FAST TRAINING FOR OBJECT RECOGNITION WITH STRUCTURE-FROM-MOTION

M. Grzegorzek, I. Scholz, M. Reinhold, H. Niemann

Chair for Pattern Recognition, University of Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany, {grzegorz,scholz,reinhold,niemann}@informatik.uni-erlangen.de

In this paper we present a system for statistical object classification and localization, which applies a simplified image acquisition process for the learning phase. Instead of using complex setups to take training images in known poses, which is very time-consuming and not possible for some objects, we use a hand-held camera. The pose parameters of objects in all training frames which are necessary for creating the object models are determined using a structure-from-motion algorithm. The local feature vectors we use are derived from wavelet multiresolution analysis. We model the object region as a function of 3-D transformations and introduce a background model. Experiments made on a real data set taken with a hand-held camera with more than 2500 images show that it is possible to obtain good classification and localization rates using this fast image acquisition method.

Introduction

For many tasks the recognition of objects in images is very useful, sometimes even necessary. Possible applications in this area are for example: face recognition [3], localization of obstacles on the road with a camera mounted on a driving car, service robotics [12], and so on. The learning process in most object recognition systems begins with the image acquisition of all possible object classes in many known poses. In the laboratory environment, the images can be taken with a special setup like a turntable with a camera arm (Fig. 1, left).

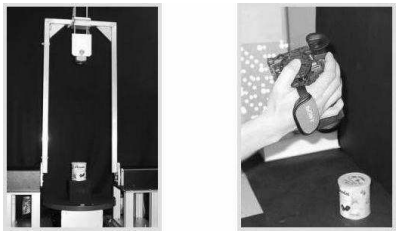

Fig. 1. Left: turntable with camera arm. Right: hand-held camera.

In real problems of object recognition in images, it is much easier to record the objects using a hand-held camera (Fig. 1, right). For this reason we propose a new approach for object recognition, where the image acquisition is done in this way. The goal of our algorithm is to optimize the training process with respect to execution time and ease of image acquisition while still getting satisfying classification and localization rates. The poses of the objects in all training frames are computed using a structure-from-motion algorithm [5]. The whole learning process is therefore independent of environment assumptions, but we have to deal with an additional training inaccuracy.

Two main approaches exist to solve the problem of object recognition in images: the model- and the appearance-based methods. The model-based systems use a segmentation step to extract features of objects [6]. The appearance-based approaches compute the feature vectors directly from pixel intensities in the images [3, 10]. There are appearance-based systems that use one global feature vector for the whole image (e.g. eigenspace approach [2]), and those that use more local feature vectors (e.g. neural networks [8]). In the present work, local feature vectors with two components are applied, which are computed with a wavelet multiresolution analysis [7] and statistically modeled by density functions.

In the next section we introduce the pose parameter reconstruction using a structure-from-motion algorithm, which yields the training pose parameters needed for object modeling. Then the system for statistical

object recognition is presented. After that we describe experiments, and discuss the results. We close our contribution with a conclusion.

Pose Parameter Reconstruction

Suppose an image sequence is given which was taken by moving a hand-held camera around an object and showing it from different directions (Fig. 1, right). In order to train the object recognition system it is necessary to estimate internal and external object pose parameters for all frames. The internal pose parameters denote two translations and a rotation inside the image plane. The external pose parameters are two rotations outside the image plane and a translation along the optical axis of the camera. Only four of these six pose parameters, internal translations $\boldsymbol{u} = (u_1, u_2)^{\mathrm{T}}$ and external rotations $\boldsymbol{\Phi} = (\theta, \varphi)^{\mathrm{T}}$, are used in our experiments, therefore only the computation of these parameters will be explained in the following.

The first step is to compute a 3-D reconstruction of the camera motion and scene structure using a structure-from-motion algorithm [5]. This requires the knowledge of point correspondences in the images, which are retrieved by feature detection and tracking as explained in [11]. By applying a factorization method, in this case the paraperspective factorization introduced by Poelman and Kanade [9], the camera motion parameters and 3-D point positions corresponding to the tracked 2-D features are reconstructed for a relatively short initial sub-sequence. The results are refined by a non-linear optimization as proposed in [4]. The remaining camera and point positions are estimated by a similar optimization image by image, a method which is explained in detail in [5].

At this point the cameras are given as projection matrices $\boldsymbol{P}_i = \boldsymbol{K}(\boldsymbol{R}_i^T \,|\, -\boldsymbol{R}_i^T \boldsymbol{t}_i)$, where $\boldsymbol{K}$ contains the camera intrinsic parameters, and $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ denote the rotation and translation of the camera. The object recognition system on the other hand requires an entirely different parameter representation. Therefore, the parameters are transformed as follows. First, the origin of the coordinate system is translated into the center of mass of the object $\bar{\boldsymbol{p}}$. Since the object was placed on a black background, the feature tracking algorithm is only able to track features on the object itself. Thus, the centroid of the reconstructed 3-D points is used as an approximation to the center of mass of the object. The calculated translation is applied to all camera and 3-D point positions.

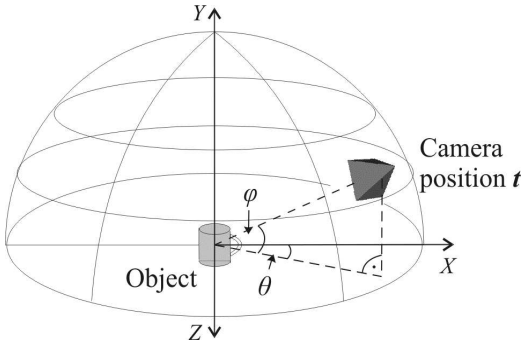

Fig. 2. Calculating θ_i and φ_i using the camera pose. The camera is depicted as a pyramid with its tip being the optical center and its base being the image plane.

The external rotations in polar coordinates for the training image $\boldsymbol{f}_i$ can now be calculated easily, as depicted in Figure 2. For a given translation $\boldsymbol{t}_i = (t_{i,x}, t_{i,y}, t_{i,z})^{\mathrm{T}}$ of the camera in world coordinates, the angle θ_i computes as

$$\theta_i = \arcsin(t_{i,x} / \sqrt{t_{i,x}^2 + t_{i,z}^2}) \qquad (1)$$

and the angle φ_i as

$$\varphi_i = -\arcsin(t_{i,y} / \sqrt{t_{i,x}^2 + t_{i,y}^2 + t_{i,z}^2}) \qquad (2)$$

The internal translation is estimated by back-projecting the center of mass of the object into image coordinates, i.e. $\boldsymbol{u}_i' = \boldsymbol{P}_i \bar{\boldsymbol{p}}'$ where $\boldsymbol{u}_i'$ and $\bar{\boldsymbol{p}}'$ denote $\boldsymbol{u}_i$ and $\bar{\boldsymbol{p}}$ in homogeneous coordinates.

Statistical Object Recognition

At the beginning of the statistical modeling we select one of the training images for each object class as a reference image. With the pose of an object in the image $\boldsymbol{f}_i$ we denote the 3-D transformation (translation and rotation) that maps the object in the reference image to the object in $\boldsymbol{f}_i$. In all of these images, feature vectors are computed using a wavelet transformation [1]. A grid with size

$\Delta r = 2^{-s}$, where s is the scale of the wavelet transformation, is laid over each training image. At each grid point a feature vector with two components is calculated:

$$\boldsymbol{c}_m = \begin{pmatrix} \ln\left(2^s \left| b_{s,m} \right|\right) \\ \ln\left(2^s \left(\left| d_{0,s,m} \right| + \left| d_{1,s,m} \right| + \left| d_{2,s,m} \right| \right)\right) \end{pmatrix} \quad (3)$$

$b_{s,m}$ is a low-pass coefficient and $d_{0\ldots2,s,m}$ are high-pass coefficients. For each feature vector $\boldsymbol{c}_m$ we define a function that assigns it to the object or to the background:

$$\xi_m(\boldsymbol{\Phi},\boldsymbol{u}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{c}_m \in O \\ 0 & \text{if} \quad \boldsymbol{c}_m \notin O \end{cases} \quad (4)$$

This function is interpolated using $(\boldsymbol{\Phi}_i, \boldsymbol{u}_i)$ and defined on a continuous domain $(\boldsymbol{\Phi},\boldsymbol{u})$. O denotes the object area.

The computed feature vectors are interpreted as random variables. Their components are modeled as normally distributed. The density function for the object class κ is:

$$p(C \mid \boldsymbol{B}_\kappa, \boldsymbol{\Phi},\boldsymbol{u}) = \prod_{\{m \mid \xi_{m,\kappa}=1\}} p(\boldsymbol{c}_m \mid \boldsymbol{\mu}_{m\kappa}, \boldsymbol{\sigma}_{m\kappa}, \boldsymbol{\Phi},\boldsymbol{u}) \quad (5)$$

where $\boldsymbol{B}_\kappa$ comprehends the trained mean vectors $\boldsymbol{\mu}_{m\kappa}$ and standard deviation vectors $\boldsymbol{\sigma}_{m\kappa}$ of the feature vectors $\boldsymbol{c}_{m\kappa}$. C is the set of feature vectors that belong to the object.

After an object model was created for each object class, the system is able to classify and localize objects. The recognition algorithm is described by the following equation:

$$\left(\hat{\kappa}, \hat{\boldsymbol{\Phi}}_\kappa, \hat{\boldsymbol{u}}_\kappa\right) = \underset{\kappa}{\operatorname{argmax}} \left\{ \underset{(\boldsymbol{\Phi},\boldsymbol{u})}{\operatorname{argmax}}\, p\left(C_{O_\kappa} \mid \boldsymbol{B}_\kappa, \boldsymbol{\Phi},\boldsymbol{u}\right) \right\} \quad (6)$$

For each pose hypothesis $(\boldsymbol{\Phi},\boldsymbol{u})$ we determine the set of feature vectors C_{O_κ} that belong to the object area. The parameters $(\kappa,\boldsymbol{\Phi},\boldsymbol{u})$ with the highest probability p are taken as the recognition results. More details on the whole recognition system are given in [10].

Experiments and Results

We tested our approach on a data set that consists of 8 objects which are illustrated in Figure 3.

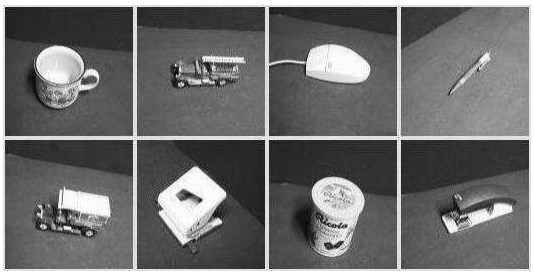

Fig. 3. Used object classes. In the first row from left: cup, toy fire engine, mouse, pen. In the second row from left: toy passenger car, hole puncher, candy box, stapler.

In the training phase sequences with more than 200 frames of each object class were taken with a hand-held camera (Fig. 1, right), which accelerates the image acquisition process compared to the common methods. The recording of 200 training images of objects located on a turntable (Fig. 1, left) takes about 20 minutes. Using the hand-held camera we get a video with 200 frames in about 5 seconds. Next, we preprocessed the original images by converting the 512×512 color images to gray level images sized 128×128 pixels, and created the object models. The preprocessing of 100 training frames and creation of one object model takes 27s on a Pentium 4, 2.66 GHz.

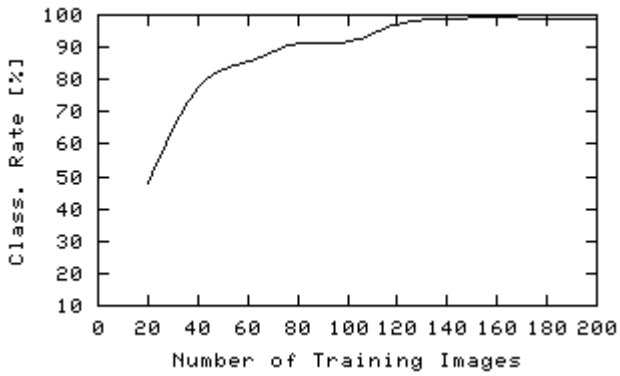

Fig. 4. Classification rate depending on the number of training images.

For the recognition phase we took 8 image sequences with about 120 frames on homogeneous background. The recognition time in 100 test images amounts to 72s for the 128×128 pixel images. The classification rates as a function of the number of training images are presented in Figure 4. A very good classification result (98.8%) with a relatively short execution time (training of one object class: 38s, recognition in 100 test images: 72s on a Pentium 4, 2.66 GHz) was obtained using 140 training images.

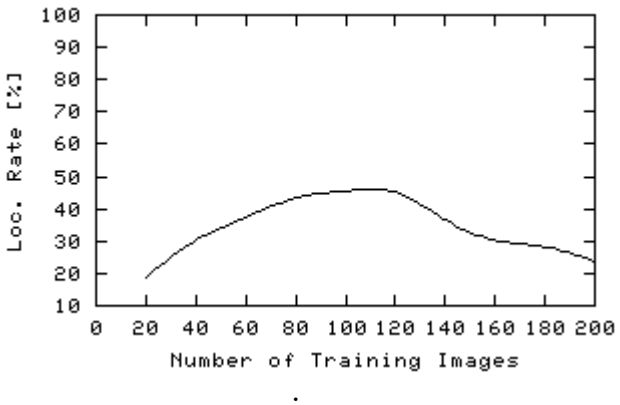

Fig. 5. Localization rate depending on the number of training images. Evalutation criteria: 10 pixels (translations), 15° (rotations).

The best localization rate amounts to 45.5% and was obtained for 100 training frames.

Conclusion

In this paper we presented an approach for the statistical object classification and localization of 3-D objects where the image data acquisition was made using a hand-held camera. This innovation accelerated, simplified, and universalized the learning process compared to most other object recognition systems. The pose parameters of the training frames, which are needed for creating the object models, were calculated using a structure-from-motion algorithm. For robustness of the system we applied a statistical framework which includes both object and background models.

In the experiments we showed that it is possible to get excellent classification and good localization rates in relatively short execution time.

In the future we will work on the algorithm for pose parameter reconstruction and the system for statistical object recognition in order to improve the localization rates.

References

1. C. Chui. *An Introduction to Wavelets.* Academic Press, San Diego, 1992.
2. C. Gräßl, F. Deinzer, and H. Niemann. Continuous parametrization of normal distribution for improving the discrete statistical eigenspace approach for object recognition. In *Pattern Recognition and Information Processing 03,* pages 73-77, Minsk, Mai 2003.
3. R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light fields. In *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26(4):449-465, April 2004.
4. R. Hartley. Euclidean reconstruction from uncalibrated views. In *Applications of Invariance in Computer Vision,* volume 825 of *Lecture Notes in Computer Science,* pages 237-256. Springer-Verlag, 1994.
5. B. Heigl. *Plenoptic Scene Modeling from Uncalibrated Image Sequences.* ibidem-Verlag, Stuttgart, 2004.
6. J. Kerr and P. Compton. Toward generic model-based object recognition by knowledge acquisition and machine learning. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,* pages 9-15, Acapulco, August 2003.
7. S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11(7):674-693, July 1989.
8. S. Park, J. Lee, and S. Kim. Content-based image classification using a neural network. *Pattern Recognition Letters,* 25(3):287-300, February 2004.
9. C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 19(3):206-218, March 1997.
10. M. Reinhold. *Robuste, probabilistische, erscheinungsbasierte Objekterkennung.* Logos Verlag, Berlin, 2004.
11. T. Zinßer, C. Gräßl, and H. Niemann. Efficient feature tracking for long video sequences. In *Pattern Recognition, Proceedings of 26th DAGM Symposium,* Springer-Verlag, August 2004. To appear.
12. M. Zobel, J. Denzler, B. Heigl, E. Nöth, D. Paulus, J. Schmidt, and G. Stemmer. MOBSY: Integration of vision and dialogue in service robots. *Machine Vision and Applications,* 14(1):26-34, 2003.