(t)\end{bmatrix} + \nu \\ \phi_{M.L.}(t+1) &= \begin{bmatrix} d(t+1) \\ n(t+1)\end{bmatrix} + \eta \qquad (3)\end{aligned}$$

where the process noise ν is related to the noise in the estimates $\hat{T}$ and $\hat{R}$ of the motion parameters T and R, while η is the error in the maximum likelyhood estimates of the plane parameters whose variance may be approximated with the Cramèr-Rao bound.

Some results of simulations on real image sequences are presented in the figures shown in the last page.

3 TEXTURE ESTIMATION AND IMAGE MOSAICING

One of the peculiarities of electronic publishing most worth exploiting is the possibility of representing an object at predetermined image/object ratio. This is achieved by "mosaicing" together a number of high resolution images representing each a small piece of the object [?]. The algorithm presented in the previous section produces an estimate of the 3-D structure of a multiplanar scene together with the position of the videocamera relative to it. This allows to register all images by back projecting them on the estimated structure. We merge, therefore, *texture with structure data*. By computer graphics techniques it is, then, possible to generate a virtual image taken from any point of view.

By projective transformation techniques, see, for example, [4, 9], images of a single wall (planar environment) can be registered withouth knowing the parameters of the plane modeling the wall and even without camera calibration. However, in multiplanar environments, back projection on the 3-dimensional structure simplifies both the generation of virtual images and the definition of the trajectory of a virtual visitor. This redundancy may be used to estimate the error in the reconstruction of the wall.

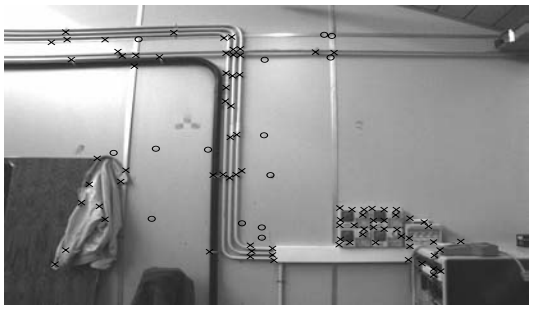

Figure 1: Some features extracted in a sample image of a videosequence taken in our laboratory. The circled ones are manually selected to estimate, with standard photogrammetric techniques, the true plane parameters.

References

[1] M. Cossi. Stima ricorsiva di ambienti planari a tratti da sequenze di immagini. Tesi

di Laurea, DEI Università di Padova, Italy, 1994 (in Italian).

[2] O.D. Faugeras and D. Lustman. Motion and structure from motion in a piecewise planar environment. *Int. J. of Pattern Recognition and Artificial Intelligence*, 2(3):485–508, 1988.

[3] S. Mann, R. W. Picard. 'Video Orbits': Characterizing the coordinate transformation between two images using the projective group. Proc. ICCV'95, 1995.

[4] H. Maître and W. Luo. Using models to improve stereo reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-14(2):269–277, 1992.

[5] S. Soatto, P. Perona. Motion estimation from point features; unified view. *Proc. of the IEEE Int. Conf. on Image Processing ICIP '95,* Oct. 1995.

[6] M. Straforini, C. Coelho, and M. Campani. Extraction of vanishing points from images of indoor and outdoor scenes. *Image and Vision Computing*, 11(2):91–99, 1993.

[7] M. Straforini, C. Coelho, M. Campani, and V. Torre. The recovery and understanding of a line drawing from indoor scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-14(2):298–303, 1992.

[8] R. Szeliski. Image mosaicing for tele-reality applications. Tech. Rep. CRL 94/2, DEC, May 1994.

[9] J. Weng, N. Ahuja, T. S. Huang. Motion and structure from point correspondences with error estimation: Planar surfaces. *IEEE Trans. on Signal Processing,* SP-39(12):2691–2717,1991.

[10] J. Weng, N. Ahuja, T. S. Huang. Optimal motion and structure estimation. *IEEE Trans. on Pat. Anal. Mach. Intel.,* PAMI-15(9):864–884, 1993.

[11] G. M. Cortelazzo, C. Doignon, R. Frezza. High resolution visualization of paintings and frescoes. *Proc. Int. Conf. on Environment and Climate,* Roma, March 1996.

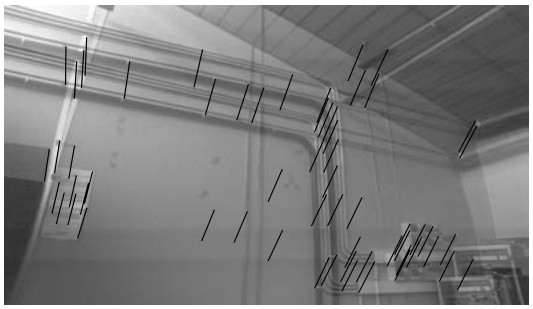

Figure 2: Example of feature tracking. Two successive frames are superimposed.

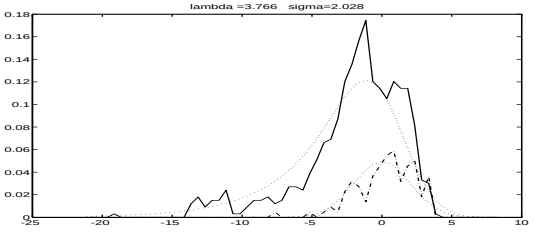

Figure 3: Estimated probability densities against experimental distribution of the features. The smaller one is relative only to manually selected features belonging to the plane and it representative of the error in the estimates of the motion parameters.

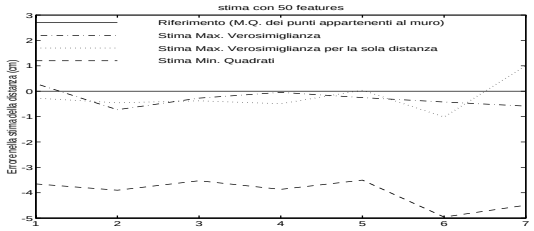

Figure 4: Error in the estimate of the d parameter in the wall model. Unlike the least sqaure estimate (M.Q.), the maximum likelyhood estimate (M.V.) does not present a significant bias.