Learning feature representations for an object recognition system

Kai Welke[†‡] , Erhan Oztop[*†] , Ales Ude[§] , Rüdiger Dillmann[‡] and Gordon Cheng[*†]
[*]JST, ICORP, Computational Brain Project
4-1-8 Honcho, Kawaguchi, Saitama, Japan
[†]ATR Computational Neuroscience Lab., Dept. of Humanoid Robotics and Computational Neuroscience
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
[‡]University of Karlsruhe (TH), IAIM, Institute of Computer Science and Engineering (CSE)
P.O. Box 6980, 76128 Karlsruhe, Germany
[§]Jožef Stefan Institute, Dept. of Automatics, Biocybernetics, and Robotics
Jamova 39, 1000 Ljubljana, Slovenia

***Abstract*— For humanoid robots to be part of our daily lives, not only mobility and their manipulation capability being essential, but their ability to represent and recognize objects in an adaptable manner is also crucial. To this end, we propose an object representation scheme that fits well with the view-based cortical representation of objects found in the primate Inferotemporal cortex (IT). We derive our proposal from the simple observation that a single object may exhibit very different sets of visual features when transformed in space. Nonetheless, there are some fixed (object dependent) views of an object, which even with small transformations would not lead to large feature changes. We refer to these views as keyframes. With this in mind, an object is represented with a set of keyframes. The changes in the features around a keyframe are nullified with a neural network that learns to represent the keyframe, and its rotational variations in a compact and rotation invariant form. To evaluate the proposed representation scheme, 100 real life objects are tested in a recognition task. Furthermore, a method for minimizing the number of keyframes for a given object is proposed, which we suggest must yield optimal generalization and computational efficiency. The proposed representation scheme is ideal for building humanoid cognitive architectures as it is decoupled from the recognition system.**

I. INTRODUCTION

The main target of our work is to develop object representations and a learning scheme, which is suitable for learning in humanoid robots, e.g. via action-perception coupling, much the same way as humans learn about objects in their environment. In our earlier work [1], we have shown how to make use of foveated vision to realize object recognition on a humanoid robot. Having the fovea centered on an object simplifies the task of object recognition, as the object is already located and we only have to concentrate on object learning and identification. In this paper, we will introduce an approach, which is able to learn representations from experienced images based on artificial neural networks (ANN). This enables us to train new inputs when they occur, and permits to refine object representations in the active vision process.

Figure 1 provides a conceptual overview of the structure of object representations as used thoughout this paper. M individual feature components form one representation. Each stored representation describes a different aspect of the object. To describe the whole object, K representations have to be stored. Throughout this paper we will refer to these representations as *keyframes*.

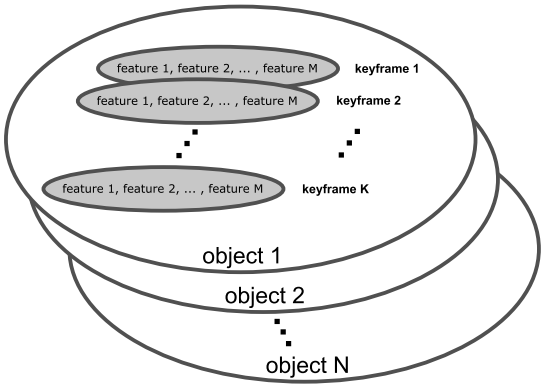


Fig. 1. Every object is described with a set of K representations. We call these representations keyframes. For every keyframe, we store M features. Our approach reduces the number of features M per keyframe and the number of keyframes K required to describe an object.

For high level object recognition tasks it is crucial to rely on a set of features and a set of keyframes per object that reflect the main object properties. In our opinion, optimal performance of the recognition task can only be achieved when the object representation consists of the minimum number of features and the minimum number of keyframes per object without sacrificing from recognition performance. Keyframes highlighting intrinsic object properties allow similarities of objects to be revealed and benefit in building object classes. We will introduce an approach which allows us to investigate and minimize the number of features and keyframes required to describe an object.

The literature offers a variety of recognition systems which use methods of feature compression to reduce the dimensionality of the feature space. One of the most popular techniques for feature compression uses the principal component analysis (PCA), which finds a linear subspace of the image space for a given number of dimensions with maximum variance

[2] [3]. In more recent work, non-linear dimensionality reduction methods and their application to feature compression have been studied [4]. Prominent methods comprise ISOMAP, which has been evaluated in face recognition [5] and NLPCA, which has been applied to a wide range of dimensionality reduction problems [6] [7]. In contrast to these methods that focus on feature compression alone, we introduce a non-linear approach that combines dimensionality reduction and the generalization for depth rotational invariance. We will show that this approach reduces the dimension of the feature space, as well as the number of keyframes required to describe an object.

Other approaches introduce representations that incorporate invariances to affine transformation of the object in the scene. Prominent methods use the "Scale invariant feature transform" (SIFT) [8] [9] or complex moments [10] to retrieve invariant feature representations. Although these approaches achieved satisfactory results in practical systems, however, our own objective is to produce a system that can lead to biological plausible object representations. Our motivations are presented in the following section.

II. Learning feature representations

A. Relation to human visual processing

Psychophysical experiments on humans [11] and on monkeys [12] have lead to a view-based model of how the visual system achieves consistent identification of 3D objects. Contrasting object-based approaches like the "Recognition-by-Components" theory [13], view-based methods model the object by selected views rather than by constructing a 3D-model to match the object in the scene. Logothetis *et al.* found cells in the macaque Inferotemporal cortex (IT) that are tuned to specific views of an object [14]. Psychophysical studies carried on with human subjects indicate, that object recognition is performed around views presented while training [15].

The human visual system achieves invariance to object transformations in multiple stages in the visual processing. Invariances comprise affine transformations of the object, like shifting, scaling and rotation of the object in the viewing plane and non-affine transformations like rotation in depth and deformation. Oram and Perrett developed a computational model for shift and size invariance based on the human visual system. They introduced constraints derived from neurophysiology and psychophysics and propose a model matching these constraints [16]. Following some of the results of this study, Olshausen *et al.* developed the dynamic routing circuit model which achieves invariance to shift and scaling [17]. The proposed system generates a normalized representation of a region in the visual field to which attention is guided. Aspects of the human visual processing are modeled, considering the ventral pathway from lateral geniculate nucleus (LGN) to IT [1]. This model suggests that invariance to shift and scaling are solved along the path to IT, while invariance to depth rotation is achieved by the view-tuned cells in IT, as proposed by Logothetis *et al.*, lending support to our approach.

[1]following the path: LGN - primary visual cortex (V1) - V2 - V4 - posterior inferotemporal cortex (PIT) - central inferotemporal cortex (CIT) - anterior inferotemporal cortex (AIT)

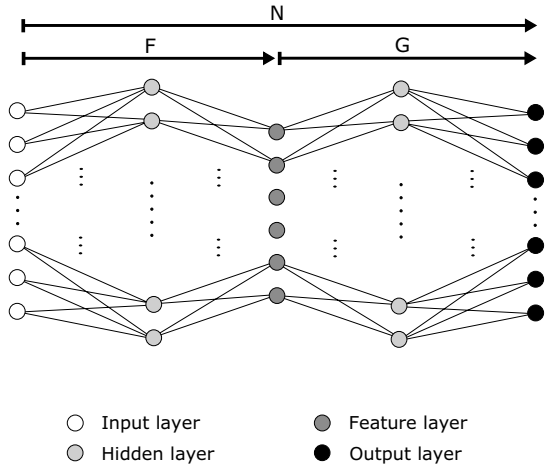


Fig. 2. An example network architecture is shown. We use a symmetrical five layered network structure. All units are sigmoidal and fully connected. The network performs a combination of the mapping function F and the demapping function G. Different layer sizes were applied during the experiments.

The response of the view-tuned cells can be used to achieve embodiment for the task of determining the presence of an object. Logothetis *et al.* found view-tuned cells which respond with a significant spike rate for different ranges of depth rotation around a frame. For distractor objects, small or no spiking was observed. The combination of view-tuned cells responses for a specific object gives a good prediction of the presence for that object. As will be detailed later, our learning scheme leads to representations that have similar properties as the view-tuned cells. Our approach keeps classification and representation steps seperate so that any classifier can be applied to the representation learned by the network. For simulations presented in this paper we have used the nearest neighbor classifier. Like the view-tuned cells, our representations are valid for a certain amount of depth rotation around a learning view. Object labels are added to the representations, to allow embodiment of the objects.

B. Network structure

Figure 2 shows the basic network architecture. We use a fully connected feedforward network. The network consists of 5 layers (including the input layer) and has symmetrical structure. The number of neurons in the input layer equals the number of neurons in the output layer. Both hidden layers consist of the same amount of neurons. The middle layer has the smallest number of neurons and serves as bottleneck for the network. All units of the network have a sigmoidal activation function.

The basic network layout follows an approach by Kramer [18]. In his work, he analyzed the ability of the network to perform non-linear principal component analysis (NLPCA) on the input. The network is trained in an autoassociative manner. During training the network learns to reconstruct the input

patterns on the output layer by backpropagation learning. We also exploit this capability of extracting principal components in the feature layer, but also apply a different training mode to generalize for depth rotations.

The function $N(I)$ that is performed by the network can be decomposed into two functions F and G with $N(I) = G(F(I))$. As shown in Figure 2, F describes the function performed by neurons from input layer to the feature layer output and is called mapping function. The demapping function G describes network calculations from feature layer to output layer. If we consider the sub-networks that perform the functions F and G, we can describe them with two three layered non-linear networks. Both sub-networks have hidden layers of the same size, input and output layer sizes are switched between F and G. Kolmogorov proved in 1957, that any continuous n-dimensional function can be represented by superposition and sum of one-dimensional continuous functions [19]. Applied to feedforward networks the theorem guarantees, that a network with one hidden sigmoidal layer of infinite size can represent any non-linear n-dimensional function [20]. The lower bounding for the size of the hidden layer depends on the function to represent by the network. For the mapping function F and demapping function G we can conclude, that both functions can represent any non-linear continuous function, if the hidden layer size is sufficiently large and the weights are chosen accordingly.

C. Image preprocessing

For the human visual system, an object is represented by the combination of several cues, experienced with two eyes. To simplify the analysis of the behaviour, we restrict our approach to the intensity of the image of one single camera. However, the approach can be adapted to other cues like opponency color maps or disparity maps and to a stereo camera setup.

We segment the objects from the black background and normalize the intensity image in size to 80×80 pixels. The normalized intensity image is preprocessed with a two dimensional Gabor filter. It has been shown, that the response of primary visual cortex (V1) simple cells can be modeled well by the Gabor filter response [21]. The line and edge selectivity of the simple cells responses can be derived from an information maximization process [22]. This indicates that Gabor filter responses encode important properties of the intensity image.

A two dimensional Gabor filter is defined by the following function:

$$\psi_{\vec{k}}(\vec{x}) = \frac{||\vec{k}||^2}{\sigma^2} e^{\frac{-||\vec{k}||^2||\vec{x}||^2}{2\sigma^2}} (e^{i\vec{k}\vec{x}} - e^{\frac{-\sigma^2}{2}}) \quad (1)$$

The parameter vector $\vec{k}$ defines the properties of the filter:

$$\vec{k} = k_v \begin{pmatrix} \sin\phi_\mu \\ \cos\phi_\mu \end{pmatrix} \text{ where } k_v = \alpha\pi 2^{-\frac{\beta v+2}{2}} \; ; \; \phi_\mu = \frac{\mu\pi}{D} \quad (2)$$

The parameter v defines different scaling factors and the parameter μ defines different orientations of the filter.

A Gabor jet is a collection of Gabor filter reponses at different orientations and scaling. To retrieve the feature vector for an intensity image, we apply the filter with a common set of parameters where $\mu = 0, \ldots, 7$ and $v = 0, \ldots, 4$ as proposed by Wiskott *et al.* [23]. Each component of the Gabor jet is calculated by the signal energy of the complex part of the Gabor filter response corresponding to one parameter combination. With the chosen set of parameters the calculations result in a 40-dimensional Gabor jet.

Besides representing important properties of the image, Gabor filters have advantages for real camera image processing. Gabor filters exhibit invariance to changing illumination conditions. Along with the representation as Gabor jet, invariance to small shifts is also achieved which helps to cope with segmentation inaccuracies.

In the literature Gabor filters are often used as a filter array with reduced receptive fields, each responsive for a subregion of the image. While this implementation is more biological plausible, we use only one Gabor jet, extracted at the center of the image, to reduce the processing time of the system (note that the dimensionality of the preprocessing output will be linear in the number of filters applied).

D. Learning details

The network is trained with backpropagation. We use a backpropagation implementation with momentum to speed up the convergence of the network. The weights are updated using the following equation:

$$\Delta w^t = \eta \frac{\delta E}{\delta w} + \beta \Delta w^{t-1} \quad (3)$$

During training, different rotations in depth i of the object n are presented to the network. The Gabor jet of each rotated image is denoted with $\vec{r}_{n,i}$. For every object there exists a set of rotations $R_n = \{\vec{r}_{n,0}, \ldots, \vec{r}_{n,I}\}$. We only consider rotations relative to the vertical axis.

For training a subset of rotations is selected as keyframes. For every object we choose a set of keyframes $K_n = \{\vec{r}_{n,i_0}, \ldots, \vec{r}_{n,i_K}\}$ which cover all rotations of the object. We choose the indices $i_0, \ldots, i_K$ to separate R_n in intervals of equal size. For each keyframe $\vec{r}_{n,i_k}$, we select a set of training views $V_{n,k}$ which are rotated a small amount relative to the keyframe. The network is trained to reconstruct the keyframe for all selected rotations around the keyframe.

The backpropagagtion algorithm reduces the error $E = 0.5||N(V_{n,k}) - \vec{r}_{n,i_k}||$ for all objects n and all keyframes k. Figure 3 shows example images used to calculate the training views $V_{n,k}$ and the corresponding keyframe $\vec{r}_{n,i_k}$. The training forces the network to nullify rotations around the keyframe $\vec{r}_{n,i_k}$.

The trained network is tested with all available rotations in $R = R_0 \cup \ldots \cup R_n$. For all rotations we test if the Gabor jet of the closest keyframe is reproduced at the output layer. All views with closest keyframe k are denoted with $T_{n,k}$. The training is stopped, if the mean square error is below the tolerated error ϵ. If the network does not converge after a

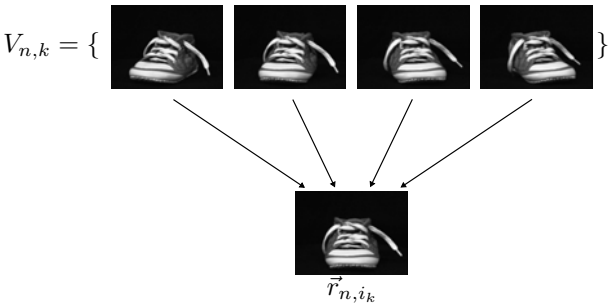


Fig. 3. The network is trained to associate the training views $V_{n,k}$ with the corresponding keyframe $\vec{r}_{n,i_k}$.

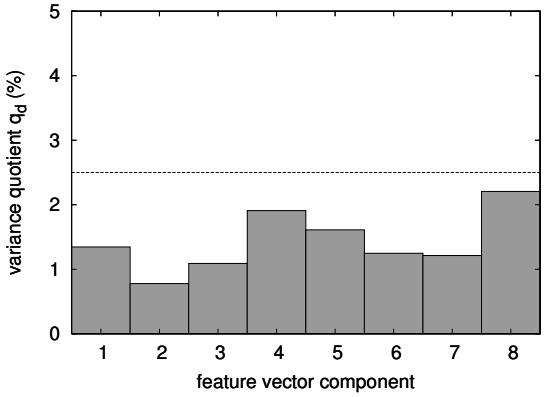


Fig. 4. Quotient between variance among views trained to reconstruct the same keyframe and all ten views in the experiment for every individual feature component. The plot shows the mean values for ten different networks. For all components, the quotient lies below 2.5%, indicating that the network learns an invariant representation in the feature layer.

fixed number of training steps, we assume that the network is stuck in a local minimum and reinitialize the weights.

III. Experiments

A. Generalization for depth rotations

In our first experiment, we demonstrate that the network is able to generalize for rotations in depth. We show that for some amount of depth rotation around a keyframe, the network can find a representation for the keyframe, which is invariant to the rotation.

We used 10 objects from the Amsterdam library of object images (ALOI) [24] during the experiment. The network was trained and tested with rotations around the vertical axis. For this type of rotation, the ALOI offers object images in 5° steps. For testing the network all 9 rotations in the interval $[-20^\circ; 20^\circ]$ were used. During the training, four Gabor jets were presented to the network, extracted from views with the rotations $i = -15^\circ, -5^\circ, 5^\circ$ and $i = 15^\circ$. The network was trained to reconstruct the Gabor jet corresponding to the keyframe at $i = 0^\circ$.

In this experiment we apply a network structure with a hidden layer size of 90 neurons and a feature layer size of 8 neurons. We use a learning rate of $\eta = 0.05$ and a momentum rate of $\beta = 0.1$. The error threshold for termination is set to $\epsilon = 0.05$. Both, layer sizes and learning parameters were empirically determined. Layer sizes were evaluated by probing with different parameters until the training error converged. The error threshold was chosen to match the convergence of a network which learned generalization. For the described network structure, the training converged after about 1,000,000 training steps. After training, the network found a compressed representation in the feature layer, which allows the reconstruction of the keyframe Gabor jet at the output layer.

To analyze the learned representations, we measured the variance of the individual feature components in the representation for all rotations of one keyframe from the test set. We found that the network is able to reduce the variance significantly. We calculated the average variance per output of feature unit number d over all objects following the equation:

$$var_{avg,d} = \frac{\sum_{n=0}^{N} var(F_d(T_{n,0}))}{N} \tag{4}$$

A successful training leads to an average variance of about $var_{avg,d} = 0.001$. To relate the mean average per unit among one keyframe to the variance among all objects we calculate the quotient of the average variance $var_{avg,d}$ and the variance of the feature component d among all keyframes:

$$q_d = \frac{var_{avg,d}}{var(F_d(R))} \tag{5}$$

Figure 4 shows the quotient q_d for all individual feature components. The quotient lies below 2.5% for all components. The test proves, that the remaining variance in the features of views corresponding to one keyframe is significantly smaller than the variance of features among the different keyframes. Based on this observation, we conclude that the mapping function F learns to encode an invariant representation of views around a keyframe in the feature layer.

Figure 5 visualizes the feature layer output for four out of the ten objects in this experiment. The outputs of the feature layer neurons are displayed for all rotations from the test set. To retrieve an invariant representation for a keyframe, we calculate the mean of the feature layer output for the four training set views for each keyframe. With this result we found a representation, which is invariant to rotations around the vertical axis in the interval of $[-20^\circ; 20^\circ]$ around the keyframe.

B. Distinctiveness

In the second experiment we show that we can reduce the dimension of the feature space and the number of keyframes per object without sacrificing the ability to distinguish objects. We use the first 100 objects of the Amsterdam library of object images. For each object we select eight keyframes in distances of 45° from the set of rotations. With this selection of keyframes we cover the complete rotation in depth around the vertical axis. We apply the same method as used in the first experiment for each keyframe of an object. Four Gabor jets of the views at angles $i = -15^\circ, -5^\circ, 5^\circ$ and $i = 15^\circ$

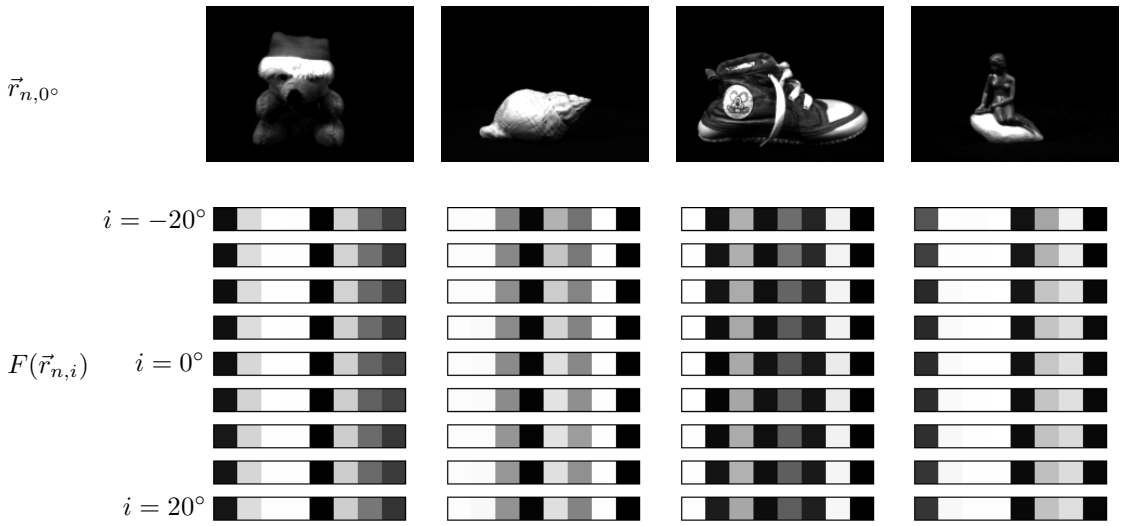


Fig. 5. After training 4 views at the rotations $i = -15^\circ, -5^\circ, 5^\circ$ and $i = 15^\circ$ relative to the keyframe $\vec{r}_{n,0^\circ}$ the mapping function F of the network has learned an invariant representation for rotations around the keyframe. The output of the eight feature neurons in the feature layer $F(\vec{r}_{n,i})$ is displayed for four of ten trained objects. Strong responses of neurons are denoted with white fields.

relative to the keyframe rotation are presented to the network. The network is trained to reconstruct the Gabor jet of the keyframe, corresponding to the view at i_k.

For this experiment we apply a hidden layer size of 400 neurons and a feature layer size of 20 neurons. We use a learning rate of $\eta = 0.01$ and a momentum rate of $\beta = 0.01$. The error threshold for termination is set to $\epsilon = 0.08$. Again these parameters were determined empirically. The network converges after about 5,500,000 training steps. For all keyframes we build the corresponding representation by calculating the mean of the four representations derived by computing the output of the feature layer for the test set views. The representations are labelled with the object names.

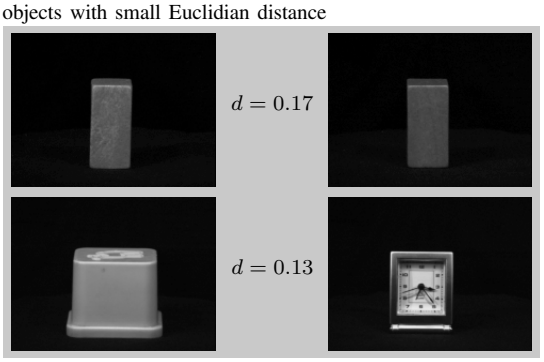
For object identification, we apply a nearest neighbor classifier to associate a view from the test set to the representation with a minimal Euclidian distance. For rotations in the test set that lie between two keyframes, we cannot determine which keyframe of the object is valid. We define a correct classification, if the test view is associated with a keyframe which has the same object label. Remember that an object is represented with eight keyframes in this experiment.

During evaluation, all rotations of each of the first 100 objects from ALOI, totaling $100 \times 72 = 7200$ views were presented to the network. The network was able to associate 83% of the test views to the correct object. For the remaining 17% of mismatches we found two major reasons.

The first fact, which leads to mismatches is the non-uniqueness of Gabor jets among the ALOI objects. As explained in Section II-C the extraction of a single Gabor jet for the whole object is a very rough feature for object appearance. For reduction of computational complexity, we abstained from extracting an array of Gabor jet responses. To verify our reasoning, we measured the Euclidian distance of all jets for views of objects used in this experiment. Indeed we found that Gabor jets of two different objects can be very close. Figure 6 shows pairs of objects, with small Euclidian distance and with well apart Euclidian distance in between. We do not introduce a constraint, that restricts the learning set to a subset of ALOI images to preserve the comparability with other approaches. Regarding the first example in Figure 6 the behaviour of mismatching both objects can be considered correct, because the intensity image of both objects is very similar.

Another reason for mismatches originates from our method to choose keyframes for the objects. We selected keyframes in equal rotational distances around the object. In our approach of keyframe selection every object is learned with the same amount of keyframes. This approach does not take into account, that specific views of the objects allow invariance to depth rotation for a large amount of rotation. For optimal results, we would choose only those keyframes with fairly similar images across rotations in depth to describe the object. The number of keyframes required for one object and the amount of invariance which can be introduced with this keyframe are object properties which our system is not aware of. During learning, this can lead to two training sets (I_1, O_1) and (I_2, O_2), where different outputs O_1, O_2 are trained for very similar inputs I_1, I_2. The output trained for the input I_1 and I_2 will be a combination of both target outputs and cannot be classified correctly to either class. To solve this conflict, we need a mechanism to determine the minimum amount of keyframes for the object. For keyframes of one object, the selection method should assure, that similar views of the object are covered by the same keyframe. When adding keyframes of new objects to the training set, similar keyframes have to be handled by labeling the representation with both object labels.

objects with small Euclidian distance



objects with distinctive Gabor jets

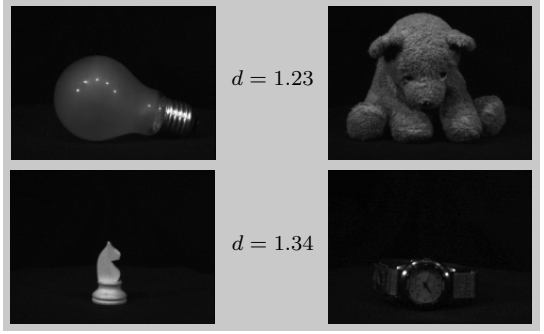


Fig. 6. Objects from the ALOI and the Euclidian distance d of the corresponding Gabor jets. The first two object pairs lead to misclassification, while the second two pairs are classified correctly for all rotations.

IV. Conclusion

The proposed biologically plausible object representation provides a general and robust scheme for learning object features, making it a prime candidate as a building block for the construction of humanoid cognitive architectures. We introduced a learning scheme which is able to generalize for depth rotational invariant, compact representations for object views. We showed that the proposed representations are valid for a certain amount of rotation, like the cortical representation with view-tuned cells in IT. In further experiments we described all rotations around the vertical axis of 100 objects with eight keyframes per object. We applied a nearest neighbor classifier to recognize these objects. We analysed the representations for cases where the recognition leads to mismatches. From this analysis we conclude, that the selection criteria only based on the distance between keyframes is not sufficient to describe objects. The number of keyframes and the number of views associated with a keyframe are object properties and have to be determined automatically to achieve an optimal representation for the objects.

Acknowledgment

The work described in this paper was partially conducted within the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

References

[1] A. Ude, C. Gaskett, and G. Cheng, "Support vector machines and gabor kernels for object recognition on a humanoid with active foveated vision," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 668–673.

[2] S. Nayar, S. Nene, and H. Murase, "Real-time 100 object recognition system," in *Proceedings of the IEEE Conference on Robotics and Automation*, vol. 3, 1996, pp. 2321–2325.

[3] V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua, "Fully automated and stable registration for augmented reality applications," in *Proceedings of the Second IEEE and ACM Symposium on Mixed and Augmented Reality*, 2003, pp. 93–102.

[4] S. de Backer, A. Naud, and P. Scheunders, "Nonlinear dimensionality reduction techniques for unsupervised feature extraction," *Pattern Recognition Letters*, vol. 19, no. 8, pp. 711–720, June 1998.

[5] R. Souvenir and R. Pless, "Isomap and nonparametric models of image deformation," in *IEEE Workshop on Motion and Video Computing*, 2005, pp. II: 195–200.

[6] P.-J. Wang and C. Cox, "Study on the application of auto-associative neural network," in *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, vol. 5, 2004, pp. 3291–3295.

[7] D. Tzovaras and M. Strintzis, "Use of nonlinear principal component analysis and vector quantization for image coding," *IEEE Transactions on Image Processing*, vol. 7, no. 8, pp. 1218–1223, August 1998.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 1150–1157.

[9] ——, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] S.-H. Kim, I.-C. Kim, and I.-S. Kweon, "Probabilistic model-based object recognition using local zernike moments," in *IAPR workshop on Machine Vision Applications*, Nara , Japan, Dec. 2002.

[11] M. Tarr, P. Williams, W. Hayward, and I. Gauthier, "Three-dimensional object recognition is viewpoint dependent," *Nature Neuroscience*, pp. 275–277, 1998.

[12] N. Logothetis, J. Pauls, H. Bülthoff, and T. Poggio, "View-dependent object recognition by monkeys," *Current Biology*, vol. 4, pp. 401–414, 1994.

[13] I. Biedermann, "Recognition-by-components: A theory of human image understanding," *Psychological Review*, vol. 94, pp. 369–384, 1987.

[14] H. Bülthoff and S. Edelmann, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," in *Proceedings of the National Academy of Sciences*, vol. 89, 1992, pp. 60–64.

[15] N. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," Current Biology 5, pp. 552–563, 1995.

[16] M. W. Oram and D. I. Perrett, "Modeling visual recognition from neurobiological constraints," *Neural Networks*, vol. 7, no. 6-7, pp. 945–972, 1994.

[17] B. A. Olshausen, C. H. Anderson, and D. C. V. Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *The Journal of Neuroscience*, vol. 13, no. 11, pp. 4700–4719, 1993.

[18] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AICHE Journal*, vol. 37, pp. 233–243, 1991.

[19] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superpositions of continuous functions of one variable and addition," *Akademii Nauk USSR*, vol. 114, no. 5, pp. 953–956, 1957.

[20] V. Kurková, "Kolmogorov's theorem and multilayer neural networks," *Neural Networks*, vol. 5, no. 3, pp. 501–506, 1992.

[21] T. S. Lee, "Image representation using 2d gabor wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[22] A. Bell and T. Sejnowski, "The independent components of natural scenes are edge filters," in *Vision Research*, vol. 37, 1997, pp. 3327–3338.

[23] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, July 1997.

[24] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 103–112, 2005.