ON THE APPLICATION OF LIGHT FIELD RECONSTRUCTION FOR STATISTICAL OBJECT RECOGNITION

B. Heigl, J. Denzler, H. Niemann
Universität Erlangen–Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstr. 3, D–91058 Erlangen, Germany
email: heigl@informatik.uni-erlangen.de
www: http://www5.informatik.uni-erlangen.de

ABSTRACT

In this paper we apply light field reconstruction and rendering of object views to the problem of automatic generation of training material for a statistical object recognition system. The advantages of using a light field instead of real images are shown. We evaluate with respect to the error rate of the classifier, whether the reconstructed light field can be applied to the training step. We also show how the recognition rate of the classifier trained by the light field depends on its resolution.

1 Motivation

Statistical methods become more and more important in the field of pattern recognition. First impressive results have been reported in the mid 80's in speech recognition, where Hidden Markov Models have been applied. Today, almost all systems for speech recognition are based on this kind of statistical framework. It turns out that for a suited configuration of a statistical classifier an enormous amount of representative training data is necessary [8].

In the past years in computer vision statistical methods have been of increasing interest, too. Examples are 2–D and 3–D object recognition and pose estimation using Bayesian classifier [4], or motion segmentation with Markov–Random fields [3]. Where in speech recognition large data sets can be recorded in a relatively short time, in computer vision it is a very time consuming task to record a sufficiently large representative training set. One of the reasons is the number of free parameters. In the case of one 3–D object, the pose and the illumination must be varied. Assuming Lambertian reflection six parameters for the pose and three parameters for the light source need to be varied.

Thus, in the case of 3–D object recognition a large amount of different views of the object must be recorded. In practice, a certain amount of images of the object is recorded manually, semi–automatically, or automatically using a robot arm with a mounted on camera. Of course, manual recording is not feasible for such a task, but even in the case of an automatic strategy the movement of the robot takes a lot of time. Additionally, the variation of the light source direction increases the effort. Finally, for real applications a non trivial number of different objects needs to be distinguished.

What might be a solution to these problems? For recognition algorithms which are based on segmented primitives in images (lines, corners) CAD models for synthetic view generation could be a solution. Sophisticated lighting simulations and texture mapping increase the quality of the images, but up to now, such a strategy does not result in data which is close to real images taken with a camera. In the case of appearance based strategies [9], this way fails, because an adequate modelling of the surface reflectance of complex objects is actually impossible.

In this paper we investigate a new concept in computer graphics regarding the suitability for statistical pattern recognition. The *light field* [6] — also cited as *lumigraph* [2] — is a new image–based representation of arbitrary complex objects. The light field allows a photo–realistic image rendering, without explicitly modelling the geometry of the object. In Sect. 2 we shortly summarize the concept. We automatically build light fields by a robot arm with a mounted on camera moving around the object (Sect. 2). With the light field, new photo–realistic views are rendered and they are then used as training material to configure a statistical object recognizer. The statistical approach [9] is described in Sect. 3. Sect. 4 gives a discussion of theoretical suitability of the light field for the training of statistical classifiers. In Sect. 5 the configured classifier is applied to recognize and localize real objects in 3–D. The results are compared to those of a classifier which is trained with real images taken with a camera. The results will show, whether the rendered views reflect the reality accurate enough to be applied to statistical training, and if an increasing number of rendered views also increases the recognition rate. Sect. 6 summarizes the results and gives ideas for future work.

2 Recording the Light Field

In correspondence to [6], one way to represent a light field, is to define each viewing ray by two points on two

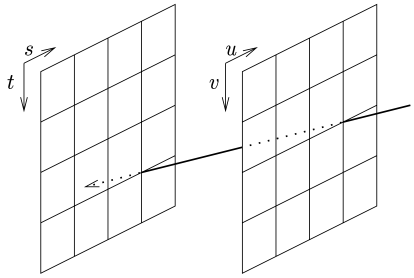

Fig. 1: Representation of a viewing ray in the lumigraph.

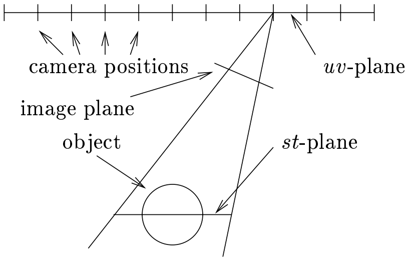

Fig. 2: Side view of the recording situation.

specified planes, as illustrated in Fig. 1. Fig. 2 shows a constellation, used to get the necessary conditions to create a light field in this representation. The images are recorded automatically by a moving camera, mounted on a robot arm [1]. The optical center of the camera is moved to discrete positions of the uv–plane, whereas the optical axis always intersects the object center. This requires a good hand–eye calibration. The recorded images are resampled, so that each image plane is transformed to the corresponding st–plane. Therefore, each transformed image represents a bundle of viewing rays, determined by one single uv–position and all positions of the st–grid. Now the light field consist of a collection of viewing rays with the corresponding color values.

To render new unknown views with the light field, the coordinates u,v,s and t of the required light rays have to be determined. The resulting color values are calculated by bipolar interpolation between the values of the neighboring grid points. It is possible to render images from all arbitrary positions, if they are within the viewing angle. Combining more of those pairs of planes, the viewing angle can be varied in a more wide range.

3 Statistical Object Recognition

When analyzing a 2–D image with multiple 3–D objects two major problems have to be solved: the pose estimation and classification of each object in the scene. Most publications in the field of statistical object modeling use geometric information of segmentation results as random variables. Lines or vertices, for example, can be used for the construction of statistical models. There are two major disadvantages of using segmentation results. When restricting the recognition process to this level of abstraction a lot of information contained in the image is lost. Another disadvantage are the errors made by segmentation.

These are the reasons for using the gray–level information of an image [9]. There exists a lot of work based on segmentation free object recognition, for example, correlation, appearance based modeling [7], maximization of mutual information [10], or methods based on mixture densities of the gray values of object images.

All approaches in statistical pattern recognition need a huge amount of training data to configure the classifier which depends on the number of parameters of the statistical model as well as on the number of objects, which shall be distinguished.

For the experiments in this paper we apply a statistical 3–D object localization and classification approach described in [9] extended for using color images. In Fig. 3 an overview of the system is given. The parameter space is six–dimensional for this task. The 3–D transformation consists of the rotation $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ determined by three parameters and the translation $\boldsymbol{t} \in \mathbb{R}^3$

In a first step of the localization process a multiresolution analysis of the image is done to derive feature values on different scales and resolutions (sampling rates) at the locations of rectangular sampling grids.

We define a statistical measure for the probability of those features under the assumption of an object transformation. The complexity of the pose estimation is high if all features on the different scale levels are combined into one measure function. Therefore, a hierarchical solution is used. Measures are defined for each scale. The analysis starts on a low scale and a rough resolution. The resolution is then increased step by step. The transformation estimation becomes more exact with each step. Let $\tilde{\boldsymbol{c}}_s$ be the vector of the concatenated feature vectors detected in an image on scale s, $\boldsymbol{B}_s$ the model parameters of an object class and $\boldsymbol{R}, \boldsymbol{t}$ be parameters for rotation and translation. The model parameters $\boldsymbol{B}_s$ consist of geometric information like probability density locations and other density parameters. The density $p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$ is then used for localization. The maximum likelihood estimation results in $(\widehat{\boldsymbol{R}}_s, \widehat{\boldsymbol{t}}_s) = \mathrm{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$.

As here is not enough room for a detailed description, we refer to [9], including feature extraction, model formulation and the parameter estimation.

4 Light Field Rendering for Statistical Object Recognition

The problem, for which we propose a solution, is the requirement of a huge amount of training material, i.e. sample views of each object. In the training step the pose of the object in 3–D is needed. For this, usually

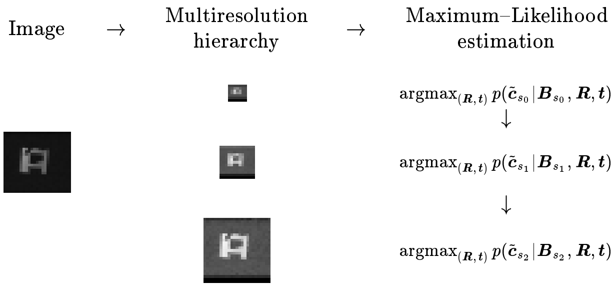
Image
→
Multiresolution hierarchy
→
Maximum–Likelihood estimation
$\mathrm{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(\tilde{\boldsymbol{c}}_{s_0}|\boldsymbol{B}_{s_0}, \boldsymbol{R}, \boldsymbol{t})$
↓
$\mathrm{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(\tilde{\boldsymbol{c}}_{s_1}|\boldsymbol{B}_{s_1}, \boldsymbol{R}, \boldsymbol{t})$
↓
$\mathrm{argmax}_{(\boldsymbol{R},\boldsymbol{t})}\, p(\tilde{\boldsymbol{c}}_{s_2}|\boldsymbol{B}_{s_2}, \boldsymbol{R}, \boldsymbol{t})$

Fig. 3: System overview: Probability density maximization on multiresolution hierarchy of images (from [9]).

a known constellation of object and camera is achieved, by moving a camera by a robot's arm around the object. This is a time consuming task, because for the robot's position high precision is needed.

An alternative approach to this problem is to reconstruct the light field of an object with a certain resolution and to render views for training. The advantages of this approach are:

1. If we have recorded a light field once, we are able to render arbitrary many views at arbitrary positions with arbitrary camera parameters like for focus or radial distortion.

2. We are also able to render views, even at positions which can not be reached by the robot's arm or be recorded sharply because of the limitations of the camera lens.

3. The rendering process is much faster (app. one second per image) than the movement of the robot (app. twelve seconds per image).

4. You can use one special constellation for recording the light field which is calibrated exactly.

Therefore we have a virtual environment for generation of scene views.

It may sound strange to use one object model (the light field) to create another one (the statistical). So it seems to be possible to use the light field itself as object model. But the light field just is able to represent a special recording situation with a certain illumination whereas the statistical object model also includes the variance of illumination. Therefore we have to use several light fields recorded at different illuminations to render views for the statistical training.

One basic assumption is, that the rendered images reflect the reality with a sufficient accuracy to configure a statistical model. This depends on the resolution of the light field, the exactness of the camera positioning during recording and the accuracy of interpolation. These preliminaries are examined in the next section, where the verification is based on the results of the 3–D object localization described in Sect. 3, taking into account just the coarsest resolution.

5 Experiments and Results

For configuring the statistical classifier, a training set has to be recorded. For our special classifier, the recording positions must have a constant distance to the object, so they have to lie on a sphere. Also their exact pose in 3–D must be known. Now we achieve these views by two ways: we directly record them with a robot and we reconstruct them with a previously recorded light field. We have used two objects for evaluation. Both original objects can be seen in Fig. 4 as well as the corresponding views rendered by the reconstructed light field.

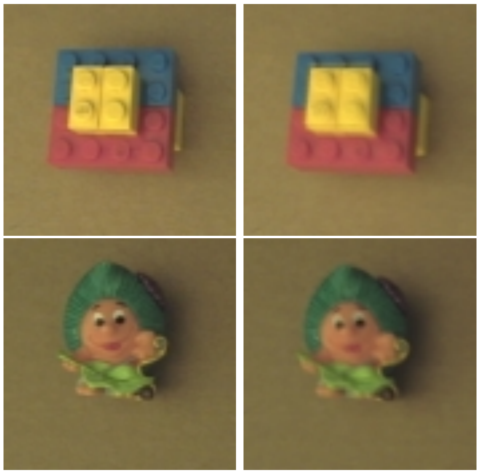

Fig. 4: Two objects for experiments. Left: original image. Right: rendered image using the reconstructed light field. First row: object 1. Second row: object 2.

The image data in the training step for a certain illumination on the one hand consists of 17^2 original color images, and on the other hand of 33^2 and 17^2 rendered color images of a light field, which has been reconstructed from 17^2 original images, as described in Sect. 2. For training, three different illuminations have been used. For testing both classifier — trained by original images as well as the classifier configured by rendered images — they have to localize 289 different directly recorded views of the objects in 3–D, i.e. three angles and the displacement in the 2–D image plane must be estimated. The distance to the object is assumed to be known and fixed. At present, this is a limitation by the statistical model described in Sect. 3. For evaluation, we decide, that an object is not correctly localized, if the reconstructed pose for at least one angle differs more than 10 degrees from the real pose. In Tab. 1 the results can be seen for the classifiers trained with original images (ORIG) and trained with rendered images (LIF).

It is worth noting, that the information which is used for the light field reconstruction always consists of 17^2 images, independently of the number of images rendered for training.

object	# training images	error rate ORIG	error rate LIF
1	3×33^2	n.a.	21.9
1	3×17^2	12.1	22.8
2	3×33^2	n.a.	15.3
2	3×17^2	7.6	14.1

Tab. 1: Localization results for the objects using a threshold of 13 degrees for one of the three rotation axes and 10 pixels for x– and y–translation.

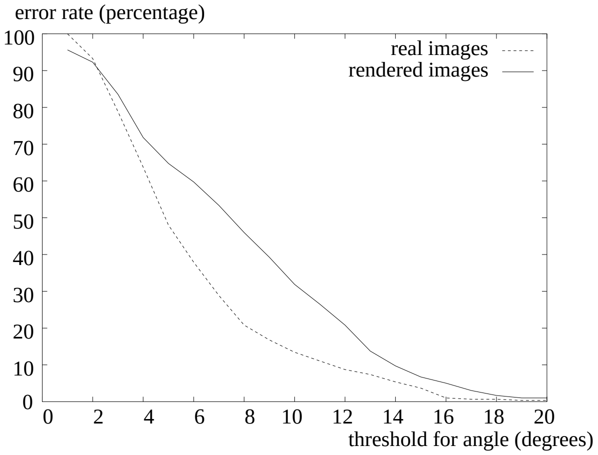

Fig. 5: Comparison of error rates for object 2 in dependence on the threshold for angles.

The results show, that the rendered images still lack the necessary reality to result in the same recognition rates as for real images. Increasing the number of rendered training images does not result in a better recognition rate generally. The classifier LIF has an up to two times as large error rate. Fig. 5 shows the dependency of the error rates on the threshold for angles.

A detailed inspection of the results have shown, that for rendered images at the border of the field of view the perspective distortions increase significantly. This is caused by small errors in the camera calibration and the hand–eye calibration of the robot's arm. These problems are not applicable for the classifier ORIG, because the statistical model is optimized with respect to the camera and the accuracy of hand–eye calibration. The exact calibration parameters play no part in the parameter estimation step as long as the test images are taken with the same camera and the same robot configuration. Thus, to run a fair comparison between ORIG and LIF, test images should be taken with another camera and robot arm.

Nevertheless, the results also show, that the concept of light fields is a promising image based rendering technique for statistical object recognition. Including geometric information in the reconstruction process as already mentioned by [6] will improve the quality. Also, more accurate camera calibration and hand–eye calibration of the robot arm will fix some errors in the reconstruction process.

6 Conclusion

In this paper we have presented a comparison between the quality of real and rendered images with respect to the suitability of light fields for statistical object recognition. The experiments have shown that although the rendered images from the light field look quite photo–realistic, the reality is not fully covered. Also increasing the number of rendered images does not result in a better recognition rate. On the one hand, that is no surprise, because this does not change the information, which the light field contains, on the other hand, in contrast to the statistical model, the light field contains all information on the 3D–constellation of the viewing rays.

In our future work we will optimize the reconstruction process by a more accurate camera calibration, more sophisticated interpolation processes for rendering, and the integration of geometric information. Also experiments are actually conducted for training of different distances to the object, for which the robot's arm is not able to move completely around the object.

References

1. R. Beß, D. Paulus, and H. Niemann. 3D recovery using calibrated active camera. In *Proceedings of the International Conference on Image Processing (ICIP)*, volume II, pages 855 – 858. IEEE Computer Society Press, Lausanne, Schweiz, September 1996.
2. S. J. Gortler, R. Grzeszczuk, R. Szelinski, and M. F. Cohen. The lumigraph. *Computer Graphics (SIGGRAPH '96 Proceedings)*, pages 43–54, August 1996.
3. F. Heitz and P. Bouthemy. Multimodal motion estimation and segmenation using markov random fields. In *Proceedings of International Conference on Pattern Recognition*, pages 378–383, 1990.
4. J. Hornegger and H. Niemann. Statistical learning, localization, and identification of objects. In ICCV 95 [5], pages 914–919.
5. *Proceedings of the* 5^{th} *International Conference on Computer Vision (ICCV)*, Boston, Juni 1995. IEEE Computer Society Press.
6. M. Levoy and P. Hanrahan. Light field rendering. *Computer Graphics (SIGGRAPH '96 Proceedings)*, pages 31–45, August 1996.
7. H. Murase and S. K. Nayar. Visual learning and recognition of 3–D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, Januar 1995.
8. H. Niemann. *Klassifikation von Mustern.* Springer, Berlin, 1983.
9. J. Pösl and H. Niemann. Statistical 3–D object localization without segmentation using wavelet analysis. In *Computer Analysis of Images and Patterns (CAIP)*, pages 440–447, Kiel, September 1997. Springer.
10. P. Viola and W. Wells III. Alignment by maximization of mutual information. In ICCV 95 [5], pages 16–23.