Biol Cybern (2005) 93: 79–90
DOI 10.1007/s00422-005-0585-8

ORIGINAL PAPER

Biological Cybernetics

**Wolfgang Einhäuser · Jörg Hipp · Julian Eggert
Edgar Körner · Peter König**

Learning viewpoint invariant object representations using a temporal coherence principle

Received: 19 February 2004 / Accepted: 23 May 2005 / Published online: 13 July 2005

Abstract Invariant object recognition is arguably one of the major challenges for contemporary machine vision systems. In contrast, the mammalian visual system performs this task virtually effortlessly. How can we exploit our knowledge on the biological system to improve artificial systems? Our understanding of the mammalian early visual system has been augmented by the discovery that general coding principles could explain many aspects of neuronal response properties. How can such schemes be transferred to system level performance? In the present study we train cells on a particular variant of the general principle of temporal coherence, the "stability" objective. These cells are trained on unlabeled real-world images without a teaching signal. We show that after training, the cells form a representation that is largely independent of the viewpoint from which the stimulus is looked at. This finding includes generalization to previously unseen viewpoints. The achieved representation is better suited for view-point invariant object classification than the cells' input patterns. This property to facilitate view-point invariant classification is maintained even if training and classification take place in the presence of an – also unlabeled – distractor object. In summary, here we show that unsupervised learning using a general coding principle facilitates the classification of real-world objects, that are not segmented from the background and undergo complex, non-isomorphic, transformations.

W. Einhäuser · J. Hipp · P. König
Institute of Neuroinformatics, University & ETH Zürich, Zürich, Switzerland

J. Eggert, E. Körner
HONDA Research Institute Europe GmbH, Offenbach/Main, Germany

P. König
Institute of Cognitive Science, Department Neurobiopsychology, University of Osnabrück, Osnabrück, Germany

W. Einhäuser (✉)
California Institute of Technology, Division of Biology, 139–74, 1200 E. California Blvd., Pasadena, CA 91125, USA
E-mail: wet@klab.caltech.edu
Tel.: +1-626-3952878
Fax: +1-626-7968876

1 Introduction

Humans readily recognize an object independently of its position in space, its illumination, its size or their viewpoint. This is particularly remarkable considering that the signal on their retinae changes dramatically under such transformations. Invariant recognition implies discarding information that is irrelevant for the task, while at the same time keeping the relevant information. This raises the question whether there are general principles that allow a system to decide which part of the input signal it has to neglect? For contemporary machine vision, on the other hand, invariant recognition remains one of the main challenges, fostering the interest in the mechanisms of the visual system.

In the mammalian visual system, the issue of invariance arises already at its earliest stages: the retina achieves invariance to absolute illumination by adaptation of photoreceptors. A prominent example of an early invariance is found in the primary visual cortex: complex cells, one of the two major cell classes in primary visual cortex, are orientation selective, but at the same time largely position (phase) invariant (Hubel and Wiesel 1962). A number of recent studies address the principles underlying this invariance. One popular principle is that of temporal coherence. It is based on the idea to separate different aspects of a visual stimulus according to their time-scale, neglecting fast variations while keeping slower features. Phase invariance of complex cells in this scheme follows from local orientation being slower (i.e., having a longer correlation time-constant) than local position as first shown in Földiak's (1991) trace rule implementation. This relation between orientation and position also holds for natural visual stimuli (Kayser et al. 2003a; Betsch et al. 2004). Hence, the invariant response properties of complex cells can also be learnt by applying the temporal coherence principle to sequences of natural scenes. A number of recent studies exploited this possibility by using different implementations of the temporal coherence principle, such as 'stability' (Kayser et al. 2001, 2003b; Körding et al. 2004), 'slow feature analysis' (Berkes and Wiskott 2003) and a physiologically inspired learning rule (Einhäuser et al. 2002). Concluding,

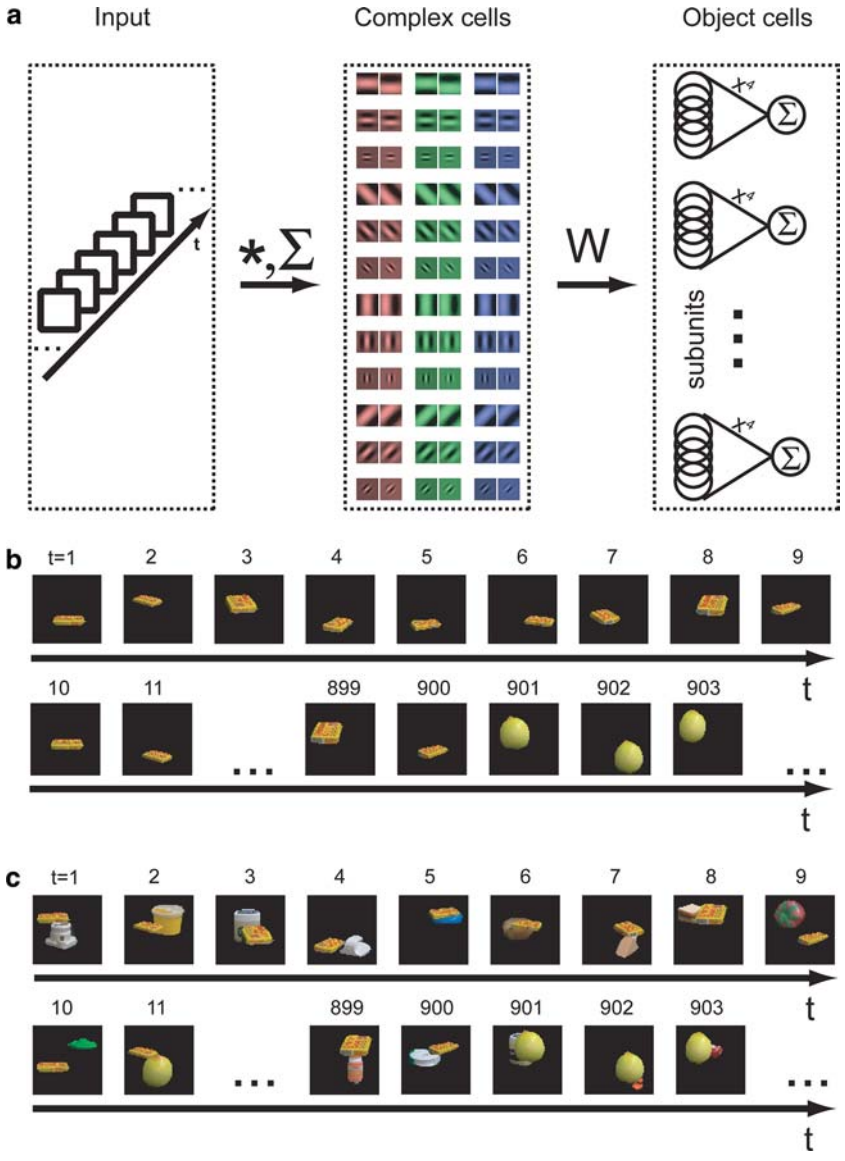

Fig. 1 Network and Stimuli. **a** Schematic of the network architecture. The stimulus sequence is fed into a set of colored *CCs*, whose properties are fixed. On the pooled output of these cells, *OCs*, which consist of multiple linear subunits, are trained to optimize the stability objective, **b** Plain stimuli. Several training views of each object are shown repeatedly at different positions; the square retina has 64×64 pixel resolution in three color channels. **c** In the cluttered condition a random object is added as distractor to the background of the target object at a random location

the general principle of temporal coherence explains important aspects of invariance properties in primary visual cortex.

Temporal coherence has also been used to investigate other properties of the visual system: optimising stability over natural stimuli may explain the segregation of colour from orientation in neuronal representations (Einhäuser et al. 2003). Repeated application of the same optimization in a hierarchical scheme leads to texture selective cells (Franzius et al. submitted). Proceeding further up through the hierarchy of the visual system, an iterative application of the trace rule can even learn to recognize faces (Wallis and Rolls 1997). The responses of cells trained in this network thus resemble those of neurons found in inferotemporal cortex. Multiple complex transformations including changes in viewpoint, have also been successfully addressed using simplified stimuli (Wiskott and Sejnowski 2002; Stringer and Rolls 2002). In summary, the principle of temporal coherence explains invariance properties of various parts of the visual system.

In spite of the analogies of networks trained with the principle of temporal coherence with the biological system, it remains to be investigated in how far the resulting systems support complex recognition tasks. We like to single out three aspects that make classification a difficult task: *real world stimuli* undergoing *complex transformations* in *cluttered scenes* during training. Machine learning approaches have reached a high degree of sophistication, and given a large number of training examples can solve these problems. However, large sets of labeled training examples are hard

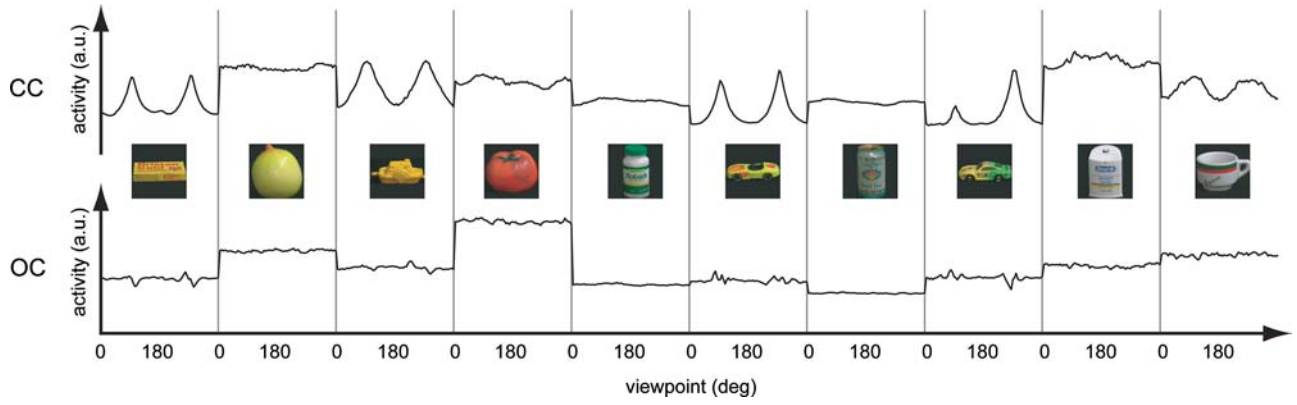

Fig. 2 Single-cell activity. Single-cell activity during presentation of ten objects rotating over all 72 viewpoints during testing. *Top row*: complex cell activity, *bottom row*: object cell activity, *middle*: presented object (for illustration depicted centrally in full database resolution and from 0° viewpoint only). Data of each viewpoint are averaged over 100 test presentations at different locations. In the used example, OCs were trained with 12 views (30° steps) of each of the ten shown objects

to obtain. Therefore it is attractive to divide the classification task into separate steps: first, the stimuli are processed by a network trained on unlabeled representative data with a general principle. Exploiting the statistical nature of the data this shall lead to a transformation of the input pattern into a representation that facilitates subsequent classification. Here we follow this approach and apply the temporal coherence principle to real world stimuli. The stimuli are subject to a complex transformation, namely change in viewpoint, and learned and classified in the presence of a distractor.

2 Methods

2.1 Network

Recently we demonstrated that applying a temporal coherence objective to natural videos yields complex cell-type receptive fields (Körding et al. 2004). Here we aim at extending this scheme into a homogeneous hierarchical network for invariant object recognition. Although complex cells might not be the only output stage of primary visual cortex, they are the dominant cell type in the supragranular layers where the projections to higher areas originate. To fulfil the requirement of homogeneous network architecture we hence add on top of a layer of modeled neurons with complex cell properties a second layer of "object cells" (OCs) that will be trained to optimize the stability objective on the output of complex cells (Fig. 1a). Since we and others have already demonstrated that the stability objective can be used to learn complex cell-type receptive fields, here – for computational efficiency, we fix the complex cells (CCs) instead of learning their properties. Complex cells are modeled as complex Gabor filters V:

$$V_{k,\varphi}(x,y) = \exp\left[-2k\left((x-x_0)^2+(y-y_0)^2\right)\right] \times \exp\left[-2\pi i k\big((x-x_0)\cos(\varphi) - (y-y_0)\sin(\varphi)\big)\right]$$

$$1 \le x, y \le 32;\ x_0 = y_0 = 16.5 \qquad (1)$$

where k denotes the spatial frequency and φ the orientation of the Gabor; x and y the pixel coordinate in a 32×32 pixel wide patch and x_0, y_0 the patch centre. To compute the CCs' activities on each input image $I(x, y, t)$ the colour channels I_C of I are first convolved separately with the Gabors:

$$\underset{c,k,\varphi}{\tilde{I}}(x,y,t) = \sum_{x'}\sum_{y'} V_{k,\varphi}\left(x'-x, y'-y\right) I_c\left(x', y', t\right). \qquad (2)$$

Accordingly the CC activity is then given by the absolute value of $\tilde{I}$ pooled over all locations:

$$A_i^{(CC)}(t) = \sum_x \sum_y \left|\tilde{I}_i(x,y,t)\right|, \qquad (3)$$

where the index i summarizes the subscripts k, φ, c. While this pooling over all locations is not meant to model any specific cortical projection, it mimics the increasing receptive field size, when proceeding from primary visual cortex to higher visual areas. By this pooling a large-scale translation invariance is achieved.

Each CC's activity is normalized to zero mean and unit standard deviation. We use four orientations (0°, 45°, 90°, 135°), three spatial frequencies (1/32, 1/16, 1/8 pixel^{-1}) and three colour channels (R, G, B), resulting in $4 \times 3 \times 3 = 36$ different CCs.

The CCs project on a second layer of N neurons, which will be referred to as OCs throughout this paper. Each OC consists of S linear subunits akin to functional subunits observed in real cortical cells (Touryan et al. 2002). The activities of these subunits $\sum_{i=1}^{36} A_i^{(\mathrm{CC})} W_{ijs}$ are added nonlinearly:

$$A_j^{(\mathrm{OC})} = \sqrt[4]{\left(\sum_{s=1}^{S}\left(\sum_{i=1}^{36} A_i^{(\mathrm{CC})} W_{ijs}\right)^4\right)} \quad 1 \le j \le N. \qquad (4)$$

Since Touryan et al. (2002) observed up to five significant subunits for cortical CCs, we – for the OCs – chose a

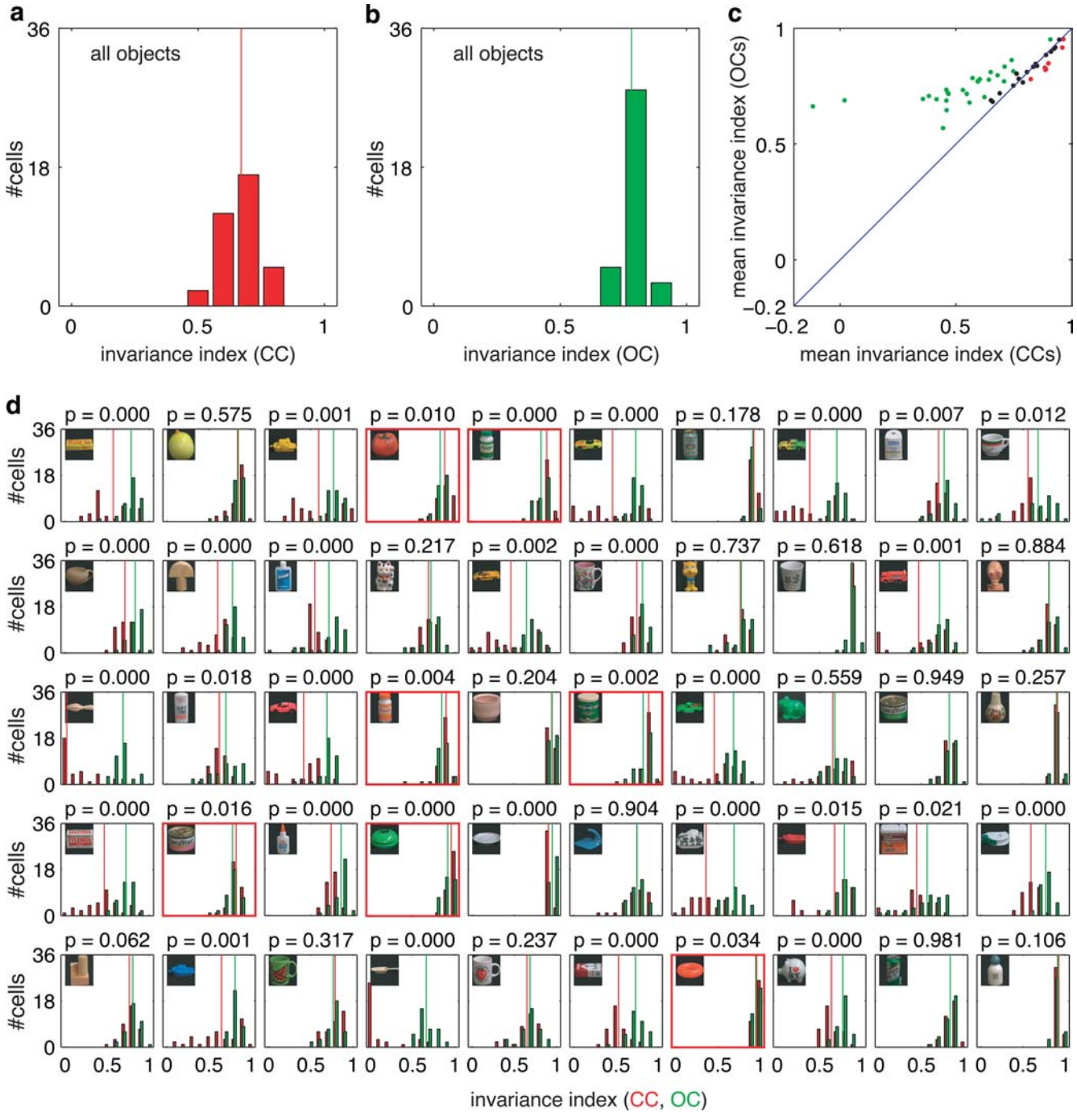

Fig. 3 Population data: viewpoint invariance. **a** Histogram of invariance index of all CCs (mean over all objects). *Vertical line* indicates mean. **b** Histogram of invariance index of all OCs (mean over all objects). *Vertical line* indicates mean. **c** Invariance index mean over all CCs (x-axis) versus mean over all OCs (y-axis). Each datapoint corresponds to one object. Objects for which mean invariance indices are significantly different (at $p < 0.05$) are colored (*red* for larger CC invariance, *green* for larger OC invariance). **d** Histograms of invariance index for all CCs (*red*) and OCs (*green*) for each object (insets). *Red* and *green vertical lines* mark mean invariance index of CCs and OCs respectively. Significance values for uncorrected t-tests on the difference of this means are given on top of the respective panels. *Red boxes* mark objects for which CC invariance is significantly larger than OC invariance

somewhat larger number for our baseline simulation ($S = 8$) and test in additional simulations, the effect of reducing this number. We chose unless otherwise stated the number of OCs equal to the number of CCs ($N = 36$). The exact choice of the non-linearity (4-norm) is not critical; in a separate study we demonstrate in the context of texture discrimination that there are no qualitative differences between using the 2-norm and the 4-norm (Franzius et al. unpublished observations).

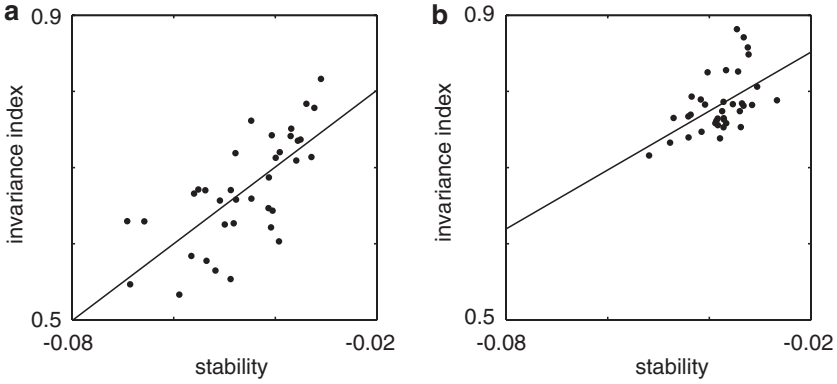

Fig. 4 Stability and viewpoint invariance. Invariance index for each cell (averaged over objects) plotted versus individual stability of each cell. Solid lines: regression lines for best linear fit. **a** CCs, **b** OCs

2.2 The stability objective function

We use the 'stability' formulation of the temporal coherence principle as introduced in Kayser et al. (2001): the squared-temporal derivative of the neurons' activity is minimized, while the trivial solution of constant zero activity is avoided by normalization by the temporal variance:

$$\Psi^{\text{stable}} = \sum_i \psi_i^{\text{stable}} = \sum_i -\frac{\left\langle \left(\frac{\mathrm{d}}{\mathrm{dt}} A_i\,(t) \right)^2 \right\rangle_t}{\mathrm{var}_t\,(A_i\,(t))}, \tag{5}$$

where $<.>_t$ denotes temporal averaging. The derivative is implemented as a finite difference $A_i(t+\Delta t) - A_i(t)$, where Δt is the step size between subsequent stimuli, as in the definition that follows. ψ_i^{stable} measures the stability of a single neuron and will be referred to as "individual stability". The stability objective does not include interaction between neurons in the network. Consequently, maximizing Ψ^{stable} alone would lead to a population of identical neurons. An additional de-correlation term forces neurons to acquire dissimilar receptive fields, by penalizing correlated and anti-correlated activity of different neurons:

$$\Psi^{\text{decorr}} = -\frac{1}{(N-1)^2} \left\langle \sum_i \sum_{i \neq j} \left(\sigma_{ij}^2\,(t) \right) \right\rangle_t, \tag{6}$$

where $\sigma_{ij} \equiv (A_i - \langle A_i \rangle_t)\,(A_j - \langle A_j \rangle_t) / \sqrt{\mathrm{var}_t\,(A_i)\,\mathrm{var}_t\,\left(A_j\right)}$ and N denotes the number of neurons as described previously. Averaging over the square of σ instead of using the correlation coefficient as such, ensures that anti-correlated *activity* is also penalized to ensure dissimilar *receptive fields*. The total objective is then defined as

$$\Psi^{\text{total}} = \Psi^{\text{stable}} + \Psi^{\text{decorr}}. \tag{7}$$

This total objective function is optimized for $A^{(\text{OC})}$ with respect to W by scaled gradient ascent.

All simulations were performed using MATLAB software (Mathworks, Natick, MA, USA).

2.3 Stimuli

All stimuli used in this study were derived from the Columbia University Object Image Library (COIL-100, Nayer et al. 1996). This database provides photographs of 100 objects from 72 different viewpoints each (5° steps). Objects are located in front of a black background at a original resolution of 128×128 pixels and three-channel (RGB) colour representation. Stimuli were used in two ways: in a plain condition, where one object ('target') was present on the retina, and in a 'cluttered' condition, where another object ('distractor') was placed on the retina in the background of the target. To allow for partial overlap in the cluttered condition from all 7200 stimuli, the background was removed before processing. This was done by pixel-wise applying of a threshold to the luminance (weighted-sum of all color channels) and such that rather parts of the object boundary were cut than leaving black fringes around the object. In both conditions objects were placed at a random position on a 256×256 pixel wide retina. The random position mimics the relatively fast change of position in comparison to other transformations such as changes in viewpoint. However, since this study does not deal with learning translation invariance, which is built-in by pooling over CCs, we did not mean to realistically model this relative timing and thus chose the extreme case of random positions. The 256×256 pixels are finally down-sampled to 64×64 using bi-cubic interpolation (Fig. 1b), which in combination with the random position adds additional variation to the input signal for each object.

For training only certain views of the objects were used, the number of training views per object being a divider of the total 72 views. From one time-step to next the viewpoint of each object was changed by 360°/(number of training views), while the position was randomly selected independently across time-steps (Fig. 1b). All views of each object were presented 100 times. In the cluttered condition the distractor was selected and placed independently for each step (Fig. 1c). Distractor objects were always taken from the whole database even if the network was trained only on a subset.

For testing, all 72 views of each trained object were presented to the converged network 100 times at random position. For testing in the cluttered condition a distractor was added for each test presentation. The object used as this testing distractor was chosen randomly from a subset of objects disjoint from the trained objects. That means, if objects #1–#10 were used as targets for training, objects #11–#100 were used as distractors; for cells trained with object #1–#50 as target, testing distractors were objects #51–#100. To ensure a testing distractor set of at least equal size as the trained target set, the maximum number of trained objects was restricted to 50.

For testing generalization of the converged network to previously unseen objects, we performed the testing on objects #1–#10, while the training was performed on various numbers (10, 20, ...) of objects. In one set of these simulations, the tested stimuli were included, i.e., we trained on stimuli #1–#10, #1–#20, etc. In a second set of simulations, the tested objects were not part of the training set, i.e., we trained the network on objects#11–#20, #11–#30, etc.

2.4 Quantifying viewpoint-invariance

In order to quantify viewpoint-invariance, we probed CCs as well as converged OCs with all views of the objects used for training (note that for the training usually only a subset of views was used). Each view of each object is presented 100 times at different locations and – for simulations including distractors – with changing random distractors. The cell activity was averaged over these 100 presentations and this average will be referred to as the *response* of the cell to a certain view of a certain object. The response of each cell over this complete set of objects and views was then normalized to zero mean and unit standard deviation. To quantify a cell's dependence on the viewpoint we take the standard deviation of the response over all views of each object. One minus this standard deviation defines a measure of *viewpoint invariance* for each cell and object. Note that viewpoint invariance can be calculated for each object separately to distinguish effects that result from object inherent rotational symmetries from those that are an acquired property of a cell.

2.5 Classification performance

In order to analyse how well OC and CC outputs can be used for object classification, we performed a clustering on the cells' activities during testing. We repeatedly applied the k–means algorithm (implementation from MATLAB's statistics toolbox) with ten different random initial settings. The number of clusters was chosen to match the number of objects. To assess the number of correctly classified objects, each cluster had to be assigned an object, which was done by the procedure described in the appendix. The fraction of correctly classified test presentations was computed independently for each repetition of the k-means algorithm and the mean across repetitions is referred to as unsupervised classification performance.

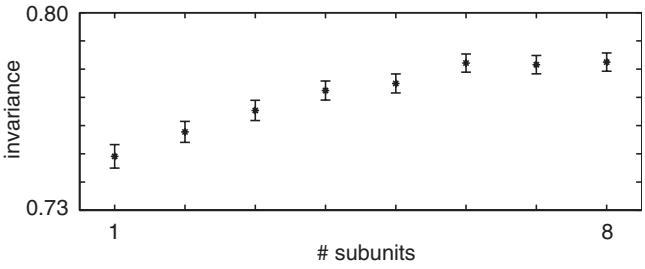

Fig. 5 Dependence on network parameters. Dependence of invariance on the number of subunits of each OC. Mean and standard error over objects and OCs

3 Results

3.1 Single cell properties

We train OCs to optimize the stability objective on a subset of viewpoints of several objects of the COIL-100 database. After convergence we probe both OCs and their inputs (CCs) on all trained objects from all viewpoints. The response trace of one example cell of each cell type is depicted in Fig. 2: For objects, whose appearance change considerably with changing viewpoint, the response of the CC dramatically depends on the viewpoint (Fig. 2, top row). The response of the OC, on the other hand, does not exhibit a strong modulation while the same object is presented, but responds rather independently of the viewpoint (Fig. 2, bottom row). Consequently the OC can be characterized view-point invariant cell.

We now investigate to what extent the discussed example is typical for the whole population of cells. Unless otherwise stated, all data presented in the successive paragraphs refer to a simulation, in which OCs were trained on 12 views (30° steps) of 50 objects in the absence of a distractor. We first measure viewpoint invariance – as defined in the methods – for all objects. Averaged over all objects, the mean OC invariance index (0.78, Fig. 3b) is significantly larger than the mean CC invariance index (0.67, Fig. 3a; $p < 10^{-10}$, t-test). Next we test to what extent this result also holds for individual objects (Fig. 3c). For 27 objects the mean OC invariance index is indeed significantly larger (at $p < 0.05$, t-test) than the mean CC invariance index (green points in Fig. 3c), while the reverse is the case for only seven objects (red points in Fig. 3c). Since the distribution for CC invariance indices is not normal for all objects (Fig. 3d), we in addition perform a Wilcoxon test on the difference of medians. At a significance level of $p = 0.05$ this test yields the same result as the t-test for each individual object. All of the seven objects, for which OC invariance is significantly worse than CC invariance, already have a very high invariance index for CC (>0.82). These objects themselves are inherently rotation invariant, as they resemble cylinders, toroids or spheres (marked by red boxes in Fig. 3d). On the contrary, objects whose CC invariance is rather low, exhibit especially improved OC invariance. In total there are significantly more objects for which OC invariance is larger than CC invari-

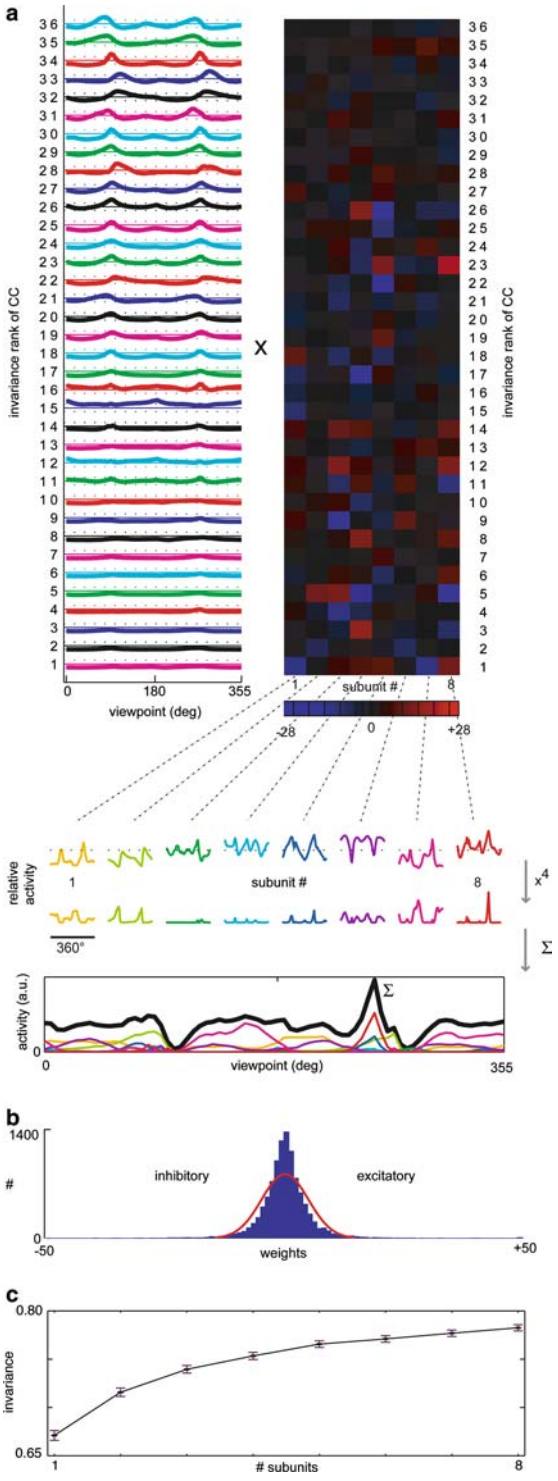

Fig. 6 Improvement of invariance from CCs to OCs. **a** Example of how the improved invariance in OC activity arises. Example OC is the same as in Fig. 2, response to object #1 is depicted. *Left column*: Response of all CCs sorted by viewpoint invariance to this object. Spacing between baselines corresponds to three standard deviations (over all objects) in activity. Colors are used to easily assign each activity trace to its baseline. *Right column*: Weight matrix W of projections of CCs (rows) to an OC's subunits (columns). Weights are pseudo-color-coded as indicated by the color-bar beneath W (red–excitatory, blue–inhibitory). *Third row from bottom*: activity traces of the subunits before fourth power non-linearity ($A^{(CC)} * W_{\ldots,s}$). *Second row from bottom*: forth power of subunit activity (non-normalized as used for OC activity computation). *Bottom row*: Forth power of OC activity (*thick black line*) and of subunit activities (*thin lines* – colors as in rows above). Forth root and normalization have been omitted to show how subunit activities combine to OC activity. **b** Distribution of weights for all subunits of all OCs (*blue histogram*). For comparison a normal distribution with identical mean and variance is depicted in *red*. **c** Invariance versus number of subunits used for computing OC activity. Mean and standard error over objects and OCs. Note, that all data of this panel is from a simulation trained with eight subunits, while Fig. 5 shows data of eight simulations *trained* with different number of subunits

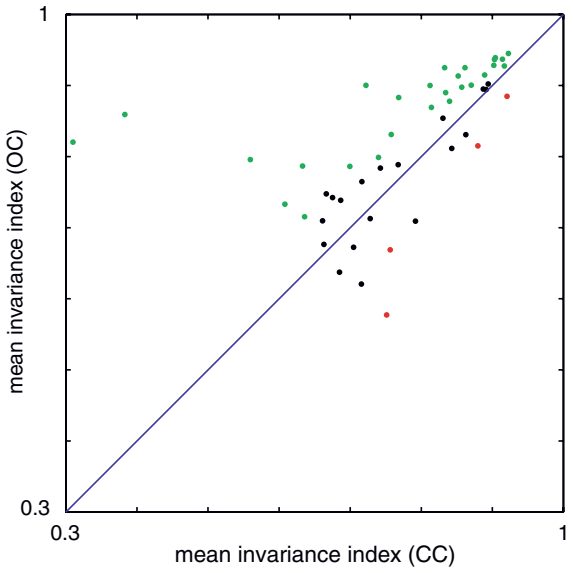

Fig. 7 Results with distractor. Analogous to Fig. 3c for cluttered condition: Mean OC invariance for individual objects plotted versus mean CC invariance. OCs are trained in the presence of a distractor, both cell types are tested in presence of distractor. All other parameters unchanged as compared to Fig. 3c

ance (36 objects) than there are objects for which the reverse holds (14 objects $p = 0.005$, sign-test). In summary, optimizing the stability objective leads to an improvement of viewpoint invariance for most objects and this improvement is especially prominent for objects that show little intrinsic invariance.

To further investigate how the stability objective relates to viewpoint invariance, we directly compare the individual stability values of each cell to its mean viewpoint invariance (averaged over objects). As expected – since OCs are trained to optimize stability, the individual stability values of CCs (mean: -0.046) are significantly smaller ($p < 10^{-4}$, t-test) than those of OCs (mean: -0.038). The same is true for the mean viewpoint invariance (CC: 0.671, OC: 0.783, $p < 10^{-11}$, t-test). Nevertheless, there is a highly significant correlation between viewpoint invariance and individual stability for both CCs ($r = 0.71$, $p < 10^{-5}$ Fig. 4a), and OCs ($r = 0.53$, $p < 10^{-3}$, Fig. 4b). Note, that the difference in correlation values might be partly due to a lack of low invariance/stability values for OCs. Indeed the correlation for the combined population of CCs and OCs also yields a highly significant result ($r = 0.74$, $p < 10^{-10}$). This shows that stability is indeed a good correlate of viewpoint invariance, and the optimization of stability is the key to the improved viewpoint invariance of the OCs as compared to CCs.

3.2 Dependence on the number of subunits

All simulations reported so far used eight subunits ($S = 8$). Do the observed invariance properties critically depend on this choice? We perform additional simulations varying the number of subunits from 1 to 8 (Fig. 5). We find that invariance depends only slightly – but significantly – on the number of subunits ($p < 10^{-15}$, ANOVA). However, invariance increases monotonically only up to six subunits and then saturates. This shows that increasing the number of subunits further does not affect invariance anymore. Consequently eight subunits is a reasonable choice, although this choice does not critically affect performance (see Sect. 3.5 below).

3.3 Improvement of invariance from CCs to OCs

In order to achieve a sound understanding as to how OCs achieve an improved invariance by appropriately combining their CC afferents, we more closely analyse these projections. Fig. 6a illustrates how OC activity derives from CC activity. Most CCs exhibit a strong modulation with viewpoint (Fig. 6a, left). However, only few of the CCs have strong projections to a given OC subunit (Fig. 6a, right). As the viewpoint-dependent modulation is strong for inhibitory as well as for excitatory projections to a subunit, the peaks get flattened out (Fig. 6a, third row from the bottom). After passing through the even non-linearity the negative troughs become peaks (Fig. 6a, second row from the bottom). Consequently the sum over subunits only shows a weak and biphasic modulation with viewpoint (Fig. 6a, bottom), and thus increased viewpoint invariance. The chosen OC exemplifies three major factors that shape the OC response: (1) only few CCs have strong connections to a given subunit, i.e., the connectivity is sparse; (2) subunit activities get more invariant by excitatory and inhibitory projections from CCs, that antagonize each other; (3) subunits that receive mainly inhibitory and excitatory projections antagonize each other and flatten the response further; while the absolute integral, and thus the object-specific response, is persevered by the even non-linearity. To quantify these observations for the whole population of cells we first analyse connectivity weights over all cells and subunits, and find that this distribution is indeed sparse, i.e., heavy-tailed (Fig. 6b). We measure subunit invariance analogously to CC and OC invariance. We find that the mean invariance of the subunits (0.72) is inbetween those of CCs (0.67) and OCs (0.78), and significantly different from either ($p < 0.005$ for both pair-wise t-tests). To test how many subunits are actually needed to achieve invariance, we compute each OC's activity only including its n most invariant subunits. We find that invariance increases gradually with an increasing number of subunits (Fig. 6c). Adding an additional subunit leads to a significant increase compared to leaving that subunit out until five subunits are included (pair-wise t-tests: $1 \rightarrow 2$: $p < 10^{-11}$; $2 \rightarrow 3$: $p < 10^{-4}$; $3 \rightarrow 4$: $p < 0.01$; $4 \rightarrow 5$: $p < 0.05$; ANOVA over all n: $p < 10^{-15}$). This shows that the mutually antagonizing effect of two or more subunits is a factor in the improvement of invariance. In conclusion, the three factors leading to the improved OC invariance described in the example of panel 6a (sparse connectivity, antagonizing CCs→subunit projections, antagonizing subunit→OC projections), are typical for the invariance properties of all OCs.

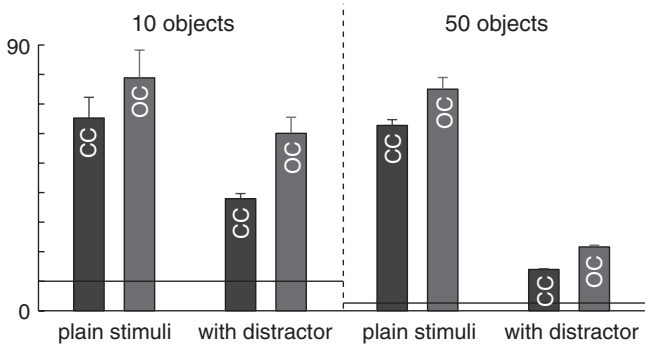

Fig. 8 Unsupervised clustering. Classification performance for k-means clustering on CCs and OCs using 10 and 50 objects in presence and absence of distractor. *Horizontal lines* indicate chance level (10 and 2% respectively); *error-bars* denote standard deviation over ten repetitions of k-means clustering

3.4 Training and testing in the presence of a distractor

We have shown so far that optimizing the stability objective leads to viewpoint invariant cells when trained on plain stimuli. Do these results persist if a distractor is present as well during training and testing? We train and OCs with the same stimuli and parameters as in the baseline situation (50 objects, 12 views), but add to each presentation a randomly chosen second stimulus as distractor. For testing, the distractors are taken only out of the 50 objects, that are not trained. In the simulation with distractor, viewpoint invariance is significantly larger (at $p < 0.05$, t-test or Wilcoxon-test) for OCs than for CCs in 27 objects, while the reverse only holds for four objects (Fig. 7). In total, OC invariance exceeds CC invariance for a significant number of images (39 out of 50 objects, $p < 10^{-4}$, sign-test). This implies that the difference between OCs and CCs in viewpoint invariance is not only conserved for training and testing with distractors but gets even more pronounced. This shows that the stability objective is also well suited to suppress distractors on the basis of input statistics.

3.5 Unsupervised clustering

The stability objective increases viewpoint invariance for individual cells, but does this also facilitate invariant object classification on the population level? In order to test this issue, we perform an unsupervised clustering on both the CC and OC activities. As before, OCs are trained on a subset of 12 views per object (30° steps). After convergence k-means clustering is performed on the activities in response to all 72 views of all trained objects. In the absence of a distractor, classification performance reaches 79% (ten objects, Fig. 8 left) and 75% (50 objects, Fig. 8 right) for OCs, which is larger than the 65% and 62% reached for CCs. These values are obtained as mean over ten different random initial conditions for the clustering algorithm and the observed difference between CCs and OCs is highly significant ($p < 0.005$ and $p < 10^{-7}$, t-test).

In order to ensure that this result does not depend on the specific design of the network, we tested the performance for different numbers of subunits. While OC performance for $S = 1$ is slightly smaller (72%) than for higher numbers of subunits, there is no significant dependence of classification performance on subunit number ($p = 0.27$, ANOVA). This demonstrates that nearly optimal performance is already reached with one subunit. The fact that performance does not worsen for higher number of subunits (and invariance even increases – see 3.2), in turn shows that this increase in the number of optimized parameters does not lead to a loss in generalization performance. Hence – even for large numbers of subunits – there is no "overfitting" of training views. In summary, the choice of the exact number of subunits is not critical.

When training and testing is performed in the presence of a distractor the difference of classification performance between OCs and CCs becomes even more remarkable: both for 10 and 50 objects, the relative increase in classification performance is over 50% (from 38 to 60% and from 14 to 22%) and are both highly significant ($p < 10^{-9}$, t-test). In the simulations discussed so far, a rather large number of views (12) was presented for training the OCs. Do the results persist if this number is reduced? We tested classification performance for OCs trained with 50 objects and three different views per object (120° steps). The results do not differ remarkably, still OCs classification performance reaches 71% (50 objects) in the absence of a distractor and 20% in the presence of a distractor. In both conditions the OC performance is still significantly higher than CC performance ($p < 10^{-5}$, t-test).

Concluding, even for a small number of training views and in the presence of a distractor, unsupervised classification performance is significantly increased for cells trained with the stability objective (OCs) as compared to their afferents (CCs). Consequently, the stability objective is a valuable pre-processing step for viewpoint-invariant object classification, especially if the object to be classified cannot readily be segmented from the background and can be trained from few viewpoints only.

3.6 Generalization to novel objects

So far we have tested classification performance only on the set of objects the OCs had also been trained with. While the network generalizes well to previously unseen viewpoints, we may not expect an equally good generalization to novel objects, since the object sample is small and rather heterogeneous. We train OCs with different number of objects (10, 20, ...) and test their classification performance on the first ten objects only. In one set of simulations the test objects are included in the training set, in another set they are excluded (Fig. 9). We perform a two-way ANOVA to test whether performance depends on the number of training objects and/or on whether or not the test objects are part of the training set. First we find no difference between the two types of simulations ($p = 0.33$, ANOVA), OCs' performance does not

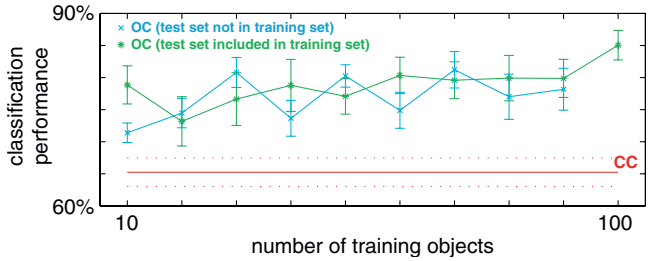

Fig. 9 Generalization over objects. Classification performance on ten objects for OCs trained with different numbers of objects. *Cyan crosses*: test objects were not included in training set; *green stars*: test objects were included in the training set. *Solid red line*: CC classification performance. *Errorbars* and dashed lines denote standard errors of the mean over ten repetitions of k-means clustering

depend on whether or not the tested objects are included in the training set. Second, we do not observe a decrease in performance when more objects are added to the training set ($p = 0.19$, ANOVA); if anything, the performance tends to be better for larger training sets: OCs trained with 100 objects show a performance of 85% correct compared to the 79% when only the test objects are used for training. In conjunction with the independence of whether or not tested objects are in the training set at all, this demonstrates that OCs do not only generalize over unseen viewpoints but also to some extent over unseen objects.

4 Discussion

We show that unsupervised training with a general optimization principle facilitates the classification of real-world stimuli. The facilitation still holds, if stimuli undergo a complex, non-isomorphic, transformation – in our case the change of viewpoint – and only few examples are seen during training. The degree of facilitation is furthermore maintained in the presence of a distractor during training and classification. This result rests on the fact that cells trained to optimize a temporal coherence objective represent the input more invariant to viewpoint than their afferent cells. This result is especially remarkable as it is achieved in a homogenous architecture: already the input neurons, modeled as complex cells, can be understood as the temporally coherent features of a natural stimulus sequence (Kayser et al. 2001; Berkes and Wiskott 2003; Körding et al. 2004). Nevertheless, the degree of invariance still increases for the second layer trained here. Hence, the arguably most difficult step in object classification, namely achieving a representation of the input, which is invariant under the desired transformations, does not require a large set of labeled examples. Instead, training on unlabeled examples according to a general principle like temporal coherence, generates such a representation. A simple classification algorithm operating on the resulting representation then suffices for successful classification.

The results of the present study are rather insensitive to the choice of the exact network parameters. In particular, there is no dependence of classification performance, and only a slight – though significant – dependence of the invariance measure, on the number of subunits. This shows, that – for the dataset used – the generic principle of temporal coherence combined with a non-linear operation suffices to achieve increased performance with a linear classifier. It is important, however, that generalization does not worsen with an increasing number of subunits. Therefore the system can readily be extended to more complex sets of inputs. Having multiple subunits is especially desirable, when pooling over a topographically organized input space shall be learnt. In the present study, we implemented this pooling already in the input (CC) representation, as previous studies had demonstrated that such representations can be learnt by optimizing the stability objective (Körding et al. 2004). Such a requirement for multiple functional subunits at early levels of processing has also received recent experimental support by electrophysiological data of CCs in cat visual cortex (Touryan et al. 2002). While the present study does not include any topographical organization of cells, the robustness of the results to the number of subunits makes the use of the stability objective in a topographically organized hierarchical multilayer network a promising approach for future research.

A large body of recent studies applies the temporal coherence principle to achieve invariant representations and to explain some properties of the visual system (Födiak 1991; Stone 1996; Wallis and Rolls 1997; Rolls and Milward 2000; Kayser et al. 2001; Einhäuser et al. 2002; Stringer and Rolls 2002; Wiskott and Sejnowski 2002; Berkes and Wiskott 2003; Hurri and Hyvärinen 2003; Wiskott 2003; Körding et al. 2004). These studies as well as those on similar coding schemes (see Olshausen 2002, for review) are often taken as evidence, that the functional organization of large parts of the visual cortex can be derived from such general principles. The important criterion during evolution and in development of technical applications, however, is whether such a coding principle boosts system's performance. Invariant object classification of real-world objects under complex, i.e., non-isomorphic, transformations is thus a critical test case for the above conjuncture. Only few of the aforementioned studies, however, fully address this issue: Wallis and Rolls (1997) show that cells of their trace-rule based 'VISNET' architecture can acquire translation and viewpoint invariance for face stimuli. Thereby they delivered the proof of concept, that temporal coherence-based learning rules in principle can learn viewpoint invariant representations. The analysis of their results is mainly based on a so-called "discrimination factor", which does not exclude that information on the stimulus is lost across their hierarchical levels. While in case of translation invariance the authors control for this possibility by adding a simple supervised classifier to the output of their network, they do not report a similar control for viewpoint invariance. Furthermore, the number of stimuli to be discriminated in their viewpoint invariance experiment is very small (3). Although more recent studies on a similar architecture ("VISNET2") use more stimuli, the authors either restrict themselves to translation invariance (Rolls and Milward 2000) or do not use real-world stimuli (Stringer and

Rolls 2002). Consequently the present study is the first to apply the temporal coherence principle to viewpoint invariant classification of a large number of real-world objects. In addition – unlike Wallis and Rolls (1997) – we demonstrate the suppression of distractors and generalization to unseen viewpoints. Hence our study is the first to show the functional usefulness of a general coding principle like temporal coherence on the system level.

Learning invariance to complex transformations has also been addressed in other physiologically inspired network models. Mel's (1997) carefully designed "SEEMORE" architecture allows very reliable recognition of the trained objects from novel viewpoints and is robust to various other distortions. SEEMORE uses a large variety of adhoc pre-defined feature channels to reach high classification rates. Here we restricted ourselves to very few simple input 'feature' channels (color, orientation and spatial frequency preferences of the CCs), in conjunction with an adaptive network structure. Hence, although engineered solutions like SEEMORE might reach higher classification rates for specific problems, the appropriate combination with general learning principles like temporal coherence therefore reduces design and computation costs.

Feature-based recognition systems often rely on the precise segmentation of the object from its background. Wersing and Körner (2003) show that a network trained by the principle of sparse-coding achieves some degree of viewpoint invariance for classifying real-world objects in cluttered scenes. Here we show that the principle of temporal coherence generates viewpoint invariant representations. We do not use a fully cluttered background but just a single distractor. It is hence likely that most information extracted by our system is still based on the outline of each object, especially since the retinal resolution is low and using fine internal object structure thus difficult. These considerations could be addressed by a complementary segregation based on independent cues, like motion, which are presently not used. Concluding the present study provides an unsupervised learning scheme that is insensitive to the background during both classification and training.

Two dominant views exist with respect to human recognition of objects at novel viewpoints. So-called "structural-description" or "object-centred" approaches, such as Biederman's (1987) "Recognition-by-Components" model, postulate an object to be defined by the arrangement of a small set of generic features ("geons"). Viewpoint invariant recognition in this view is achieved, since the relative configuration of these geons do not change when the object is rotated in depth. "Image-based" or "view-based" models in turn postulate the existence of a small number of two-dimensional templates for objects and propose several mechanisms that could allow generalization to novel viewpoints, such as mental rotation (Tarr and Pinker 1989), mapping to a standard ("canonical") view generated from few example views (Poggio and Edelman 1990) or linear combination of few familiar views of the object (Ullman and Basri 1991). The controversy between both types of theories has remained pretty much alive and has prepared the ground for considerable theoretical and experimental progress (Tarr and Bülthoff 1998; Biederman 2000 for reviews). Here we do not make use of any structural description, instead our network model extracts the relevant features from only a few views to generalize novel views and – to some extent – novel objects. While we hereby demonstrate that a general optimization scheme can generate a representation, as it is needed for "view-centred" models, it will be an interesting issue for further research, to what extent relevant structural information could also be extracted by similar principles.

When learning takes places in the presence of distractor, learning of the specific target–distractor combination needs to be avoided. Instead the isolated target has to be learnt. This can be achieved by a variety of segregation mechanisms. Human and non-human primates can achieve segregation by selectively directing attention to the target (Desimone and Duncan 1995). When the target to be trained is presented with a sufficient number of different distractors, this statistical property can be exploited. Our network model uses the higher temporal coherence of the target relative to the distractor. While the cluttered condition surely exaggerates the difference in temporal coherence between target and distractor compared to real-world situations, it directly relates to the definition of an object in the tradition of Gestalt psychologists. The law of common fate binds different features of one object and sets them apart from features of other objects. The temporal stability defines an object and is contrasted by the faster changes of the arrangement of different objects to a visual scene. Therefore the present study provides the proof of concept that bottom–up mechanisms, i.e., mechanisms based on the input statistics, may contribute to suppression of distractors. It will be a fruitful approach for future research to incorporate top–down interactions in the present bottom–up scheme, which learns relevant representations unsupervised from the natural statistics of the input.

Acknowledgements This work was supported by the EU-AMOUSE (JH) project and the Swiss National Science Foundation (PK, grant-no. 31-61415.01). We are grateful to K.P. Körding for making his MATLAB code for optimizing the stability objective available.

Appendix

Since the k-means clustering is an unsupervised algorithm, determining the fraction for correctly classified object presentations requires assigning each cluster to an object. This is done by the following procedure:

Generate a hit-matrix $\mathrm{H}^{(0)}$ by assigning each cluster an object at random.

For all n from 1 to the number of objects (= number of clusters)

- Find the maximum entry $(k_n, \mathrm{l}_n) = \arg\max_{(i,j)}(\mathrm{H}^{(n-1)}(i, j))$ in this hit matrix,
- Assign cluster k_n to object l_n.

- delete row k_n and column l_n from hit matrix: $H^{(n)}(k_n,.) = 0$, $H^{(n)}(.,l_n) = 0$; for all other i, j: $H^{(n)}(i, j) = H^{(n-1)}(i, j)$

Create the final hit-matrix H by rearranging the columns of $H^{(0)}$: For all n $H(.,k_n) = H^{(0)}(.,l_n)$.

The fraction of correctly classified test classifications is now given by the sum over the diagonal entries of H by the sum over all entries.

References

Berkes P, Wiskott L (2003) Slow feature analysis yields a rich repertoire of complex-cell properties. Cognit Sci EPrint Arch (CogPrints) 2804, http://cogprints.ecs.soton.ac.uk/archive/00002804/

Betsch BY, Einhäuser W, Körding KP, König P (2004) The world from a cat's perspective – statistics of natural videos. Biol Cybern 90:41–50

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94(2):115–147

Biederman I (2000) Recognizing depth-rotated objects: a review of recent research and theory. Spat Vis 13:241–253

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222

Einhäuser W, Kayser C, König P, Körding KP (2002) Learning the invariance properties of complex cells from their responses to natural stimuli. Eur J Neurosci 15:475–486

Einhäuser W, Kayser C, Körding KP, König P (2003) Learning distinct and complementary feature-selectivities from natural colour videos. Rev Neurosci 14:43–52

Földiak P (1991) Learning Invariance from Transformation Sequences. Neural Comput 3:194–200

Franzius M, Einhäuser W, König P, Körding KP (2005) Learning a hierarchical model of cortical function from natural stimuli. (submitted)

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Hurri J, Hyvärinen A (2003) Simple-Cell-Like Receptive Fields Maximize Temporal Coherence in Natural Video. Neural Comput 15(3):663–691

Kayser C, Einhäuser W, Dümmer O, König P, Körding KP (2001) Extracting slow subspaces from natural videos leads to complex cells. In: Dorffner G, Bischoff H, Hornik K (eds) Artificial neural networks – (ICANN) LNCS 2130, Springer, Berlin Heidelberg New York, pp 1075–1080

Kayser C, Einhäuser W, König P (2003a) Temporal correlations of orientations in natural scenes. Neurocomputing 52:117–123

Kayser C, Körding KP, König P (2003b) Learning the nonlinearity of neurons from natural visual stimuli. Neural Comput 15:1751–1759

Körding KP, Kayser, C, Einhäuser W, König P (2004) How are complex cell properties adapted to the statistics of natural stimuli? J Neurophysiol 91:206–212

Mel BW (1997) SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. Neural Comput 9(4):777–804

Nayer SK, Nene SA, Murase H (1996) Real Time 100 object recognition system. In: Proceedings of ARPA Image Understanding Workshop. Morgan Kaufmann, San Matteo

Olshausen BA (2002) Principles of image representation in visual cortex. In: Chalupa LM, Werner JS (eds) The visual neurosciences, MIT Press, Cambridge

Poggio T, Edelman S (1990) A network that learns to recognize three-dimensional objects. Nature 343(6255):263–266

Rolls ET, Milward T (2000) A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures. Neural Comput 12:2547–2572

Stone JV (1996) Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. Neural Comput 8:1463–1492

Stringer SM, Rolls ET (2002) Invariant object recognition in the visual system with novel views of 3D objects. Neural Comput 14:2585–2596

Tarr MJ, Pinker S (1989) Mental rotation and orientation-dependence in shape recognition. Cognit Psychol 21(2):233–282

Tarr MJ, Bülthoff HH (1998) Image-based object recognition in man, monkey and machine. Cognition 67:1–20

Touryan J, Lau B, Dan Y (2002) Isolation of relevant visual features from random stimuli for cortical complex cells. J Neurosci 22:10811–10818

Ullman S, Basri R (1991) Recognition by linear combinations of models. IEEE Trans Pattern Anal Mach Intell 13(10):992–1006

Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual systems. Prog Neurobiol 51:167–194

Wersing H, Körner E (2003) Learning optimized features for hierarchical models of invariant object recognition. Neural Comput 15:1559–1588

Wiskott L, Sejnowski T (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comput 14:715–770

Wiskott L (2003) Slow feature analysis: a theoretical analysis of optimal free responses. Neural Comput 15(9):2147–2177