OBJECT LOCALIZATION WITH MIXTURE DENSITIES OF WAVELET FEATURES

Josef Pösl *and Heinrich Niemann*

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

ABSTRACT

In this paper we address the localization of a single 2–D or 3–D object in scenes with complex background. A wavelet transform is applied to the 2–D gray level image for the extraction of local features. We use discrete tensor product wavelets and compute feature values for all image positions from the wavelet coefficients. A statistical object model of these features is defined. The object model is combined with a statistical background model in a mixture density for the image. The object localization is performed with an Expectation Maximization (EM) approach.

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 abstraction level. 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 of an object directly. The simplest method in this area is correlation with an object template. [3] describe a recognition method based on the singular value decomposition of a vector space spanned by the gray–level data of several images. Thereby a large number of images is approximately encoded by a small number of basis images. The projection parameters of an image into this eigenspace can be used for recognition. Maximization of the mutual information between an object model and an object in a scene is a further approach [6].

[2] describe a method based on mixture densities of the gray level values of object images. To incorporate local coherence into their model they use the POEM (Perceptually Organgamized EM) algorithm for recognition which adds a heuristic (quadratic) weighting factor to the EM energy term. As all possible object positions are modeled as hidden variables in a mixture density and without an hierarchical solution, this approach tends to be very complex.

We have developed a statistical appearance based 3–D object model for local wavelet features in [5]. The model — as most of the others — is suitable to detect one or more appearances of one object in scenes with homogeneous background. If there is no object occlusion and the image data contains no extreme outliers this model is applicable for the localization of a 2–D object with heterogeneous background. This is also true for 3–D objects if the model is trained with different types of background. Due to the amount of training data needed this is no practical solution. Therefore a background model is trained independently of the object and combined with the object model in a mixture density. In contrast to the POEM algorithm, which adds a heuristic weighting term to the expectation, we consider statistical dependencies for a better local coherence.

2. FEATURES

Instead of using the image gray–level data directly for modeling, we apply a wavelet transform to the image in order to extract multiscale features. Local feature vectors are constructed by either only the features on one scale or the features of different scales. By this we either can take an hierarchical approach for pose search as described in [5] or we get a multidimensional feature description of each image location, which contains information about different neighborhood regions.

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

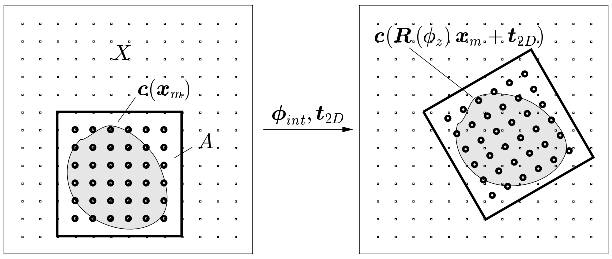

Figure 1: Object covered with grid for feature extraction

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 and resolution $r_s \in \mathbb{R}^+$ 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 logarithmic coefficients of a discrete Johnston wavelet transform.

3. STATISTICAL MODEL

3.1. Overview

The aim of the presented system is the pose estimation of a rigid 2–D or 3–D object in a single 2–D image. We assume that the objects do not vary in scale.

Let $\tilde{\boldsymbol{c}}_s$ be the vector of the concatenated feature values for 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:

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

3.2. Object density

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 in this section 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 $\sigma_{m,\bar{m},n} = \operatorname{cov}(c_{m,n}, c_{\bar{m},n})$.

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

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

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 [4]). The value of L is limited mainly by the computation time for the density and the size of the training set for estimation.

3.3. Mixture densitiy for arbitrary background

In the previous section we have shown the definition of the statistical model for a single object. This model already contains a implicit background model. We now consider background explicitly. First we model arbitrary background. This type of background occurs if multiple objects are contained in the background but we do not know their position or class.

Let Ω_1 denote the object class and Ω_0 the background class which has the same normal distribution at each image location and therefore no positional parameters. We assume that each image location belongs

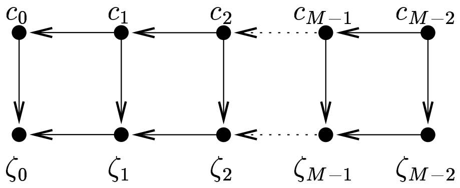

Figure 2: Dependency structure of mixture

either to background or object. Let $\zeta : X \to \{0,1\}$ denote the assignment function which bears the hidden information to which class $\Omega_{\zeta(\boldsymbol{x}_m)}$ location $\boldsymbol{x}_m$ belongs. For a shorter notation we write $\zeta_m = \zeta(m) := \zeta(\boldsymbol{x}_m)$. Then the mixture density is

$$p(\boldsymbol{c}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \quad = \quad \sum_{\boldsymbol{\zeta}} p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$$

with $\boldsymbol{\zeta} = (\zeta(m))_{\boldsymbol{x}_m \in X}$. In [4] we have shown that better localization results can be achieved by considering local dependencies in the object model instead of assuming all local features as independent. Therefore we assume row dependencies with respect to the image locations in this paper. The theory can easily be generalized to other neighborhood systems.

We describe the density for one–dimensional feature vectors $\boldsymbol{c}(\boldsymbol{x}_m) = c_m$. Let the feature locations be ordered by their row dependency. Then we get

$$p(\boldsymbol{c}|\Omega_1,\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \quad = \quad p(c_0|\Omega_1,\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \prod_{\boldsymbol{x}_m \in X\setminus\{\boldsymbol{x}_0\}} p(c_m|c_{m-1},\Omega_1,\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$$

for the object density, where the features outside the object window A are modeled as background. The background density is built of independent and locally invariant components $p(c_m|\Omega_0,\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) = p(c|\Omega_0,\boldsymbol{B})$:

$$p(\boldsymbol{c}|\Omega_0,\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in X} p(c_m|\Omega_0,\boldsymbol{B}).$$

The mixture density with the same type of dependencies extended to both, features and assignment (see Figure 2), is

$$p(\boldsymbol{c} \mid \boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) = \sum_{\boldsymbol{\zeta}} p(c_0,\zeta(0)|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) \prod_{\boldsymbol{x}_m \in X\setminus\{\boldsymbol{x}_0\}} p(c_m,\zeta(m)|c_{m-1},\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}).$$

The mixture considers all possible assignments of the local feature vectors. The product terms can be written as:

$$p(c_m,\zeta(m) \mid c_{m-1},\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) = \frac{p(c_m,c_{m-1}|\zeta(m),\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})}{p(c_{m-1}|\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})} \frac{p(\zeta(m),\zeta(m-1)|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})}{p(\zeta(m-1)|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})}$$

If two neighbor locations are assigned to the same class the term $p(c_m,c_{m-1}|\zeta(m),\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$ is equal to $p(c_m,c_{m-1}|\Omega_{\zeta(m)},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$. Otherwise it is assumed as $p(c_m|\Omega_{\zeta(m)},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})p(c_{m-1}|\Omega_{\zeta(m-1)},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$.

In order to localize the object in an image we use an expectation maximization approach. The following expectation term (Kullback–Leibler) is maximized:

$$\mathcal{E}_{\boldsymbol{\zeta}}\ (\log\ p(\boldsymbol{c},\boldsymbol{\zeta}|\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})|\,\boldsymbol{c},\boldsymbol{B},\boldsymbol{R},\boldsymbol{t}) = \sum_{\boldsymbol{x}_m \in X} \mathrm{h}_m(c_m,c_{m-1})$$

with

$$\mathrm{h}_0\ (c_0,c_{-1}) = \sum_{\zeta(0)} p(\zeta(0)|c_0,\boldsymbol{B},(\boldsymbol{R},\boldsymbol{t}))\ \log\,(p(c_0|\zeta(0),\boldsymbol{B},(\boldsymbol{R},\boldsymbol{t}))$$

and

$$\mathrm{h}_m\ (c_m,c_{m-1}) = \sum_{\zeta(m-1),\zeta(m)} p(\zeta(m-1),\zeta(m)|c_{m-1},c_m,\boldsymbol{B},(\boldsymbol{R},\boldsymbol{t}))\ \log\,(p(c_m,\zeta(m)|c_{m-1},\zeta(m-1),\boldsymbol{B},\boldsymbol{R},\boldsymbol{t})$$

for $m > 0$. In contrast to the standard EM-approach we perform no alternating estimation of assignment probability and parameters. We maximize the estimation function directly. The expectation is a sum of local functions h_m. If we subtract the logarithmic background density which we assume as constant with respect to the parameters, we only have to consider local functions $\tilde{\mathrm{h}}_m$ for the features inside the object area A:

$$\sum_{\boldsymbol{x}_m \in A} \tilde{\mathrm{h}}_m(c_m,c_{m-1}).$$

Therefore its calculation has the same time complexity as the calculation of the single object density.

The observation, that many of the simple constituent functions are similar, suggests to approximate the expectation term by a small subset $\left\{\tilde{\mathrm{h}}_k\right\}_{k=0\ldots K-1}$ of those functions by

$$\tilde{\mathrm{h}}_m \approx \sum_{k=0}^{K-1} \tilde{\mathrm{h}}_k \sum_{r=0}^{L-1} d_{m,k} v_r.$$

This leads to a filter technique for the global pose search. First the image features are transformed by each of the functions $\tilde{\mathrm{h}}_k$ separately. The transformation results are then combined by convolutions.

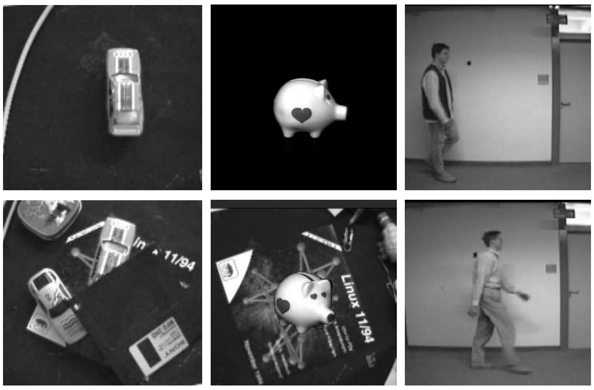

Figure 3: Objects *car* and *pig* and walking person for training (top) and test (bottom)

3.4. Mixture density with static background

If we know the appearance of the background we introduce a third class in the mixture model. This background model assumes different densities for all feature locations. The densities do not depend on the object transformation parameters. This type of model is applicable for example if a person moves in front of a static background and we want to determine the position of the head. The expectation term is similar to the case with arbitrary background only. A speed up of the global search by a filter technique is also possible.

4. RESULTS

Figure 3 shows the objects used in this work. The images are 256 pixels in square. The local assignment probabilities were assumed as independent.

For the experiments with varying background the localization was performed on one rough scale level s_0 with resolution $r_{s_0} = 8$ pixels. The logarithmic low pass coefficient of a Johnston 8 transform was chosen as local feature resulting in one–dimensional feature vectors for each location.

The 2–D object *car* was trained with 20 images with homogeneous background and different object positions and lighting. The test images contain heterogeneous background and object occlusions. The localization with the mixture was correct for 10/20 images. Without the mixture density we got 9/20 correct results. This shows that the single object density can cope with occlusions because of its local structure but still can be improved by a mixture model.

For object *pig* from the Columbia Object Image Library (COIL) one image sequence with a complete object rotation with respect to ϕ_y in 72 equidistant steps and homogeneous background was available. Half of the images were used for training. The other half were mixed with a background image for testing. The range of $\phi_z, \boldsymbol{t}$ was searched completely, resulting in a four-dimensional search. The localization with independence assumption gave 18/36 correct results, where the mean error was one pixel for translation, 1^o (6^o) for internal (external) rotation and the maximum allowed rotational error was 15^o. The translation was estimated correctly for all images. The rotation was partially incorrect due to similar appearance on the rough scale.

The experiments with static background were performed on a fine scale with resolution $r_{s_0} = 2$ pixels and independence assumption. The finest three resolution levels of a Johnston transform were used for feature extraction. Each four–dimensional local feature vector is composed of the logarithm of the sums of the high pass coefficients for each analysis scale and the logarithmic low pass coefficient of the roughest scale. A general head model was trained with 140 images of 20 different people and combined with a general background model and a static background trained on all images. The static model was trained for each test sequence. The model was tested for two sequences of walking people with 100 images altogether. A global search for the head without object rotations resulted in 5% incorrect results for the mixture density and 20% for the single object density. If the search is restricted to a local area after a global search in the first image of the sequence, no errors occur. The localization thereby is considered as correct if the main part of the head is inside the object window.

5. REFERENCES

[1] *Proceedings of the* 5^{th} *International Conference on Computer Vision (ICCV)*, Boston, Juni 1995. IEEE Computer Society Press.

[2] 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.

[3] H. Murase and S. K. Nayar. Visual learning and recognition of 3–D objects from appearance. *Int. Journal of Computer Vision*, 14(1):5–24, Januar 1995.

[4] 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.

[5] 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.

[6] P. Viola and W. Wells III. Alignment by maximization of mutual information. In ICCV 95 [1], pages 16–23.