in: Proceedings of the Fifth International Conference on Signal and Image Processing (SIP 2003),
ACTA Press, pp. 411-416, Honolulu, USA, 2003.

OBJECT STRUCTURE FROM NOISY IMAGES

Gabriele Peters
Informatik VII, Graphische Systeme
Universität Dortmund
Otto-Hahn-Str. 16
D-44227 Dortmund, Germany
email: peters@ls7.cs.uni-dortmund.de

ABSTRACT
We describe the establishment of a compound object model for object recognition purposes which provides the frame for the extraction of object structure from images degraded by noise. Our vision system is inspired by cognitive principles. From a set of sample views we automatically generate a sparse and view-based object representation, which contains enough information to represent the object for all poses. To verify this property we apply it in a pose estimation task with noisy and unfamiliar test views of the object. With an appropriate number of views in the object representation the proposed method shows a good selectivity and is able to distinguish views with a distance of only 3.6°, even if they are degraded considerably by Gaussian noise.

KEY WORDS
Computer Vision, Noise, Pose Estimation, Tracking, 3D Object Recognition

1 Introduction

Each object in our environment can cause considerably different patterns of excitation in our retinae depending on the observed viewpoint of the object. Despite this we are able to perceive that the changing signals are produced by the same object. It is a function of our brain to provide this constant recognition from such inconstant input signals by establishing an internal representation of the object. There are uncountable behavioral studies with primates that support the model of a view-based description of three-dimensional objects by our visual system. If a set of unfamiliar object views is presented to humans their response time and error rates during recognition increase with increasing angular distance between the learned (i.e., stored) and the unfamiliar view [1]. This angle effect declines if intermediate views are experienced and stored [2]. The performance is not linearly dependent on the shortest angular distance in three dimensions to the best-recognized view, but it correlates with an "image-plane feature-by-feature deformation distance" between the test view and the best-recognized view [3]. Thus, measurement of image-plane similarity to a few feature patterns seems to be an appropriate model for human three-dimensional object recognition. Experiments with monkeys show that familiarization with a "limited number" of views of a novel object can provide viewpoint-independent recognition [4]. In a psychophysical experiment subjects were instructed to perform mental rotation, but they switched spontaneously to "landmark-based strategies", which turned out to be more efficient [5].

Numerous physiological studies also give evidence for a view-based processing of the brain during object recognition. Results of recordings of single neurons in the inferior temporal cortex (IT) of monkeys, which is known to be concerned with object recognition, resemble those obtained by the behavioral studies. Populations of IT neurons have been found which respond selectively to only some views of an object and their response declines as the object is rotated away from the preferred view [6].

The capabilities of technical solutions for three-dimensional object recognition still stay far behind the efficiency of biological systems. Summarizing, one can say that for biological systems object representations in form of single, but connected views seem to be sufficient for a huge variety of situations and perception tasks.

2 Sparse Object Representation

In this section we introduce our approach of learning an object representation which takes these results about primate brain functions into account. We automatically generate sparse representations for real-world objects, which satisfy the following conditions:

a1 They are constituted from *two-dimensional* views.

a2 They are *sparse*, i.e., they consist of *as few views as possible*.

a3 They are capable of *performing perception tasks*, especially pose estimation, even from degraded images.

Our system consists of a *view representation builder* and an *object representation builder*. They are shown, together with their input and output data, in the diagram in figure 1, which depicts a one-directional flow of information. Of course, feedback from higher levels of processing to lower ones would allow for, e.g., unsupervised system

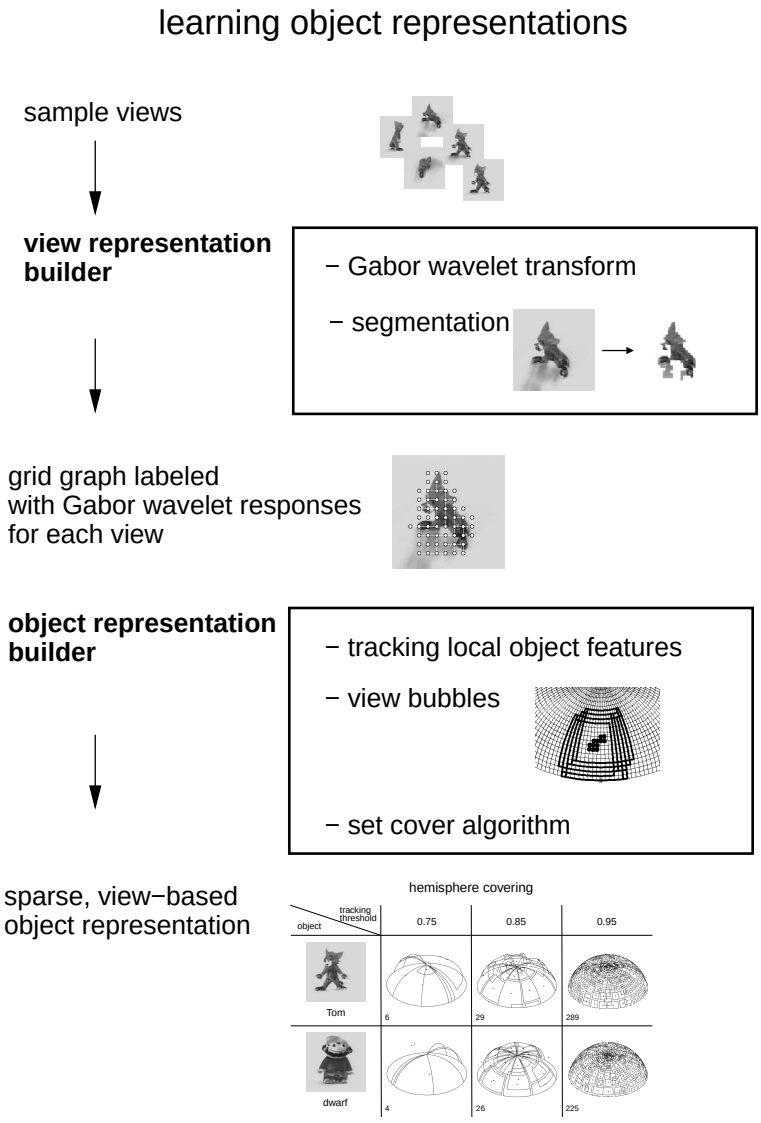

Figure 1. The system for learning sparse object representations consists of a view and an object representation builder. The resulting object representation consists of single but connected views. The numbers next to the resulting partitionings of the viewing hemisphere are the numbers of view bubbles which constitute the representation.

tuning or an improved segmentation, but this is not subject of this contribution. We start with the recording of a densely sampled set of views of the upper half of the viewing sphere of a test object. In the following we aim at choosing only such views for a representation which are representative for an area of viewpoints as large as possible.

Each of the recorded views is preprocessed by a *Gabor wavelet transform*, which is biologically inspired because Gabor wavelets approximate response patterns of neurons in the visual cortex of mammals [7, 8]. A *segmentation* based on gray level values [9] follows. It separates the object from the background. This results in a representation of each view in form of a *grid graph labeled with Gabor wavelet responses*. The graph covers the object segment. Each vertex of such a graph is labeled with the responses of a set of Gabor wavelets, which describe the local surroundings of the vertex. Such a feature vector is called *jet*. To facilitate an advantageous selection of views for the object representation a surrounding area of similar views is determined for each view. This area is called *view bubble*. For a selected view it is defined as the largest possible surrounding area on the viewing hemisphere for which two conditions hold:

b1 The views constituting the view bubble are *similar* to the view in question.

b2 *Corresponding object points* are known or can be inferred for each view of the view bubble.

The similarity mentioned in **b1** is specified below. Condition **b2** is important for a reconstruction of novel views as,

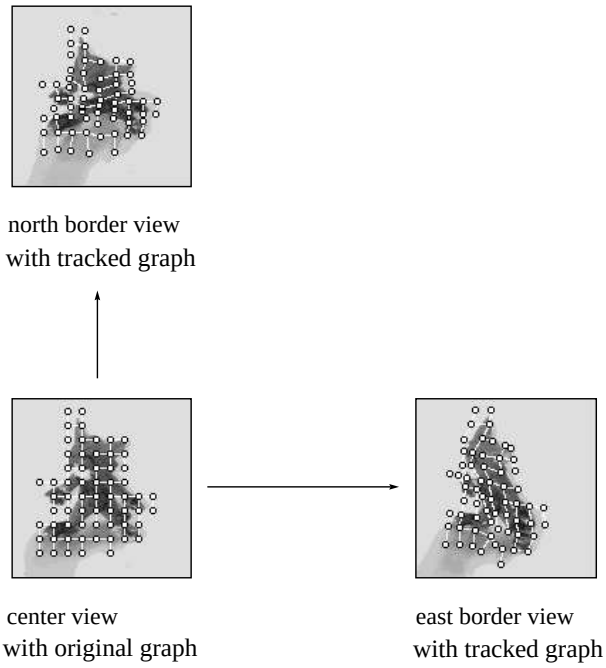

Figure 2. This figure shows a graph of the center view of a view bubble tracked to its east and north border views.

e.g., needed by our pose estimation algorithm. A view bubble may have an irregular shape. To simplify its determination we approximate it by a rectangle with the selected view in its center, which is determined in the following way.

The object representation builder starts by *tracking local object features*. Jets can be tracked from a selected view to neighboring views [10]. A similarity function $S(\mathcal{G}, \mathcal{G}')$ is defined between a selected view and a neighboring view, where $\mathcal{G}$ is the graph which represents the selected view and $\mathcal{G}'$ is a tracked graph which represents the neighboring view. Utilizing this similarity function we determine a *view bubble* for a selected view by tracking its graph $\mathcal{G}$ from view to view in both directions on the line of latitude until the similarity between the selected view and either the tested view to the west or to the east drops below a threshold τ, i.e., until either $S(\mathcal{G}, \mathcal{G}^w) < \tau$ or $S(\mathcal{G}, \mathcal{G}^e) < \tau$. The same procedure is performed for the neighboring views on the line of longitude, resulting in a rectangular area with the selected view in its center. The representation of a view bubble consists of the graphs of the center and four border views

$$\mathcal{B} := \langle \mathcal{G}, \mathcal{G}^w, \mathcal{G}^e, \mathcal{G}^s, \mathcal{G}^n \rangle \,, \tag{1}$$

with *w, e, s,* and *n* standing for *west, east, south,* and *north*. As this procedure is performed for each of the recorded views, it results in view bubbles overlapping on a large scale on the viewing hemisphere (see figures 1 and 2).

To meet the first condition **a1** of a sparse object representation we aim at choosing single views (in the form of labeled graphs) to constitute it. To meet the second condition **a2** the idea is to reduce the large number of overlapping view bubbles and to choose as few of them as possible which nevertheless cover the whole hemisphere. For the selection of the view bubbles we use the *greedy set cover algorithm* [11]. It provides a set of view bubbles which covers the whole viewing hemisphere. We define the *sparse, view-based object representation* by

$$\mathcal{R} := \{\langle \mathcal{G}_i, \mathcal{G}_i^w, \mathcal{G}_i^e, \mathcal{G}_i^s, \mathcal{G}_i^n \rangle\}_{i \in R} \tag{2}$$

where R is a cover of the hemisphere. Neighboring views of the representation are "connected" by known corresponding object points (the correspondences between center and border views), which have been provided by the tracking procedure. Figure 1 shows different covers of the hemisphere for two test objects.

We suppose that this representation contains enough information about the object to extract object structure from unfamiliar views even if they are degraded by noise. We varify this ability in a pose estimation task described in the next section.

3 Pose Estimation from Noisy Images

We recorded views of toy objects as gray level images of size 128×128 pixels with 256 gray levels with a distance of 3.6° on the upper viewing hemisphere. These images are then degraded independently by adding Gaussian white noise of zero mean and variance 0.06. Examples for noisy images are depicted in figure 3-2.

Given the sparse representation of the object in question, which is generated from the original images as described in section 2, and given a noisy test view of the object, the aim is now the determination of the object's pose displayed in the noisy view, i.e., the assignment of the test view to its correct position on the viewing hemisphere.

We extract a grid graph $\mathcal{G}_T$ from the noisy image of view T (figure 3-2)). This means that no a priori knowledge about the object is provided. Our pose estimation algorithm proceeds in two steps.

First, we match $\mathcal{G}_T$ to the center image of each view bubble using an elastic graph maching algorithm [12]. As a *rough estimate* of the object's pose we choose that view bubble the center image of which provides the largest similarity to $\mathcal{G}_T$.

In a second step we generate a virtual graph $\widehat{\mathcal{G}}$ for each unfamiliar view inside the chosen view bubble by

(1) an interpolation of corresponding Gabor wavelet responses and

(2) a linear combination of corresponding vertex positions

of representing graphs of neighboring views in the sparse object representation [13].

From each virtual graph $\widehat{\mathcal{G}}$ we reconstruct a virtual view $\widehat{V}$ using an algorithm which reconstructs the information contained in Gabor wavelet responses [14]. Accordingly, we reconstruct a virtual test view $\widehat{V}_T$ from $\mathcal{G}_T$

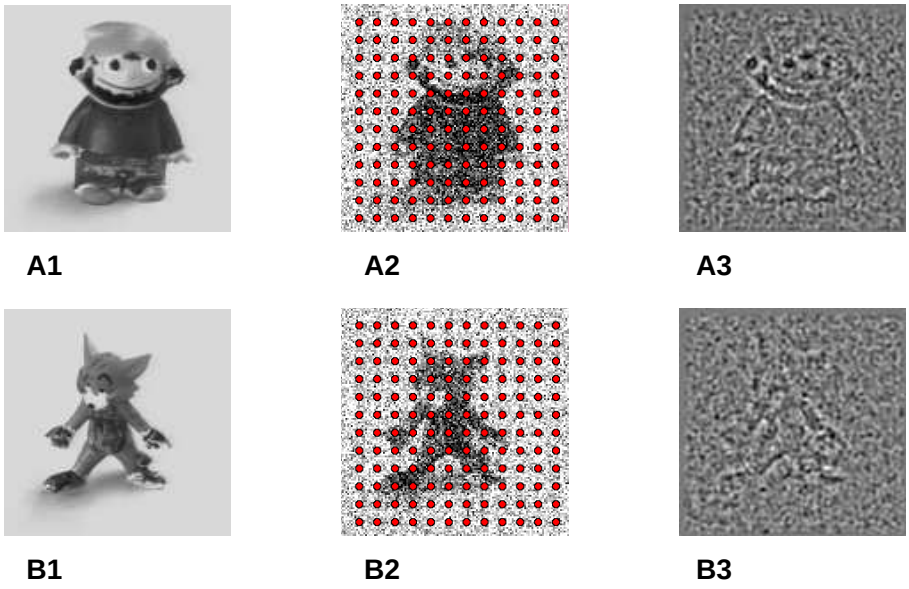

Figure 3. Original, Noisy, and Reconstructed Object Views. A) object "dwarf". B) object "Tom". 1) original image. 2) noisy image T with its representing graph $\mathcal{G}_T$. 3) image $\widehat{V}_T$ reconstructed from $\mathcal{G}_T$.

(figure 3-3)). The estimated pose of the test view T is the position on the viewing hemisphere of that virtual view $\widehat{V}$ which provides the smallest error $\epsilon(\widehat{V}, \widehat{V}_T)$ in a pixelwise comparison between $\widehat{V}_T$ and each $\widehat{V}$ [13].

For the evaluation of the algorithm 30 test views have been degraded by noise. The positions of some of them on the viewing hemisphere are displayed in figure 4. For two different toy objects and for five different partitionings of the viewing hemisphere, which have been derived by applying different tracking thresholds τ, the poses of these noisy test views have been estimated.

4 Results

The illustrations in figure 4 indicate that pose estimation becomes more precise with an increasing number of sample views in the object representation. This result has been expected and is confirmed by an inspection of the mean estimation errors taken over the 30 test views for each object and each partitioning of the hemisphere separately. They are summarized in table 1. The mean errors are decreasing with an increasing value of τ, i.e., with an increasing number of views in the object representation. For individual samples the proposed estimation method is capable to provide errors less than 4° even if the used object representation contains only few sample views.

5 Discussion

Figure 3-3) can be regarded as visualization of the amount of information on the object's structure in a test image. It reveals a small amount of information on details of the object. Only this information is available in the representation of a degraded image and thus can be utilized by our algorithm. In view of this fact results such as that for object "dwarf" depicted in figure 4 are remarkable, because in this example the object representation contains 45 views only. Also the mean estimation errors provide a satisfying result, especially for object "dwarf", for which the mean errors are less than 15° for a reasonable partitioning of the viewing hemisphere ($\tau = 0.85$), taking into account that humans are hardly able to recognize a difference of 10° between two object poses from non-degraded images. In general, the quality of the pose estimation from test views degraded by Gaussian noise can be regarded as fairly good. It supports a good quality of our sparse object representation and allows the conclusion that the view-based approach to object perception with object representations that consist of only single but connected views is suitable for performing perception task.

The sparse object representations can be used for data compression and they can be applied to object recognition even under degraded conditions.

References

[1] S. Edelman and H. H. Bülthoff. Orientation Dependence in the Recognition of Familiar and Novel Views of Three-Dimensional Objects. *Vision Research*, 32(12):2385–2400, 1992.

[2] M. J. Tarr. *Orientation Dependence in Three-Dimensional Object Recognition*. Ph.D. Thesis, MIT, 1989.

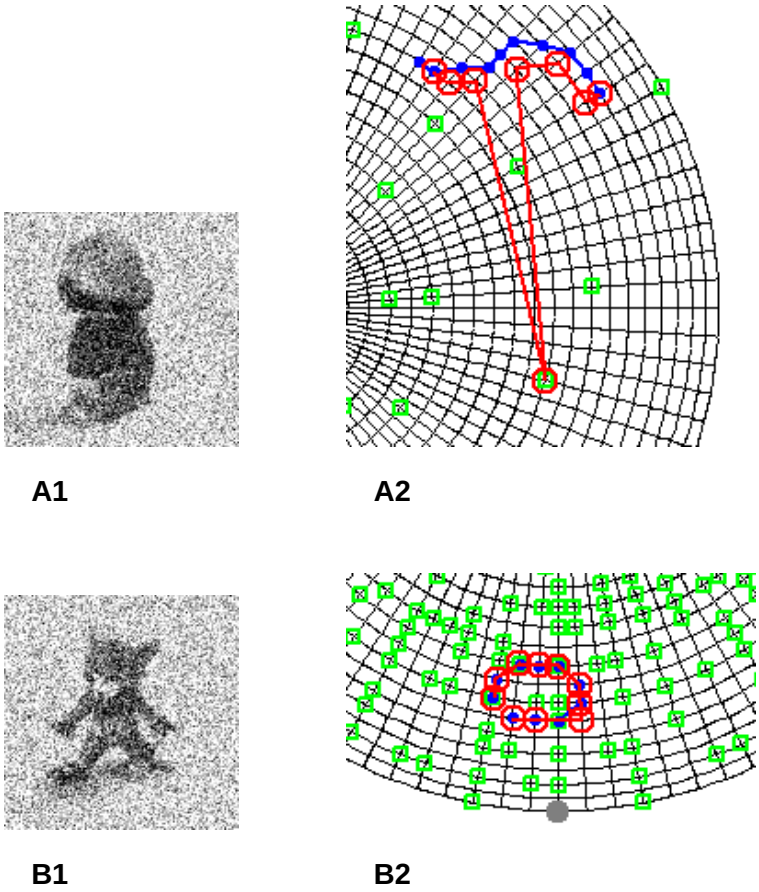

Figure 4. Noisy Test Views and their Estimated Poses. A) object "dwarf", $\tau = 0.8$, 45 views in object representation, mean estimation error $9.39°$. B) object "Tom", $\tau = 0.95$, 1445 views in object representation, mean estimation error $0.36°$. 1) one sample of the noisy test views depicted in 2). 2) section of the viewing hemishere with positions of views constituting the sparse object representation (light gray squares), positions of some of the noisy test views (black dots), and their estimated positions (dark gray circles). Each crossing of the grid stands for one of the original, recorded images.

[3] F. Cutzu and S. Edelman. Canonical Views in Object Representation and Recognition. *Vision Research*, 34:3037–3056, 1994.

[4] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and Poggio T. View-Dependent Object Recognition by Monkeys. *Current Biology*, 4:401–414, 1994.

[5] M. Wexler, S. M. Kosslyn, and A. Berthoz. Motor processes in mental rotation. *Cognition*, 68:77–94, 1998.

[6] N. K. Logothetis, J. Pauls, and Poggio T. Shape Representation in the Inferior Temporal Cortex of Monkeys. *Current Biology*, 5(5):552–563, 1995.

[7] D. C. Burr, M. C. Morrone, and D. Spinelli. Evidence for Edge and Bar Detectors in Human Vision. *Vision Research*, 29(4):419–431, 1989.

[8] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[9] C. Eckes and J. C. Vorbrüggen. Combining Data-Driven and Model-Based Cues for Segmentation of Video Sequences. In *Proc. WCNN96*, pages 868–875, San Diego, CA, USA, 1996.

[10] T. Maurer and C. von der Malsburg. Tracking and Learning Graphs and Pose on Image Sequences of Faces. In *Proc. Int. Conf. on Automatic Face- and Gesture- Recognition*, pages 176–181, Killington, Vermont, USA, 1996.

[11] V. Chvatal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[12] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Trans. Comp.*, 42:300–311, 1993.

mean estimation errors, noisy images					
τ	0.75	0.8	0.85	0.9	0.95
object “Tom”	$64.42°$	$40.02°$	$34.69°$	$29.64°$	$11.28°$
object “dwarf”	$41.79°$	$30.98°$	$13.48°$	$3.9°$	$2.09°$

Table 1. Mean estimation errors, noisy images.

[13] G. Peters and C. von der Malsburg. View Reconstruction by Linear Combination of Sample Views. In *Proc. BMVC 2001*, pages 223–232, Manchester, UK, 2001.

[14] M. Pötzsch. Die Behandlung der Wavelet-Transformation von Bildern in der Nähe von Objektkanten. Technical Report IRINI 94-04, Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany, 1994.