in: Proceedings of the 12th British Machine Vision Conference (BMVC:2001),
Vol. 1, pp. 223-232, Manchester, UK, 2001.

View Reconstruction by Linear Combination of 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
http://www.neuroinformatik.ruhr-uni-bochum.de/ini/PEOPLE/gpeters/

Abstract

Ullman and Basri [1] have shown theoretically, that a three-dimensional object can be represented by a linear combination of two-dimensional images of the object. But they have applied their calculations to artificially created images only, like line drawings of cars. The application to images of real objects turns out to be difficult, because a crucial point of their algorithm is the knowlegde of correspondences in the sample views. In this article we describe a biologically inspired system which automatically provides correspondences between views of a three-dimensional object. This enables us to apply Ullman and Basri's linear combination approach to images of arbitrary, real objects. We give detailed formula of our linear combinations and examples for reconstructed object views.

1 Introduction

In three-dimensional object recognition two-dimensional, *viewer-centered* object representations gained acceptance in recent years, because of psychophysical and neurophysiological findings which support these models (for a survey see [2]). According to these models humans and other primates can achieve viewpoint-invariant recognition of objects by a system that can generalize unfamiliar views from a small number of stored sample views, which have been experienced previously. One well-known approach has been proposed by Ullman and Basri [1] who proved the theoretical possibility of novel view creation by linear combinations of sample views. As the application of their calculations to images of real objects bears some difficulities, they applied their algorithm to artificially created images only, like line drawings of cars. Difficulties derive from the facts that the nature of the used features has to be considered, correspondences in sample views have to be known, and singularities have to be avoided by an appropriate choice of sample views.

In the next sections we describe a biologically inspired system, which is able to deal with these difficulties and which applies the linear combination approach to images of real objects. We represent a single view of an arbitrary, three-dimensional object by a

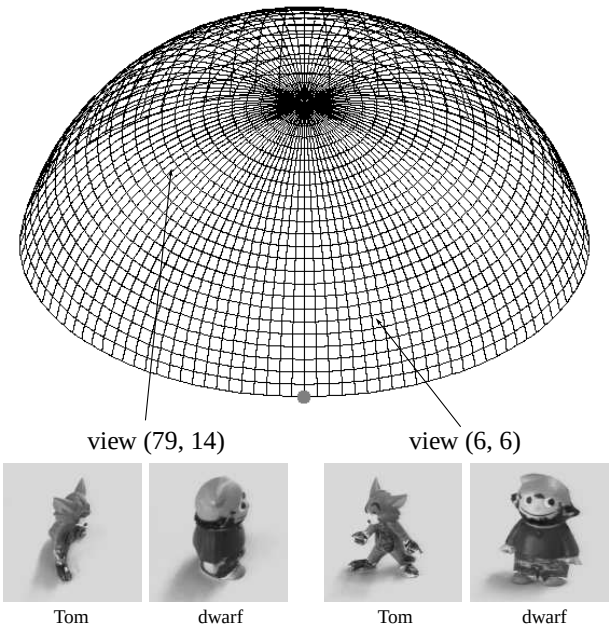

Figure 1: Viewing hemisphere with two sample views of two objects. Each crossing of the grid stands for one view, thus the hemisphere consists of 100×25 views. The dot in front marks view (0, 0).

graph which is labeled with Gabor wavelet responses described in section 2. Corresponding points in sample views are obtained from a tracking algorithm, described in section 3. Given some graphs which represent single sample views of the object, we calculate an *interpolated* graph of an unfamiliar view by calculating its vertex *positions* as linear combinations of the vertex positions in the sample views (described in section 4), and by interpolating its vertex *features* as weighted sums of the Gabor features in the sample views (described in section 5). To evaluate the quality of the interpolated graph, we reconstruct a novel view from it and compare the resulting virtual view with one that is reconstructed from a directly recorded graph (described in section 6). In addition, we are able to generate a morphed, unfamiliar view directly from one original sample view and the vertex *positions* of the interpolated graph, i.e., without utilizing interpolated features (described in section 7).

2 Representation of Single Object Views

We recorded views of small toy objects at increments of 3.6° in both longitude and latitude on the upper viewing sphere, resulting in 2500 views per object (see figure 1). Each view of an object 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. The vector is called a *jet*. The graphs are generated automatically from the images: first, the object is separated from the background by a segmentation algorithm described in [3] based on the gray level values of the image. Then a grid graph is put on the resulting object segment.

3 Tracking of Object Features

For the calculation of the position of an object point in an unfamiliar view from the positions of the point in a small number of neighbouring sample views it is necessary to have the exact correspondences of the point between the sample views. It has been shown that the *tracking* of object features over a dense sequence of intervening images gives much more precise correspondences than graph matching between the images directly. This is also supported by human object recognition, where the temporal context which is provided by tracking seems to be important to perceive three-dimensional form (see [4]).

We use a tracking procedure which provides subpixel precision. It is described in [5]. Given a sequence of a moving object and the pixel position of an object point in frame n, the aim is to find the corresponding position of the point in frame $n+1$. We extract a Gabor jet at the same pixel position in the frames n and $n+1$. From a similarity function between the two jets we can calculate the displacement vector between them, and thus estimate the new position of the object point 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 can track its representing graph to the right and above it on the hemisphere (see figure 2). We stop tracking if 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.

4 Linear Combination of Positions of Object Points

In this section we describe the calculation of the position of an object point in a novel view from the corresponding positions of the point in two and three sample views. In the case of interpolation of novel views from *two* sample views we use the start view and the right view as samples (see figure 2), in the case of *three* sample views we use the start, the right and the upper view. For our simulations we use a projection algorithm, which is proved to be equivalent to Ullman and Basri's linear combination approach. In the following our calculations are described in their context.

Let $\varphi_1, \varphi_2, \varphi_3$, and $\hat{\varphi}$ denote the pan angles and $\lambda_1, \lambda_2, \lambda_3$, and $\hat{\lambda}$ denote the tilt angles of the start, right, upper, and novel view, respectively (see figure 3). The angles are given, as well as the coordinates of the object point in the sample views. They are provided by the tracking procedure and 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. The coordinates $(\hat{x}, \hat{y})$ of the object point in the unfamiliar view are to be found.

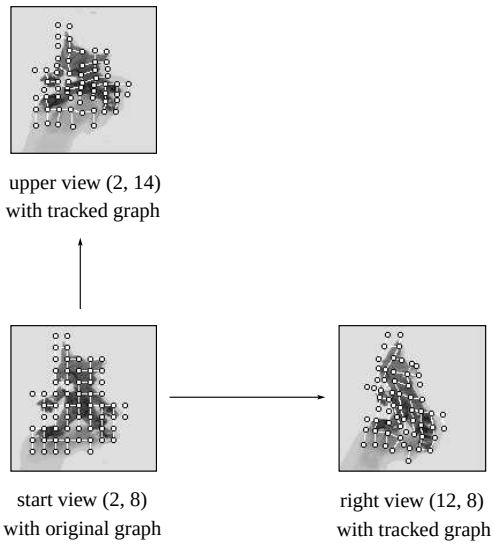

Figure 2: Example of tracked graphs. The graph which represents view (2, 8) is tracked to the view (12, 8) at the right and to the view (2, 14) above it.

4.1 Three Sample Views

If φ_1 and φ_2 are not equal and not $180°$ apart, the x-coordinate $\hat{x}$ is a linear combination of x_1 and x_2 only. The coefficients of this linear combination are simple functions in φ_1, φ_2, and $\hat{\varphi}$. In detail: $\hat{x} = \sum_{i=1}^{2} a_i x_i$ with

$$a_1 = -csc(\varphi_1 - \varphi_2) \cdot sin(\varphi_2 - \hat{\varphi}), \quad (1)$$

$$a_2 = csc(\varphi_1 - \varphi_2) \cdot sin(\varphi_1 - \hat{\varphi}). \quad (2)$$

The y-coordinate $\hat{y}$ is a linear combination of y_1, y_2, and y_3, if, again, φ_1 and φ_2 are not equal and not $180°$ apart, if λ_1 and λ_3 are not equal, and if λ_1 is not zero (i.e. if the start view is not positioned on the equator). All of these conditions can easily be met by the choice of the positions of the sample views in relation to each other. The coefficients of this linear combination are more complex, and they depend on $\varphi_1, \varphi_2, \hat{\varphi}, \lambda_1, \lambda_3$, and $\hat{\lambda}$. In detail: $\hat{y} = \sum_{i=1}^{3} b_i y_i$ with

$$\begin{aligned} b_1 = {} & csc(\lambda_1 - \lambda_3) \cdot csc(\varphi_1 - \varphi_2) \cdot \\ & \Big[cos(\hat{\varphi}) \cdot sin(\hat{\lambda}) \cdot \Big(cos(\lambda_3) \cdot sin(\varphi_2) - cot(\lambda_1) \cdot sin(\lambda_3) \cdot sin(\varphi_1)\Big) + \\ & sin(\hat{\lambda}) \cdot sin(\hat{\varphi}) \cdot \Big(cos(\lambda_3) \cdot cos(\varphi_2) - cos(\varphi_1) \cdot cot(\lambda_1) \cdot sin(\lambda_3)\Big) - \\ & cos(\hat{\lambda}) \cdot sin(\lambda_3) \cdot sin(\varphi_1 - \varphi_2)\Big], \end{aligned} \quad (3)$$

$$b_2 = csc(\lambda_1) \cdot csc(\varphi_2 - \varphi_1) \cdot sin(\hat{\lambda}) \cdot sin(\varphi_1 + \hat{\varphi}), \quad (4)$$

$$\begin{aligned} b_3 = {} & csc(\lambda_1 - \lambda_3) \cdot csc(\varphi_1 - \varphi_2) \cdot \\ & \Big(cos(\hat{\lambda}) \cdot sin(\lambda_1) \cdot sin(\varphi_1 - \varphi_2) + \end{aligned}$$

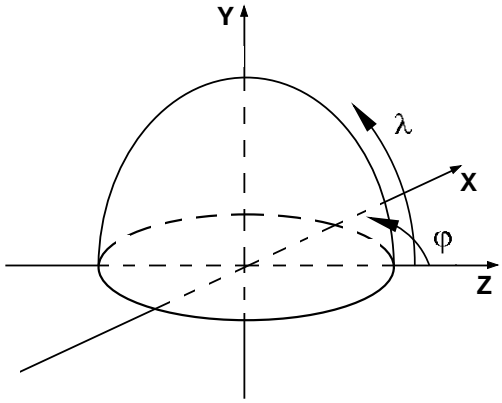

Figure 3: Pan and tilt angles. φ and λ are measured for all views as depicted in this scheme: $0^\circ \leq \varphi < 360^\circ$ and $0^\circ \leq \lambda \leq 90^\circ$.

$$cos(\lambda_1) \cdot sin(\hat{\lambda}) \cdot \Big(sin(\varphi_1 + \hat{\varphi}) \; - \; sin(\varphi_2 + \hat{\varphi}) \Big) . \tag{5}$$

4.2 Two Sample Views

For two sample views the linear combination for the x-coordinate $\hat{x}$ is the same as for three sample views.

The y-coordinate $\hat{y}$ is a linear combination of x_1, x_2, and y_1, if φ_1 and φ_2 are not equal and not 180° apart and if λ_1 is not $\frac{\pi}{2}$ (i.e. if the start view is not positioned in the north pole of the viewing hemisphere). Two of the coefficients depend on $\varphi_1, \varphi_2, \hat{\varphi}, \lambda_1$, and $\hat{\lambda}$. The third of them depends on λ_1 and $\hat{\lambda}$ only. In detail: $\hat{y} = \sum_{i=1}^{2} b_i \cdot x_i + b_3 \cdot y_1$ with

$$b_1 \;=\; cos(\varphi_2 - \hat{\varphi}) \cdot csc(\varphi_1 - \varphi_2) \cdot sin(\hat{\lambda}) \; - \; cos(\hat{\lambda}) \cdot cot(\varphi_1 - \varphi_2) \cdot tan(\lambda_1), \tag{6}$$

$$b_2 \;=\; csc(\varphi_1 - \varphi_2) \cdot \Big(cos(\hat{\lambda}) \cdot tan(\lambda_1) \; - \; cos(\varphi_1 - \hat{\varphi}) \cdot sin(\hat{\lambda}) \Big) , \tag{7}$$

$$b_3 \;=\; cos(\hat{\lambda}) \cdot sec(\lambda_1). \tag{8}$$

5 Linear Combination of Features

Besides calculating the vertex positions of the interpolated graph as linear combination, the features of the vertices have to be adapted to the interpolated positions, too. We determine a jet of a vertex of the interpolated graph as a weighted sum of the jets of the corresponding vertices in the sample views. The weights of the sum depend on the relative position of the novel view with respect to the sample views. A smaller distance between the novel view and a sample view leads to a stronger weight of the jets of the sample view than a larger distance.

Let (x_i, y_i) be the position of the i-th sample view, (m, n) the position of the novel view, and $d_i := d\left(\binom{m}{n}, \binom{x_i}{y_i} \right)$ the Euclidean metric for $i = 1, \ldots, N$ with $N =$ number of sample views (see figure 4).

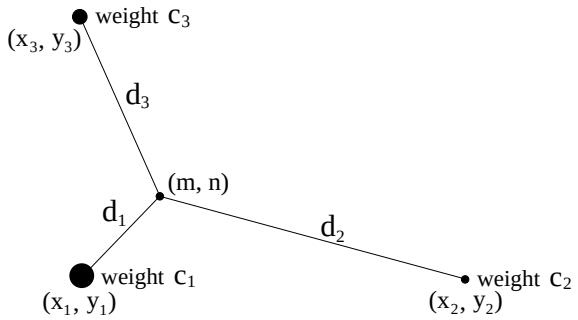

Figure 4: Weighting of feature vectors for three sample views. In this example c_1 is the strongest weight, because the novel view (m, n) is closer to the sample view (x_1, y_1) than to the sample views (x_2, y_2) and (x_3, y_3).

Let $\mathcal{J}(x, y)$ be a jet of view (x, y). Then following equation should hold true

$$\mathcal{J}(m, n) = \sum_{i=1}^{N} c_i \mathcal{J}(x_i, y_i) \tag{9}$$

for all jets $\mathcal{J}(m, n)$, $\mathcal{J}(x_i, y_i)$ of corresponding vertices, $i = 1, \ldots, N$ with $\sum_{i=1}^{N} c_i = 1$.

5.1 Three Sample Views

In the case of three sample views the weights c_i are calculated according to following three equations

$$c_1 = d_2 d_3 / (d_1 d_2 + d_1 d_3 + d_2 d_3), \tag{10}$$

$$c_2 = d_1 d_3 / (d_1 d_2 + d_1 d_3 + d_2 d_3), \tag{11}$$

$$c_3 = d_1 d_2 / (d_1 d_2 + d_1 d_3 + d_2 d_3). \tag{12}$$

5.2 Two Sample Views

For two sample views the weights are $c_1 = d_2/(d_1 + d_2)$ and $c_2 = d_1/(d_1 + d_2)$.

6 Evaluating the Quality of an Interpolated Graph

Now we can assess the quality of an interpolated graph. For that purpose we reconstruct the novel image from its original graph on the one hand, and from the interpolated graph on the other. Using the reconstruction from the original graph derived directly from an image 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 5). We apply the algorithm described in [6] to reconstruct a gray level image from its Gabor transform. For the comparison of both reconstructed images we calculate a relative error E 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.

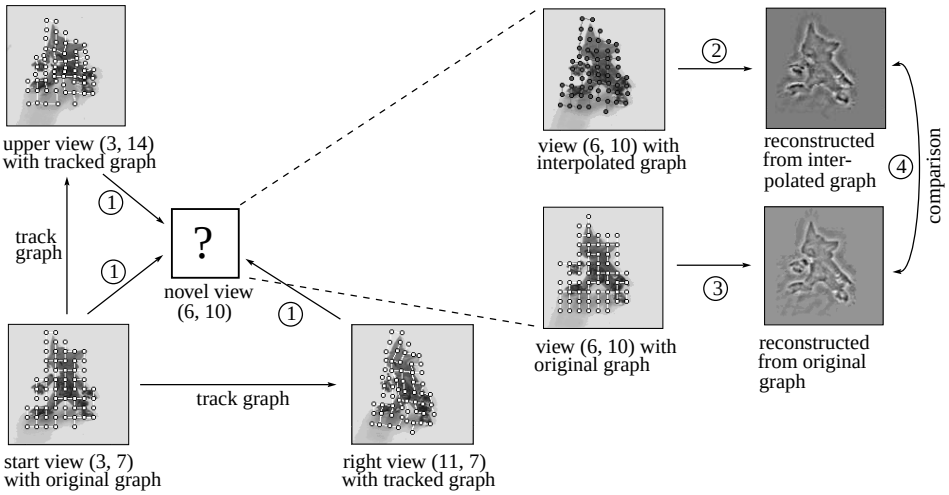

Figure 5: Evaluation of an interpolated graph. Step 1) Calculate the interpolated graph of the novel view (6, 10) from three sample views as described in sections 4 and 5. Step 2) Reconstruct the novel view from the interpolated graph. The upper image on the right is the resulting virtual view. Step 3) Reconstruct the novel view from its original graph described in section 2 (ground truth). Step 4) Compare both reconstructed images by calculating the relative error E.

The maximum error of approximating a pixel of 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 the error E between Z and $\tilde{Z}$ relative to e_{max}:

$$E := \frac{1}{N} \cdot \frac{1}{e_{max}} \cdot \sum_{i=1}^{N} |z_i - \tilde{z}_i| \tag{13}$$

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] We evaluate the quality of the interpolated graphs by calculating the relative errors for a large set of reconstructed images depending on the distances between the sample views. For that purpose we partition the viewing hemisphere of an object into areas of similar views, called *view bubbles* [4]. They are determined by tracking graphs along varying viewpoints as described in section 3. This results in a partitioning of the viewing hemisphere into a smaller or larger number of view bubbles, depending on the used tracking threshold. Figure 6 shows five partitionings for object "Tom". We use border views of a view bubble as sample views to reconstruct views inside the view bubble. This is performed for each view bubble of each of five different partitionings of the viewing hemisphere (depending on five different tracking thresholds) resulting in a set of reconstruction errors for each partitioning. Thus, we can specify a maximum and a mean reconstruction error for each partitioning, i.e. the relative errors depend on the number of view bubbles. A logarithmic function seems to

[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.

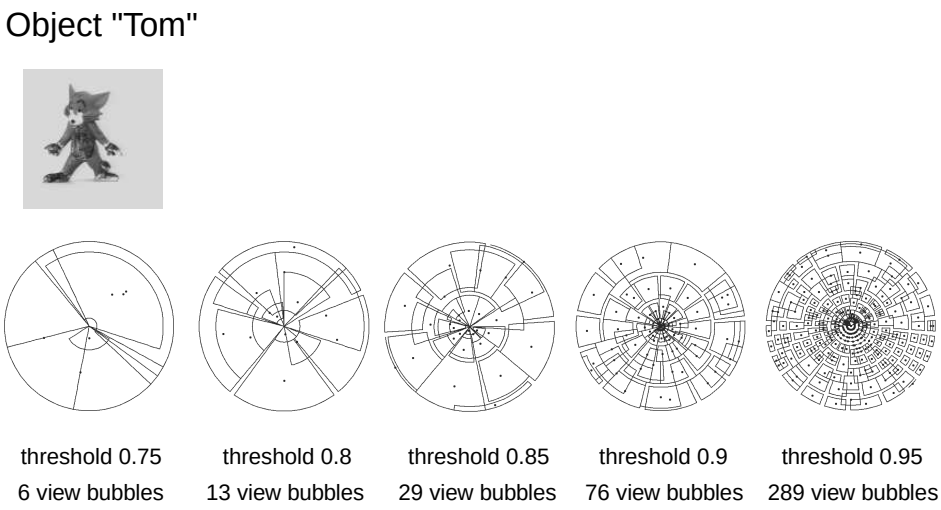

Figure 6: Partitioning of the Viewing Hemisphere for Five Different Tracking Thresholds.

be an appropriate description of the dependence of the mean reconstruction error on the number of view bubbles used to cover the hemisphere (see figure 7).

E is a measure for the quality of our interpolated graph. It has been shown in many studies, e.g., in [7,8], 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 the relative errors E are small enough to justify the assumption of comparable recognition capabilities of an interpolated graph, if the number of sample views is chosen appropriately.

7 Morphing of Unfamiliar Views

By evaluating the quality of an interpolated graph in the way we described in the last section we cannot distinguish the influence of the interpolated positions from the influence of the interpolated features. But it is possible to estimate the quality of the interpolated positions independently. For that purpose we generate a simple, morphed view from one original object view and an interpolated graph. We warp the gray level values of the original view by a triangulation of its original graph vertices and a linear mapping of the resulting triangular patches to the corresponding positions given by the interpolated graph. Figure 8 shows an example for a view morphed from three sample views.

8 Conclusions

The union of Ullman and Basri's theory of linear combination of views on the one hand, and a biologically inspired system which is known for its object representation and recognition capabilities on the other hand leads to a tool which can represent any three-dimensional object by only a few sample views. If the sample images of the object cover the

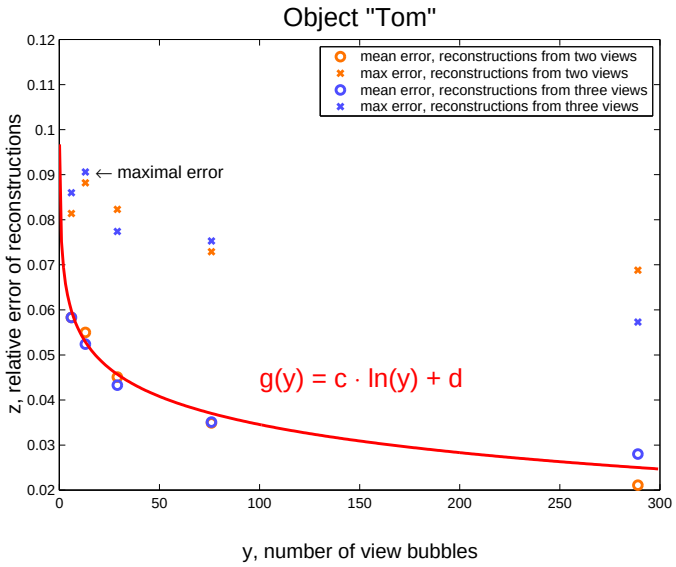

Figure 7: Correlation Between Relative Reconstruction Errors and the Number of View Bubbles. The fitting function for the mean errors of reconstruction has the same parameters $c = -0.009$ and $d = 0.076$ for object "Tom" and the other tested object "dwarf".

viewing sphere in an appropriate way a representation in form of a labeled graph as well as a morphed version of any unfamiliar view can be generated.

Acknowledgements

This work was partially supported by ONR, contract No. N00014-98-1-0242. We thank Pervez Mirza for proof-reading this article.

9 References

[1] S. Ullman and R. Basri, *Recognition by Linear Combinations of Models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 10, pp. 992-1006, 1991.

[2] G. Peters, *Theories of Three-Dimensional Object Perception - A Survey*, Part-I, Vol. 1, pp. 179–197, Transworld Research Network, 2000.

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

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

[5] T. Maurer and C. v. d. Malsburg, *Tracking and Learning Graphs and Pose on Image Sequences of Faces*, Proceedings of the 2nd International Conference on Automatic Face-

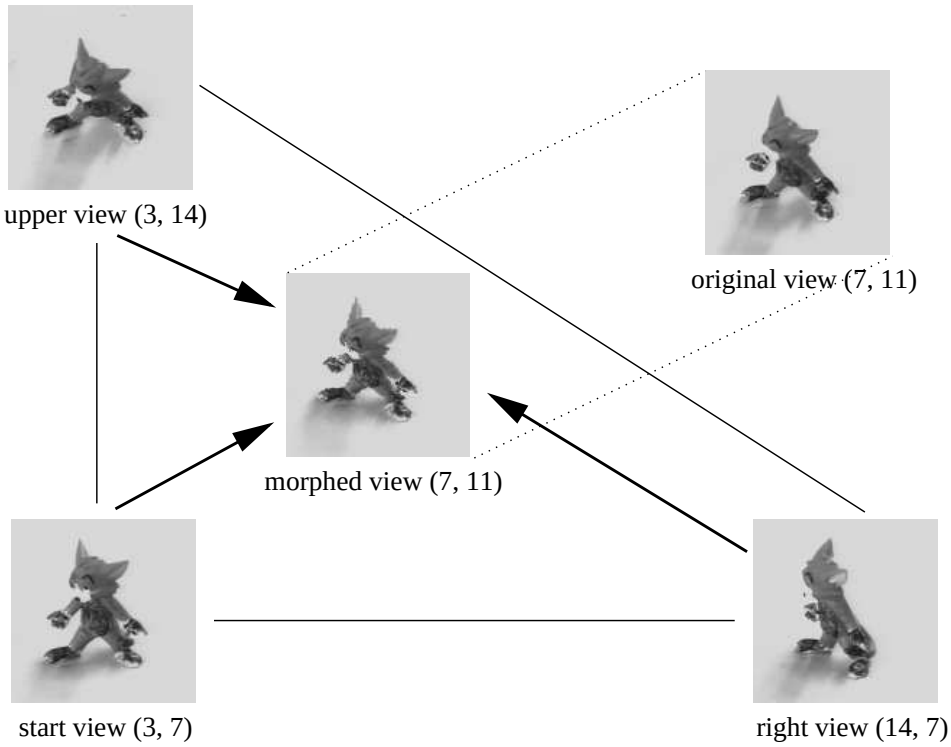

Figure 8: View Morphing. First, the vertex positions of the interpolated graph are calculated from the start, upper, and right view as described in section 4. Then the original start view and the interpolated positions are used to calculate the morphed version of the unfamiliar view. It has to be compared with the original view shown in the upper image on the right. The angle between the start and the right view is 39.6 degrees, between the start and the upper view 25.2 degrees, and between the start and the morphed view about 20.3 degrees.

and Gesture- Recognition, pp. 176–181, Killington, Vermont, USA, 1996.
[6] M. Pötzsch, T. Maurer, L. Wiskott, and C. v. d. Malsburg, *Reconstruction from Graphs Labeled with Responses of Gabor Filters*, Proceedings of the ICANN 1996, pp. 845–850, Bochum, Germany, 1996.
[7] 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, Vol. 42, pp. 300–311, 1993.
[8] L. Wiskott, J.-M. Fellous, N. Krüger, and C. v. d. Malsburg, *Face Recognition and Gender Determination*, International Workshop on Automatic Face- and Gesture-Recognition, pp. 92–97, Zürich, Switzerland, 1995.