Proc. IEEE International Workshop on Biologically Motivated Computer Vision (BMCV2000), 15-17 May 2000 Seoul, Korea, to appear in Springer Lecture Notes in Computer Science.

# Distinctive Features Should Be Learned

Justus H. Piater and Roderic A. Grupen

University of Massachusetts, Amherst MA 01003, USA
{Piater, Grupen}@cs.umass.edu
http://www.cs.umass.edu/~piater

**Abstract.** Most existing machine vision systems perform recognition based on a fixed set of hand-crafted features, geometric models, or eigen-subspace decomposition. Drawing from psychology, neuroscience and intuition, we show that certain aspects of human performance in visual discrimination cannot be explained by any of these techniques. We argue that many practical recognition tasks for artificial vision systems operating under uncontrolled conditions critically depend on incremental learning. Loosely motivated by visuocortical processing, we present feature representations and learning methods that perform biologically plausible functions. The paper concludes with experimental results generated by our method.

## 1 Introduction

How flexible are the representations for visual recognition, encoded by the neurons of the human visual cortex? Are they predetermined by a fixed developmental schedule, or does their development depend on their stimulation? Does their development cease at some point during our maturation, or do they continue to evolve throughout our lifetime? For some of these questions, the answers have been well established. For example, the development of receptive fields in the early visual pathways is influenced by stimulation of the visual system. Some visual functions do not develop at all without adequate perceptual stimulation during a maturational *sensitive period*. Higher-order visual functions such as pattern discrimination capabilities are also subject to a developmental schedule. It is still debated to what extent feature learning for pattern discrimination continues throughout adulthood. Recent psychological studies indicate that humans are able to form new features if required by a discrimination task [11].

In contrast to the human visual system, most work on machine vision has not used learning at the level of feature detectors. In the following section, we briefly discuss visual object recognition by humans and machines, and we argue that low-level learning is an essential ingredient of a robust and general visual system. The remainder of the paper presents our experimental system for learning discriminative features for recognition.

This work was supported in part by the National Science Foundation under grants CISE/CDA-9703217, IRI-9704530 and IRI-9503687, by the Air Force Research Labs, IFTD (via DARPA) under grant F30602-97-2-0032, and by Hugin Expert A/S through a low-cost Ph.D. license of their Bayesian network library.

## 2 Feature Learning in Humans and Machines

How do humans learn recognition skills? Two principal hypotheses can be identified [9]: According to the Schema Hypothesis, sensory input is matched to internal *representations of objects* that are built and refined through experience. On the other hand, the Differentiation Hypothesis postulates that *contrastive relations* are learned that serve to distinguish among the items. Psychological evidence argues for a strong role of Differentiation learning [9, 13]. What exactly the discriminative features are and how they are discovered is unclear. It appears that feature discovery is a hard problem even for humans and takes a long time to learn [5]:

- Neonates can distinguish certain patterns, apparently based on statistical features like spatial intensity variance or contour density.
- Infants begin to note simple coarse-level geometric relationships, but perform poorly in the presence of distracting cues. They do not consistently pay attention to contours and shapes.
- At the age of about two years, children begin to discover fine-grained details and higher-order geometric relationships. However, attention is still limited to “salient” features [15].
- Over much of childhood, humans learn to discover distinctive features even if they are overshadowed by more salient distractors.

There is growing evidence that even adults learn new features when faced with a novel recognition task. In a typical experiment, subjects are presented with computer-generated renderings of unfamiliar objects that fall into categories based on specifically designed but unobvious features. After learning the categorization, the subjects are asked to categorize other objects that exhibit controlled variations of the diagnostic features, which reveals the features learned by the subjects. Schyns and Rodet [12] employed three categories of “Martian cells.” The first category was characterized by a feature called $X$, the second by a feature $Y$, and the third by a feature $XY$, which was a composite of $X$ and $Y$. Subjects were divided into two groups that differed in the order they had to learn the categories. Subjects in one group first learned to discriminate categories $X$ and $Y$ and then learned category $XY$, whereas the other group learned $XY$ and $X$ first, then $Y$. Subjects of the first group learned to categorize all objects based on two features ($X$ and $Y$), whereas the subjects of the second group learned three features, not realizing that $XY$ was a compound consisting of the other two. Evidently, feature generation was driven by the recognition task. For a summary of evidence for feature learning in adults, see a recent article [11].

Feature learning does not necessarily stop after learning a concept. Tanaka and Taylor [14] found that bird experts were as fast to recognize objects at the subordinate level (“robin”) as they were at the basic level (“bird”). In contrast, non-experts are consistently faster on basic-level discriminations as compared to subordinate-level discriminations. Gauthier and Tarr [4] trained novices to become experts on unfamiliar objects and obtained similar results. These findings indicate that the way experts perform recognition is qualitatively different than novices. We suggest that experts have developed specialized features, facilitating rapid and reliable recognition in their domain of expertise.

General theories of vision such as those by Marr [6] and Biederman [2] have sparked extensive research efforts in both human and machine vision, and have contributed substantially to our understanding of how visual processes may operate. However, they have not led to artificial vision systems of noteworthy generality. Why is this so? From our point of view, a key reason is that most theories of vision do not address adaptation and learning. The real world is very complex, noisy, nonstationary – too variable for any fixed visual system, too unpredictable for its designer. Today's functional vision systems are highly specialized and operate under well-controlled conditions. They break if the built-in assumptions about task and environment do not hold.

Consider visual recognition. It is easy to see that there is no particular representation that can express all perceivable distinctions between objects or object categories that may later be required of a recognition system. Most existing machine vision systems perform recognition either based on a fixed set of handcrafted features, eigen-subspace decomposition, or geometric model matching. In the first case, features are chosen in a best effort to express the distinctions required, but not too much more to avoid overfitting. The same is true of geometric models. How much detail should be encoded in the models? On the one hand, the level of detail should be kept low to increase generalization and efficiency; on the other hand, models should contain sufficient detail to express the distinctions required by a given task. Thus, both these methods are restricted to tasks that are well-defined at design time. We call such tasks *closed.* In contrast, almost all human visual learning takes place in *open* settings, where tasks are open-ended and evolve over time. While eigen-subspace representations (or related subspace methods that optimally separate instances by class label) are to some extent consistent with certain aspects of human visual mechanisms (e.g. face recognition), it appears unlikely that such methods can account for all of biological discrimination learning since they can tolerate only a limited degree of occlusion and object variability.

Humans can learn an impressive variety of distinctions ranging from miniscule local features such as a tiny scratch to abstract global features such as symmetry. In light of the evidence cited above, it seems clear that humans are capable of forming new representations of global and local appearance characteristics in a task-driven way. Thus, a key concept for building artificial vision systems of substantially increased generality and robustness is task-driven learning or adaptation. An adaptive system should be able to

- optimize its performance on-line with respect to individual tasks,
- expand its functionality incrementally,
- optimize its performance on-line under the actual working conditions, and
- track a nonstationary environment by adapting its parameters

by building new representations and adapting parameters. In the following sections, we describe our current work on a model of feature learning for recognition that addresses all of these issues, building on our previous work [8].

## 3 An Infinite Feature Space

We argued above that any fixed object representation is insufficient for learning arbitrary distinctions. Instead, we begin by specifying a small set of *primitive* features that can be combined into *higher-order* features according to a small number of rules that will be discussed below. All features that can be represented in this way form an infinite feature *space*. The structural complexity of a feature, i.e. the number of primitive features that form a compound, naturally provides a partial ordering of this space. Our learning procedure searches the feature space beginning with structurally simple features, and considers more complex features as needed [1]. The underlying assumption is that structurally simple features are easier to discover and have less discriminative potential than complicated features, but are still useful for some aspects of the learning problem.

### 3.1 Primitive Features

In our current system, primitive features are local appearance descriptors represented as vectors of local filter responses. The filters are oriented derivatives of 2-D Gaussian functions, with orientations chosen such that they form a steerable basis [3]. Here, the steerability property permits the efficient computation of filter responses of Gaussian-derivative kernels at any orientation, given $d + 1$ measured filter responses for the $d$th derivative at specific orientations. Specifically, our system currently uses two specific variants of such descriptors:

- An *edgel* is encoded as a 2-vector containing the filter responses to the two first-derivative basis filters. These values encode the local intensity gradient of horizontal ($G_x$) and vertical ($G_y$) orientation. Using the steerability property, the magnitude of gradients in any orientation can be computed.
- A *texel* is represented as an 18-vector comprising the responses to the basis filters of the first three derivatives at two scales. This represents a local texture signature. Like edgels, texels have an associated orientation that is defined by the two first-derivative filter responses. When the orientation of a texel is steered, the entire vector containing all derivatives is rotated rigidly with reference to the first derivative computed at the largest scale [10].

This choice of low-level representations is plausible of biological early vision. While it is unlikely that any biological visual systems exploit steerability, this is an attractive computational alternative in the absence of massively parallel hardware. Steerability leads to rotational invariance which simplifies artificial vision systems at essentially no extra cost. We are not aware of any conclusive evidence for or against the biological faithfulness of our texel representation.

### 3.2 Higher-order Features

Primitive features by themselves are not very discriminative. However, spatial combinations of these can express a wide range of shape and texture characteristics at various degrees of specificity or generality. We suggest the following four complementary types of feature composition:

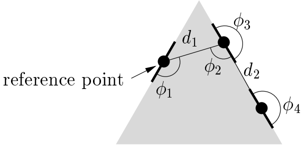

**Fig. 1.** A geometric feature of order 3, composed of three primitives. The feature is defined by the angles $\phi$ and the distances $d$. Each primitive is either an edgel or a texel.

- *Geometric* relations are given by the relative angles and distances between the participating lower-order features (Fig. 1). As long as these are rotation-invariant, so is their geometric composition. Geometric features are useful for representing e.g. corners, angles, and collinearity.
- *Topological* relations here refer to relaxed geometric relationships between component features that allow some degree of variability in angles and distances. Topological compound features are more robust to viewpoint changes than are geometric features, at the expense of specificity.
- *Conjunctive* features assert the presence of all component features without making any statement about their geometric or topological relationship.
- *Disjunctive* features are considered to be present in a scene if at least one component feature is detected. This can express statements such as "If I see a dial *or* a number pad, I may be looking at a telephone."

In contrast to other work, our features can be composed into increasingly complex and specific descriptors of 2-D shape, which is consistent with current models of the inferotemporal cortex. Features are computed at various scales, generated by successively subsampling images by factor two. This achieves some degree of scale invariance. Moreover, many compositions of edgels are inherently tolerant to changes in scale. For example, the arrangement shown in Fig. 1 applies equally to triangles of any size. Another desirable property of these features is that no explicit contour extraction or segmentation is required. This avoids these two difficult open problems in computer vision and should provide robustness to various kinds of image degradation. In contrast, the human visual system detects meaningful contours with remarkable robustness. This capability can probably not be explained entirely as a low-level visual process, but is supported by pre-segmentation recognition and task-dependent top-down processes.

## 4 Bayesian Networks for Recognition

The presence of a given feature $\mathbf{x}^*$ at a point $i$ in the image is denoted by its *strength* $s \in [0, 1]$. For primitive features, $s = \max\{0, r(\mathbf{x}^*, \mathbf{x}(i))\}$, where $r$ is the normalized cross correlation function. The vector quantity $\mathbf{x}^*$ is a model feature, and the function $\mathbf{x}(i)$ returns the corresponding feature at location $i$. A geometric feature is described by the concatenation of the constituent feature values $\mathbf{x}$. In the case of topological and conjunctive features, the strength of the compound feature is the product of the strengths of its constituents; for disjunctive features, the maximum is used. Recognition is based on the maximum strengths of features found in the scene (or within a region if interest). Mapping feature

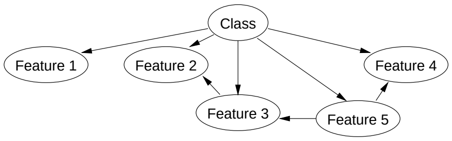

**Fig. 2.** A Bayesian network for one class. Note some interdependent features. A network such as this is created for each class.

vectors to class (or object) labels is the problem of classification, for which many algorithms exist. We chose Bayesian networks for their attractive properties that are desirable for open-domain recognition problems. In our system, each class is modeled as a separate Bayes net. The presence of an object is modeled as a discrete random variable with two states, *true* and *false*. The presence of an object gives rise to observable features, represented by random variables whose distributions are conditional on the presence of an object of this class. Assuming that the features are conditionally independent given the class, the resulting Bayes net has the topology of a star, with arcs connecting the class node to each of the feature nodes.

If some features are not independent, corresponding arcs must be inserted between the appropriate feature nodes. For example, in Fig. 2, Feature 3 may be a geometric composition with Feature 2, that is also in the feature set. Then, the presence of Feature 3 in an image implies the presence of Feature 2. Thus, in the Bayes net there is an arc from node 3 to node 2. An analogous argument holds for topological and conjunctive features, such as Feature 5 in Fig. 2, that combines Features 3 and 4. In the case of disjunctive features, the direction of the argument (and that of the additional arrows) is reversed.

Our feature strengths are continuous. We split each feature variable into two bins, corresponding to "present" and "not present", using a threshold. This threshold is determined individually for each feature variable such that its discriminative power between its own class and a misrecognized class is maximized. The discriminative power of a feature variable given a threshold is measured in terms of the Kolmogorov-Smirnoff distance (KSD). The KSD between two conditional distributions of a random variable is the difference between the cumulative probabilities at a given value of this variable under the two conditions. This separates the instances of the two conditions optimally, in the Bayesian sense, using a single cutpoint.

To perform recognition, we compute feature values one by one and update the Bayesian network after incorporating each feature. The class with the highest posterior probability gives the recognition result. Features are processed in decreasing order of informativeness. The informativeness of a feature is defined by the mutual information between a feature and the class node, i.e. its potential to reduce the entropy in the class random variable. In practice, only a fraction of all features are computed, because the entropy in the class nodes diminish before all features have been queried. This phenomenon suggests a straightforward, but very effective *forgetting* procedure: We delete any features that cease to be used during recognition.

## 5 Adaptive Feature Generation

As an agent (e.g. an animal, a human or a robot) interacts with the world, it uses vision (and maybe other sensory modalities) to acquire state information about the world, and performs actions appropriate in this state. This requires that the agent's visual features discriminate relevant aspects of the state of the world. We posit that such features are generated in response to feedback received during interaction with the world. For simplicity, we restrict the following discussion to a conventional supervised-learning scenario: The actions of the agent consist of naming class labels, the sensory input is an image, and the feedback received from the world consists of the correct class label. We further assume that the agent can retrieve random *example images* of known classes. This assumption is realistic in many cases. For example, an infant can pick up a known object and view it from various viewpoints; or a child receives various examples of letters of the alphabet from a teacher.

Initially, the agent does not know about any objects or features. When it is presented with the first object, it simply remembers the correct answer given by the teacher. When it is shown the second object, it will guess the only category it knows about.

When the agent gives a wrong answer, it needs to learn a new feature to discriminate this object category from the mistaken category (or categories). This is done by random sampling, with a bias for structurally simple features. We employ the following heuristic procedure, where each step is iterated up to a constant number of times:

1. Pick a random feature from some other Bayes net (corresponding to another class) that is not yet part of this Bayes net (corresponding to the true class). This promotes the usage of general features that are characteristic of more than one class.
2. Sample a new feature directly from the misrecognized image by either picking two edgels and turning them into a geometric compound, or by picking a single texel.
3. Pick a random feature that is already part of this Bayes net, find its strongest occurrence in the current image, and expand it geometrically by picking an additional edgel or texel close-by.
4. Pick two random features and combine them into a conjunctive feature.
5. Pick two random features and combine them into a disjunctive feature.

After each new feature is generated, it is evaluated on a small set of example images, retrieved from the environment, that contains examples of the true class and the mistaken class(es). If it has any discriminative power, it is then added to the Bayes net of the true class using the conditional probabilities estimated using a small set of example images, randomly chosen from the training set. If the image is now recognized correctly by the expanded Bayes net, the feature learning procedure stops; if not, the feature is removed from the net, and the learning procedure continues. This procedure may terminate without success.

During operation of the learning system, an instance list of all classes encountered and features queried is maintained. Periodically, all feature cutpoints, class priors and conditional probabilities in the Bayes nets are updated according to this list.

Feature learning does not have to stop after learning a training set perfectly. The system can continue to search for better features. The quality of a feature is its discriminative power at a given stage during a recognition procedure, given by the KSD between its own class and the combined set of all other classes. We can train our system to develop better features by imposing a minimum KSD on all features that are used during a recognition procedure. If a feature does not meet this requirement, the system has to learn a new and better feature. The minimum KSD can iteratively be raised, until the system fails to find adequate features. As a consequence, fewer (but superior) features will be queried while recognizing a given image, and many of the inferior features will become obsolete. We suggest this procedure, called *feature upgrade*, as a crude model of expert learning, as outlined in Section 2.

## 6 Experiments

To illustrate that our algorithm is able to produce discriminative features, we performed pilot experiments on two example tasks (Fig. 3). In the COIL task, the images of the first four objects from the COIL-20 database [7] were split into two disjoint sets such that no two neighboring viewpoints were represented in the same set. As a result, each image set contained 36 images, spaced 10 degrees apart on the viewing sphere, at constant elevation. We performed a 2-fold cross-validation on these two sets: In one run, one set served as a training set and the other as the test set; in a second run, the roles were reversed. In the PLYM task, there were eight geometric objects on 15 artificially rendered images each, covering a small section of the viewing sphere[1]. We performed a 10-fold stratified cross-validation on this data set, with random subdivision of the 15 images of each class into 10 subsets of 1 or 2 images each.

The results of the experiments are summarized in Table 1. While the recognition results fall short of current machine recognition technology, they were achieved by an uncommitted visual system with a strong bias toward few and simple features that had access only to a small number of random training views at any given time during an incremental training procedure. Most of these properties are contrary to current computer vision technology, but are characteristic of biological vision systems.

In accord with our biased search strategy, most learned features were isolated texels and simple geometric compounds of edgels and/or texels. Smaller numbers of the other compound types of features were also found. In most cases, the training set was not learned perfectly. This is because our system currently gives up after 10 iterations through the training set. Clearly, more effective techniques for finding distinctive features are called for.

[1] http://www.cis.plym.ac.uk/cis/levi/UoP_CIS_3D_Archive/8obj_set.tar

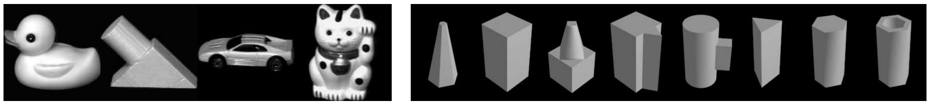

**Fig. 3.** Objects of the COIL task (left) and the PLYM task (right).

**Table 1.** Summary of experimental results. The "expert level" column gives the number of feature upgrade iterations. The "other" columns contain cases where the system returned an ambiguous answer, or no answer at all.

| Task | expert | avg. # features | Training Set: | | | Test Set: | | |
|---|---|---|---|---|---|---|---|---|
| | level | queried | correct | wrong | other | correct | wrong | other |
| COIL | 0 | 44 | 0.98 | 0.02 | | 0.81 | 0.19 | |
| | 1 | 36 | 0.85 | 0.11 | 0.04 | 0.73 | 0.23 | 0.05 |
| | 2 | 23 | 0.97 | 0.03 | | 0.83 | 0.16 | 0.01 |
| | 3 | 11 | 0.83 | 0.14 | 0.03 | 0.67 | 0.27 | 0.06 |
| PLYM | 0 | 19 | 1.00 | | | 0.72 | 0.28 | |
| | 1 | 21 | 1.00 | | | 0.76 | 0.21 | 0.03 |
| | 5 | 13 | 0.95 | 0.03 | 0.02 | 0.71 | 0.09 | 0.20 |

As the minimum KSD required of a feature is increased during feature upgrade, it is increasingly difficult to find appropriate features in order to learn the training set perfectly. However, feature upgrade has the desired effect of decreasing the number of features queried during recognition, and where the training set is learned well, it also tends to reduce the number of false recognitions while marginally increasing the correct recognition rate on the test set.

## 7 Conclusions

There is overwhelming evidence that humans learn features for recognition in a task-driven manner. Biological learning is on-line and incremental. We have presented an artificial vision system that follows these characteristics, based on an infinite combinatorial feature space and a generate-and-test search procedure for finding discriminative features. Our method successfully learns to discriminate objects. We also proposed that developing visual expertise involves the construction of better features. Our system models this by increasing the minimum KSD required of features during recognition. While our system reflects certain aspects of human vision, it is not a complete model in that it focuses on appearance-based discriminative features. Biological vision systems are probably composed of several complementary algorithms.

As a model of feature learning for discrimination, the main limitation of our system is the undirected search for features in images that is only guided by a few simple heuristics. A more faithful (and more practical) model requires a developmental schedule that initially constrains the search for features to increase the likelihood of finding useful features fast, while temporarily restricting generality. Over time, these restrictions should be relaxed, while the system learns better heuristics from experience. This is an area of further research.

Another critical limitation of our current system is the restricted expressiveness our feature space that encodes only high-contrast edge, corner and texture information. As such, our model roughly corresponds to the human visual system during early infancy [5]. A more complete model should at least encode color and blob-type features. In addition, more sophisticated recognition requires higher-level features such as qualitative ("Gestalt") features (e.g. parallelism, symmetry, continuity, closure) and multiplicity (a triangle has three corners; a bicycle wheel has many spokes). We hope to address these in future work.

## References

[1] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1300–1305, 1997.

[2] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[3] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

[4] I. Gauthier and M. J. Tarr. Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682, 1997.

[5] E. J. Gibson and E. S. Spelke. The development of perception. In J. H. Flavell and E. M. Markman, editors, *Handbook of Child Psychology Vol. III: Cognitive Development*, chapter 1, pages 2–76. Wiley, 4th edition, 1983.

[6] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Freeman, San Francisco, 1982.

[7] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (COIL-20). Technical Report CUCS-005-96, Columbia University, New York, NY, Feb. 1996.

[8] J. H. Piater and R. A. Grupen. Toward learning visual discrimination strategies. In *Proc. Computer Vision and Pattern Recognition (CVPR '99)*, volume 1, pages 410–415, Ft. Collins, CO, June 1999. IEEE Computer Society.

[9] A. D. Pick. Improvement of visual and tactual form discrimination. *J. Exp. Psychol.*, 69:331–339, 1965.

[10] R. P. N. Rao and D. H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505, 1995.

[11] P. G. Schyns, R. L. Goldstone, and J.-P. Thibaut. The development of features in object concepts. *Behavioral and Brain Sciences*, 21(1):1–54, 1998.

[12] P. G. Schyns and L. Rodet. Categorization creates functional features. *J. Exp. Psychol.: Learning, Memory, and Cognition*, 23(3):681–696, 1997.

[13] J. R. Silver and H. A. Rollins. The effects of visual and verbal feature-emphasis on form discrimination in preschool children. *J. Exp. Child Psychol.*, 16:205–216, 1973.

[14] J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23:457–482, 1991.

[15] J.-P. Thibaut. The development of features in children and adults: The case of visual stimuli. In *Proc. 17th Annual Meeting of the Cognitive Science Society*, pages 194–199. Lawrence Erlbaum, 1995.