ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

A view-based approach for object recognition from image sequences

Andreas Zehender, Pierre Bayerl, Heiko Neumann

University of Ulm, Germany
Department of Neural Information Processing

Abstract. We describe a view-based approach for object recognition that is able to learn and detect object representations from image sequences. We introduce a stability value for individual views by tracking its features over a small range of consecutive views. Based on the stability a set of key frames is extracted to represent a specific object. Consistent with psychophysical findings canonical views of objects, e.g. the sides of a cube, seem to be most stable and therefore suitable for object representation. The second novelty of our approach is to recognize objects from short image sequences by combining the information provided by the temporal varying input data, which is caused by small rotations of the observed objects. We show that the performance of recognition is substantially improved if a sequence of views is processed instead of a single frame. The detected features are integrated over time and stored in a short-term memory biasing the recognition stage via feedback. The feedback mechanism is adapted from a previous model of cortical boundary processing to describe the temporal dynamics of the interaction between the short-term memory and the detection stage.

1 Introduction

One of the major tasks in our every-day life is the recognition of three-dimensional objects. Models, which have been proposed to achieve this task, can be divided into object-centered and view-centered approaches. There is psychophysical and physiological evidence that the human brain uses a view-centered representation of objects. However, most of these approaches try to recognize objects from *single* images [6]. We present an object recognition system, which incorporates temporal information from short image sequences of objects rotating around a vertical axis. The creation of object representations is motivated by a psychophysical experiment in which subjects were allowed to view objects from various directions [1]. Based on computational investigations it turned out that planar views (e.g. the sides of a cube) seem to be most relevant to accurately learn and recognize objects. The existence of such views

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

is due to geometrical singularities in the visual projection of objects, which are invariant for most vantage points [2]. We extended the model proposed by Wallraven and Bülthoff [8] to acquire such a robust representation of objects from image sequences. We combined this approach with a feedback mechanism [4] to integrate the information of consecutive frames during object recognition.

During the learning stage objects are rotated around a vertical axis presenting views from 0° to 360°. The selected set of views (key frames) to represent individual objects is consistent with the psychophysical findings of Humphrey et al. [1]. To test the recognition capabilities we first present single frames under various viewing conditions (feedforward processing). Then we show how the performance is improved when short image sequences are processed using a feedback mechanism [4] to bias the current recognition with previous recognition results stored in short-term memory.

2 Object Processing and Learning Architecture

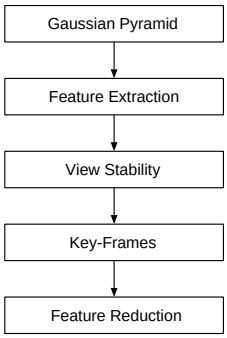

Figure 1: Overview of the learning part

The main task of the learning part is the extraction of stable views of an object. We use computer generated images of three dimensional objects for system evaluation. For each object, a sequence of 36 views is calculated, showing the object rotating in depth around it's vertical axis, one view per step of 10°.

First, a Gaussian pyramid with 3 scales is calculated for each view. Each of the following steps is applied separately to each scale, if not mentioned otherwise. After the stage of feature extraction, the stability of views is calculated by tracking the features in adjacent views. The stability value depends on the number of features that can be successfully tracked over a range of views. Based on these stability values a subset of all views is chosen as key-frames representing the investigated object.

Corners are used as features because of their high significance in the human vision system. Corners are extracted with the structure tensor using image gradients in a neighborhood around each point [7]. A sub-image of 9×9 pixels around each corner is then used as feature.

2.1 View Stability

The features of all views are tracked in both directions over a range of 5 views. The mean value of the percentage of successfully tracked features over all views is denoted as *coverage rate*. Since feature are lost over time during tracking the coverage rate is a monotonically decreasing function $T(\varphi)$ for each direction (see figure 4). Features that could not be tracked in at least one direction are removed from the view.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

Figure 2: The traces of features when rotating a textured cube

Features are tracked in adjacent views by an algorithm described in [8]. This paper uses a matching algorithm for corner features in gray-scale images based on the singular value decomposition (SVD) [5] (see figure 2 for an example).

If the tracking rates are calculated for a view A in both directions, the view stability is calculated as $S_A = \sum_{\varphi=-5}^{5} T_A(\varphi)$ by integrating the coverage rates. Figure 3 shows the stability values for a textured cube.

Views whose features can be tracked better or longer in adjacent views get a higher stability value than views whose features are lost after a small rotation (edge views).

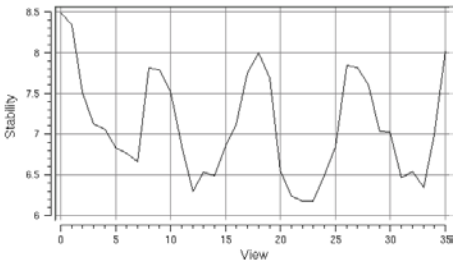

Figure 3: Stability values for a textured cube. Views 0, 9, 18 and 27 are the planar views

Computational results show that the stability functions for a three-sided prism, a cube and a five-sided prism had 3, 4 and 5 maxima respectively for planar views and corresponding minima for edge views. Even the stability function for more complex objects like a chair had two maxima at the front and back views. Objects can be better represented by planar views because they have a high similarity to adjacent views and their features are visible in a greater range of views.

These results concur with psychophysical experiments about the significance of planar views in the human object learning and recognition [1].

2.2 Key Frames

Based on the coverage rates a coverage interval I is calculated for each view. We consider a view V_i to be represented sufficiently by a view V_k, if the coverage rate $T_k(i)$ is greater than a chosen minimum track rate T_{min}. The coverage interval I_k for view V_k contains all views that are represented sufficiently by V_k, i.e. whose coverage rate T_A is greater than T_{min} (see figure 4).

The task is now to choose a subset of all views as key frames, so that all views are represented at least by one key frame. This problem is known as the "minimum circle cover problem on weighted intervals" [3]. When using $w_i = 1 - S_i/2 \cdot \sum_{j=1}^{n} S_j$ as weight for the intervals, the algorithm in [3] determines the set of key frames with minimum size whose intervals cover the whole viewing circle while maximizing the sum of all key frame stability values. Figure 5 displays the intervals for a cube and the chosen key frames.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

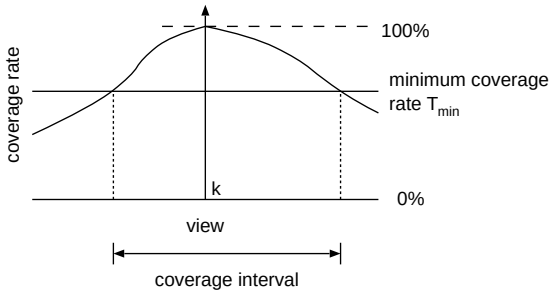

Figure 4: An example of the coverage rate function when tracking the features of a view k in both rotation directions. The coverage interval for view k contains all views whose coverage rate is greater than the minimum coverage rate T_{min}

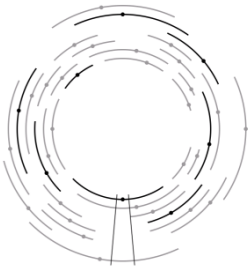

Figure 5: Coverage intervals for the textured cube and chosen key frames (black)

2.3 Feature Reduction

The features in the chosen key frames are reduced to a maximum number of 64, 32 and 16 for the three scales. The views are subdivided into 64 sub-images. Then the feature with the lowest structure tensor value from the sub-image with the maximal number of features is removed repeatedly, until the maximum number of features is reached. This ensures evenly distributed features while biasing features with a high 2D structure.

3 Recognition Mechanism

The recognition mechanism uses the same matching algorithm already used in the tracking process [5]. A presented view is matched against all stored key frames for all objects. The similarity is determined by the ratio $matched\ features/features\ in\ key\ frame$ and the winner object is determined by simple maximum selection.

Recognition performance was tested with several computer generated objects: A 3-sided prism, a cube, a 5-sided prism, a little toy, an ant and three types of chairs. The two prisms and the cube shared a common texture on one side.

Recognition results for various minimal coverage rates T_{min}:

T_{min}	average number of key frames	recognition rate
0.50	4.6	88.5%
0.65	7.5	95.5%
0.70	9.6	95.5%
0.75	12.5	97.6%

Recognition results under various viewing conditions ($T_{min} = 0.65$):

modification	recognition rate
scale 0.8	87.9%
contrast -25%	96.2%
contrast -50%	82.6%
gaussian noise, $\sigma = 0.1$	95.1%
gaussian noise, $\sigma = 0.2$	89.2%
shearing $x' = x + y \cdot 0.2$	96.2%

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

4 Feedback

A key characteristic of cortical architecture is the existence of feedforward and feedback connections with localized foci of projection. Although the functional role of feedback processing is still a matter of intensive investigations, there are a number of models that describe possible functions.

We adopted a feedback mechanism from a previous model of cortical boundary processing [4]. In our presented work feedback processing is used to handle sequences of views, where detected views are stored in short-term-memory to subsequently bias the detection process.

4.1 Computational Mechanism of Gain Control

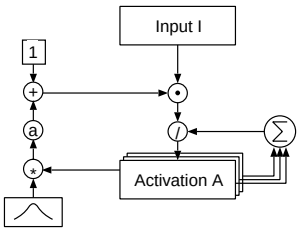

Figure 6: Scheme of the feedback model

The input from the recognition stage at time t is denoted as I^t representing the similarity matrix of the presented image to the key frames. A^t denotes the biased detection result and F^t the feedback at time t. The activation A of the detection layer at time $t+1$ is then computed as follows:

$$A^{t+1} = I^t \cdot \frac{1 + a \cdot F^t}{1 + \sum_{views} A^t},\ F^t = A^t * \Lambda_{views}$$

The weighted average of activities over the views of each object is computed by a convolution operation ($*$) utilizing the weighting function Λ. Because the detection layer is only sparsely activated by key frames, the coefficients of the smoothing mask are normalized, so that the sum of all coefficients at key frames is 1. The inhibitory term $1 + \sum_{views} A^t$ limits and normalizes the activities in the detection layer. We used a Gaussian smoothing mask as Λ_{views} with $\sigma = 5$ and the scaling factor $a = 50$. The winner object is then determined by the maximum in the detection layer.

4.2 Results

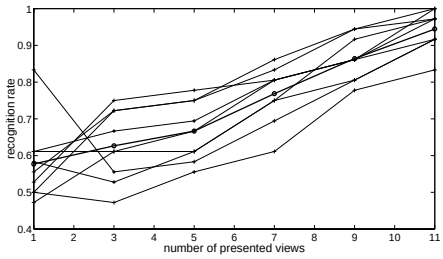

Figure 7: Recognition results with the feedback model. '+': Recognition results for each object, '•': The average recognition rate

We tested the feedback mechanism with 9 computer generated cubes with textures on the front, back and the two side faces.
The textures were chosen randomly out of a set of 18 textures (6 types with 3 variations) so that each texture appears on two cubes. By using the method described above, we created an object database that contained 8.1 key frames per object on average.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 457-462

We presented a sequence of n successive views, $n = 1, 3, 5, \ldots, 11$. An object is correctly recognized if one of the object's key frames is the most activated in the detection layer after the last image was presented.

As can be seen in figure 7 the recognition result is somewhat greater than 50% when presenting only one view, since each texture can be found exactly on two cubes. The recognition rate increases with the number of presented views up to 94.4% with 11 views.

5 Conclusion

We presented a view-based object recognition system that is able to recognize objects from short image sequences. The novelty of our approach is to combine a model of feedback processing [4] with a view-based object recognition system [8]. This object recognition system was modified in order to extract an optimal selection of key frames. Our results show that the recognition performance using single frames is substantially improved when image sequences are presented. The biologically motivated feedback that biases the recognition process further distinguishes our approach from simple feedforward architectures [6].

The current work only considers object rotations around one axis but could be extended to handle arbitrary rotations. Considering the displacement of detected features would create additional constraints, which could be included in our model feedback processing to further improve the results.

References

[1] G. K. Humphrey K. H. James and M. A. Goodale. Manipulating and recognizing virtual objects: Where the action is. *Canadian Journal of Experimental Psychology*, 55:2:113–122, 2001.

[2] J. J. Koenderink and A. J. Van Doorn. The singularities of visual mapping. *Biological Cybernetics*, 24:51–59, 1976.

[3] D. Z. Chen M. J. Atallah and D. T. Lee. An optimal algorithm for shortest paths on weighted interval and circular-arc graphs, with applications. *Algorithmica*, 14:5:429–441, 1995.

[4] H. Neumann and W. Sepp. Recurrent v1–v2 interaction in early visual boundary processing. *Biological Cybernetics*, 81:425–444, 1999.

[5] M. Pilu. A direct method for stereo correspondence based on singular value decomposition. In *CVPR'97*, pages 261–266, 1997.

[6] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3:1199–1204, 2000.

[7] E. Truco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.

[8] C. Wallraven and H. Bülthoff. Acquiring robust representations for recognition from image sequences. In *DAGM-Symposium München 2001*, pages 216–222. Springer, 2001.