Automatic Visual Alignment Using Planar Regional Features and Stereo Vision

Forster, Carlos H. Q.
UNICAMP - Universidade Estadual de Campinas
FEEC, DCA, CP6101
13081-970, Campinas, SP Brazil
forster@dca.fee.unicamp.br

Tozzi, Clésio L.
UNICAMP - Universidade Estadual de Campinas
FEEC, DCA, CP6101
13081-970, Campinas, SP Brazil
clesio@dca.fee.unicamp.br

ABSTRACT

This paper addresses the determination of the rigid transformation between camera and object reference frames from a pair of intensity images and a known scene model. Two difficult parts of this problem that deserve particular attention are the matching between image and model features and the matching of image-features between stereo views. We propose the use of planar regions as features, what make both problems simpler. The former is handled by an invariant-based approach, for which a less complex base can be adopted, and the latter, by applying the epipolar constraint for inferior and superior bounds of region coordinates. The presented approach may be useful in many applications where camera-based tracking requires automatic initialization.

Keywords

Invariant-based matching, stereo vision, regional features, tracking, object recognition.

1. INTRODUCTION

In general terms, alignment is the determination of a transformation which can be used to relate and convert measurements between two reference frames. The relationship between camera and object reference frames is represented by a rigid or isometric transformation, composed of transformations of translation and rotation and represented by a six-parameter vector, representing position and orientation information, called pose. When camera intrinsic parameters are known, one can easily find the relative pose between camera and object provided that a set of correspondences between points from the camera image and from the object model are established. The difficulty arises when it is necessary to discover automatically the correspondences between these points.

Short paper proceedings ISBN 80-903100-9-5
WSCG'2005, January 31-February 4, 2005
Plzen, Czech Republic.

The objective of this paper is to propose an approach to visual alignment based on the design of a set of invariants for matching, a matching algorithm and complementary techniques. Matching based on invariants is recognized as a successful means for creating efficient and robust matching algorithms [Forsyth *et al.*, 1991].

Our proposal for automatic visual alignment consists in considering regional features as a source of information for the alignment and in obtaining these features from a binocular image acquisition system for which camera relative orientation and intrinsic parameters are known. The combination of stereo-vision and regional-features presents the following synergistic benefits:

(1) The matching between image- and model-features is simplified through the adoption of the rigid 3D transformation for the alignment model. It is then possible to avoid the inconveniences of the planar projective transformation in 2D, such as non-linearity, deformation and lack of simple invariants. In the adopted case, it is easier to find invariants due to the preservation of distances and angles and due to its linearity.

(2) Solution does not depend on region shape. As a result, symmetric regions are not problematic, a less strict segmentation procedure can be adopted and the contour of the convex hulls can be used instead of the original region contours.

(3) Stereo matching is simplified. We show that region-matching between 2 rectified views is very simple using the epipolar constraint in contrast to point-matching approaches. A straightforward algorithm can reconstruct features in 3D.

Possible applications for the proposed approach are robotic navigation, tangible interfaces and augmented reality. In these scenarios, it is necessary to discover the pose of known objects or scenes without user intervention, responding in interaction time and robust against occlusion and abrupt motion.

2. REVIEW OF USUAL APPROACHES

The usual approaches to automatic determination of correspondences which avoid high complexity are user intervention (the user marks the correspondences interactively), incremental solution (an image sequence with little variation between frames and known initial state are required) [Koller, 1993] [Simon and Berger, 1997] [Cornelis *et al*., 2000], predictive multiple-hypothesis solution (less sensitive to abrupt motion and clutter) [Fox *et al,*. 1999], the hybrid sensorial approach (additional non-optical sensors) [Azuma, 1999], artificial landmarks (for environments where objects can be tagged) [Bajura and Neumann, 1995], appearance approaches (use pixel neighbourhood information to compare images to a view-based model) [Se *et al.,* 2002], shape matching and search in the spaces of parameters or correspondences with the assumption of a fixed parametric form (usually approached by some form of clustering) [Haustler and Ritter, 1999] [Olson, 1994] [Wolfson and Rigoutsos, 1997].

3. ELEMENTS OF OUR APPROACH

We propose a simple 3D reconstruction technique for image-features that allows the extraction of feature attributes to form invariants. These invariants are then used to match model-features to image-features through a recognition algorithm.

Reconstruction

The 3D reconstruction of the contours based on a stereo image pair demands matching of features from the right-hand and the left-hand images and also finding homologous contour points.

As the relative orientation of the camera pair is known, a rectification step can be applied to polygon vertex coordinates in order to have the same *y*-coordinate for homologous points. With rectified coordinates, stereo matching of regions becomes a simple problem as each polygon representing a region is described by a pair of values corresponding to its maximal and minimal *y*-coordinate.

Homologous polygons must present the same pair of values except for their imprecision. Due to the additional imposed constraints, ambiguities are less frequent in the region matching case than when matching points. A hash based implementation of the stereo matching is illustrated in figure 1, the table is constructed with information from features from the left-hand image and it is indexed by the attributes from the right-hand image. In this example, feature F matches feature C as their set of attributes collides in the hash table.

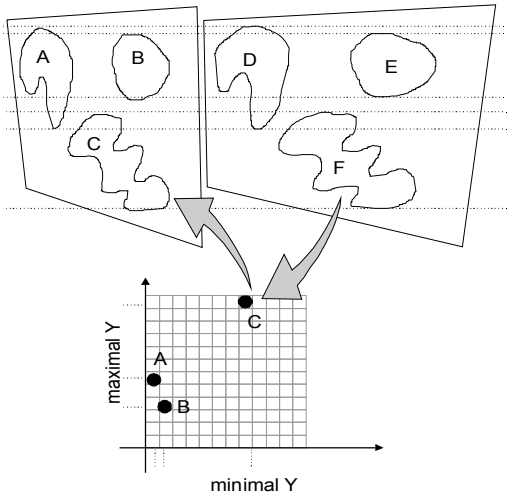

Figure 1. Stereo matching of regional features.

The intersection between a horizontal line and the contours of the convex hulls form two pairs of homologue points. The whole contour can then be reconstructed using a sequence of horizontal lines.

Base and invariants

A base is an ordered tuple of model-features associated to their corresponding image-features whose size is minimal, but enough to determine the geometric transform that maps model-features to image-features. A measure over feature tuples is considered invariant to a particular type of geometric transformation if it remains unchanged independently of the parameters of the transformation applied to the features. Aiming lower complexity, a compromise between search in correspondence- and pose-spaces is established: the search for correspondences can be restricted to tuples of image-features as large as the base.

From each 3D-reconstructed polygon, the barycentre coordinates and plane normal vector direction are extracted. A plane can be adjusted through the reconstructed vertices of a feature convex hull contour and its orientation can be estimated. The barycentre coordinates can then be computed by projecting the reconstructed points onto the adjusted plane and integrating along the contour according to Green's theorem.

It is necessary 2 features to form a base because the information provided by these attributes (barycentre and normal direction) allows a remaining degree of freedom for each feature. Defining a coordinate system with the origin in the barycentre of the first feature and the *x* axis oriented as its normal and the *y* axis oriented such that the barycentre of the second feature lie on the plane *xy,* The resulting *x* and *y* coordinates of the barycentre of the second feature and its normal direction defined in this new reference frame constitute a 4-value vector invariant to the rigid transformation.

Recognition model and algorithm

Three types of attributes are obtained from the observed geometry: (1) invariant attributes of individual features (colour, area); (2) attributes of individual features that depend on the transformation (barycentre and normal direction); (3) invariant attributes of feature-pairs (distance between barycentres, angle between planes etc.).

We develop a recognition algorithm using type (3) attributes considering the model illustrated in figure 2, where I_i and M_m are image- and model-features respectively and the Boolean C_{im} represents the correspondence between them.

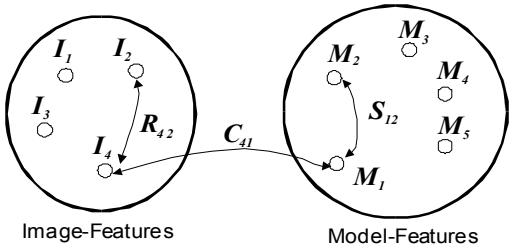

Figure 2. Recognition model

The proposed algorithm is based on abductive reasoning (diagnostics reasoning). If the measured attribute vector R_{ij} of the pair of image-features (I_i, I_j) is close enough to the attribute vector S_{mn} of model-features (M_m, M_n), we can raise the hypothesis C_{im}, i.e., I_i matches M_m. As more pairs of image-features that include I_i are analysed, more clues about I_i are collected, allowing reconsideration of the existing hypotheses. A voting scheme is then used to select the most reliable hypothesis to recognize the model-feature corresponding to I_i. An equivalent probabilistic approach is the estimation of the correspondences by maximum likelihood.

The equality $R_{ij} = S_{mn}$ is verified using a hash table whose granularity models the imprecision. The less dense are hash entries, the less strict or precise is the verification of the equality. Attributes are considered equal if they index the same cell of the hash table.

The hash table is built from the previously known geometric model. For each pair of model-features (M_m, M_n), with an attribute vector S_{mn}, the table index is computed based on the value of the attribute and a reference to the pair (*m*, *n*) is appended to the indexed cell. A description of the algorithms follows.

Building the hash table, off-line

For each model-feature *m*
  For each other model-feature *n*
    Measure the attributes *S(m,n);*
    Compute the index to the hash table given *S(m,n);*
    Append to the indexed cell in the table a reference to *(m,n).*

Recognition of image-features, on-line

For each image-feature *i*
  For each other image-feature *j*
    Measure the attributes *R(i,j);*
    Compute the index to the hash table given *R(i,j);*
    For each item in the indexed cell
      Find the identity of the pair *(m,n)* referenced in the item.
      Verify the compatibility of the remaining attributes of *(i,j)* and *(m,n);*
      If successful, cast a vote for *C(i,m)* and *C(j,n);*
For each image-feature *i*
  Find *X*, model-feature with most votes
  so that $C(i,X) \geq C(i,m), \ \forall m$;
  State that *i* matches *X*.

4. IMPLEMENTATION

The first steps of image-feature extraction are segmentation, labelling and tracing of region convex hull contours. A set of convex polygons representing image-features is returned. A simple segmentation strategy based on colour normalization and Euclidean distance in RGB-space is employed. Labelling is implemented as described in [Gonzalez and Woods, 1992]. The contours of features are traced placing a cursor in any point of the contour and examining each 2x2-pixel window to decide the direction the cursor must be moved to follow the contour. The collected contour coordinates are sorted by *y*-coordinate and the vertices not belonging to the convex hull are removed by a linear-time algorithm.

Pose is estimated by building and solving a system of linear equations for parameters from the information provided by the image-model feature matching. A least-squares technique is employed due to the availability of a more correspondences than the minimum required. The orthonormal property of the rotation matrix is fixed afterwards by factorisation.

We illustrate the proposed approach with an implemented case. A virtual model of the scene was created and a physical scene was constructed based

on this model. For a JPEG 1280x960-pixel image, the best 7 features ranked by maximal number of votes were correctly recognized and the resulting alignment is depicted in figure 3. We detected that 6 identified features were enough for a good alignment result in this experiment. Additional features contributed little to a better result because we lack a model for the dispersion of pose error and a means to evaluate the quality of individual pose measures.

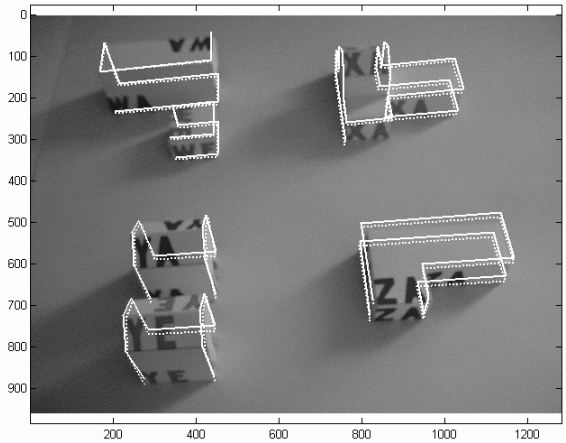

Figure 3 Resulting alignment. The continuous lines represent the computed alignment, while the dashed lines represent the nominal alignment.

5. Complexity Analysis

Let M be the number of model-features, I, the number of image-features and $O(H)$, the average complexity of a single access to the hash-table. Vote collection considers an average of $O(H)$ votes obtained from the hash-table for each of the $O(I^2)$ pairs of image-features, resulting a cost of $O(I^2H)$. Vote table analysis consists of a search in each row of the vote table for the column with greatest number of votes for which the cost is $O(IM)$. Total cost of the on-line phase is then $O(I^2H + IM)$. If the number of features is kept small, as in the expected usual case where each feature is represented by a considerable area of the image, the cost of sweeping the whole image for segmentation and labelling turns out to be the bottleneck.

6. CONCLUSION

We address in this paper the problem of automatic visual alignment from images. Our approach is comprehensive integrating several Computer Vision techniques around the assumptions of planar regional features and stereo vision. We show that under these assumptions, most of the involved components of the vision process become simple. The method is appropriate for images with regional features and designed to work in conditions of near perspective and partial occlusion. Being not incremental, it is designed to work independently of additional sensors and can be used to estimate initial state pose without user intervention. Regional features are typically found in cityscapes, traffic, office and home scenes.

7. ACKNOWLEDGMENTS

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

8. REFERENCES

Azuma, R. T., 1999. The Challenge of Making Augmented Reality Work Outdoors. In *Mixed Reality: Merging Real and Virtual Worlds*, Y. Ohta and H. Tamura, eds. Springer-Verlag, pp 379-390.

Bajura, M. and Neumann, U., 1995. Dynamic Registration Correction in Augmented Reality Systems, *IEEE VRAIS proc.*, pp 189-196.

Cornelis, M. V. K., Pollefeys, M. and Gool, L. V. Augmented reality from uncalibrated video sequences. *3D Structure from Images SMILE 2000.*

Forsyth, D.A., Mundy, J.L., Zisserman, A., Coelho, C., Heller, A.J. and Rothwell, C.A., 1991, Invariant descriptors for 3D object recognition and pose, *PAMI(13)*, no 10, pp 971-991.

Fox, D., Burgard, W., Dellaert, F. and Thrun, S., 1999. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. *Proc. of the 16th National Conference on Artificial Intelligence*.

Gonzalez, R.C. and Woods, R.E., 1992, *Digital Image Processing*, Addison-Wesley, Reading.

Haustler, G. and Ritter, D., 1999. Feature-Based Object Recognition and Localization in 3D-Space using a Single Video Image. *CVIU* 73(1) pp 64-81.

Koller, D. 1993. Moving Object Recognition and Classification based on Recursive Shape Parameter Estimation. In *Proc. of the 12th Israeli Conf. on Artificial Intelligence, Computer Vision, and Neural Networks*, pp 359-368, Tel-Aviv, Israel.

Olson, C. F., 1994. Time and Space Efficient Pose Clustering, IEEE *Conference on Computer Vision and Pattern Recognition*, pp 251-258.

Se, Lowe, Little 2002. Global Localization using Distinctive Visual Features. *Proc. of the Intl. Conference on Intelligent Robots and Systems*, Lausanne, Switzerland, pp 226--231.

Simon, G. and Berger, M. O. 1997, *A two-stage robust statistical method for temporal registration from features of various type*. INRIA TR 3235.

Wolfson, H.J. and Rigoutsos, L., 1997, Geometric hashing: an overview, *CalSE(4)*, no 4, pp 10-21.