Machine learning of visual object categorization: an application of the SUSTAIN model.

Giovanni Sirio Carmantini (giovanni.carmantini@plymouth.ac.uk)
School of Computing and Mathematics,
Plymouth University,
Plymouth PL4 8AA, United Kingdom.

Angelo Cangelosi (a.cangelosi@plymouth.ac.uk)
School of Computing and Mathematics,
Plymouth University,
Plymouth PL4 8AA, United Kingdom.

Andy Wills (andy.wills@plymouth.ac.uk)
School of Psychology,
Plymouth University,
Plymouth PL4 8AA, United Kingdom.

Abstract

Formal models of categorization are psychological theories that try to describe the process of categorization in a lawful way, using the language of mathematics. Their mathematical formulation makes it possible for the models to generate precise, quantitative predictions. SUSTAIN (Love, Medin & Gureckis, 2004) is a powerful formal model of categorization that has been used to model a range of human experimental data, describing the process of categorization in terms of an adaptive clustering principle. Love et al. (2004) suggested a possible application of the model in the field of object recognition and categorization. The present study explores this possibility, investigating at the same time the utility of using a formal model of categorization in a typical machine learning task. The image categorization performance of SUSTAIN on a well-known image set is compared with that of a linear Support Vector Machine, confirming the capability of SUSTAIN to perform image categorization with a reasonable accuracy, even if at a rather high computational cost.

Keywords: Categorization; SUSTAIN; Object recognition.

Introduction

The problem of categorization is a classic one in the cognitive sciences, approached from many perspectives by many disciplines (Cohen and Lefebvre, 2005). One powerful approach is represented by the formal modelling of categorization in psychology. The strength of the approach comes from the precise formulation of the theories it generates, which enables those theories to produce precise predictions, model human experimental data, and give a lawful account of the process of categorization (Pothos and Wills, 2011).

The Supervised and Unsupervised STratified Adaptive Incremental Network, or SUSTAIN (Love, Medin & Gureckis, 2004), has been used to model a range of inference tasks, and it has confirmed many times to be a flexible formal model of categorization. Its core feature is the use of an adaptive clustering principle: SUSTAIN forms a clustered representation of the input data which is continuously adjusted to accommodate new information. A cluster can be thought as a bundle of features, initially centered on some input stimulus, but that is modified with learning so that similar inputs can be explained as well. In this way SUSTAIN is able to form very flexible representations.

SUSTAIN doesn't passively represent the input data, it also selectively weighs the dimensions that seem to bear the greatest information content; this is possible thanks to the implementation of attentional learning in the model. Clusters compete to explain the input, so that only the cluster which is most similar to the input (i.e. it is most activated) wins the competition and can be updated to accommodate the new data.

The activation of cluster j is given by

$$H_j^{act} = \frac{\sum_{i=1}^{m} (\lambda_i)^r e^{-\lambda_i \mu_{ij}}}{\sum_{i=1}^{m} (\lambda_i)^r}, \quad (1)$$

where m is the number of dimensions of the description of the stimuli, λ_i is the tuning of the receptive field for the i^{th} dimension, μ_{ij} is the absolute difference between the stimulus and the j^{th} cluster for that dimension, and r is a non-negative parameter. The receptive field tuning can be thought of as an attentional weight. If λ_i is large, then small differences between stimulus and cluster in the i^{th} dimension contribute more to its activation. The r parameter accentuates this effect; for large r, the dimensions which are most attended to dominate the activation function. The winning cluster updates its values on each dimension towards that of the input stimulus, and the λ tunings are updated as well so that dimensions with smaller differences get more attention. The learning of the dimensions' values and the tunings is controlled via a learning rate parameter.

The formal model also has an output layer whose purpose is to learn to mirror the input. The mathematical details of the output layer implementation in SUSTAIN are omitted here, as its role in this study is negligible. The output layer has weighted connections to every recruited cluster, and the win-

ning cluster for each trial updates its weights. The activation of the winning cluster and the weights determine the activation of the units in the output layer, each corresponding to an input dimension. This way it's possible to model inference tasks in SUSTAIN: if a dimension of the input stimulus is hidden to the model, the output layer can infer its value even in its absence. The category membership is also represented on the input and output layer. In supervised learning, output units corresponding to category membership are compared with the corresponding units on the input layer. If their values are different, then SUSTAIN recruits a new cluster centred on the incorrectly categorized input. With this feedback mechanism, SUSTAIN can integrate exceptions to the similarity: even if two stimuli are similar, they can belong to different categories (Gureckis & Love, 2002). However, the forming of a clustered representation of the input data from SUSTAIN is not bound to feedback signals. SUSTAIN can learn in an unsupervised fashion, guided only by the similarity between the input and the stored representations.

Love et al. (2004) suggested that it might be possible to apply SUSTAIN to the domain of image categorization, and made predictions about the model's behaviour in performing this particular task. The purpose of the present study is to explore the feasibility of using SUSTAIN to perform image categorization in supervised learning, and test the predictions advanced by its authors. Naturally stemming from these specific goals, the study is also, more generally, an investigation of the utility of applying formal models of categorization to machine learning tasks such as image categorization.

Specifically the three predictions tested in the current paper are:

1. SUSTAIN doesn't need to store each image presented to be able to perform categorization, unlike other view-based approaches (for an example, see Poggio & Edelman, 1990). That's the power of adaptive clustering, which permits a compressed representation of the input data.

2. Categories that are associated with more variability in their members' appearance will be represented with more clusters than needed for categories which vary less. As clustering is driven by similarity, less similarity should be associated to the recruitment of more clusters.

These predictions about the behaviour of SUSTAIN in image categorization were made by Love et al. (2004, p.328). Each presented input is compared by SUSTAIN with the stored clusters. The difference between the input and each cluster is computed. The lower the difference, the higher the activation of a given cluster. A threshold parameter controls how much similarity is considered enough for the input to be considered explained by one or more of the clusters. If a cluster's activation is over the threshold, then the cluster is considered similar enough to be taken into account as a possible explanation of the input. If no cluster is found with a sufficient activation, then the input is considered too dissimilar to the stored representations and a new cluster is recruited. A higher threshold is more difficult to reach, leading to a larger number of clusters recruited to represent the same set. This behaviour leads to a third prediction, original of this study:

3. A higher number of clusters stored is associated with less information loss. Thus, a higher activation threshold is predicted to lead to higher categorization accuracy, relative to a lower activation threshold.

Method

The methods used in this study are an extension of those used in Eichhorn and Chapelle (2004) and Grauman and Darrell (2005). The authors used Support Vector Machines, machine learning models widely employed in image categorization, and evaluated their categorization performance using different descriptions of the images. This study will make use of the same images and image descriptions: Images from the ETH-80 image set (Leibe & Schiele, 2003) will be used. For each image, its features will be detected using a Harris detector (Harris & Stephens, 1988), and described using a SIFT descriptor (Lowe, 1999). The variable number of features extracted this way will be mapped to a fixed dimensional representation and used in the context of a cross-validation paradigm. Each of these steps will be explained in detail in the next sections.

Image description

A description must be used for the images in order to extract their informative content. A powerful and well-established approach in computer vision is that of finding interest points in an image like corners and edges. These interesting points are called "features". Different objects will tend to be associated with different features. The features are in turn analyzed using one of many specific procedures, in order to find a compact description for each feature. The type of analysis and description carried out depend on the descriptor used.

The image set used in this study is a subset of the ETH-80 image set (Leibe & Schiele, 2003), which contains 80 objects from 8 categories. The categories are apple, pear, tomato, cow, dog, horse, cup and car. The subset selected for the experiment is composed by 5 widely separated views for each object (see Figure 1). For each image, interest points are found by the use of a Harris corner detector (Harris & Stephens, 1988). The detector has been chosen for continuity with the methods of Eichhorn and Chapelle (2004) and Grauman and Darrell (2005). The authors decided to use this detector because of its robustness and its ease of implementation. The SIFT descriptor (Lowe, 1999) is used for its robustness to rotation, affine transformations and changes in contrast and illumination, consistent with the already mentioned literature. For each category the average number of features extracted is reported in Figure 2. The number associated with each category in the Figure represents the average number of features per image for that category, averaged over a total of 50 images per category (10 objects for category x 5 images for each object).

(a)

(b)

Figure 1: In (a), an example of the 5 view angles for each object used in this study. In (b) a view for each object in the set is reported. Images from Leibe and Schiele (2003).

The variable number of features is mapped to a fixed-dimensional representation with the use of a Bag-Of-Keypoints method similar to that of Csurka et al. (2004). The method involves using a k-mean clustering algorithm on a training pool of features to compute a vocabulary of k words, where each of the k words can be thought as a "mean-feature", capturing the central tendency of a family of similar features in the training pool. The images in the image set are then described using the vocabulary, so that each feature of each image is brute-force matched to the most similar word in the vocabulary. The vocabulary used in this study comprises 250 words, and the training set used to form the vocabulary is represented by 300 randomly picked images from the entire ETH-80 image set. The final representation for each image is a frequency histogram, where each bin represents the frequency with which a given word is matched to a feature in the image. The histograms computed this way are composed of 250 bins, one for each word in the vocabulary, and the sum of the values of the bins is equal to 1. This is the kind of image representation that is used as an input for SUSTAIN in this study.

A toy example will help summarizing the description process. An image, let's say a tomato, is to be presented to SUSTAIN. The image is searched for features, by the Harris detector, which finds 4 (to keep it simple). Each one is processed by the SIFT algorithm to obtain a compact description. The features so described are mapped to the most similar ones in a pre-computed vocabulary of 250 words. Let's say that 2 features are mapped to the last word in the vocabulary, 1 to the first word and 1 to the third. Then the image description will look as a vector of the form $\langle 1,0,1,0,...,0,2\rangle$. SUSTAIN needs each dimension to be described by a number between 0 and 1, so each entry is divided by the total number of features in the image, obtaining this way a vector of frequencies. The label "tomato" is associated to this vector and used as an input for SUSTAIN.

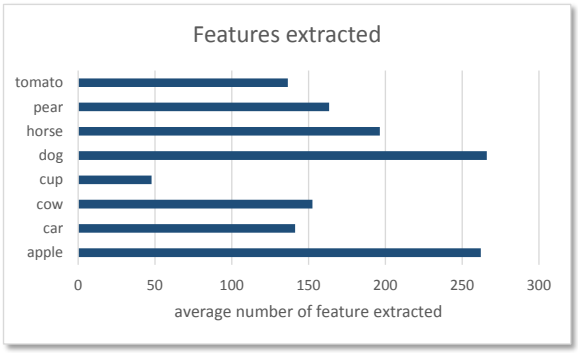

Figure 2: Average number of features extracted for each category

Experimental design

A cross-validation experiment was conducted to test the predictions presented in the introduction and to evaluate the capability of SUSTAIN to perform image categorization. In a cross validation experiment, the training takes place using all the objects in the image set with the exception of one, which is used for testing. The procedure is repeated for each object in the image set. For each image, its Bag-Of-Keypoints representation was presented with a label stating the category membership.

SUSTAIN was trained in a supervised fashion. As the model, like humans, is susceptible to ordering effects, the order of presentation of the images was randomized for each cross-validation. The experiment had two conditions. In the first condition, the activation threshold parameter, τ from now on, was set to 0: a cluster was recruited only as a result of an incorrect categorization. This is the baseline condition, showing the minimum amount of clusters SUSTAIN can recruit to represent the set, given the other parameters used and the order of presentation of the images. In the second condition, τ was set to 0.97. In this case a cluster was recruited both in the case of an incorrect categorization, and when the categorization was correct but the winning cluster showed an activation lower than 0.97, where a cluster activation of 1 means no difference between the cluster values and the input image values. A higher τ value forces SUSTAIN to recruit more clusters. This way the prediction stating that a higher number of recruited clusters should be associated with a higher categorization accuracy could be tested. Note that a τ value of 1 would have meant for SUSTAIN to save each presented image as a cluster. SUSTAIN would have been in this case

just like a bruteforce matching algorithm.

With the exception of the learning rate parameter, the other parameters used in this implementation of SUSTAIN were the ones provided in Love et al.(2004,p.313). The learning rate was set to 0.065, this value being the result of a greedy search over a reasonable range of values. A search for the optimal parameters could have yielded better results than those of this study. The effects of lateral inhibition, together with the action of the output layer, won't be analysed here. Consistently, the associated parameters are not reported as they have no effect on what is investigated in the study. During the testing phase, SUSTAIN could only compare the input image representation with the clusters stored during training, so to find the most similar one. No learning or cluster recruitment was permitted. If the category label of the cluster was the same as the image presented, then the prediction was considered correct, incorrect otherwise.

In addition to examining SUSTAIN, we also examined a linear multiclass Support Vector Machine (Csurka et al. 2004). A linear multiclass SVM represents each image in the training set as a point in a n-dimensional space (250 dimensions for this study), and computes for each category an optimal separating hyperplane that divides points belonging to the category to points which don't. This way, when a new point is presented, i.e. an image from the test set, the model predicts a category membership by looking at its position in respect to the hyperplanes. The model is widely used in Computer Vision for image categorization, and was used here for the sole purpose of putting into context the performance, both in accuracy of categorization and in computation time, of SUSTAIN.

The OpenCV (Bradski, 2000) implementations of the Harris detector, SIFT descriptor, Bag-Of-Keypoints method and linear SVM were used. The SUSTAIN implementation was programmed by the authors of the present study. SUSTAIN and the SVM were trained and tested on a PC (2.4 Ghz Intel Core 2 Duo). Data were collected on the categorization accuracy for both SUSTAIN and the SVM, together with the computation times. To test the predictions on the categorization behaviour of SUSTAIN, data on the number of clusters stored for each category were collected. As reported, a cluster can be recruited by supervised SUSTAIN for two reasons: the input stimulus is not sufficiently similar to any stored representation, or it is but the category of the input and the category of the similar clusters are different. If a category tends to be heterogeneous between its members, then we expect to see a higher number of dissimilarity-driven cluster recruitments for that category. For this reason, we also collected data on the modality of recruitment for each cluster, i.e. recruitment caused by dissimilarity versus recruitment caused by prediction error. From the ETH-80 dataset, the animal categories were expected to be associated to a higher number of clusters recruited, as they seem to vary more both within views and between members. For the same reason, more recruitments because of dissimilarity were expected for the animal categories in respect to the object categories.

Results

Categorization performance of SUSTAIN and the SVM.

The SUSTAIN categorization accuracy was comparable to that of the SVM when τ was set to 0, and better than that of the SVM when τ was set to 0.97. A higher τ value leads to a higher number of cluster recruited, as it becomes more difficult for every stored cluster to reach the τ threshold and thus be considered similar enough to explain the input. When τ is set to 0, SUSTAIN just considers the input explained by whatever cluster has the highest activation, so that new clusters are recruited only because of a prediction error and never because of dissimilarity. The SVM correctly categorized the input from the testing set 60% of the time on average. SUSTAIN scored an average of 60% of correct categorizations when τ was set to 0, and an average of 66% of correct categorizations when τ was set to 0.97. However, the computational cost for the increase in accuracy was high.

SUSTAIN and SVM computation time

Training and testing SUSTAIN seem somewhat more computationally intensive than training and testing an SVM. For each cross-validation, SUSTAIN spent an average of 1.41 seconds in training and 0.30 seconds on average in testing when τ was set to 0, which became 2.53 in training and 0.71 in testing when τ was set to 0.97 In comparison, for each cross-validation the SVM was trained in an average of 0.22 seconds, and tested in an average of 0.26 seconds. It must be specified, however, that this implementation of SUSTAIN wasn't coded with regard to efficiency. One of the reasons why SUSTAIN seem slow in relation to the SVM is that both in training and in testing each input must be compared with all the stored clusters. That's also the reason why a higher τ value is associated with longer times: more clusters are stored so more comparisons must be made, as shown in the next section.

SUSTAIN cluster recruitment

When τ was set to 0, the number of cluster recruited was equal to the 30% of the presented images. Thanks to the use of the adaptive clustering principle, SUSTAIN compressed the information presented by two thirds. A higher value of τ, set to 0.97, was associated with an increase of the number of clusters stored, which reached an average of 58% of the number of images presented. The information presented was compressed only by a little more than a third. Less information was lost, so we expect a higher categorization accuracy to be associated with the higher threshold.

Looking at Figure 3, it's possible to see that, for both the conditions, more clusters were stored overall for the animal categories, whereas less were stored for the object categories. The cup category, curiously, seem to be associated with a rather high number of recruited clusters, given that all cups

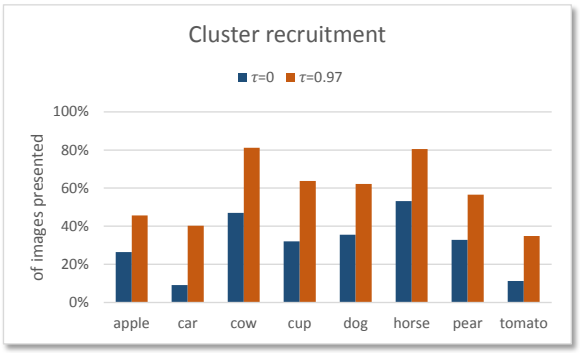

Figure 3: Number of clusters recruited by SUSTAIN, expressed as a percentage of the number of images presented, for $\tau = 0$ and $\tau = 0.97$.

look similar. To better understand why this is, data are reported on the modalities of cluster recruitment for each category. When τ was equal to 0.97, SUSTAIN could recruit a new cluster both in response to an incorrect categorization and in case the presented image was too dissimilar to the stored clusters. In Figure 4, the percentage of clusters recruited in respect to the number of images presented is shown divided for modality of recruitment. The cup has a rather high percentage of clusters stored due to dissimilarity, while the percentage of clusters stored for incorrect categorization is consistent with the ones from the other object categories. This result is rather surprising, given that the cups don't look very different between one another, and it seems to stem from the way the images are described, as discussed in the next section.

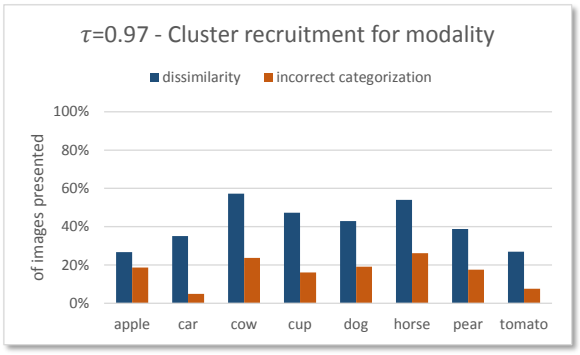

Figure 4: Percentage of clusters recruited in respect to image presented for SUSTAIN with $\tau = 0.97$, divided for modality of recruitment.

Discussion

The first prediction, that SUSTAIN wouldn't need to store each presented input in order to perform categorization, was confirmed. In fact, when τ was set to 0, SUSTAIN recruited a cluster in response to an input image only 30% of the time. When the higher threshold was used, SUSTAIN recruited a cluster 58% of the time. Our results also confirm the third prediction: a higher number of clusters recruited is associated with a higher accuracy in recognition, as more information is available to perform categorization. When τ was equal to 0, SUSTAIN correctly categorized 60% of the testing set. With the higher threshold of 0.97, the correct categorizations went up to 66%. It must be said that high thresholds could lead to fitting noise in certain situations, e.g. when large heterogeneous training sets are used. The possibility is worth exploring in future research. The second prediction seems to be supported by the data on the cluster recruitment for each category and for the two modalities, i.e. dissimilarity and incorrect input categorization during training. Specifically, it seems that the categories that were expected to show more variation, i.e. the animal categories, were the ones associated with the higher number of clusters stored. The only exception was the cup category, which was surprisingly associated with a higher number of clusters stored than a heterogeneous category like dog. A possible explanation stems from the fact that the cup category was described with a very low number of features (see Figure 2). This meant that even small changes between the images from the category were associated with comparatively big changes in the values of the Bag-Of-Keywords frequency histogram. These were in turn interpreted as large dissimilarities by SUSTAIN, resulting in the creation of new clusters to accommodate these dissimilarities. Such an interpretation is consistent with the fact that the cup category elicited an high amount of cluster recruitments for reason of dissimilarity, higher than the dog category ones (see Figure 4).

Surprisingly, when τ was set to the higher value, SUSTAIN outperformed the SVM in relation to the categorization accuracy. This relatively high accuracy came at a cost, though, as SUSTAIN was approximately ten times slower than the SVM in training, and approximately three times slower in testing. For bigger sets of images, the slowdown would have been even more dramatic, as SUSTAIN would have had to iteratively compare the input to each stored cluster; the more the clusters stored, the slower SUSTAIN becomes.

Future directions

This exploratory study certainly showed that it's possible to use SUSTAIN to perform image categorization. Establishing this fact opens many possibilities for future research. For example, can the computational complexity of SUSTAIN by reduced without sacrificing accuracy? And to what extent are each of the components of SUSTAIN important in producing the level of accuracy it can achieve? One reason for SUSTAIN's computational burden is that every time an input is presented, SUSTAIN has to compare it with all the stored clusters. Further research could be focused on finding a way to simplify the process. If an ordering can be found for the clusters so that SUSTAIN can narrow down the best clusters for the comparison, then maybe SUSTAIN could become a more viable option for image categorization. Another way

to speed up the processing could be that of parallelizing the comparison between the input and the clusters. This seems likely to be true for a range of approaches inspired by formal psychological models, as processing in the human brain is widely regarded to be massively parallel.

More work could also be done to explore further the possibilities of using SUSTAIN for image categorization. For example, to address what is the effect of the various forms of learning in SUSTAIN on categorization accuracy, issue which was not explored in the present study. Moreover, order effects are known to affect SUSTAIN's performance, but they were not investigated here. It would be interesting to study them in relation to categorization performance for the different categories. SUSTAIN comprises a mechanism of lateral inhibition between clusters, and an output layer. The presence of lateral inhibition permits a measure of how ambiguous the input is to the model. In fact the activation of a cluster is damped if many clusters are highly activated by the presented input. The output layer, instead, has the function to learn to mirror the input, so that if an input value over some dimension is hidden, the model is able to predict its presence. This could be used for example to predict which features are hidden in a partially occluded image. These SUSTAIN features may be very useful in image categorization, and it's worth investigating their value.

Conclusion

SUSTAIN, a psychological formal model of categorization, was not only able to perform image categorization, it outperformed the SVM (a standard approach in Machine Learning), although this was at the cost of increased computational intensiveness. One of the limitations of the approach used in this study to describe the images is that it doesn't ensure a homogeneous quality of the description. For example, the features extracted from the image of a cup can be less in number and less distinctive than the features extracted from the image of a horse. For this reason, the differences calculated between clusters and input are not normalized, while SUSTAIN expects them to be so. Thus, a homogeneous category like cup, which was expected to need a low number of clusters to be represented, actually needed a rather high one. The results of this study do support the prediction that heterogeneous categories need more clusters to be represented than homogeneous ones, but the problem of the quality of the description must be assessed in future research as not to confound the interpretation of the results.

On the issue of the practical utility of using a formal psychological model of categorization for machine learning tasks, the results of this exploration suggest that SUSTAIN is somewhat computationally inefficient. In fact, it is not built for efficiency: its original purpose is to model human data to better understand them and to make predictions. In addition to this, the specific implementation of SUSTAIN used in this study wasn't coded with regard to efficiency, as the study was thought as a proof of concept. Perhaps a more efficient implementation could make SUSTAIN a more viable option for image categorization. In conclusion, even though SUSTAIN could perhaps not be efficient enough to be used in computer vision as is, it can inform novel models inspired by its principles, which proved applicable to image categorization. The adaptive clustering of image features, together with the attentional learning and the competition between representations, could prove to be powerful principles on which to build better computer vision models.

Acknowledgments

This work was partially supported by the FP7 project Poeticon++ (FP7-STREP-288382) and by an EPSRC grant (EP/I001433/1).

References

Bradski, G. (2000). The OpenCV library. *Dr Dobbs Journal*, *25*(11), 120.

Cohen, H., & Lefebvre, C. (2005). *Handbook of categorization in cognitive science* (Vol. 4). Elsevier.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *1*, 22.

Eichhorn, J., & Chapelle, O. (2004). Object categorization with svm: kernels for local features. In *Advances in neural information processing systems.*

Grauman, K., & Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* (Vol. 2, pp. 1458–1465).

Gureckis, T., & Love, B. C. (2002). Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions. In *Proceedings of the 24th annual conference of the cognitive science society* (pp. 399–404).

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference* (Vol. 15, p. 50).

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition.* (Vol. 2, pp. 2–409).

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological review*, *111*(2), 309–332.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 1150–1157).

Poggio, T., & Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature*, *343*(6255), 263–266.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization.* Cambridge University Press.