Learning Invariant Visual Shape Representations from Physics

Mathias Franzius and Heiko Wersing

Honda Research Institute Europe GmbH

Abstract. 3D shape determines an object's physical properties to a large degree. In this article, we introduce an autonomous learning system for categorizing 3D shape of simulated objects from single views. The system extends an unsupervised bottom-up learning architecture based on the slowness principle with top-down information derived from the physical behavior of objects. The unsupervised bottom-up learning leads to pose invariant representations. Shape specificity is then integrated as top-down information from the movement trajectories of the objects. As a result, the system can categorize 3D object shape from a single static object view without supervised postprocessing.

1 Introduction

Invariant shape recognition is a hard problem because many three-dimensional objects undergo extreme appearance variations when they rotate in-depth or light conditions change. Slowness learning can be used to learn this invariance: even if views of the same object are very different, they more often occur in close temporal relationship during object interaction than views of distinct objects [1,7,3]. Such models allow object identification since different views of the same object form clusters in the learned feature space and views of visually distinct objects typically cluster in distinct regions. However, objects of similar categories do not generally appear more often in close temporal relationship than objects of different categories. The relative distances of object clusters thus remain undetermined after SFA learning, and minimal noise can cause object clusters to permute their positions in different simulations. Here, we extend a model for invariant object recognition such that a small autonomously generated additional input signal determines the relative positions of object clusters and thus implements a meaningful shape similarity measure. Although in general the problem of deriving 3D object shape from a single view is ill-defined, humans can often already guess an object's shape from a single view and predict how it would move when it is agitated. We demonstrate here a model of this behavior for a limited object set.

2 Methods

Twelve objects of four distinct shapes (cone, cube, sphere, capsule) and three colors (red, green, blue) were dropped into a box with tilted ground plane and their

K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part III, LNCS 6354, pp. 298–302, 2010.

behavior was simulated with a physics engine. In each time step, a bounding box around the object was estimated from the visual data and a quadratic section of 50×50 RGB pixels containing the object was cut out. Thus, the visible transformations of the object consisted of in-depth and in-plane rotations, scaling, small positional jitter, and change of lighting direction. After dropping an object onto the surface, 400 video frames were recorded with a framerate of 10Hz, followed by an all-black image after which the process repeats with another object dropping into the box. The process was repeated until a total of 100,000 views and 250 trajectories, corresponding to $2\frac{3}{4}$ hours, were generated.

Two properties of the moving objects were measured: the total time until they come to a stop and the distance traveled downhill. From these 250 two-dimensional trajectory descriptors 120 were randomly selected and classified by k-means ($k = 4$). The resulting trajectory clusters T coincide to a large degree with object shape and thus the fact that the trajectories of two objects falls into the same trajectory cluster can be used as a learning signal. Subsequently, each training object trajectory was classified as belonging to one of these clusters. Proportionally to the co-occurrence of two objects (A, B) in T, additional randomly selected view pairs (V_A, V_B) were presented to the system after the first unsupervised learning phase. Thus, the system was additionally trained with sequences of object pair views, which are likely to have similar shape. While all views of the object movies were weighted equally during training, the learning weight was increased by a factor of 10 for the additional shape training views.

The slowness objective was optimized using Slow Feature analysis (SFA) [9]. As the image dimensionality of 7500 is too large to perform nonlinear SFA in a single step, we employed a hierarchical network architecture as described in [3].

For analyzing learned representations, we performed unsupervised k-means clustering. Afterwards, the feature representations of all training views are assigned to the closest cluster center. Since k-means randomly permutes cluster identities in each run, we always identify the permutation of cluster names with the highest overlap to ground truth shape categorization.

3 Results

First we analyze the learned representations after training the hierarchical network with videos of objects dropped into the box with a tilted floor without any further shape-related training. For this purpose, we compute the average values (i.e., the cluster centers) of each object in the slowest ten components in the highest layer of the hierarchical network. The left panel of Fig. 1 shows the distance matrix between all pairs of clusters after normalizing the highest distance to 1. As expected, most clusters are roughly equidistant (with the exception of the cone clusters). These distances fluctuate for repeated simulations due to the small amounts of noise injected into the hierarchy. In this representation, a view of a green cube, for example, is on average as similar to a red cube as to a red sphere. Except for the distances between cones, there is no evident clustering of views from objects with same shape or color.

In a second step, we investigate how well the distance matrix of object view clusters can be “programmed” explicitly when only correct pairs of views of objects from the same shape category are shown. For this purpose, 100 views of each shape training view pairs (i.e., $\{$(red,green), (red, blue), (green,blue)$\} \times \{$cone, cube, sphere, capsule$\}$) were presented to the system additionally to the videos of the dropped objects. The central panel of Fig. 1 shows the resulting distance matrix. Here, all clusters of views of objects with identical shape are very close to each other, whereas all cross-shape cluster distances are roughly equal and close to the maximum distance. As desired, in this representation, views of objects of the same shape are categorized as similar and views of objects with distinct shapes are considered distinct. This observation is quantified by performing k-means (k=4) clustering of the view feature representations and assigning each view to the closest cluster center. On average over 20 trials, 99.5% of object views were assigned to the correct shape cluster.

Finally, we characterize the full system with unsupervised top-down learning. As before, all layers are trained with the movies of objects dropped into the box. Again, the movement trajectories of all objects are measured and clustered using k-means (k=4). More than 90% of the trajectories cluster consistently with a shape class. After training all layers of the network with the dropped object movies, pairs of object views are presented with a likelihood proportional to the frequencies of co-occurrance of objects in the trajectory clusters T. The right panel of Fig. 1 shows the resulting distance matrix after the presentation of 400 such randomly selected pairs. Again, we perform k-means (k=4) clustering of the view feature representations and assigning each view to the closest cluster center. Similar to the results with the additional supervised training views above, this autonomously learned representation categorizes views by object shape. On average, 98.0% of all views are categorized as belonging to the same shape, and, on average, less than 100 view pairs are sufficient to reach 90% shape categorization performance. Note that this number is less than one view pair per object pair ($12x12 = 144$) and that many presented view pairs show the same object (i.e., same color and same shape).

For comparison, we quantified the sizes of shape and color clusters in the pixel space. The average cluster diameter of all views of objects with the same color is 5.9 times smaller than the average size of clusters of all views of objects with identical shape. Performing 100 times k-means clustering (k$\in \{3, 4\}$) in the raw pixel space never achieved a higher overlap than 82% with either shape or color clusters in 100 trials but on average the overlap with the color clusters was 41% larger.

4 Discussion

We have presented a system for unsupervised learning of a visual feature representation that clusters views of objects with similar 3D shape. The system learns invariance to pose variations of the objects and color variations of objects within categories. The shape information is autonomously derived from the movement

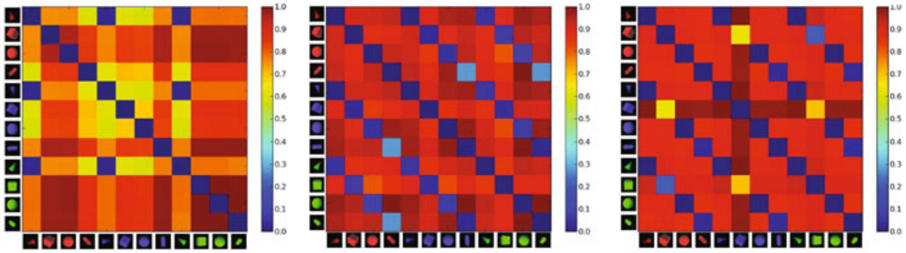

Fig. 1. Cluster center distances. Representations of single object views cluster in the learned feature space. Each subplot depicts normalized pairwise distances between the cluster centers of any given object in the learned features space. Left: Without any inter-object similarities learned (i.e., completely unsupervised), cluster center distances are undetermined and change between simulations. Center: After additional supervised presentation of 100 view pairs between objects of similar shape, features of objects with similar shape cluster but keep a similar distance to unrelated shape clusters. Right: Based on the similarity of their physical movement trajectories ("self-supervised"), most features of objects with similar shape cluster as in the supervised case.

behavior of the objects when dropped onto a tilted surface. Thus, shape information gets directly integrated into the learned visual feature representation.

We have shown that without additional self-derived training views, most object representations cluster roughly equidistantly, i.e., they are distributed as localized clusters on a hypersphere in the learned feature space. Although the SFA optimization itself is deterministic and guaranteed to find the globally optimal solution, the positions of the object clusters permute in different simulations under the addition of a small amount of noise. Theoretically, a single mini-sequence consisting of one pair of views between two distinct objects A and B already leads to a symmetry breaking such that the clusters of A and B have a smaller distance to each other than any other cluster pair. This sensitivity to small perturbations is an ideal basis for learning from few examples, either supervised (as shown in the central subplot of Fig. 1) or autonomously (as shown in the right panel of the same figure). Practically, some intermediate visual cues from the visual hierarchy will likely bias these cluster distances in simulations, even when no temporal inter-object relationship has been experienced during training (left panel of Fig. 1). However, as we have shown, object shape can not trivially be clustered in the pixel space. Instead, object color is a more prominent cue. Thus, there is a bias in the input space to cluster by color and not by shape and the fact that the resulting representations become color-invariant and shape-specific demonstrates that the input categorization bias is small and can be "overwritten" even with few examples. Additionally, we have shown some robustness to noise in the self-derived additional shape training views, since the trajectory clustering does not perfectly coincide with the shape clusters.

This work is related to slowness-based invariant object recognition. While invariant object recognition with much more complex objects has been shown earlier [1,3], we restrict the object complexity here to four shapes and three colors for the sake of simplicity. While it seems likely that our system can find invariant representations for visually more complex objects, ground truth shape categories are less evident and thus objective evaluation is harder. However, one shape representation could be evaluated as better than another if it facilitates a given task. Such an integrated approach is an interesting subject for further research. The main difference to existing slowness-based invariant object recognition systems of our approach is the systematic integration of top-down cross-object similarity of discrete objects, specifically for learning 3D shape categories. A similar hierarchical model architecture has earlier been shown to model most known functional aspects of hippocampal spatial codes (i.e., place cells, head direction cells, spatial view cells and grid cells [2]). As the hippocampus is crucial as a memory hub and well-known for time-delayed replay of previous experiences [4], we hypothesize that if replay sequences of objects with similar movement trajectories occur more often than those of different shape, i.e., out of experienced temporal context but in new task-specific context, the hippocampus could implement a mechanism similar to the one proposed here.

Additionally, this work is related to affordance learning approaches from the developmental robotics community [5,6,8]. These approaches also show autonomous learning of affordances but use sophisticated robotic actuators, whereas we have shown our approach only for simulations. However, these approaches tend to employ much simpler visual features (e.g., nearest neighbor classification in the pixel space), whereas our approach focuses on the learning of visual feature representations with invariances to strongly changing visual stimuli.

References

1. Einhäuser, W., Hipp, J., Eggert, J., Körner, E., König, P.: Learning viewpoint invariant object representations using a temporal coherence principle. Biol. Cyber. 93, 79–90 (2005)
2. Franzius, M., Sprekeler, H., Wiskott, L.: Slowness and sparseness lead to place, head-diretion and spatial-view cells. PLoS Comp. Biol. 3(8), e166 (2007)
3. Franzius, M., Wilbert, N., Wiskott, L.: Invariant object recognition with slow feature analysis. In: Kůrková, V., Neruda, R., Koutník, J. (eds.) ICANN 2008, Part I. LNCS, vol. 5163, pp. 961–970. Springer, Heidelberg (2008)
4. Gupta, A., van der Meer, M., Touretzky, D., Redish, A.: Hippocampal replay is not a simple function of experience. Neuron 65(5), 695–705 (2010)
5. Metta, G., Fitzpatrick, P.: Better vision through manipulation. In: Proc. 2nd Inter. Workshop on Epigenetic Robotics, vol. 11, pp. 109–128 (2002)
6. Ridge, B., Skočaj, D., Leonardis, A.: A system for learning basic object affordances using a self-organizing map. In: Proc. ICCS (2008)
7. Rolls, E.T., Stringer, S.M.: Invariant visual object recognition: A model, with lighting invariance. Journal of Physiology - Paris 100, 43–62 (2006)
8. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional object class detection based on learned affordance cues. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 435–444. Springer, Heidelberg (2008)
9. Wiskott, L., Sejnowski, T.: Slow feature analysis: Unsupervised learning of invariances. Neural Comp. 14(4), 715–770 (2002)