Vision-Based 3D Object Localization Using Probabilistic Models of Appearance

Christian Plagemann[1], Thomas Müller[2], and Wolfram Burgard[1]

[1] Department of Computer Science, University of Freiburg,
Georges-Koehler-Allee 79, 79110 Freiburg, Germany
{plagem, burgard}@informatik.uni-freiburg.de
[2] Fraunhofer Institute IITB, Fraunhoferstraße 1, 76131 Karlsruhe, Germany
mlt@iitb.fraunhofer.de

Abstract. The ability to accurately localize objects in an observed scene is regarded as an important precondition for many practical applications including automatic manufacturing, quality assurance, or human-robot interaction. A popular method to recognize three-dimensional objects in two-dimensional images is to apply so-called view-based approaches. In this paper, we present an approach that uses a probabilistic view-based object recognition technique for 3D localization of rigid objects. Our system generates a set of views for each object to learn an object model which is applied to identify the 6D pose of the object in the scene. In practical experiments carried out with real image data as well as rendered images, we demonstrate that our approach is robust against changing lighting conditions and high amounts of clutter.

1 Introduction

In this paper, we consider the problem of estimating the three-dimensional position and the orientation of rigid objects contained in images. This problem has been studied intensively in the computer vision community and its solution is regarded as a major precondition for many practical applications, like automatic manufacturing, quality assurance, or human-robot interaction. In this work, we are especially interested in view-based approaches, where objects are represented by 2-dimensional views. Such approaches allow to incorporate visual features directly and do not assume prior knowledge about the spatial structure of the objects. The limited localization accuracy caused by the view-based representation can be compensated for by a scene-based object tracking process, as will be demonstrated in Section 5.2.

Recently, Pope and Lowe [1] proposed the probabilistic alignment algorithm to identify two-dimensional views of objects in images. The goal of the work presented here is to investigate how this purely image-based approach can be utilized to achieve a robust estimate of the position and orientation of the object in the scene. The input to our system are either real images of an object or alternatively a volumetric model that is used to render the necessary views. We describe how the four parameters obtained from the 2D object recognition can

W. Kropatsch, R. Sablatnig, and A. Hanbury (Eds.): DAGM 2005, LNCS 3663, pp. 184–191, 2005.

be combined with the two parameters of the corresponding view to determine the pose of the object in the scene. We evaluate our approach on real images and perform simulation experiments to provide quantitative results for the localization accuracy. Experimental results carried out with free-form objects and objects with specular surfaces demonstrate the robustness of our approach.

2 Related Work

The problem of recognizing objects using two-dimensional views has been approached from many directions. Popular methods are based on eigenvector decomposition to represent and recognize 3D objects [2,3]. Alternative approaches learn networks of Gaussian basis functions [4] or build their models on wavelet-based features [5]. Several authors [6,7] apply support vector machines to object recognition. An additional approach is to train a neural network spanning the whole view sphere to classify single object views into orientation categories [8]. These methods, however, assume a segmented test image where the object instance is roughly isolated by a bounding box. The approach presented in this paper focuses on the case in which the object instance covers just a small part of the test image. To avoid the feature correspondence problem, Schiele and Pentland build position independent feature histograms [9]. To localize objects they apply a voting scheme similar to the Hough-transform. Several authors also combine view-based and model-based approaches to recognize and localize textured objects [10,11]. Finally, Lanser et al. [12] present a view-based localization system called MORAL. In their approach, the 3D object structure has to be known and is assumed to be polyhedral. The constructed object views are not clustered and generalized to achieve a more compact model.

3 View-Based Probabilistic Alignment

Pope and Lowe [1] introduced a visual object recognition approach based on probabilistic models of appearance. The appearance of an object is modeled by a relatively small set of 2-dimensional model views, each of which is assembled from discrete visual features, like edges, corners, joints, and complex combinations thereof. The features are associated with their uncertainty in presence and position as well as their distribution of attribute values. Hence, a single model view represents the appearance of an object from a whole range of view points. The scope of each model view is determined by an unsupervised learning process.

The recognition method, called probabilistic alignment, resembles Huttenlocher and Ullman's alignment approach [13] by gradually building feature pairings between model view and test image to align the view to the test image. The recognition process is guided by a probabilistic quality measure for possible alignment hypotheses.

$$g(E,T) \approx \log \mathrm{P}(H \mid E,T) \tag{1}$$

In this equation, H denotes the hypothesis that the model view is contained in the image, E stands for the set of feature pairings contained in the hypothesized

match, and T is the similarity transformation that aligns the model view with the test image. The typically large set of possible feature pairings is ordered using this measure to process likely hypotheses first.

The learning process for an unknown object starts with an empty set of model views and gradually incorporates all training views. If a new training view cannot be matched with a sufficient accuracy to an existing model view, a new model view is created. Otherwise, the training view and the matching model view are generalized to a combined model view. Therefore, the resulting view-based model is a set of generalized clusters of training views. The minimum description length principle is used to obtain the smallest model that sufficiently describes the visual appearance of the object.

4 3D Object Localization

In real world applications, it is generally not sufficient to identify an object within an image. Rather, one is often interested in its exact pose. The goal of this section is to embed the 2D recognition approach described above into a 3D object localization system, which covers object learning, view-based recognition and the calculation of the 3D pose.

4.1 Object Learning

A popular technique for the acquisition of training views for an object is to record images of the real object from different view points. This procedure requires an elaborate hardware setup to accurately measure the viewing angles and is quite time consuming. An alternative approach is to construct a 3D object model (e.g., by using photometric 3D scanning methods) and generate artificial training views. Using state of the art rendering techniques, like ray-tracing, a large training set of photo-realistic views covering different image resolutions and lighting conditions can be constructed. It may be favorable for specific applications to extract just the object silhouette, which can easily be achieved using rendering techniques. We generate photo-realistic views to keep the repertoire of visual features as wide as possible. Other systems that construct artificial views by just projecting 3D model features like lines and corners into the image plane limit themselves in that respect. As the experiments in Section 5 demonstrate, our recognition system achieves good results with real training images as well as generated ones.

The learning algorithm discussed in Section 3 clusters the training views to reflect the variance in visual appearance across the view sphere. It is therefore desirable to uniformly sample the view sphere. We achieve this by iteratively dividing a spherical triangular mesh. The individual training views are generated using ray-tracing. The optimal number of training views that have to be generated for a specific object can be determined as follows. As discussed in Section 3, the learning step builds a set of 2-dimensional model views by iteratively integrating training images. Assuming that the appearance of the object changes smoothly with small variations of the viewing angles and furthermore assuming

Fig. 1. One of the generated training images (left) and the recognition result (right) for a free-form object. It was sufficient for this object to learn and represent the silhouette.

that the views are uniformly distributed and presented in a coarse to fine order, the complete appearance of the object is covered after a certain number of training views. This number depends on the object and the learning parameters and is independent from the specific training views. In our current system, we define as the stopping criterion for generating new training views the point when the number of model views stops to increase. For example, 130 training views are sufficient to learn the 18 model views of the highly symmetric barbell object depicted in Figure 3.

4.2 Localization

For every 2-dimensional model view m, the alignment algorithm yields a 4D pose vector r (2D position (x, y), orientation α, and scale s) and a match quality measure. Assuming a calibrated camera, we can calculate the 6D pose vector of the object in the scene (3D position, 3D orientation) by utilizing information from the training step. In this section, we describe how a transformation T can be derived that maps object coordinates to scene coordinates. The localization result can be obtained directly from the transformation matrix. Let us decompose T^{-1} into four individual transformations $T^{-1} = T_4\ T_3\ T_2\ T_1$. In this equation, T_1 translates the center of the object to the origin of the scene coordinate system. It is therefore defined by the position vector $\mathbf{p}$ of the object. The distance of the object from the origin (the length of $\mathbf{p}$) can be derived from the scale factor s and the known object size from the training process. The direction of $\mathbf{p}$ is defined by the position (x, y) of the recognized view in the image plane and the camera calibration parameters. T_1 can be written as

$$T_1 = \begin{pmatrix} 1 & 0 & 0 & -p_x \\ 0 & 1 & 0 & -p_y \\ 0 & 0 & 1 & -p_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} . \tag{2}$$

The transformation T_2 rotates the vector $\mathbf{p}$ onto the z-axis. Vividly speaking, T_2 transforms the object into the same pose that its training counterpart was in

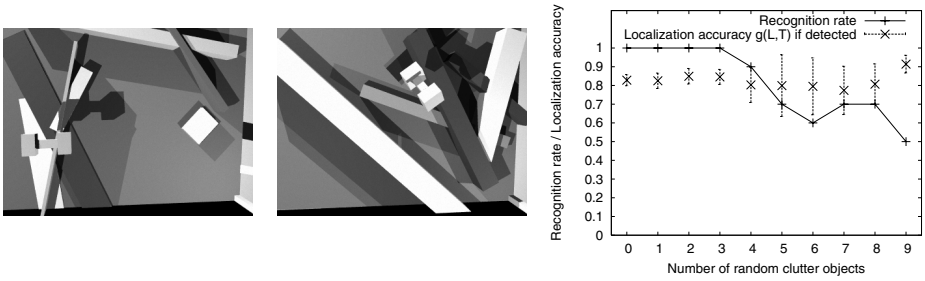

Fig. 2. Typical synthetic test data for quantitative performance analysis (left two images) and the localization results depending on the number of clutter objects (right)

when the training images for model view m were acquired. In the following, the vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ denote the three unit vectors respectively. If we define $T_{\mathbf{a},\mathbf{b}}$ as the transformation that rotates the vector $\mathbf{a}$ onto a vector $\mathbf{b}$ around the axis perpendicular to $\mathbf{a}$ and $\mathbf{b}$ we have $T_2 = T_{\mathbf{p},\hat{\mathbf{z}}}$.

Furthermore, T_3 rotates the object within the image plane to account for the angle α that was part of the 2D recognition result. If we define $T_{\mathbf{a},\alpha}$ as the transformation that rotates around the axis $\mathbf{a}$ with angle α we obtain $T_3 = T_{\hat{\mathbf{z}},\alpha}$. Finally, T_4 rotates the object according to the viewing angles ϑ (azimuth angle) and φ (polar angle) associated with the recognized view during training, thus we obtain

$$T_4 = T_{4,2}\, T_{4,1} \qquad T_{4,1} = T_{\hat{\mathbf{x}},(180^\circ - \varphi)} \qquad T_{4,2} = T_{\acute{\mathbf{z}},(\vartheta - 90^\circ)} \qquad \acute{\mathbf{z}} = T_{4,1}\, \hat{\mathbf{z}}.$$

The 3D position vector of the object can be read directly from the last column of the transformation matrix of T. We refer to Shoemake [14] for details about the derivation of the 3D orientation vector in Euler's angular notation from the transformation matrix.

5 Experimental Results

5.1 Quantitative Evaluation

We conducted simulation experiments to evaluate the localization accuracy using the known ground truth. Figure 2 depicts typical examples of a series containing 100 test images. The task was to localize the barbell object among a varying number of random clutter objects using features like edges, corners, joints and complex groupings thereof. The used rectangular clutter structures made recognition harder than seemingly more realistic curved structures, because of their similarity to the barbell object. The test images were rendered to a size of 384×288 pixels using the freely available ray-tracer Povray. To compare the true object pose $\mathbf{T}$ with the localization result $\mathbf{L}$ we use the measure

$$g(\mathbf{L}, \mathbf{T}) := \max\left\{0, 1 - (f_z\, \delta_z^2 + f_{xy}\, \delta_{xy}^2 + f_\alpha\, \delta_\alpha^2)\right\} \ . \tag{3}$$

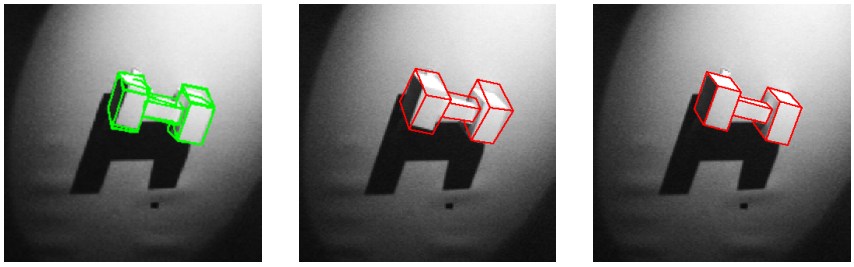

Fig. 3. View-based recognition result (left), the derived 3D pose (middle), and the 3D pose after refinement by a scene-based tracking process (right)

Fig. 4. Real training image (left) and recognition result (middle) for a highly specular object. Detected barbell (right) under extensive occlusion by fog.

In Equation 3, δ_z denotes the distance between the two poses along the z-axis (the viewing direction of the camera), δ_{xy} is the distance perpendicular to the z-axis, and δ_α is the difference in orientation. The weighting factors for the different dimensions have been set to $f_z = \frac{1}{250000}$, $f_{xy} = \frac{1}{50000}$, and $f_\alpha = \frac{1}{10000}$ to reflect the importance of the different dimensions given their scales. The quantities δ_z and δ_{xy} are measured in mm, δ_α in degrees. A typical value of $g = 0.80$ is reached for example with distances of $\delta_{xy} = 2.3$ cm, $\delta_z = 23$ cm, and $\delta_\alpha = 15°$. This displacement is small enough to initialize a scene-based tracking algorithm like the one described in the next section. The recognition rate for an increasing number of clutter objects as well as the achieved localization accuracy is plotted in Figure 2. The mean localization accuracy for recognized objects was $g = 0.82$ with a median of $g = 0.84$.

5.2 Recognizing Specular and Free-Form Objects and Refining the Estimated Object Pose

Specular and free-form objects are particularly hard to recognize visually. The appearance of a specular object varies greatly with changing lighting conditions or object movement. Therefore, feature-based recognition has to get by with only few robust feature pairings among many spurious ones. The rating of fea-

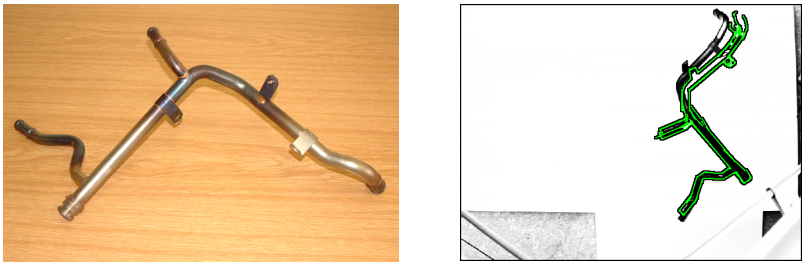

Fig. 5. Typical image containing the learned top view of a compound pipe object (left) and recognized object instance in a noticeably different pose (right)

ture pairings using statistics acquired during the training phase is especially important in such cases to reduce the amount of tested alignment hypotheses. Figure 4 shows a training image and recognition result for the highly specular coffee machine object. For this experiment, we used real training images rather than rendered ones to demonstrate that this is a viable option.

Objects of predominantly curved structure are harder to represent and recognize than purely polyhedral ones. To accurately represent views of non-polyhedral objects using curve-based features, higher order curves have to be fitted to the images. This not only increases the complexity in the feature extraction step, it also makes the feature-based representation less predictable. This points out the importance of the rating function for feature pairings to process useful pairings first.

Figure 1 shows a typical result of our experiments with a toy dinosaur. We used the optic 3D-Scanner DigiScan 2000 to obtain the 3D structure from the real object to generate the training set by simulation. A further experiment has been carried out with an object consisting of compound pipes (see Figure 5). Here, our system learned the view-based object model using only top-view images. Note that the test images contained object instances largely displaced from the image center which lead to perspectively distorted views that were not included in the training data. As shown on the right image of Figure 5, the object can still be localized.

We also applied our system to initialize a scene-based object tracking process [15]. In our experiments we found that the achieved localization accuracy was sufficient to robustly obtain a refined object pose after a few tracker iterations. Figure 3 shows the relationship between the 2D recognition result (the left image), the derived 3D pose (middle), and the refined pose after a few tracker iterations (on the right). Localization and pose refinement was also possible under extensive partial occlusion by fog as shown in the right image of Figure 4.

6 Conclusions

In this paper, we presented an approach to 3D object localization in single images using a probabilistic view-based alignment technique. Our system learns a view-based object model from a set of training views. During application, our

system combines for the recognized view the four parameters of the 2D similarity transform with the orientation of the object in the corresponding training image to extract the 3D position and orientation of the object in the scene. The system has been implemented and validated on real images and images rendered from 3D-models. Experiments with free-form objects and objects with specular surfaces in cluttered scenes demonstrate the robustness of our approach. We furthermore presented an application of our approach for the initialization of a 3D scene-based tracking system.

Acknowledgments

We would like to thank Chen-Ko Sung, Fraunhofer Institute IITB, Karlsruhe, and the Benteler Group for providing valuable support.

References

1. Pope, A.R., Lowe, D.G.: Probabilistic models of appearance for 3-d object recognition. Int. Journ. of Computer Vision **40** (2000) 149–167
2. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. Int. J. Comput. Vision **14** (1995) 5–24
3. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: CVPR. (1994)
4. Pauli, J.: Learning to recognize and grasp objects. Machine Learning **31** (1998)
5. Reinhold, M., Paulus, D., Niemann, H.: Improved appearance-based 3-D object recognition using wavelet features. In: Vision, Modeling, and Visualization. (2001)
6. Schölkopf, B.: Support Vector Learning. PhD thesis, Universität Berlin (1997)
7. Blanz, V., Schölkopf, B., Bülthoff, H.H., Burges, C., Vapnik, V., Vetter, T.: Comparison of view-based object recognition algorithms using realistic 3d models. In: ICANN. (1996) 251–256
8. Wunsch, P., Winkler, S., Hirzinger, G.: Real-time pose estimation of 3-d objects from camera images using neural networks. In: ICRA. (1997) 3232–3237
9. Schiele, B., Pentland, A.: Probabilistic object recognition and localization. In: ICCV. (1999) 177–182
10. Allezard, N., Dhome, M., Jurie, F.: Recognition of 3d textured objects by mixing view-based and model-based representations. In: ICPR. (2000) 1960–1963
11. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In: CVPR (2). (2003) 272–280
12. Lanser, S., Zierl, C., Munkelt, O., Radig, B.: Moral - a vision-based object recognition system for autonomous mobile systems. In: CAIP. (1997) 33–41
13. Huttenlocher, D.P., Ullman, S.: Recognizing solid objects by alignment with an image. Int. J. Comput. Vision **5** (1990) 195–212
14. Shoemake, K.: Euler angle conversion. In: Graphics gems IV. Academic Press Professional, Inc. (1994) 222–229
15. Plagemann, C.: Ansichtsbasierte Erkennung und Lokalisierung von Objekten zur Initialisierung eines Verfolgungsprozesses. Master's thesis, University of Karlsruhe (2004) in German.