LOCALIZATION USING COMBINATIONS OF MODEL VIEWS

Ronen Basri

Dept of Applied Math
The Weizmann Institute of Science
Rehovot, Israel 76100

Ehud Rivlin

Center for Automation Research
University of Maryland
College Park, MD 20742-3411

Abstract

A method for localization, the act of recognizing the environment, is presented. The method is based on representing the scene as a set of 2D views and predicting the appearances of novel views by linear combinations of the model views. The method accurately approximates the appearance of scenes under weak perspective projection. Analysis of this projection as well as experimental results demonstrate that in many cases this approximation is sufficient to accurately describe the scene. When weak perspective approximation is invalid, either a larger number of models can be acquired or an iterative solution to account for the perspective distortions can be employed.

The method has several advantages over other approaches. It uses relatively rich representations; the representations are 2D rather than 3D; and localization can be done from only a single 2D view.

1 Introduction

Basic tasks in autonomous robot navigation are localization and positioning. *Localization* is the act of recognizing the environment, that is, assigning consistent labels to different locations, and *positioning* is the act of computing the coordinates of the robot in the environment. Positioning is a task complementary to localization, in the sense that position (e.g., "1.5 meters northwest of table T") is often specified in a place-specific coordinate system ("in room 911").

This paper addresses the problem of localization. Positioning is addressed in [6]. Unlike existing methods, which represent the environment using $3D$ models (e.g., [1, 2, 4]), our method, based on the linear combinations scheme of [7], represents scenes by sets of their 2D images. Localization is achieved by comparing the observed image to linear combinations of model views.

The rest of the paper is organized as follows. The next section describes the method of localization using linear combinations of model views. The method assumes weak perspective projection. An iterative scheme to account for perspective distortions is presented in Section 3. An analysis of the error resulting from the projection assumption is presented in Section 4. Experimental results follow.

2 Localization

The problems of localization and object recognition are similar in many ways. Both problems require the matching of visual images to stored models, either of the environment or of the observed objects. Both problems face similar difficulties, such as varying illumination conditions and changes in appearance due to viewpoint changes. Similar methodologies therefore often are used to handle both problems.

The problem of localization is defined as follows: given P, a 2D image of a place, and $\mathcal{M}$, a set of stored models, find a model $M^i \in \mathcal{M}$ such that P matches M^i. A method for localization, based on the "Linear Combinations" (LC) scheme [7], is defined as follows. Given an image, we construct two view vectors from the feature points in the image (x and y-coordinates). The environment is modeled by a set of such views, where the points in these views are ordered in correspondence. The appearance of a novel view of the object is predicted by applying linear combinations to the stored views. The predicted appearance is then compared with the actual image, and the object is recognized if the two match.

Formally, given P, a 2D image of a scene, and $\mathcal{M}$, a set of stored models, the objective is to find a model $M^i \in M$ such that $P = \sum_{j=1}^{k} \alpha_j M_j^i$ for some constants $\alpha_j \in \mathcal{R}$. More concretely, let $p_i = (x_i, y_i, z_i)$,

0-8186-3870-2/93 $3.00 © 1993 IEEE

$1 \leq i \leq n$, be a set of n points in the environment. Under weak perspective projection, the position $p'_i = (x'_i, y'_i)$ of these points in the image are given (in vector equation) by

$$\begin{aligned} \mathbf{x}' &= sr_{11}\mathbf{x} + sr_{12}\mathbf{y} + sr_{13}\mathbf{z} + t_x\mathbf{1} \\ \mathbf{y}' &= sr_{21}\mathbf{x} + sr_{22}\mathbf{y} + sr_{23}\mathbf{z} + t_y\mathbf{1} \end{aligned} \quad (1)$$

where r_{ij} are the components of a 3×3 rotation matrix, s is a scale factor. Notice that $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{x}', \mathbf{y}' \in R^n$. Consequently,

$$\mathbf{x}', \mathbf{y}' \in span\{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{1}\} \quad (2)$$

or, in other words, $\mathbf{x}'$ and $\mathbf{y}'$ belong to a four-dimensional linear subspace of $\mathcal{R}^n$. (Notice that $\mathbf{z}'$, the vector of depth coordinates of the projected points, also belongs to this subspace. This fact is used in Section 3 below.) A four-dimensional space is spanned by any four linearly independent vectors of the space. Two views of the scene supply four such vectors [5, 7]. Denote by $\mathbf{x}_1$, $\mathbf{y}_1$ and $\mathbf{x}_2$, $\mathbf{y}_2$ the location vectors of the n points in the two images; then there exist coefficients a_1, a_2, a_3, a_4 and b_1, b_2, b_3, b_4 such that

$$\begin{aligned} \mathbf{x}' &= a_1\mathbf{x}_1 + a_2\mathbf{y}_1 + a_3\mathbf{x}_2 + a_4\mathbf{1} \\ \mathbf{y}' &= b_1\mathbf{x}_1 + b_2\mathbf{y}_1 + b_3\mathbf{x}_2 + b_4\mathbf{1} \end{aligned} \quad (3)$$

(Note that the vector $\mathbf{y}_2$ already depends on the other four vectors.) Since R is a rotation matrix, the coefficients satisfy the following two quadratic constraints:

$$\begin{gathered} a_1^2 + a_2^2 + a_3^2 - b_1^2 - b_2^2 - b_3^2 = \\ 2(b_1b_3 - a_1a_3)r_{11} + 2(b_2b_3 - a_2a_3)r_{12} \\ a_1b_1 + a_2b_2 + a_3b_3 + (a_1b_3 + a_3b_1)r_{11} + \\ (a_2b_3 + a_3b_2)r_{12} = 0 \end{gathered} \quad (4)$$

To derive these constraints the transformation between the two model views should be recovered. This can be done under weak perspective using a third image. Alternatively, the constraints can be ignored, in which case the system would confuse rigid transformations with affine ones. This usually does not prevent successful localization since generally scenes are fairly different from one another.

The scheme therefore is the following. The environment is modeled by a set of images with correspondence between the images. For example, a spot can be modeled by two of its corresponding views. The corresponding quadratic constraints may also be stored. Localization is achieved by recovering the linear combination that aligns the model to the observed image. The coefficients are determined using four model points and their corresponding image points by solving a linear set of equations. Three points are sufficient to determine the coefficients if the quadratic constraints are also considered. Additional points may be used to reduce the effect of noise.

The LC scheme uses viewer-centered models, that is, representations that are composed of images. It has a number of advantages over methods that build full three-dimensional models to represent the scene. First, by using viewer-centered models that cover relatively small transformations we avoid the need to handle occlusions in the scene. If from some viewpoints the scene appears different because of occlusions we utilize a new model for these viewpoints. Second, viewer-centered models are easier to build and to maintain than object-centered ones. The models contain only images and correspondences. By limiting the transformation between the model images one can find the correspondence using motion methods. If large portions of the environment are changed between visits a new model can be constructed by simply replacing old images with new ones.

One problem with using the LC scheme for localization is due to the weak perspective approximation. In contrast with the problem of object recognition, where we can generally assume that objects are small relative to their distance from the camera, in localization the environment surrounds the robot and perspective distortions cannot be neglected. The limitations of weak perspective modeling are discussed both mathematically and empirically in the next two sections. It is shown that in many practical cases weak perspective is sufficient to enable accurate localization. The main reason is that the problem of localization does not require accurate measurements in the entire image; it only requires identifying a sufficient number of spots to guarantee accurate naming. If these spots are relatively close to the center of the image, or if the depth differences they create are relatively small (as in the case of looking at a wall when the line of sight is nearly perpendicular to the wall), the perspective distortions are relatively small, and the system can identify the scene with high accuracy.

By using weak perspective we avoid stability problems that frequently occur in perspective computations. We can therefore compute the alignment coefficients by looking at a relatively narrow field of view.

When perspective distortions are relatively large and weak perspective is insufficient to model the environment, two approaches can be used. One possibility is to construct a larger number of models so as to keep the possible changes between the familiar and the novel views small. Alternatively, an iterative computation can be applied to compensate for these

distortions. Such an iterative method is described below.

3 Handling Perspective Distortions

The linear combination scheme presented above accurately handles changes in viewpoint assuming the images are obtained under weak perspective projection. Error analysis and experimental results demonstrate that in many practical cases this assumption is valid. In cases where perspective distortions are too large to be handled by a weak perspective approximation, matching between the model and the image can be facilitated in two ways. One possibility is to avoid cases of large perspective distortion by augmenting the library of stored models with additional models. In a relatively dense library there usually exists a model that is related to the image by a sufficiently small transformation avoiding such distortions. The second alternative is to improve the match between the model and the image using an iterative process. In this section we consider the second option.

The suggested iterative process is based on a Taylor expansion of the perspective coordinates. As described below, this expansion results in a polynomial consisting of terms each of which can be approximated by linear combinations of views. The first term of this series represents the orthographic approximation. The process resembles a method of matching 3D points with 2D points described recently by DeMenthon and Davis [3]. In this case, however, the method is applied to 2D models rather than 3D ones. In our application the 3D coordinates of the model points are not provided; instead they are approximated from the model views.

An image point $(x, y) = (fX/Z, fY/Z)$ is the projection of some object point, (X, Y, Z) in the image, where f denotes the focal length. Consider the following Taylor expansion for $F(Z) = 1/Z$ around some depth value Z_0:

$$\begin{aligned}\frac{1}{Z} &= \sum_{k=0}^{\infty} \frac{F^{(k)}(Z_0)}{k!}(Z - Z_0)^k \\ &= \frac{1}{Z_0} + \sum_{k=1}^{\infty} \frac{(-1)^k}{(k-1)!}\frac{(Z - Z_0)^k}{Z_0^{k+1}} \\ &= \frac{1}{Z_0}\left[1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{(k-1)!}\left(\frac{Z - Z_0}{Z_0}\right)^k\right]\end{aligned} \quad (5)$$

The Taylor series describing the position of a point $x = fX/Z$ is therefore given by

$$x = \frac{fX}{Z_0}\left[1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{(k-1)!}\left(\frac{Z - Z_0}{Z_0}\right)^k\right] \quad (6)$$

Notice that the zero term contains the orthographic approximation for x. Denote by $\Delta^{(k)}$ the kth term of the series:

$$\Delta^{(k)} = \frac{fX}{Z_0}\frac{(-1)^k}{(k-1)!}\left(\frac{Z - Z_0}{Z_0}\right)^k \quad (7)$$

A recursive definition of the above series is given below.

Initialization:

$$x^{(0)} = \Delta^{(0)} = \frac{fX}{Z_0}$$

Iterative step:

$$\begin{aligned}\Delta^{(k)} &= -\frac{Z - Z_0}{(k-1)Z_0}\Delta^{(k-1)} \\ x^{(k)} &= x^{(k-1)} + \Delta^{(k)}\end{aligned}$$

where $x^{(k)}$ represents the kth order approximation for x, and $\Delta^{(k)}$ represents the highest order term in $x^{(k)}$.

According to the orthographic approximation both X and Z can be expressed as linear combinations of the model views (Eq. (3)). We therefore apply the above procedure, approximating X and Z at every step using the linear combination that best aligns the model points with the image points. The general idea is therefore the following. First, we estimate $x^{(0)}$ and $\Delta^{(0)}$ by solving the orthographic case. Then, at each step of the iteration we improve the estimate by seeking the linear combination that best estimates the factor

$$-\frac{Z - Z_0}{(k-1)Z_0} \approx \frac{x - x^{(k-1)}}{\Delta^{(k-1)}} \quad (8)$$

Denote by $\mathbf{x} \in \mathcal{R}^n$ the vector of image point coordinates, and denote by

$$P = [\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{1}] \quad (9)$$

an $n \times 4$ matrix containing the position of the points in the two model images. Denote by $P^+ = (P^T P)^{-1} P^T$ the pseudo-inverse of P (we assume P is overdetermined). Also denote by $\mathbf{a}^{(k)}$ the coefficients computed for the kth step. $P\mathbf{a}^{(k)}$ represents the linear combination computed at that step to approximate the X or the Z values. Since at every step Z_0, f, and k are constant they can be merged into the linear combination.

Denote by $\mathbf{x}^{(k)}$ and $\mathbf{\Delta}^{(k)}$ the vectors of computed values of x and Δ at the kth step. An iterative procedure to align a model to the image is described below.

Initialization:
Solve the orthographic approximation, namely

$$\begin{aligned} \mathbf{a}^{(0)} &= P^{+}\mathbf{x} \\ \mathbf{x}^{(0)} = \mathbf{\Delta}^{(0)} &= P\mathbf{a}^{(0)} \end{aligned}$$

Iterative step:

$$\begin{aligned} \mathbf{q}^{(k)} &= (\mathbf{x} - \mathbf{x}^{(k-1)}) \div \Delta^{(k-1)} \\ \mathbf{a}^{(k)} &= P^{+}\mathbf{q}^{(k)} \\ \Delta^{(k)} &= (P\mathbf{a}^{(k)}) \otimes \mathbf{\Delta}^{(k-1)} \\ \mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)} + \mathbf{\Delta}^{(k)} \end{aligned}$$

where the vector operations $\otimes$ and $\div$ are defined as

$$\begin{aligned} \mathbf{u} \otimes \mathbf{v} &= (u_1 v_1, \ldots, u_n v_n) \\ \mathbf{u} \div \mathbf{v} &= (\frac{u_1}{v_1}, \ldots, \frac{u_n}{v_n}) \end{aligned}$$

4 Projection Model—Error Analysis

In this section we develop an error term for the LC scheme assuming that both the model views and the incoming image are obtained by perspective projection.

The error obtained by using the LC scheme is given by

$$E = |x - ax_1 - by_1 - cx_2 - d| \tag{10}$$

Since the scheme accurately predicts the appearances of points under weak perspective projection, it satisfies

$$\hat{x} = a\hat{x}_1 - b\hat{y}_1 - c\hat{x}_2 - d \tag{11}$$

where accented letters represent orthographic approximations. Assume that in the two model pictures the depth ratios are roughly equal:

$$\frac{Z_0^M}{Z^M} = \frac{Z_{01}}{Z_1} \approx \frac{Z_{02}}{Z_2} \tag{12}$$

(This condition is satisfied, for example, when between the two model images the camera only translates along the image plane.) Using the fact that

$$x = \frac{fX}{Z} = \frac{fX}{Z_0}\frac{Z_0}{Z} = \hat{x}\frac{Z_0}{Z} \tag{13}$$

we obtain

$$\begin{aligned} E &= |x - ax_1 - by_1 - cx_2 - d| \\ &\approx \left|\hat{x}\frac{Z_0}{Z} - a\hat{x}_1\frac{Z_0^M}{Z^M} - b\hat{y}_1\frac{Z_0^M}{Z^M} - c\hat{x}_2\frac{Z_0^M}{Z^M} - d\right| \\ &= \left|\hat{x}\frac{Z_0}{Z} - (a\hat{x}_1 - b\hat{y}_1 - c\hat{x}_2)\frac{Z_0^M}{Z^M} - d\right| \\ &= \left|\hat{x}\frac{Z_0}{Z} - (\hat{x} - d)\frac{Z_0^M}{Z^M} - d\right| \qquad (14) \\ &= \left|\hat{x}(\frac{Z_0}{Z} - \frac{Z_0^M}{Z^M}) - d(\frac{Z_0^M}{Z^M} - 1)\right| \\ &\leq |\hat{x}|\left|\frac{Z_0}{Z} - \frac{Z_0^M}{Z^M}\right| + |d|\left|\frac{Z_0^M}{Z^M} - 1\right| \end{aligned}$$

The error therefore depends on two terms. The first gets smaller as the image points get closer to the center of the frame and as the difference between the depth ratios of the model and the image gets smaller. The second gets smaller as the translation component gets smaller and as the model gets close to orthographic.

Following this analysis, weak perspctive can be used as a projection model when the depth variations in the scene are relatively low and when the system concentrates on the center part of the image. We conclude that, by fixating on distinguished parts of the environment, the linear combinations scheme can be used for localization.

5 Experiments

The LC method was implemented and applied to images taken in an indoor environment. Images of several offices were taken. Semi-static objects, such as heavy furniture and pictures, were used to distinguish between the offices. Due to lack of space we limit ourselves here to a presentation of a match between one model and an image, where the image was taken after a large motion forward and to the left. The views were taken at a distance of about 5m from the wall. Correspondences were picked manually. Figure 1 shows the two model views, and Figure 2 shows the results of matching a linear combination of the model views to an image of the same office. In this case, because the image was taken from a relatively close distance, perspective distortions cannot be neglected. Perspective effects were reduced by using the iterative process presented in Section 3. The results of applying this procedure after three iterations are shown in Figure 2.

The experimental results demonstrate that the LC method achieves accurate localization in many cases,

Figure 1: Two model views of an office.

Figure 2: Matching the model to an image obtained by a relatively large motion (left). Perspective distortions can be seen in the table, the board, and the hanger at the upper right. The results of applying three iterations in the process for reducing the perspective distortions is presented in the right.

and that when the method fails because of large perspective distortions an iterative computation can be used to improve the quality of the match.

6 Conclusions

A method of localization was presented. The method is based on representing the scene as a set of 2D views and predicting the appearance of novel views by linear combinations of the model views. The method accurately approximates the appearances of scenes under weak perspective projection. Analysis of this projection as well as experimental results demonstrate that in many cases this approximation is sufficient to accurately describe the scene. When the weak perspective approximation is invalid, either a larger number of models can be acquired or an iterative solution can be employed to account for the perspective distortions.

The method presented in this paper has several advantages over existing methods. It uses relatively rich representations; the representations are 2D rather than 3D, and localization can be done from a single 2D view only. Application of the same basic method to the positioning problem is given in [6].

References

[1] N. Ayache and O. D. Faugeras. Maintaining representations of the environment of a mobile robot, *IEEE Trans. on Robotics and Automation,* Vol. 5, pp. 804–819, 1989.

[2] D. J. Braunegg. Marvel—A system for recognizing world locations with stereo vision. *AI-TR-1229, MIT,* 1990.

[3] D. F. DeMenthon and L. S. Davis. Model-based object pose in 25 lines of code. *Proc. 2nd European Conf. on Computer Vision,* Genova, Italy, 1992.

[4] C. Fennema, A. Hanson, E. Riseman, R. J. Beveridge, and R. Kumar. Model-directed mobile robot navigation. *IEEE Trans. on Systems, Man and Cybernetics,* Vol. 20, pp. 1352–1369, 1990.

[5] T. Poggio. 3D object recognition: on a result by Basri and Ullman. *Technical Report 9005-03, IRST,* Povo, Italy, 1990.

[6] E. Rivlin and R. Basri, Localization and Positioning using Combinations of Model Views. A.I. Memo 1376, M.I.T., 1992.

[7] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 13, pp. 992–1006, 1991.