Color and Depth in Appearance Based Statistical Object Localization

Josef Pösl *, **Benno Heigl** ** and **Heinrich Niemann**

Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
email: {poesl,heigl,niemann}@informatik.uni-erlangen.de

Abstract

Appearance based approaches use the intensity information of images directly for object recognition without an intermediate step of low level segmentation. In this paper we show how pose estimation results can be improved by additional color and depth data. We define a statistical 3–D object model and describe a method for depth calculation based on a dense disparity map. The experiments show that the additional data helps to solve for ambiguities caused by object symmetries and similar object and background intensity values.

1 Introduction and Motivation

Object recognition in images can be accomplished by the two main approaches in computer vision: based on low level segmentation or on object appearance. The segmentation approach suffers from errors in the segmentation and the loss of information when restricting the recognition process to a higher level of abstraction. As the segmentation is an isolated first step in the recognition task this step does not use knowledge about the concerned objects to adapt to different object parts.

Appearance based approaches avoid the intermediate step of low level segmentation. They rather model the pixel intensities or derived local features originating of an object directly. The simplest method in this area is correlation with an object template. [2] describe a recognition method based on the singular value decomposition of a vector space spanned by the gray–level data of several images. [1] present a method with mixture densities of the gray level values of object images.

We have developed a statistical appearance based 3–D object model for local wavelet features in [4, 3]. There the model is evaluated for gray–level images. In this paper we will evaluate the gain of additional information for the recognition task. We will consider color as well as depth in the experiments. The depth map is calculated from a dense disparity map computed by a phase based method [5].

2 Statistical Model

2.1 Overview

The aim of the presented system is the pose estimation of a rigid 3–D object in a single 2–D image with multi channels for color and depth. We assume that the objects do not vary in scale.

In a first step of the localization process a multi–resolution analysis of the image is used to derive feature values on a rough scale $s \in \mathbb{Z}$ and resolution (sampling rate) $r_s \in \mathbb{R}^+$ at the locations of rectangular sampling grids. In this paper we evaluate the approach only for one rough scale. The estimation can be refined with further scale levels as described in [4]. Given an image $f(x, y)$ with $x \in \{0, 1, \ldots, D_x - 1\}$, $y \in \{0, 1, \ldots, D_y - 1\}$ the observed feature values at scale s are denoted by $\boldsymbol{c}_s(x, y) = (c_{s,0}, \ldots, c_{s,N-1})^{\mathrm{T}}$ ($x \in \{0, r_s, \ldots, r_s D_x - 1\}$, $y \in \{0, r_s, \ldots, r_s D_y - 1\}$). In the experiments of this paper the features $\boldsymbol{c}_s$ are chosen as the logarith-

*The author is member of the Center of Excellence 3-D Image Analysis and Synthesis sponsored by the „Deutsche Forschungsgemeinschaft“ (DFG).

**This work was partially funded by the DFG under grant number SFB 182.

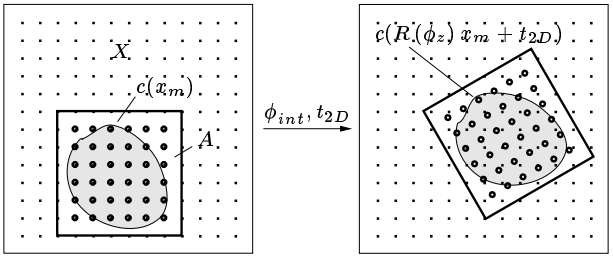


Figure 1: Object covered with grid for feature extraction

mic coefficients of the scaling functions — that are the low pass coefficients — of a discrete Johnston wavelet transform ($N = 1$).

Let $\tilde{\boldsymbol{c}}_s$ be the vector of the concatenated feature values detected in an image on scale s, $\boldsymbol{B}_s$ the model parameters and $\boldsymbol{R}, \boldsymbol{t}$ be the 3–D rotation matrix and translation vector. The rotation $\boldsymbol{R}$ is defined by the rotation angles ϕ_x, ϕ_y and ϕ_z round the x–, y– and z–axis respectively.

The model parameters $\boldsymbol{B}_s$ consist of geometric information like probability density locations and other density parameters. The density $p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t})$ can then be used for generating object location hypotheses with a maximum likelihood estimation:

$$(\widehat{\boldsymbol{R}}_s, \widehat{\boldsymbol{t}}_s) = \underset{(\boldsymbol{R},\boldsymbol{t})}{\operatorname{argmax}}\, p(\tilde{\boldsymbol{c}}_s|\boldsymbol{B}_s, \boldsymbol{R}, \boldsymbol{t}).$$

2.2 Model formulation

This section shows the definition of a probability density function for a single object. To simplify the notation the index s is omitted.

The model object is covered with a rectangular grid of local feature vectors (see Figure 1). The grid resolution is the same as the image resolution on the actual scale. Let $A \subset \mathbb{R}^2$ be a small region (e.g. rectangular) which contains the object projection to the image plane for all possible rotations $\boldsymbol{\phi}_{ext} = (\phi_y, \phi_x)$ outside the image plane. Let $X = \{\boldsymbol{x}_m\}_{m=0,\ldots,M-1}$, $\boldsymbol{x}_m \in \mathbb{R}^2$ denote the grid locations and $\boldsymbol{c}(\boldsymbol{x})$ the feature vector at location $\boldsymbol{x}$. We assume the background features as distibuted uniformly and independent of the object features. Then it is sufficient to consider $p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t})$, where $\boldsymbol{c}_A$ is the subset of $\boldsymbol{c}$ which is covered by A. The grid positions and the model area A are part of the model parameters $\boldsymbol{B}$.

The feature vectors are assumed to be normally distributed with independent components. Let $\mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the normal density, where $\boldsymbol{\mu}$ is the mean vector with concatenated local feature mean vectors $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}$ is the covariance matrix with elements $\sigma_{m,\bar{m},n} = \operatorname{cov}(c_{m,n}, \boldsymbol{c}_{\bar{m},n})$.

The density parameters are a function of the rotation parameters ϕ_y, ϕ_x for 3–D objects, so that:

$$\begin{aligned} & p(\boldsymbol{c}_A|\boldsymbol{B}, \boldsymbol{R}, \boldsymbol{t}) \\ &= p(\boldsymbol{c}_A|(\boldsymbol{\mu}(\phi_y, \phi_x), \boldsymbol{\Sigma}(\phi_y, \phi_x)), \boldsymbol{R}, \boldsymbol{t}) \\ &= \mathcal{N}(\boldsymbol{c}_A(\phi_z, \boldsymbol{t}_{2D})|\boldsymbol{\mu}(\phi_y, \phi_x), \boldsymbol{\Sigma}(\phi_y, \phi_x)), \end{aligned}$$

with $\boldsymbol{c}_A(\boldsymbol{R}(\phi_z), \boldsymbol{t}_{2D})$ as the concatenated feature vectors $\boldsymbol{c}(\boldsymbol{R}(\phi_z)\,\boldsymbol{x}_m + \boldsymbol{t}_{2D})$, the 2–D rotation matrix $\boldsymbol{R}(\phi_z)$ for the rotation and the translation $\boldsymbol{t}_{2D}$ in the image plane. The image feature vectors at the transformed 2–D locations are calculated by linear interpolation. Assuming continuous functions $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}$ they can be rewritten using a basis set for the domain of two-dimensional functions $\{v_r\}_{r=0,\ldots,\infty}$ with coordinates $a_{m,n,r}, b_{m,\bar{m},n,r} \in \mathbb{R}$ $(r = 0, \ldots, \infty)$ and the elements $\tilde{\sigma}_{m,\bar{m},n}$ of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$:

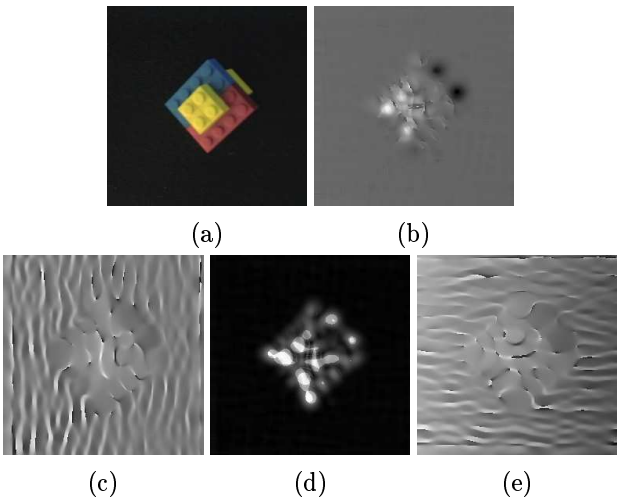
$$\mu_{m,n} = \sum_{r=0}^{\infty} a_{m,n,r} v_r, \quad \tilde{\sigma}_{m,\bar{m},n} = \sum_{r=0}^{\infty} b_{m,\bar{m},n,r} v_r.$$

The functions are approximated by using only part of the complete basis set $\{v_r\}_{r=0,\ldots,L-1}$. The Taylor decomposition shows, that the approximation error can be made as small as possible by choosing L large enough. With this approximation a fast computation of the density function and a maximum likelihood estimation of the basis coefficients is possible. The estimation results in closed estimation terms if $\tilde{\sigma}_{m,\bar{m},n}$ is assumed as constant (see [3]). The value of L is limited mainly by the computation time for the density and the size of the training set for estimation.

3 Depth Information

We recorded different views of an object with a camera mounted on a robot arm. The camera moved around the object within a partial sphere always pointing to one point inside the object. The single positions are arranged in a grid–like manner.

To compute the displacement vectors between two images, we applied the technique proposed in [5] providing a dense displacement vector field and additionally a dense reliability map. Other methods also could be applied.



(a) (b) (c) (d) (e)

Figure 2: (a) one input image; (b) corresponding combined weighted depth map from four neighbors; (c) and (e) depth maps with different neighbors; (d) reliability map

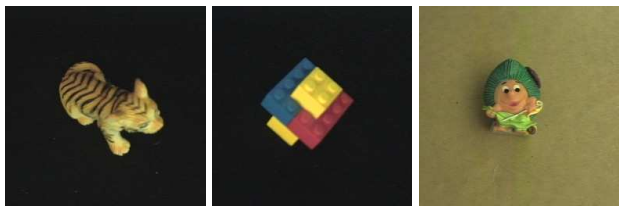
Using two adjacent views we can calculate the homogeneous matrix $\boldsymbol{H}$ describing the transformation from one position to the other ($H_{44} = 1$). Knowing this transformation and assuming orthogonal projection the following dependency between the projection $\boldsymbol{p} = (p_1, p_2)^{\mathrm{T}}$ of an object point $\boldsymbol{x} = (x, y, z, 1)^{\mathrm{T}}$ in the first view and the displacement vector $\boldsymbol{d} = (d_1, d_2)^{\mathrm{T}}$ to the second view holds: $\boldsymbol{p} + \boldsymbol{d} = \tilde{\boldsymbol{H}} \cdot \boldsymbol{x}$, where $\tilde{\boldsymbol{H}}$ consists of the first two rows of $\boldsymbol{H}$. The depth value z then can be calculated at every position $\boldsymbol{p}$, describing the distance in units of pixel size.

If we model depth data with the appearance based approach proposed above, two main problems have to be solved:

- The first problem is how to consider the reliabilities of depth values at every point,
- and the second is how to handle the dependency of the depth values on the locations of the different views to calculate the depth image.

The first problem can be seen in Figure 2 (c,d). In areas of high reliability, the calculated depth structure is according to the real one, but in areas of low reliability, the errors are obviously great. To solve this problem, we chose the way of weighting the depth values by their reliability, ranging from 0 to 1. Therefore in homogeneous regions, the depth values become nearly zero whereas in structured regions, where the depth values are significant, they are nearly equal the real values.

To see the second problem, imagine the ideal stereo configuration where two positions are translated only horizontally. At horizontal gray–level edges no depth values can be calculated, But if you choose a configuration with vertically translated views, the depth of vertical edges cannot be calculated. As described before, our recording positions are arranged in a grid–like manner, so for one viewing position, we get several neighbors, which all can be used to calculate a weighted depth image. Figure 2 (c) and (e) show two different depth maps for two different neighbors. To combine these views, we calculated the average weighted depth value. The argument for this approach is the same as for the solution of the first problem. The result is shown in Figure 2 (b), where four weighted depth maps were combined.



(a) (b) (c)

Figure 3: Objects (a), (b) and (c)

Following equation describes the resulting solution of the two problems:

$$\hat{z}(x, y) = \frac{1}{n} \sum_{k=1}^{n} r_k(x, y) \cdot z_k(x, y)\,,$$

where $r_k(x, y)$ denotes the reliability value at the position $(x, y)^{\mathrm{T}}$ using the k–th neighbor. The value $z_k(x, y)$ denotes the corresponding depth value computed as described above. It is not a depth value in the usual sense, but it combines information about inhomogeneity and depth and therefore gives new relevant information about the appearance.

4 Results

Figure 3 shows the objects used in this work. The images are 256 pixels in square. The localization was performed on one scale level s_0 with resolution $r_{s_0} = 8$ pixels, $L = 21$, constant $\boldsymbol{\Sigma}$ and row dependencies. Only the best localization result of level s_0 is evaluated. The Downhill Simplex algorithm was used for the local parameter search following the global grid search. The computation

Object	Data	Fail [%]	Error					
			Transl. (Pix)		int.Rot. (^{o})		ext.Rot. (^{o})	
			mean	max	mean	max	mean	max
(a)	Gray	0	1.1	2.5	1.3	3.2	1.3	6.8
	Col+Depth4	0	1.1	2.5	1.3	4.0	1.2	3.4
(b)	Gray	33	2.5	5.3	0.4	2.4	3.5	8.6
	Color	2	1.4	3.4	0.9	3.2	2.0	8.3
	Gray+Depth4	0	1.4	2.6	0.7	2.9	1.2	4.3
	Col+Depth1	0	1.1	2.9	0.9	3.1	1.9	7.2
	Col+Depth4	0	1.1	2.5	1.0	2.9	1.3	4.5
(c)	Gray	55	1.8	3.5	1.3	5.9	3.7	8.9
	Color	37	1.3	3.2	2.1	6.7	2.8	8.9
	Gray+Depth4	29	1.5	3.3	1.9	7.1	2.5	8.8
	Col+Depth1	39	1.3	3.8	2.3	7.9	3.0	8.9
	Col+Depth4	22	1.3	3.6	2.3	6.8	2.5	8.5

Table 1: Results for different data types

time on a SGI O2 (R10000) is about 16 seconds for feature extraction and localization of one of the objects.

Each object is available in four image sequences with 256 images each. The background is homogeneous and the external rotation parameters are restricted to $-30^o < \phi_y, \phi_z < 30^o$. Three sequences were taken for training, the other for testing. The range of $\phi_z, \boldsymbol{t}$ was considered completely, resulting in a five-dimensional search. The depth values of an image were calculated from four adjacent views to get a more accurate depth map (Depth4) for training. Tests were performed for these accurate maps as well as depth maps (Depth1) from one neighboring image with test sets of 196 images.

Table 1 shows experimental results for different combinations of the available color and depth information. An estimation result is classified as failure if the translation error is more than 10 Pixels and the rotation error is more than 9^o.

The experiments demonstrate the improvement in position estimation as well as failures. Object (a) bears enough gray–level information to perform the recognition quite well without additional data. Nevertheless color and depth information improve the position estimation. Object (b) appears almost symmetrical in gray–valued images but is asymmetrical in color images. This leads to typical rotation errors of 180^o for the gray–data which disappear with color. The main part of object (c) has almost the same intensity values as the background. Furthermore the lighting was changed more than for the other objects. The gain in failure rate also is significant there.

All experiments confirm the expectation that additional color or depth information will give better recognition results. This is especially true if object symmetries can only be distinguished or the object is only separable from background with this additional data.

References

[1] V. Kumar and E. S. Manolakos. Unsupervised model–based object recognition by parameter estimation of hierarchical mixtures. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 967–970, Lausanne, Schweiz, September 1996. IEEE Computer Society Press.

[2] 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, January 1995.

[3] J. Pösl. Statistical pose estimation with local dependencies. In H.-P. Seidel, B. Girod, and H. Niemann, editors, *3D Image Analysis and Synthesis '97*, pages 147–154, Erlangen, November 1997. Infix.

[4] 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, Germany, September 1997. Springer.

[5] W.M. Theimer and H.A. Mallot. Phase–based binocular vergence control and depth reconstruction using active vision. *Computer Vision, Graphics, and Image Processing*, 60(3):343–358, 1995.