The formation of context in artificial object recognition

Anthony Knittel (aek@cse.unsw.edu.au)
University of New South Wales, Sydney, Australia

Abstract

Context plays an important role in the recognition of objects, allowing the general content of a scene to influence identification of individual parts. An autonomous learning system is presented that examines processes involved in the formation of context between multiple co-occurring objects, under the task of identifying abstract objects in a scene. Learning is performed using a form of Learning Classifier System, that builds representations of features autonomously under reinforcement. The feature identification system is used in combination with an associative network, used for finding co-occurrence relationships for establishing context. Experiments show the influence of the associative network to resolve ambiguous observations through the use of context. This approach involves the interaction of a reinforcement system, analogous to dopaminergic processes, with an associative system, based on associative Hebbian learning processes, and demonstrates the ability of a recurrent associative network for establishing context relationships.

Keywords: context; hebbian learning; reinforcement; object recognition

Introduction: recognition of scenes

Recognition of objects in a scene requires the interaction of a number of processes, such as identification of primitive features from sensory input, a general idea of the "gist" of the scene and context of the observation, and influence from the recognition of previously known objects in the observed environment (Oliva & Torralba, 2007).

A review by Bar (2004) highlights a number of influential factors in the role of context in interpreting an observed object, including the role of co-occurring objects and scenes, abstract properties of an object, as well as facilitation from visual interpretation of a scene at different spatial frequencies. Bar (2007) proposes that multiple possible interpretations of an object are activated, that a contextual frame is identified separately, before the combination of information allows the scene and objects to be recognised. Oliva and Torralba (2007) have studied the properties of co-occurrence between common types of objects in a scene, and the manner in which the presence of an object influences the likely presence of other objects, and the relative areas they may occur. There are many factors involved, including semantic interpretation and visual and spatial information, as well as other statistical properties of observation. This may allow informed guesses about the presence and location of objects in a scene, as well as biasing perceptual interpretation of individual objects (Oliva & Torralba, 2007).

A study by Auckland, Cave, and Donnelly (2007) examined the role of the presence of non-target objects in a scene on recognition of a target object. This showed improved identification of the target through a form of priming effect, , even when other objects are not involved in the task.

Developing understanding of the manner in which context acts is essential for developing effective artificial systems, and for understanding the manner in which important semantic and visual processes operate.

The manner in which context is formed, and how contextual processes between semantic elements are activated is not well understood (Bar, 2004). An important aspect for studying this topic is to implement artificial learning processes that capture the behaviour being studied, allowing examination of the process in detail and in an objective manner, to capture properties essential to creating the observed behaviour. Artificial learning and applied models contribute to the development of artificial systems for practical purposes, to take advantage of aspects of human visual and semantic processes, which are able to handle large amounts of knowledge in a fast and effective manner.

Artificial implementations

A range of approaches have been used to study recognition of objects in scenes. The most successful methods use a series of processing steps including edge detection filters, transformations and classifiers, such as (Mutch & Lowe, 2008). This approach is biologically inspired and makes use of statistical properties of the training data sets, however does not address identification of multiple objects or contextual influences between them.

A number of approaches have addressed the identification of multiple objects in a scene and composition of objects from a hierarchical arrangement of parts (Zhu, Lin, Huang, Chen, & Yuille, 2008; Parikh, Zitnick, & Chen, 2009; Kokkinos & Yuille, 2009). These address factors such as the unsupervised formation of hierarchical clusters, and the use of interaction between bottom-up and top-down processing, to aid identification.

Recent deep learning (Bengio, 2009) methods provide a means of identifying intermediate features in a neural network configuration, which have been used in object recognition (Ranzato, Huang, Boureau, & LeCun, 2007), competitive with state of the art image classifiers. This approach highlights unsupervised identification of intermediate structures, however does not allow interaction of context from a number of objects in a scene comparable with contextual processes in human cognition.

Hoiem, Efros, and Hebert (2008) address the interaction between relative positions and sizes of objects in a scene, and (Oliva & Torralba, 2006) address recognition of the "gist" of a scene based on global features, for use in aiding identification of a scene.

These address a number of factors involved in the role of context in object recognition, however processes underlying the development of features and development of contextual properties of learned features, such as contextual interactions between learned objects, requires further study.

Ambiguous objects

An important topic in visual processing is the resolution of ambiguous low-level observations, which require the interaction of high-level interpretations of the scene in order to be resolved effectively. Yuille and Kersten (2006) discuss the use of Bayesian inference to resolve ambiguous lower features, by using potential high-level objects identified to verify low level features in a top-down manner.

Implementation of an artificial learning system that captures the development of contextual information from observations, and is able to use such context to bias the interpretation of features, offers insight into practical considerations of such processes.

Learning Classifier Systems

Learning Classifier Systems (LCS) (Bull & Kovacs, 2005) are machine learning systems related to Reinforcement Learning (RL) (Sutton & Barto, 1998), which learn to recognise features of the task autonomously, guided by external reinforcement. The key difference between LCS systems and RL is that the representations used for learning incorporate a degree of generalisation, such that rules, or classifiers, developed by the system may be applicable for a wider range of states (Drugowitsch, 2008). The reinforcement process used by reinforcement systems is related to dopaminergic processes in the brain, and provides a robust and flexible method of learning (Samson, Frank, & Fellous, 2010).

The representations used in LCS tend to capture abstract concepts and symbols, related to perceptual observations, and as such provides a higher level of abstraction than artificial neural networks, which model physical properties of individual neurons. This allows a broader perspective of the processes being studied, and address properties that may require a larger scale than that which can be captured in a model based on representation of individual neurons. LCS acts autonomously, and is able to extract features relevant to the task simply based on reinforcement received.

General design of Learning Classifier Systems

Learning Classifier Systems are composed of a population of rules or classifiers, where each can be compared with the currently observed environment, and propose an action that the system can perform, or an interpretation of the observation. The system is presented with an observation from the environment, and each of the rules is tested to see which match and are relevant at the current time. From the set of matching rules an action or classification is chosen, according to the relative weight of each rule.

When external reinforcement is received by the system, the rules which have recently acted receive a reward, which influences the measure of expected reward for using the rule, and also the likelihood that the rule will be maintained in the population.

New rules are regularly created, either as copies of observations, or as modifications of existing rules. Most LCS systems follow a genetic paradigm and use a genetic algorithm to create new rules, based on combinations of pairs of other rules. The population is maintained at a fixed or maximum size, so when new rules are added, the weakest rules are removed from the population.

There are a number of different implementations of LCS systems, which handle the rule selection process, reinforcement and population selection processes in different ways. XCS (Wilson, 1995) is currently the most established method, and maintains rules according to the accuracy of predicting future rewards received by a rule.

Activation-Reinforcement based Classifier System

ARCS (Knittel, 2010) is a form of Learning Classifier System designed to maintain links with cognitive processes. Two properties are maintained for each rule, a measure of *accessibility*, based on the degree of reinforcement of the rule, and *quality*, representing the expected reward when a rule is used. The accessibility property is based on an analogy with memory traces, which are strengthened through use (A. D. Baddeley, 1997), and decline either with time or as a result of competition with newer traces (A. Baddeley, Eysenck, & Anderson, 2009). The method of reinforcement is comparable to that used in ACT (Anderson, 1996). This approach has shown to be effective at balancing generalised and specialised rules (Knittel, Bossomaier, & Snyder, 2007), and provides closer comparison with cognitive processes than other LCS systems.

Effects of context can be incorporated by maintaining associative relationships between objects observed, allowing the context of observations to influence the weight of rules used in the system.

Associative relationships

There are a number of factors influencing context, however one common factor is based on the associative relationships between objects, resulting from co-occurrence relationships. This provides a priming effect, where the activation of an object or concept promotes the activation of associated elements (Auckland et al., 2007).

A number of models are available that study the formation and activation of associative relationships. Anderson's ACT(*/R) models are based on abstract concepts, where properties of interaction between concepts is identified from behavioural studies (Anderson, 1996; Anderson et al., 2004). In this model a network is created with links between associated concepts, and when a concept is activated, the activation is passed along weighted links, following a "leaky capacitor" model.

Another model that is used to capture associations is Hebbian learning, which is based on physiological properties of the strengthening or weakening of synaptic connections between neurons (Haykin, 1998). Connections between a pair of neurons are strengthened if they fire in a correlated manner, and weakened otherwise.

Hebbian learning has been recognised as resulting in weights based on the conditional probability of firing between neurons (O'Reilly, 2001), such as with the CPCA learning rule (O'Reilly & Munakata, 2000):

$$\Delta w_{ij}(t+1) = \eta y_j(t)(x_i(t) - w_{ij}(t)) \tag{1}$$

Conditional probability and Bayesian statistics play an important role in studies of the recognition of objects in context (Torralba, 2003), as well as in general models of the composition of conceptual objects from related features (Tenenbaum, Griffiths, & Kemp, 2006).

Formation of context

Existing learning models have addressed the role of context in a number of ways. Hoiem et al. (2008) use the relative positions of objects in a 3D interpretation of the scene to bias recognition of objects, for example objects in appropriate position and scale relative to other objects in the scene. Other methods include training recognition of pre-defined contextual categories, which are then used to bias interpretation of individual objects (Torralba, 2003; Torralba, Murphy, Freeman, & Rubin, 2003). Rabinovich, Vedaldi, Galleguillos, Wiewiora, and Belongie (2007) use a method where an image is segmented, and each segment is biased to select a segment label that is consistent with other labels in the scene, trained using labelled image datasets.

Context can be loosely defined as a set of consistently co-occurring features or objects. As such, recognition of associations between these features will form a kind of clique, where consistent features of a context will be more strongly interconnected with each other than with other features. This property can be captured in a recurrent network of associations, where the activation of a number of features will strengthen other features in the same context. Assessment of the likelihood of co-occurrence of objects is typically addressed using Bayesian statistics, and an associative network should reflect this property.

There are limitations with the use of a strictly Bayesian approach. For example to evaluate the likelihood of the presence of object 'x' based on the presence of other objects, measurement of the conditional dependency of pairs such as $P(x|a)$ and $P(x|b)$ is not sufficient to accurately resolve $P(x|abc)$, rather the joint distribution must be recorded independently. As the number of objects involved grows, the quantity of statistical measurements required becomes intractable, and approximate methods are needed, which allow assumptions of the environment to simplify the task.

A recurrent network based on co-occurrence pairs provides a means of approximation, and allows an assessment of the activation of element 'x' based on the activations of a number of other elements in the network. Spreading activation networks act in this manner, however connections with Bayesian statistics are not clear.

Hebbian learning allows the formation of a network related to Bayesian statistics. The use of a Hebbian network provides a heuristic for evaluating the presence of 'x' based on a range of other activations. It would be informative to examine further the similarities and differences between activations provided by such a network and known errors of human judgement, such as those described by (Kahneman, Slovic, & Tversky, 1982).

Design of associative network

Hebbian learning modeled on physiological behaviour can be detailed, such as the BCM learning rule (Cooper, 2004), however a number of simplified rules exist, such as the CPCA rule, given in Equation 1.

This rule has been shown to converge on the conditional probability of activation of the pre-synaptic neuron, given activation of the post-synaptic neuron $P(x|y)$, as no modification takes place to the weight when firing is uncorrelated and the post-synaptic neuron is inactive. When constructing an associative network to reflect conditional probability, it is desirable to capture the reverse property, $P(y|x)$. Hebbian learning theory indicates that the emphasis on post-synaptic activation in the CPCA rule is not an essential property of Hebbian learning. When describing the BCM theory, Cooper (2004) states that "plasticity will occur only in synapses that are stimulated presynaptically". This suggests that an alternative form emphasising the role of the presynaptic neuron on synaptic weight changes is appropriate:

$$\Delta w_{ij}(t+1) = \eta x_i(t)(y_j(t) - w_{ij}(t)) \tag{2}$$

As a corollary to the conditional dependency measure reached through training of the CPCA rule (O'Reilly & Munakata, 2000), this training rule converges on the conditional probability $P(y_j|x_i)$.

Nodes in the network represent objects in the environment. Training of the network is performed by setting the nodes active representing objects present, and using the above rule to train connections. Weights are adjusted such that the sum of weights output from each node sums to one.

Evaluation of contextual bias results from activating observed elements, and running a number of activation steps until the network stabilises. Activation is based on a linear neuron model, using an additional inhibitory connection to promote stability. Incorporating external activations, this produces an update rule as follows:

$$y_j = a_j + \sum_i x_i w_{ij} - \varepsilon y_j \tag{3}$$

Training an associative clique

To demonstrate the behaviour of the associative network for identifying context, training is performed on a network of 15 elements using a generative model of co-occurrence between the elements, based on a model of co-location. Each of the elements is specified to occur in each of 3 locations with a given prior probability, and instances are created by selecting a location and evaluating which elements occur according to the specified probabilities.

With training, the weights of the network are shown to converge on the conditional probability values between elements, roughly indicated by the weight of lines shown in Figure 1. The presence of associative cliques can be seen between elements A-E and F-J.

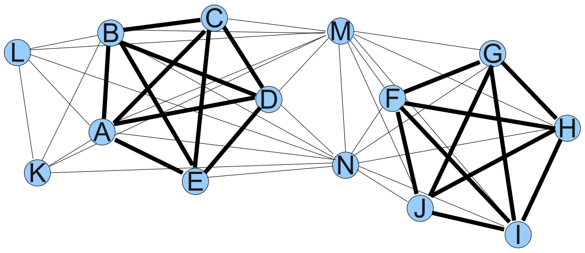

Figure 1: Arrangement of associative network after training.

Behaviour of the network is tested by simulating activation of an ambiguous observation, represented as partial activation of two elements that are possible interpretations, along with a number of co-occurring elements. This is done by applying a value of 0.6 to elements E and H, members of the two dominant cliques. Results of activation for an ambiguous pair with zero, one, two and three co-occurring elements are shown in Figure 2. Without contextual activations the ambiguous interpretations are not distinguished. With one or more parallel activations the elements of the co-occurrence clique are clearly identified, providing a preference for one of the ambiguous interpretations.

Implementation

Object identification

The task used for evaluation is an abstract form of object recognition using a generative model, where objects are represented as collections of features arranged in a two dimensional space. This allows statistical properties of the environment presented to be controlled.

Objects are constructed as a random arrangement of feature symbols in a grid. Co-occurrence relationships are created by defining a number of locations, as described previously. Scenes are produced by selecting a location, determining the objects present according to occurrence statistics, and rendering the chosen objects onto a fixed size grid representing the scene. Objects may be obscured by other objects rendered over top.

The task used to perform training, is to identify the object at a specified location. Object boundaries are not given.

To evaluate the use of context effects, ambiguous objects are created, and the task is to identify which object the ambiguous element represents, based on the co-occurring objects in the scene. To construct an ambiguous representation, first an object is created and assigned to a specific location

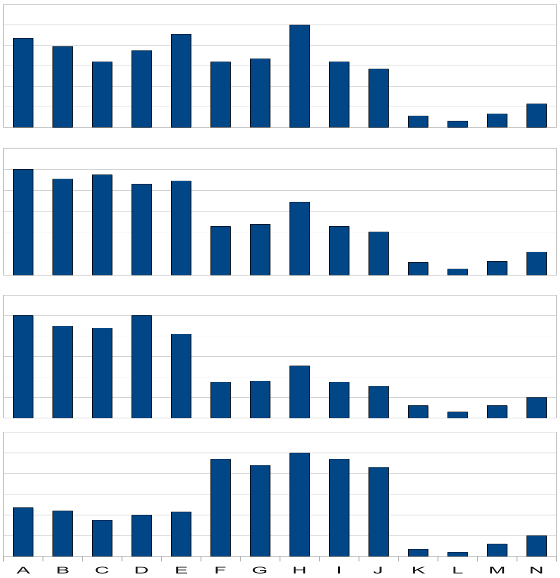

Figure 2: Stable activation level of associative network. All include partial activation of elements E and H. 1. no other activation, 2. single active (B) 3. two active (B,C) 4. three active, members of the other grouping (F,I,J).

and context. A second object is then created as a copy of the first, with a random collection of symbols altered, such that 20% of the representation is varied. The second object is attached to a different location, such that it co-occurs with a different set of objects. A third ambiguous object is constructed by changing each of the modified symbols in the second object to a third symbol, such that the ambiguous object is equally dissimilar to the first and second.

When the ambiguous object is presented as a replacement to the first or second object, it is necessary for the system to identify each of the other objects in the scene independently, and to use the presence of the objects identified to bias interpretation of the ambiguous target object.

Context effects with ARCS

Context effects can be incorporated with the ARCS classifier system, by training the associative network in parallel with the existing reinforcement process.

Rules are constructed as a two dimensional grid of symbols. Each rule is associated with a label indicating the object classification for the rule. When a state is presented, each rule is compared against the state, and a match value established by comparing the rule with each corresponding region of the state. To evaluate the match value for each rule for the assigned target position, the best match value covering the target position is used.

To select the rule to act, the product of the match value with the expected reward of the rule is used. Without using an associative bias, the rule to act is selected according to a Boltzmann distribution:

$$P_n = \frac{e^{\beta q_n m_n}}{\sum_{j=1}^{J} e^{\beta q_j m_j}} \quad (4)$$

where P_n is the probability of choosing rule n, J represents the set of matching rules, q_n is the expected reward and m_n the match level for rule n.

To introduce contextual bias, a fixed size associative network is used according to the number of object classes present. Training of the network is conducted after each classification step. When training with class labels provided, the activation level of each class present is set to 1, and those not present set to 0. A single step of the learning process described in Equation 2 is conducted.

Once the initial activation values for each class are established, the activation process proceeds according to the update rule given in 3. The resulting activation levels for each class are normalised, providing a relative value for each class. This is used to modify the selection value for each rule as follows:

$$P_n = \frac{e^{\beta q_n m_n b_x}}{\sum_{j=1}^{J} e^{\beta q_j m_j b_y}} \quad (5)$$

where b_x is the associative bias for object x, where the given rule is linked with object x.

It would be possible to capture context using a rule-based system, for example using a rule specifying that if object A is present at the same time as object B, a particular classification would be preferred; such an implementation would allow suitable classification, however involves a design that implies the solution, and is based on a discrete definition of context, requiring each combination of contextual features to be defined at the rule level. In contrast, the associative method allows contextual effects between a range of elements to be captured in the network, using a soft definition of context arising from activation properties of the network.

Results

Training is first conducted with object labels available for each position, before evaluation is performed on observations without labels. A number of training stages are used to simplify the learning process. First training is conducted on individual objects, where a single object is presented for each step. After 200,000 training steps, classification is performed with 100% accuracy on individually presented objects. The second training stage involves classification with multiple objects present, and introduces training of the associative network.

Observations are presented using a 30 x 30 grid of features, each object is constructed using on average 34 symbols, including white space. 4 locations are used to generate the environment, with 25 objects, as well as 2 ambiguous representations.

Evaluation is performed by presentation of multiple objects in a scene, including ambiguous objects, which require the use of context for identification.

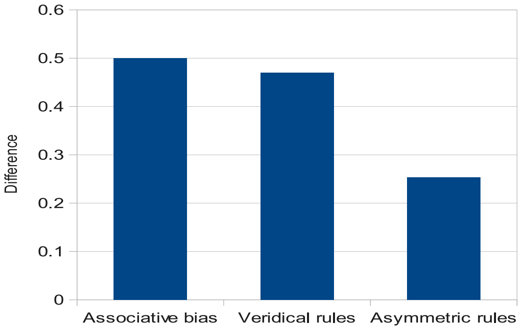

Figure 3: Result of the use of associative bias. 1. bias in match condition towards rule consistent with context, 2. performance improvement for rule set consistent with object representation, 3. performance improvement for rule set with assymetric preference to one interpretation

Results of the degree of bias and improved classification rates on ambiguous elements are shown in Figure 3. The degree of bias variation provided to classifications consistent with the associative context is +0.5, for example a bias value of 0.7 vs 0.2, providing significant discrimination. The result of classification of ambiguous elements without the use of associative bias is at chance level, at 0.47%, for a rule set consistent with the features present in the environment. With introduction of associative bias, allowing context to influence identification, identification accuracy was increased to 0.94%, an improvement of 0.47. With a rule set that has developed in a manner that is preferential to one interpretation of an ambiguous element, the improvement is 0.25.

These figures show clear discrimination between interpretations of ambiguous elements, as a result of bias introduced through association, with lesser improvement when assymmetric preference is given to one interpretation.

Discussion and Conclusions

This system provides an autonomous method for studying processes involved in formation of concepts relevant to a task, guided by reinforcement from behaviour, in tandem with the recognition of context through co-occurrence relationships captured in an associative network. The development of features autonomously according to reinforcement, acting alongside formation of contextual associative links, provides a novel means of studying formation of context, and the manner in which it can be used to influence a task.

The interaction of the two types of systems allows the use of a simplified associative implementation, acting in tandem with the reinforcement based learning process, to allow context identified between multiple elements to influence the recognition of objects being observed.

The reinforcement based system provides a useful platform for examining practical considerations of such processes in an objective, autonomous manner. Representation of context is captured in attractor basins produced from associative links in a recurrent network. The advantages provided for the interpretation of objects, and the ability to identify context in an autonomous manner, highlight the applicability of a recurrent associative network for addressing how context can be formed and utilised.

References

Anderson, J. R. (1996). *The architecture of cognition*. Cambridge, MA, USA: Harvard University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Auckland, M., Cave, K., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic bulletin & review*, *14*(2), 332.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. Hove, UK: Psychology Press.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Hove, UK: Psychology Press.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bengio, Y. (2009, January). Learning deep architectures for ai. *Found. Trends Mach. Learn.*, *2*, 1–127.

Bull, L., & Kovacs, T. (2005). *Foundations of learning classifier systems*. Berlin: Springer-Verlag.

Cooper, L. (2004). *Theory of cortical plasticity*. Singapore: World Scientific Pub Co Inc.

Drugowitsch, J. (2008). *Design and analysis of Learning Classifier Systems: A probabilistic approach*. Berlin: Springer.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Hoiem, D., Efros, A., & Hebert, M. (2008). Putting objects in perspective. *International Journal of Computer Vision*, *80*(1), 3–15.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. New York: Cambridge University Press.

Knittel, A. (2010). An activation reinforcement based classifier system for balancing generalisation and specialisation (arcs). In *Gecco '10: Proceedings of the 12th annual conference on genetic and evolutionary computation* (pp. 1871–1878). New York, NY, USA: ACM.

Knittel, A., Bossomaier, T., & Snyder, A. (2007). Concept accessibility as basis for evolutionary reinforcement learning of dots and boxes. In *IEEE symposium on computational intelligence and games.*

Kokkinos, I., & Yuille, A. (2009). HOP: Hierarchical object parsing. *CVPR*, *0*, 802-809.

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.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. In S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, & P. Tse (Eds.), *Visual perception - fundamentals of awareness: Multi-sensory integration and high-order perception* (Vols. 155, Part 2, p. 23 - 36). Elsevier.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Computation*, *13*(6), 1199-1241.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain* (1st ed.). Cambridge, MA, USA: MIT Press.

Parikh, D., Zitnick, C., & Chen, T. (2009). Unsupervised learning of hierarchical spatial structures in images. In *IEEE conference on computer vision and pattern recognition* (pp. 2743–2750).

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007, oct.). Objects in context. In *IEEE international conference on computer vision* (p. 1 -8).

Ranzato, M., Huang, F. J., Boureau, Y.-L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE conference on computer vision and pattern recognition* (p. 1 -8).

Samson, R., Frank, M., & Fellous, J.-M. (2010). Computational models of reinforcement learning: the role of dopamine as a reward signal. *Cognitive Neurodynamics*, *4*, 91-105.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309 - 318.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, *53*, 169-191.

Torralba, A., Murphy, K., Freeman, W., & Rubin, M. (2003, oct.). Context-based vision system for place and object recognition. In *IEEE international conference on computer vision* (p. 273 -280 vol.1).

Wilson, S. W. (1995). Classifier fitness based on accuracy. *Evolutionary Computation*, *3*(2), 149-175.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301 - 308.

Zhu, L., Lin, C., Huang, H., Chen, Y., & Yuille, A. (2008). Unsupervised structure learning: hierarchical recursive composition, suspicious coincidence and competitive exclusion. *ECCV*, 759–773.