Encoding Co-occurrence of Features in the HMAX Model

Sepehr Jalali (tmssj@nus.edu.sg)
Cheston Tan (cheston-tan@i2r.a-star.edu.sg)
Joo-Hwee Lim (joohwee@i2r.a-star.edu.sg)
Jo-Yew Tham (jytham@i2r.a-star.edu.sg)
Sim-Heng Ong (eleongsh@nus.edu.sg)
Paul James Seekings (mmrl@nus.edu.sg)
Elizabeth A. Taylor (tmsohe@nus.edu.sg)

National University of Singapore, Singapore 119077
Institute for Infocomm Research, A*STAR, Singapore 138632

Abstract

We introduce a method for encoding co-occurrence of features in the HMAX model of visual recognition, and conduct a series of experiments to investigate the contribution of co-occurrence towards better recognition performance. We show that classification accuracy is increased by adding a higher-order layer to the HMAX processing hierarchy, whereby co-occurrence of features is encoded as a new dictionary of features. We show that concatenation of mean pooling, max pooling and co-occurrence information results in better classification results on three datasets (Caltech101, a subset of Caltech256, and TMSI Underwater Images). Overall, we show that incorporating co-occurrence statistics into a biologically-inspired model of visual recognition provides a boost in classification performance above that produced by incorporating occurrence statistics alone.

Keywords: computer vision; HMAX; biologically inspired; co-occurrence statistics; visual cortex; image classification.

Introduction

Certain categories of visual stimuli can be characterized by the co-occurrence of multiple features. For example, images of cars frequently contain wheels, doors and windows. These co-occurring features do not occur in rigid configurations. Even for a rigid object, 3D rotations can result in inter-feature distances changing when projected as 2D images. However, co-occurring features are generally found close to each other. Using faces as an example, the exact distances between facial features (e.g. eyes, nose, mouth) vary from person to person, but these features are always relatively near to each other.

Can this particular property be exploited to achieve better visual recognition performance? This question cannot be cleanly answered through behavioral experiments unless brain cells encoding co-occurrence can somehow be "turned off"; computational modeling may be a better approach. In this paper, as a proof-of-concept, we modify the biologically-inspired HMAX model of visual recognition (Riesenhuber & Poggio, 1999) to encode co-occurrence statistics that are learnt from a training set of images, and we show that recognition performance does indeed improve.

Background

There is evidence for Max spatial pooling (finding the maximum among a set of inputs from a local spatial region) occurring at multiple levels in the visual system in the primary visual cortex of cats (Finn & Ferster, 2007; Lampl, Ferster, Poggio, & Riesenhuber, 2004), as well as in the higher visual areas of monkeys, such as areas $V4$ (Gawne & Martin, 2002) and IT (Sato, 1989). Importantly, however, each of these studies also showed evidence for "Average" pooling occurring, which can be interpreted as encoding the mean occurrence frequency of features.

Beyond just being tuned to the statistics of feature occurrences, there is strong evidence that the primate visual system is also tuned to co-occurrence statistics. This refers to either the joint or conditional probabilities of two (or more) features occurring together within images belonging to a certain object category or across categories. Since a "feature" is not always a precisely defined concept, how can the co-occurrence of two features be distinguished from the occurrence of a single feature that happens to be comprised of two simpler features? To make this distinction unambiguous, experiments were designed such that the elementary features are visually distinct, due to explicit segmentation, due to spatial separation, or from the task context. We term such features, which are the result of sensitivity to co-occurrence, as "co-occurrence features".

In some sense, mid-level features themselves can be considered as co-occurrence features, with their elementary features being simple orientation-sensitive filters (corresponding to orientation-sensitive neurons in the primary visual cortex). Since lines, curves and contours are ubiquitous in images, the presence of a short line segment of a certain orientation strongly predicts that the orientation of a neighboring line segment will be similar. This is particularly so if the relative position of that neighboring line segment is such that the two line segments have the possibility of being collinear.

Our focus here is on high-level features whose elementary features are more complex than simple oriented filters. These high-level features approach the level of semantic object parts or possibly even objects themselves. In the rest of this section, we will review the experimental evidence that the primate visual system develops sensitivity to such high-level co-occurrence features.

In the field known as visual statistical learning (VSL), it has clearly been shown that adult humans develop sensitivity to co-occurrence statistics in images (Fiser & Aslin, 2001;

Aslin & Newport, 2012). In a ground-breaking study by Fiser and Aslin (2002) it was shown that 9-month-old infants already developed sensitivity to visual co-occurrence statistics.

There is also an abundance of evidence from monkeys that their visual systems develop sensitivity to co-occurrence statistics. Miyashita (1988) and Sakai and Miyashita (1991), monkeys were trained to recognize pairs of stimuli, in a paradigm known as paired-associate learning. Neurons were found that were sensitive to such trained stimulus pairs, but not other stimulus pairs. The pairings were arbitrary, making the likelihood that such neurons had already possessed such sensitivity vanishingly small. More recently, Hirabayashi and Miyashita (2005) found that populations of IT neurons are sensitive to feature configuration within objects.

Direct evidence for sensitivity to co-occurrence (over and above sensitivity to occurrence) was found by Baker, Behrmann, and Olson (2002). Monkeys were trained to discriminate between objects that were each composed of two distinct parts linked by a line, forming "baton" objects. Compared to untrained objects, selectivity for trained objects was enhanced. This was for both the individual parts, as well as the combined "baton" objects. Crucially, selectivity for the two parts together (i.e. the whole object) was greater than the combined (summed) selectivity for each individual part.

Under what conditions does sensitivity to co-occurrence develop? In human adults, this is an implicit process that develops without awareness of the co-occurrence statistics, using a "cover task" or even through mere exposure (Turk-Browne, Jungé, & Scholl, 2005; Turk-Browne, Scholl, Chun, & Johnson, 2009; Aslin & Newport, 2012). This is also true for human infants (Fiser & Aslin, 2002; Aslin & Newport, 2012). In monkeys, most work has been done using active task learning. This is so that the neural selectivity for trained objects can be compared to the control set of untrained objects. Since neural selectivity is enhanced for features that are diagnostic for active task learning (Sigala & Logothetis, 2002), passive viewing may not be sufficient to produce selectivity that is large enough to be statistically significant when measured from electrode recordings.

How has sensitivity to co-occurrence been measured experimentally? The methods have generally been constrained by the nature of the subjects. Adult human subjects have generally been tested behaviorally, i.e. through their explicit responses (usually simple 'yes/no' tests). More recently, fMRI has been shown to be able to detect co-occurrence sensitivity (Turk-Browne et al., 2009). In human infants, due to their inability to understand or respond explicitly to verbal instruction, experiments have been constrained to using tests for novelty detection that are ubiquitous for infants. In monkeys, due to the ability to conduct invasive experiments that are not possible with humans, scientists have conducted electrophysiological experiments (i.e. using electrodes to record the responses of individual neurons). Such experiments allow for a detailed, "close-up" analysis of the effects of co-occurrence at the level of individual neurons e.g. Baker et al. (2002); Sakai and Miyashita (1991). However, there are limitations, such as the presence of noise, limited recording time, and the ability to record from at most a few hundred neurons.

Beyond just "being sensitive" to co-occurrence statistics, what are the characteristics of such sensitivity? It is specific to spatial configuration, such as the relative position of the elementary features (Hirabayashi & Miyashita, 2005). In addition, this sensitivity is reflected not in strength of neural responses *per se*, but rather in the selectivity for co-occurring features relative to non-co-occurring features (Baker et al., 2002).

One special case of sensitivity to co-occurrence of features is that of faces. The elementary features are semantic face parts such as the eyes, nose and mouth. It is very well-established that humans and monkeys are sensitive to the combination and relative configuration of face parts. Specifically, any change to the normal configuration of the face leads to reduced neural responses and poorer recognition accuracy. One manifestation of this is the Face Inversion Effect (FIE), whereby inverted faces are much more poorly recognized than upright faces (Yin, 1969). Faces with the parts in scrambled configurations are also poorly recognized. Furthermore, the sensitivity to co-occurrence seems to be unavoidable. In what is known as the Composite Face Effect, people are sensitive to the bottom halves of faces, even when they are explicitly instructed to ignore them during a discrimination task (Young, Hellawell, & Hay, 1987).

Generally, such sensitivity requires normal visual experience during infancy in order to develop (Le Grand, Mondloch, Maurer, & Brent, 2004). It also develops quickly, reaching adults levels (at least qualitatively) by age 4 (Heering, Houthuys, & Rossion, 2007); this is consistent with the notion that passive exposure is sufficient for co-occurrence sensitivity to develop (see above). Evidence for sensitivity to co-occurrence for face parts has also been found at the level of single neurons. Freiwald, Tsao, and Livingstone (2009) found that in one of the brain regions that respond selectivity to faces, neurons on average responded to combinations of two to three face parts, rather than individual parts. Co-occurrences have been studied in a series of experiments such as Edelman, Yang, Hiles, and Intrator (2002).

Use of co-occurrences of features for creating more complex features in Fidler, Boben, and Leonardis (2008) shows an improvement in classification accuracy, and bag-of-features approaches show improvements in classification results using frequency of patches in the images in (Fei-Fei & Perona, 2005). Co-occurrence information can be used to find part-part and part-whole relations of features of different receptive field sizes. If a feature is occurring too often in a class (and not likewise in other classes), it is more likely to be a discriminant feature in that class and if two features are co-occurring in a class often in a neighborhood, they may be part of a more complex feature and can have a part-part relationship and they might be more related to the object rather than the background (unless the background is also repetitive, e.g.

sky in airplane images). Also, if there exist features of different sizes and they co-occur in the same position on different scales they are likely to have a part-whole relationship.

HMAX Model

The HMAX model (Riesenhuber & Poggio, 1999) simulates the feed-forward path of the visual cortex. This model is used to find a good trade-off between invariance and selectivity. $S1$ cells provide selectivity by responding to oriented filters and $C1$ cells provide invariance by pooling over neighboring scales and positions. We use the HMAX model presented in Mutch and Lowe (2008) in the first three layers ($S1$, $C1$ and $S2$). Here we have a brief review on this model and show our modifications to it.

In this implementation, an image is fed into the structure and 10 different scales of the image are created as inputs to $S1$ layer. Gabor filters in 12 orientations are created as $S1$ layer filters:

$$G(x,y) = \exp\left(-\frac{(X^2+\gamma^2Y^2)}{2\sigma^2}\right)\cos\left(\frac{2\pi}{\lambda}X\right). \qquad (1)$$

where $X = x\cos\theta - y\sin\theta$ and $Y = x\sin\theta + y\cos\theta$. The values of x and y vary between -5 and 5, and θ varies between 0 and π. The parameters γ (aspect ratio), σ (effective width), and λ (wavelength) are all taken from Serre, Wolf, and Poggio (2005) and are set to 0.3, 4.5, and 5.6 respectively.

A fixed size of Gabor filters is implemented on different scales of the images where the smaller edge of the biggest image is set to 140 pixels while maintaining the aspect ratio (the image pyramid of 10 scales created each layer by a factor of $2^{1/4}$ smaller than the last using bicubic interpolation). The response of a patch of pixels X to a particular $S1$ filter G is given by:

$$R(x,y) = \left|\frac{\sum X_i G_i}{\sqrt{\sum X_i^2}}\right| \qquad (2)$$

These outputs are sent to the $C1$ layer, which performs a local $3D$ max operation on both scale (± 1) and position (3×3 neighborhood) of the filter responses. The output of this layer is a pyramid consisted of between 500-2000 different patches of size 4×4, 8×8, 12×12 and 16×16 in 8 scales depending on the size of the input image. In this level one or two samples are randomly sampled from each training image (from random scales and positions) and a dictionary of features of size 4096 is created. This dictionary is then made sparse by selecting the highest response from each orientation and setting the rest to 0.

The response of a patch of $C1$ units X to a particular $S2$ feature/prototype P (a dictionary feature), of size $n\times n$, is given by a Gaussian radial basis function:

$$R(X,P) = \exp\left(-\frac{\| X-P \|^2}{\sigma^2}\right) \qquad (3)$$

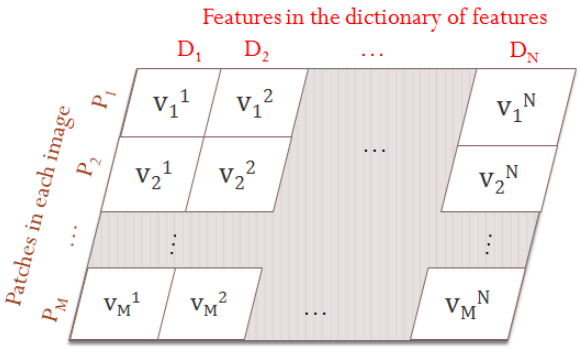

Figure 1: In HMAX, the max on the columns is taken as the response for creating $C2$ output vector. In contrast, histogram approaches based on SIFT methods use the frequency of feature occurrence, i.e. the normalized sum of the max values on the rows.

The values of R are stored as $S2$ layer. The distance of each sample from each training image with each entry on the dictionary is calculated and a local max is taken in $C2$ layer in $\pm 1 scale$ and $\pm 10\%$ spatial neighborhood (despite a global max in Serre et al. (Serre et al., 2005)). These $C2$ features are sent to the SVM for training. For testing images the same hierarchical procedure is repeated. In (Mutch & Lowe, 2008) sparse prototypes are calculated and the maximum response from all directions for each window is taken and SVM normals method (Mladenić, Brank, Grobelnik, & Milic-Frayling, 2004) is used to select the features with higher weights. In this approach, SVM is run a few times, and each time features with lower weights are dropped. In this HMAX implementation, once $S2$ features are calculated, the $C2$ layer is calculated as:

$$C2(n) = \max(V_k^n) \text{ for } \forall k \in M \quad \text{for } n = 1,...,N \qquad (4)$$

As can be seen in Figure 1 in conventional HMAX approaches, the max on the columns is taken as the value for $C2$ either in a local neighborhood of each feature or globally. Since taking the max in a local neighborhood (in ± 1 scale and $\pm 10\%$spatial neighborhood) is shown to improve the performance by about 5% in Caltech101 dataset in Mutch and Lowe (2008), in our experiments we also use a local neighborhood for calculating the responses. We also eliminate the local inhibition in $S2$ level proposed in Mutch and Lowe (2008) as it increased the performance. Once a feature belongs to the first or last scale in the pyramid, we extend the neighborhood to two neighboring scales. Same method is used for features which fall in the borders of each scale, and $+20\%$ or -20% of their neighborhood is used for comparisons.

If we take the sum of the values on rows in Figure 1 and normalize them, these are "HMean" features, which are also biologically-inspired, and significantly improve classification results when concatenated with HMAX features (Jalali, Lim,

Tham, & Ong, 2012). HMean is equivalent to the feature occurrence frequency in "bag-of-features" methods.

Encoding Co-occurrence of Features

For each class, we first find the value and index of the most-frequently occurring features (MOF). The next step is to encode the co-occurrence of these features as can be seen in Figure 1. For every class, we calculate the co-occurrence of the most frequent features and store it as a $S3$ dictionary feature. Hence a new dictionary of features is added to the model which is composed of #MOF $\times$ #MOF entries for each class, where #MOF was set as 20. In this dictionary of features, the value of each dictionary feature is calculated as:

$$C3(i,j) = C2(i)C2(j)\exp\left(-\frac{\| S_i - S_j \|^2}{\sigma^2}\right) \qquad (5)$$

where S_n represents the spatial position of the $C2$ feature and $\sigma = 0.5$.

This dictionary encodes the value of co-occurrence of every pair of features selected for each class. Hence we will have NN dictionaries where NN stands for the number of categories in the classification task. These dictionaries are concatenated to create the $C2$ dictionary of features. In the training and test phases, the respective feature to each dictionary feature is found (the most similar feature in every image) and the similarity of the values in dictionary of features are calculated for every image. This results in a #MOF $\times$ #MOF$\times$ NN feature as the $C3$ feature and it is concatenated to $C2$ feature vector and sent to the classifier for classification. The extended model for encoding the co-occurrence of features is shown in Figure 2.

Experimental Results

We evaluated our co-occurrence model on the Caltech101 dataset (Fei-Fei, Fergus, & Perona, 2004). The model was trained on 30 images per category (standard for this dataset; see Mutch and Lowe (2008)), and tested on all the other images. We also used the Caltech256 dataset (Griffin, Holub, & Perona, 2007), because it allows for more images per category than Caltech101. In particular, we considered only the 14 (out of 256) categories which had 200 or more images. We trained the model on 150 images (so that there would be at least 50 images for testing), and tested on the rest. We also examined classification accuracy as a function of number of training images for Caltech256. This was motivated by the concern that co-occurrence features could require more data for reliable co-occurrence statistics to be extracted, before the advantage of co-occurrence could be properly manifested.

We also evaluated the performance of our model on a new dataset consisting of images of underwater targets. The main challenge with underwater images is the existence of particles that limit the visibility in unclear waters and results in scattering, reflection and absorption of light, and the differential absorption of light of different wavelengths by water itself. This dataset consists of 1664 images (roughly 740×420 pixels in size) from 13 categories. Example images from this dataset are shown in Figure 3. We used 30 images per category for training, and the rest for testing.

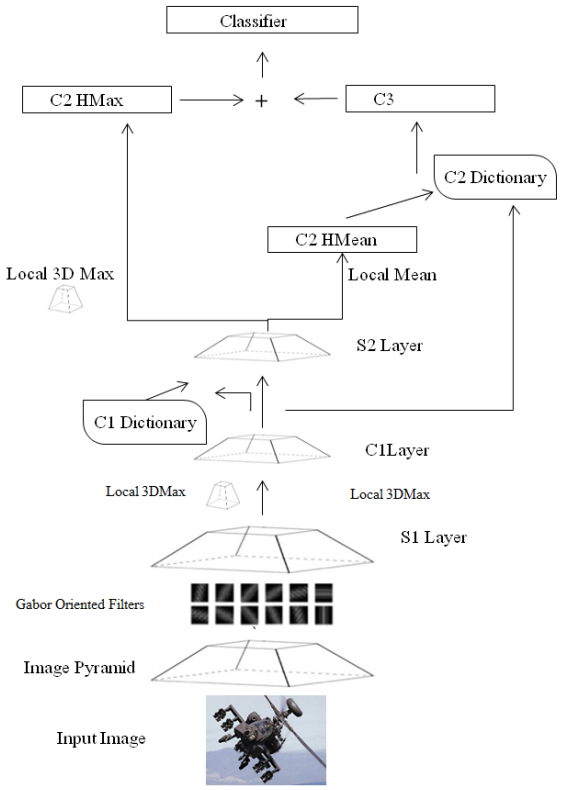

Figure 2: Diagram of model processing hierarchy.

Results are shown in Table 1. For all images, only intensity (luminance) information was used. All results were derived using 8 random train/test splits. For all three datasets, the combination of HMAX and co-occurrence features gave better results (classification accuracy) than either type of feature alone (Caltech101: 59.3% vs. 54.7% vs. 57.7%; Caltech256: 64.4% vs. 60.2% vs. 48.6%; Underwater Images: 98.7% vs. 92.9% vs. 92.2%). Since co-occurrence features were derived from the co-occurrence of HMean features, we also compared which of these two feature types (co-occurrence vs. HMean) gave better results when combined with HMAX. Again, for all three datasets, combining co-occurrence features with HMAX produced better results than combining HMean with HMAX (Caltech101: 59.3% vs. 58.9%; Caltech256: 64.4% vs. 61.3%; Underwater Images: 98.7% vs. 98.3%). Furthermore, for all datasets, the combination of all three feature types was better than just HMAX and HMean together (Caltech101: 60.1% vs. 58.9%; Caltech256: 64.1% vs. 61.3%; Underwater Images: 99.0% vs. 98.3%).

We also examined the effect of disregarding spatial distance (i.e. the exponential in Eq. 5). As seen in Table 1, for all datasets, results were better when spatial distance was taken into account (Caltech101: 57.7% vs. 55.1%; Caltech256: 48.6% vs. 44.2%; Underwater Images: 92.2% vs. 83.3%).

Figure 3: Examples from TMSI Underwater Images dataset.

Table 1: Classification performance on the Caltech101, Caltech256 (subset – see text for details), and TMSI Underwater Images datasets.

Method	Caltech101	Caltech256 (subset)	Underwater Images
HMAX	54.7	60.2	92.9
Co-occurrence (no distance)	55.1	44.2	83.3
Co-occurrence	57.7	48.6	92.2
HMAX + Co-occurrence	59.3	64.4	98.7
HMAX + HMean	58.9	61.3	98.3
HMAX + HMean + Co-occurrence	60.1	64.1	99.0

In order to evaluate the effect of number of training images for the creation of co-occurrence features, we trained the model with varying numbers of training images per category. As shown in Figure 4, the performance boost when adding co-occurrence features was greatest when using 150 training images. However, for fewer than 150 training images, the boost from adding co-occurrence features is unreliable. Nonetheless, looking at just HMAX alone, performance seems to asymptote at 150 training images, but for the combination of HMAX and co-occurrence features, performance seems to increase roughly linearly with the number of training images. While empirically, co-occurrence may help performance in all datasets, similar analyses (i.e. performance boost as a function of number of training images) for the other 2 datasets may not be meaningful, since the maximum number of training images is only 30 per category.

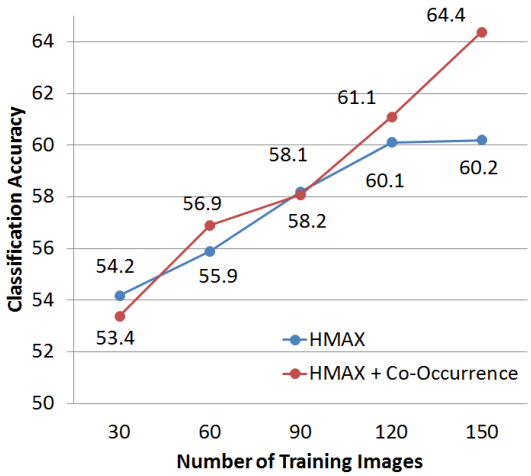

Figure 4: Classification accuracy on Caltech256 as a function of number of training images.

Discussion

In this paper, we showed that combining co-occurrence features with regular HMAX features leads to better classification performance than using either feature type alone. Furthermore, adding co-occurrence features to HMAX increases performance more than adding occurrence features. The three types of features encode different information, and therefore the combination of all three feature types gave the best overall performance. For co-occurrence, the spatial distance between the two co-occurring features also contributes to better performance. In this work, we focused solely on HMAX. However, in future work, our co-occurrence method can be applied to other vision algorithms.

In preliminary experiments not reported here, we experimented with creating co-occurrence features from HMAX features (rather than HMean features, as done in this paper). However, this resulted in either a drop in performance or no change. This will be investigated further in future work.

Fig. 4 suggests that the performance boost from using co-occurrence may be limited by the number of training images. More detailed investigation is limited by the relatively small number of images per category in these datasets. Further investigation may require utilizing or creating larger datasets.

Another prospect for further improvement is to encode co-occurrence of more than two features. However, besides possibly requiring even more training data than two-feature co-occurrence, there may be diminishing returns for such "higher-order" co-occurrences. This is because relatively fewer classes will have the underlying visual structure that will benefit from encoding such co-occurrences.

In this paper, the choice of features for encoding co-occurrence was based on their frequency. Choosing discriminative (rather than frequent) features for co-occurrence encoding may be a more direct approach to maximizing classification performance. To choose discriminative features, one approach is to train the SVM several times and remove fea-

tures with low weights, as in Mutch and Lowe (2008), or to simply use features with mean response values that differ the most between different classes.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science*, *21*(3), 170–176.

Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, *5*(11), 1210–6.

Edelman, S., Yang, H., Hiles, B., & Intrator, N. (2002). Probabilistic principles in unsupervised learning of visual structure: human data and a model. *Advances in Neural Information Processing Systems*, *1*, 19–26.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004. In *Workshop on generative-model based vision* (Vol. 2).

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 524–531).

Fidler, S., Boben, M., & Leonardis, A. (2008). Similarity-based cross-layered hierarchical representation for object categorization. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8).

Finn, I. M., & Ferster, D. (2007). Computational diversity in complex cells of cat primary visual cortex. *Journal of Neuroscience*, *27*(36), 9638–48.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(24), 15822–6.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187–96.

Gawne, T. J., & Martin, J. M. (2002). Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. *Journal of Neurophysiology*, *88*(3), 1128–1135.

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset* (Tech. Rep. No. 7694). California Institute of Technology.

Heering, A. de, Houthuys, S., & Rossion, B. (2007). Holistic face processing is mature at 4 years of age: evidence from the composite face effect. *Journal of Experimental Child Psychology*, *96*(1), 57–70.

Hirabayashi, T., & Miyashita, Y. (2005). Dynamically modulated spike correlation in monkey inferior temporal cortex depending on the feature configuration within a whole object. *Journal of Neuroscience*, *25*(44), 10299–307.

Jalali, S., Lim, J., Tham, J., & Ong, S. (2012). Clustering and use of spatial and frequency information in a biologically inspired approach to image classification. In *International Joint Conference on Neural Networks* (pp. 1–8).

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, *92*(5), 2704–13.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science*, *15*(11), 762–8.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–20.

Mladenić, D., Brank, J., Grobelnik, M., & Milic-Frayling, N. (2004). Feature selection using linear classifier weights: interaction with classification models. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 234–241).

Mutch, J., & Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, *80*(1), 45–57.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, *354*, 152–5.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Experimental Brain Research*, *77*(1), 23–30.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 994–1000).

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318–20.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552–64.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–45.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–59.