in: *Proceedings of the International NAISO Congress on Information Science Innovations (ISI'2001)*, edited by Milad Fares Sebaaly, pp. 800-805, ISBN 3-906454-25-8, ICSC Academic Press, Dubai, U.A.E., March 17-21, 2001.

Interpolation of Novel Object Views from Sample Views

Gabriele Peters **Christoph von der Malsburg**
Institut für Neuroinformatik
Ruhr-Universität Bochum
Universitätsstr. 150
D-44780 Bochum, Germany
Gabi.Peters@neuroinformatik.ruhr-uni-bochum.de

Abstract

In this article we address the problem of three-dimensional object recognition from two-dimensional views. We use a viewer-centered model of object representation and interpolate novel views from stored sample views. The sample views are represented by graphs which are labeled with Gabor wavelet responses as local descriptors of object points. The positions of the object points in a novel view are linear combinations of the corresponding point positions in the sample views, and the novel feature vectors are linear combinations of the Gabor responses in the sample views. From such an interpolated graph we reconstruct the novel view and analyse its quality for different poses of the novel view in relation to the sample views. Within the covered range of about 30° tilt and 40° pan viewing angle between the sample views we obtain good interpolation qualities. This leads to a number of about 36 views which is sufficient to represent the upper viewing hemisphere of an arbitrary object. Our results are consistent with current findings of biological and psychological research. In addition, the idea of the proposed algorithm is suitable to be applied in data compression.

1 Introduction

In the study of three-dimensional object recognition the ability of humans to recognize objects from unfamiliar views is one major subject of investigation. From a theoretical point of view, it would be possible to establish a three-dimensional, *object-centered* representation from some experienced views. To recognize novel views this representation can be aligned and projected to a hypothesized viewpoint, and matched with the input image. This approach is known as *recognition by alignment* or *mental rotation* and has been proposed early, e.g., by Shepard and Cooper [1], Lowe [2], and Ullman [3]. In recent years, however, two-dimensional, *viewer-centered* models of object representations have gained acceptance, because of psychophysical and neurophysiological findings which support these models (for a survey see [4]). One of the most supported idea is the *interpolation* of unfamiliar views. Novel views can be generalized from stored views by view approximation as described, e.g., by Poggio and Edelman [5]. According to this theory humans and other primates can achieve viewpoint-invariant recognition of objects by a system that interpolates between a small number of stored sample views, which have been experienced previously, as shown by Bülthoff and Edelman [6] in psychophysical studies with humans.

A question frequently discussed concerns the distribution of the stored sample views, i.e., their number and distances. In a study with monkeys by Logothetis et. al. [7] 10 views were found to be sufficient to achieve view-independent performance for the entire viewing sphere. But the same study claims that the number of required viewpoints may depend on the object class. The inability of monkeys to recognize objects rotated by more than approximately 40° from a familiar view is also reported. A similar result for human object recognition was published earlier by Rock and DiVita [8]. Humans become very poor in recognizing wire-like objects for view distancies larger than approximately 30°.

Inspired by these results from biological research we established a computer vision system which simulates a viewer-centered object recognition system. Unfamiliar views can be reconstructed via view interpolation. In this article we are concentrating on the quality of the reconstructed, novel view for different poses of the novel view in relation to three sample views.

2 Methods

2.1 Representation of an Object

We recorded views of small toy objects in distances of 3.6° in both, longitude and latitude direction on the upper hemisphere of the object, resulting in 2500 views per object (see figure 1). Each view is represented by a graph, which covers the object in the image. The vertices of a graph are labeled with Gabor wavelet responses, which describe the local surroundings of the vertex in the image (for an example see the graph for start view (2, 8) in figure 2). For the Gabor transform we use a set of wavelets with 8 directions and 4 frequencies, thus for each vertex we obtain a vector with 32 complex entries, which is called *jet*. The graphs are generated automatically from the images: first, the object is separated from the background by a segmentation algorithm described in [9], which is based on the gray level values of the image. Then a grid graph is put on the resulting object segment.

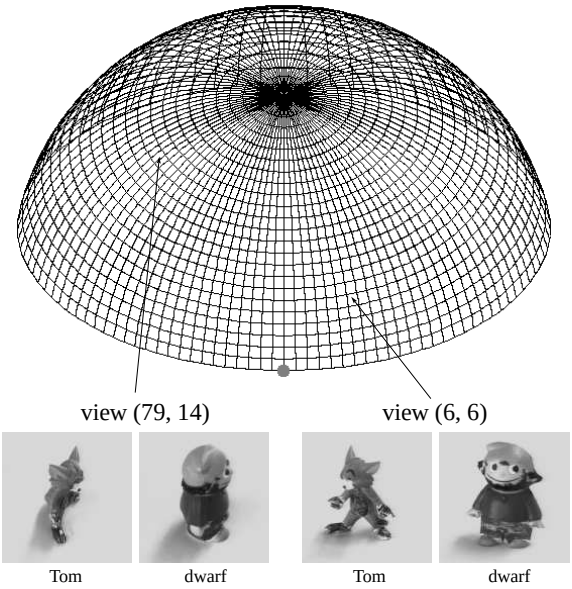

Figure 1: Viewing Hemisphere with Two Example Views of Two Objects. Each crossing of the grid stands for one view, thus the hemisphere consists of 100×25 views. The dot in front marks view (0, 0).

2.2 Tracking of Object Features

For the calculation of the position of an object point in a novel view from the positions of the point in three sample views it is necessary to have the exact correspondences of the point in the sample views. It has been shown that tracking of object features is excellently suitable for providing precise correspondences, especially in comparison with graph matching [10]. We use a tracking procedure which provides subpixel precision. It is described in [11]. Given a sequence of a moving object and the pixel position of a landmark of the object in frame n, the aim is to find the corresponding position of the landmark in frame $n + 1$. We extract the Gabor jets, described in subsection 2.1, at the same pixel positions in the frames n and $n + 1$. From a similarity function between two jets we calculate the displacement vector between them, and thus we gain the new position of the landmark in frame $n + 1$. This process can be iterated to refine the position. For each vertex of the graph of frame n the displacements are calculated for frame $n + 1$. Then a graph is created with its vertices at the new corresponding positions in frame $n + 1$, and the Gabor responses for the new vertices are extracted at the new positions.

For each recorded view of an object we track its representing graph towards the right and upwards on the hemisphere (see figure 2). We stop tracking once the similarity between the tracked graph of the current right (resp. upper) view and the original graph of the start view drops below a preset threshold. This is the case when vertices start to provide poor correspondences.

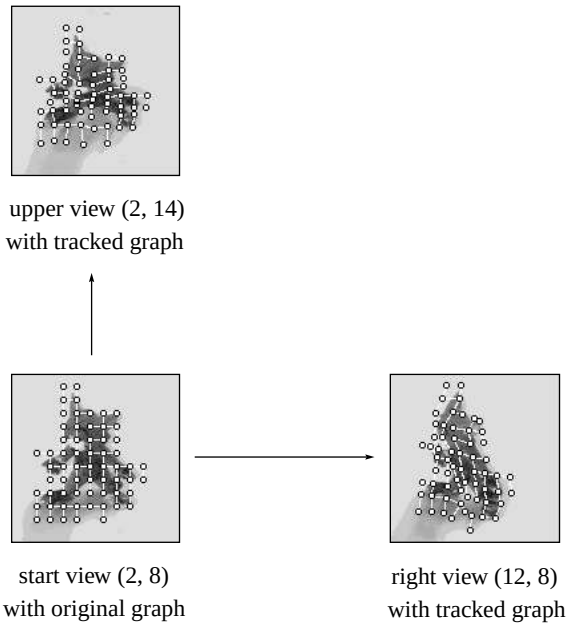

Figure 2: Example of Tracked Graphs. The graph which represents view (2, 8) is tracked to the right view (12, 8) and to the upper view (2, 14).

2.3 View Interpolation from Three Sample Views

By tracking the original graph of a start view towards the right and upwards a rectangular area is defined on the hemisphere (see figure 3). We use the start,

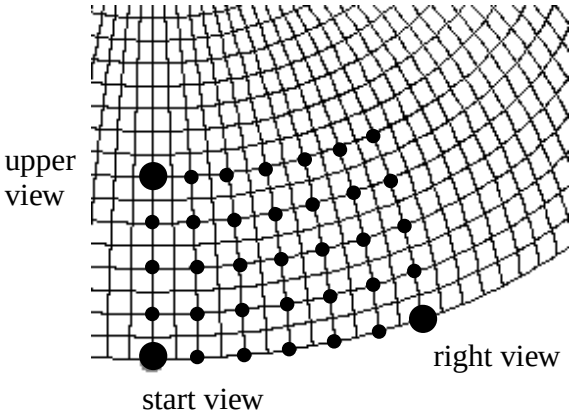

Figure 3: Example of an Interpolation Matrix on the Hemisphere. Views marked by a dot are interpolated from the start, the right, and the upper view.

the right, and the upper view as samples and interpolate the graphs of every second view which lies inside the rectangle. Our rectangles span a region of 43.2° towards the right (pan angle of right view minus pan angle of start view) and 28.8° upwards (tilt angle of upper view minus tilt angle of start view) for each view on the hemisphere. We sample the views to be interpolated in steps of 7.2° in both directions, thus we obtain a 5×7 interpolation matrix for each start view.

For creating a representation of a novel view in form of a graph we have to calculate the jets of the graph vertices, on the one hand, and their positions, on the other hand. A jet of a vertex of the interpolated graph is a weighted sum of the jets of the corresponding vertices in the sample views. The weights are calculated according to the relative position of the novel view with respect to the sample views. Detailed formula are given in [12].

The position $(\hat{x}, \hat{y})$ of a vertex in the novel view is calculated from an extended stereo algorithm. The positions $(x_i, y_i), i = 1, 2, 3$, of the corresponding vertices in the sample views are projected orthogonally to the hypothesized object point, which is approximated by the point of intersection of the projection rays. This point is projected orthogonally to the novel view.

We have proved that this procedure is equivalent to Ullman and Basri's [13] approach of generating novel views by linear combinations of sample models. In this context our new coordinates $(\hat{x}, \hat{y})$ of a vertex can be expressed as linear combination of its coordinates $(x_i, y_i), i = 1, \ldots, n, n = 2$ or $n = 3$, in the sample views, if the pan angle between start and right view is not 180° and if the start view is not situated in the north pole of the hemisphere. Because both conditions hold true for our simulations, our interpolated views can always be expressed as linear combinations of sample views. This means that our approach is purely two-dimensional, view-based.

If the coordinates in the sample views are denoted by (x_1, y_1) for the start view, (x_2, y_2) for the right view, and (x_3, y_3) for the upper view, then following equations hold true for three sample views:

$$\hat{x} = \sum_{i=1}^{2} a_i(\varphi_1, \varphi_2, \hat{\varphi}) \cdot x_i \qquad (1)$$

$$\hat{y} = \sum_{i=1}^{3} b_i(\varphi_1, \varphi_2, \hat{\varphi}, \lambda_1, \lambda_3, \hat{\lambda}) \cdot y_i \qquad (2)$$

with $\varphi_1, \varphi_2, \hat{\varphi}$ denoting pan angles of the start, right, and novel view and $\lambda_1, \lambda_3, \hat{\lambda}$ denoting tilt angles of the start, upper, and novel view, respectively. Thus the x-coordinate $\hat{x}$ is a linear combination of x_1 and x_2 only. Detailed equations for the coefficients are given in [12].

2.4 Evaluating the Interpolation by View Reconstruction

After creating the interpolated graph as described above, we have to assess its quality. For that purpose we reconstruct the novel image from the interpolated graph, on the one hand, and, on the other hand, from its original graph. Using the reconstruction from the original graph as ground truth, we can assess the quality of the interpolation by comparing the reconstruction from the interpolated graph to the ground truth reconstruction (see figure 4).

For the reconstruction we apply an algorithm, described in [14], which reconstructs a gray level image from its Gabor transform. Each jet is locally reconstructed restricted to a Voronoi area around its location. The reconstruction is done utilizing basis functions which depend on the used, linearly independent Gabor wavelets. In this way, the background of the images is reconstructed to zero values.

For the comparison of both reconstructed images we calculate a relative error between them. If Z and $\tilde{Z}$ denote the reconstruction images of the novel view from the original and the interpolated graph, respectively, we regard $\tilde{Z}$ as approximation of Z. The maximum error of approximating a pixel of

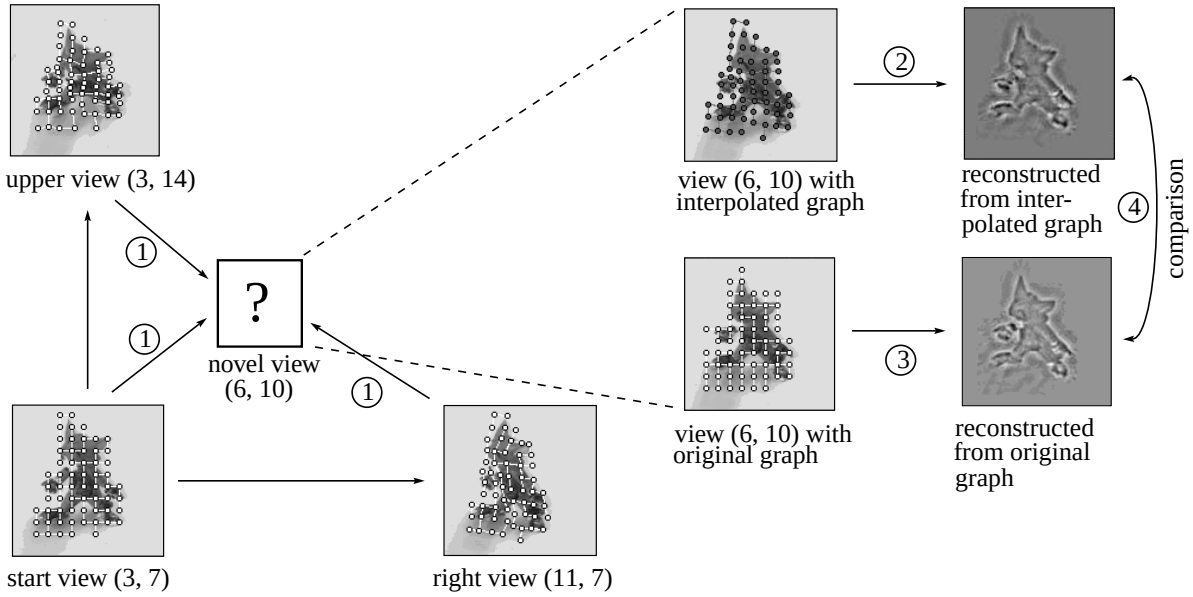

Figure 4: Evaluating the Quality of an Interpolated Graph via View Reconstruction - An Example. Step 1) Calculate the interpolated graph of the novel view (6, 10) from three sample views as described in subsection 2.3. Step 2) Reconstruct the novel view from the interpolated graph. Step 3) Reconstruct the novel view from its original grid graph described in subsection 2.1 (ground truth). Step 4) Compare both reconstructed images by calculating the relative error E.

Z by the corresponding pixel of $\tilde{Z}$ is $e_{max} = \max(\max(Z), \max(\tilde{Z})) - \min(\min(Z), \min(\tilde{Z}))$. Now we can calculate an error E between Z and $\tilde{Z}$ relative to e_{max}:

$$E := \frac{1}{N} \cdot \frac{1}{e_{max}} \cdot \sum_{i=1}^{N} |z_i - \tilde{z}_i| \qquad (3)$$

where z_i and $\tilde{z}_i, i = 1, \ldots, N$, are the pixels of Z and $\tilde{Z}$ with not both $z_i = 0$ and $\tilde{z}_i = 0$, to exclude the background from the calculations.[1]

E is a measure for the quality of our interpolated graph. It has been shown in many studies, e.g., in [15, 16], that a representation in form of a graph labeled with Gabor wavelet responses, like our original graph, can be used for a robust object recognition. We claim that a small relative error E justifies the assumption of comparable recognition capabilities of the interpolated graph.

2.5 Statistics

We use two different objects for our simulations (see figure 1), and we interpolate 5×7 novel views for each view on the hemisphere with the current view as start view, according to the scheme depicted in figure 3. For each interpolated and reconstructed view we calculate the relative error E, thus we obtain a 5×7 error matrix for each view on the hemisphere.[2]

From all entries of all error matrices we calculate the mean error, the standard deviation, and the maximum error which occurs.

3 Results

For both objects the mean value of the relative errors E in the covered range of $43.2°$ pan and $28.8°$ tilt angle is rather small (about 5%) with a small standard deviation (about 0.01). The maximum relative error which occures is about 11%. Thus we can assume, that inside this range an interpolated graph is suitable for recognition. The exact values are summarized in table 1.

4 Summary and Conclusions

We have shown that only some views of a three-dimensional object are sufficient to represent the whole object. Unfamiliar views can be interpolated from stored views. Within our tested range of about

[1] To be robust against slight translations during reconstruction we shift Z and $\tilde{Z}$ against each other in a small range and calculate E for all shifting positons. The final E is the minimum over all shifting positions.

[2] If the range which is covered by the tracked graphs is smaller than the region which is spanned by the interpolation matrix the corresponding values in the error matrices remain undefined.

	object "Tom"	object "dwarf"
mean error	0.0489	0.0522
standard deviation	0.0120	0.0110
maximum error	0.0908	0.1132

Table 1: Results of Interpolation from Three Sample Views. The set of values, from which these results are calculated, consists of one relative error for each entry of an interpolation matrix for each view on the hemisphere.

30° tilt angle and 40° pan angle between sample views the quality of the reconstructed views is sufficiently high to presume good recognition rates of the object from interpolated views. From our results we can estimate the total number of views needed to represent the upper viewing hemisphere of an object. If a mean interpolation error of about 5% and a maximum interpolation error of about 11% are tolerable, about 36 sample views are sufficient.

The range of good interpolation is probably not limited by the quality of the interpolation, rather it seems to be restricted by the tracking capabilities, which depend on the complexity of the objects. We suppose that the range of good interpolation could be much larger, if the correspondences between sample views are given.

The representation of a novel view can be derived from three sample views by a linear combination of the position of object points in the sample views and a linear combination of the corresponding feature vectors, which describe the surroundings of the points.

Our results are consistent with biological and psychological research. In the introduction the good recognition performance of humans and monkeys within a range of about 30° or 40° between sample views has already been mentioned. Although our main motivation is to learn how the brain performs three-dimensional object recognition, our results may be also relevant to technical applications. If only some views of a three-dimensional object need to be stored to represent it, our algorithms can be used for data compression.

Acknowledgements

We thank ONR, grant No. N 00014-98-1-0242, and ARO, grant No. DAAG55-98-1-0293, for support of this work. In addition, we thank Dr. Rolf P. Würtz for fruitful discussions and Pervez Mirza for proofreading this article.

References

[1] R. N. Shepard and L. A. Cooper. *Mental Images and their Transformations*. MIT Press, Cambridge, MA, 1982.

[2] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Boston, MA, 1985.

[3] S. Ullman. Aligning Pictorial Descriptions: An Approach to Object Recognition. *Cognition*, 32:193–254, 1989.

[4] G. Peters. Theories of Three-Dimensional Object Perception - A Survey. accepted for: *Recent Research Developments in Pattern Recognition*. Transworld Research Network, 2000.

[5] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional Objects. *Nature*, 343:263–266, 1990.

[6] H. H. Bülthoff and S. Edelman. Psychophysical Support for a Two-Dimensional View Interpolation Theory of Object Recognition. In *Proceedings of the National Academy of Science of the United States of America*, volume 89, pages 60–64, 1992.

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

[8] I. Rock and J. DiVita. A Case of Viewer-Centered Object Perception. *Cogitive Psychology*, 19:280–293, 1987.

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

[10] G. Peters, B. Zitova, and C. v. d. Malsburg. Two Methods for Comparing Different Views of the Same Object. In *Proceedings of the 10th British Machine Vision Conference (BMVC'99)*, pages 493–502, 1999.

[11] T. Maurer and C. v. d. Malsburg. Tracking and Learning Graphs and Pose on Image Sequences of Faces. In *Proceedings of the 2nd International Conference on Automatic Face- and Gesture- Recognition*, pages 176–181, 1996.

[12] G. Peters and C. v. d. Malsburg. View Reconstruction by Linear Combination of Sample

Views. submitted to: *Third International Conference on 3D Digital Imaging and Modeling (3DIM2001)*.

[13] S. Ullman and R. Basri. Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.

[14] M. Pötzsch, T. Maurer, L. Wiskott, and C. v. d. Malsburg. Reconstruction from Graphs Labeled with Responses of Gabor Filters. In *Proceedings of the ICANN 1996*, pages 845–850, 1996.

[15] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R. P. Würtz, and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.

[16] L. Wiskott, J.-M. Fellous, N. Krüger, and C. v. d. Malsburg. Face Recognition and Gender Determination. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 92–97, 1995.