A BIOLOGICALLY-INSPIRED CONCEPT FOR ACTIVE IMAGE RECOGNITION

Roland E. Suri  
Intelligent Optical Systems  
2520 W. 237th St., Torrance, CA 90505  
(310) 530-7130  
rsuri@intopsys.com

Abstract—A novel concept is proposed that uses active information gathering for recognizing objects on images. This novel concept mimics recent neurobiological insights on human control of eye movements (TD algorithm). TD algorithms are predictive reinforcement algorithms for learning of action sequences. In the proposed framework, standard techniques, such as template matching, are used as processing steps. Each processing step compares a template of one part of the object with image locations and computes a value that describes how well the template matches. A TD algorithm is trained on many images to optimize the sequence of processing steps by providing feedback whether the final object recognition was correct or incorrect. After training, the algorithm searches for template matches with the sequence of templates and locations that are most promising for recognition of a certain object. This object recognition strategy resembles active information gathering by saccadic eye movements.

1. INTRODUCTION

Selection of human movements, such as saccadic eye movements, is influenced by learning and planning processes, which are mediated by midbrain dopamine neurons (see for example recent papers by R.E. Suri on www.cnl.salk.edu/~suri). The activity of dopamine neurons is increased by stimuli that increase the animal’s reward expectation and is decreased below baseline levels when the reward fails to occur. These and other characteristics resemble the reward prediction error signal of the temporal difference (TD) model, which is a model of reinforcement learning. Computational modeling studies demonstrate that a dopamine-like reward prediction error can serve as a powerful teaching signal for learning with delayed reinforcement and for learning of motor sequences. In addition to acting as a teaching signal, dopamine seems also to be involved in “cognitive” processes that are not addressed by standard TD models. In particular, dopamine neuron activity seems to be crucial for planning processes, which select actions by evaluating their predicted outcomes. Such “cognitive” processes seem to be required for active information gathering.

In the framework proposed here, a sequence of actions is optimized with TD learning techniques to maximize recognition reliability, which is similar to the control of eye movements. TD learning algorithms have been successful in computer science applications, such as for learning to play backgammon on a world-class level by trial and error. In computer science terms, TD learning algorithms receive the current state and the current return as the model input, and compute actions, which maximize the return over time. In backgammon, the state corresponds to the situation on the board, the actions to the half-moves, whereas the return is positive when the game is won, negative when the game is lost, and zero during the game.

To apply the proposed TD framework to face recognition, a TD model is trained to maximize the recognition rates. Actions include for example “search nose”, “search eye near found nose,” or “search mouth of person A.” These actions are implemented by comparing a nose template with the current image for all the image locations candidates. Other actions are “scale size of image templates to be consistent with the previously found nose” or “rotate templates to be consistent with found nose.” Scaling, rotations, gray scale corrections, and view angle corrections of the templates are used to adjust the image templates to an image component that has already been found. Actions also indicate the decision to give the final result with “categorize test image as person X”, or “no face found on image.” Each state reflects the whole sequence of image analysis steps that has been performed on the current image. A state indicates for example that action “search nose was successful”, but “search mouth was not successful.” A return with positive value is given for the correct classification of an image, a return close or equal to zero is given as long as no classification has been decided, and a return with negative value is given for an incorrect classification.

This framework has significant advantages over currently popular image recognition strategies: 1) Similar to cognitive processes that direct human eye movements, the proposed framework is able to actively gather information. For example, the framework is able to learn to “look for birth mark on upper lip”, if the found nose resembles that of the person with this birthmark. 2) If the algorithm finds a certain feature, it is able to adjust the other feature templates to the corresponding size, view angle, and rotation. 3) The proposed framework automatically optimizes the processing steps, whereas in most previous frameworks a series of ad hoc processing steps have to be optimized by the algorithm developer. 4) Since the framework selects the most successful image processing steps, it serves as a research tool to compare the performance of image processing techniques that have been developed in previous studies.
2. TD LEARNING AS MODEL FOR BIOLOGICAL LEARNING AND PLANNING

In a famous experiment Pavlov (1927) trained a dog with the ringing of a bell (stimulus) that was followed by food delivery (reinforcer). In the first trial, the animal salivated when food was presented. After several trials, salivation started when the bell was rung suggesting that the salivation response elicited by the bell ring reflects anticipation of food delivery. A large body of experimental evidence led to the hypothesis that such Pavlovian learning is dependent upon the degree of the unpredictability of the reinforcer (Rescorla & Wagner, 1972; Dickinson, 1980). According to this hypothesis, reinforcers become progressively less efficient for behavioral adaptation as their predictability grows during the course of learning. The difference between the actual occurrence and the prediction of the reinforcer is usually referred to as the “error” in the reinforcer prediction. This concept has been employed in the temporal-difference model (TD model) of Pavlovian learning (Sutton & Barto, 1990). The TD model uses a reinforcement prediction error signal to learn a reinforcement prediction signal. The reinforcement prediction error signal progressively decreases when the reinforcement prediction signal becomes similar to the desired reinforcement prediction signal. Characteristics of the reinforcement prediction signal are comparable to those of anticipatory responses such as salivation in Pavlov's experiment and may guide approach behavior (Montague et al., 1995).

In Pavlov's experiment, the salivation response of the dog does not influence the food delivery. Consequently, the TD model computes predictive signals but does not select optimal actions. In contrast, instrumental learning paradigms, such as learning to press a lever for food delivery, demonstrate that animals are able to learn to perform actions that optimize reinforcement. To model sensorimotor learning in such paradigms, a model component called the Actor is taught by the reward prediction error signal of the TD model. In such architectures, the TD model is also called the Critic. This approach is consistent with animal learning theory (Dickinson, 1994) and was successfully applied to machine learning studies (Sutton & Barto, 1998).

The reinforcement prediction error signal of the TD model remained a purely hypothetical signal until researchers discovered that the activity of midbrain dopamine neurons in substantia nigra and ventral tegmental area is strikingly similar to the reward prediction error of the TD model (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998; Suri & Schultz, 1999).

**Temporal Difference (TD) Model**

The TD algorithm is popular in machine learning studies and was proven to converge to the optimal solution (Dayan & Sejnowski, 1994). Despite these successes, their development was strongly influenced by studies of animal learning (Sutton & Barto, 1990, 1998). Since animals often learn to estimate the time of the reward occurrence in Pavlovian learning paradigms, the TD model uses a time estimation mechanism (Sutton & Barto, 1990). This time estimation mechanism is implemented using a temporal stimulus representation, which consists of a large number of signals \( x_m(t) \) for each stimulus. Each of these signals \( x_m(t) \) has a value of one for one time point and is zero for all other times. Exactly one signal of the temporal stimulus representation \( x_m(t) \) peaks for each time step of the period between the stimulus and the trial end (Figure 1A). Similar hypothetical temporal stimulus representations have also been referred to as “complete serial compound stimulus” (Sutton & Barto, 1990) or “spectral timing mechanism” (Brown, Bullock, & Grossberg, 1999). The shape of these signals is not important for the algorithm, but the number of signals has to be sufficiently large to cover the duration of the intertrial interval (\( m = 1, 2, ..., 50 \) for 5 seconds interstimulus interval with time steps of 100 msec). The reward prediction \( P(t) \) is computed as the weighted sum over the temporal stimulus representation signals \( x_m(t) \) with

\[
P(t) = \sum_{m=1}^{50} V_m(t)x_m(t).
\]

The algorithm is designed to learn a “desired” prediction signal that increases successively from one time step to the next by a factor \( 1/\gamma \) until the reward \( \lambda(t) \) occurs and decreases to the baseline value of zero after the reward presentation. The prediction error signal is computed with

\[
r(t) = \lambda(t) + \gamma P(t) - P(t-1)
\]

and is zero as long as the prediction signal is equal to the desired prediction signal and nonzero otherwise. Since one time step corresponds to 100 msec, \( t-1 \) is a short hand for \( t-\)100 msec. The value of a discount factor \( \gamma \) is set between zero and one (Table 1).

The adaptive weights \( V_m(t) \) are initialized with the value zero and adapted according to the learning rule

\[
V_m(t) = V_m(t-1) + \beta r(t)x_m(t-1),
\]

with a small learning rate constant \( \beta \) (Table 1). The TD model can be represented with a neuron-like element whose weights \( V_m(t) \) correspond to synaptic conductances (Figure 1B).

When the stimulus is followed by the reward for the first time, the reward prediction is zero and the reward prediction error is phasically increased at the time of the reward (Figure 1C). After repeated presentations of the stimulus followed by the reward, the reward prediction increases before the anticipated reward. Characteristics of this reward prediction signal resemble those of reward anticipatory behaviors of animals (Sutton & Barto, 1990). The rate of this gradual increase is determined by the constant \( \gamma \), which is referred to as the temporal discount factor. We use the value \( \gamma = 0.99 \) per 100 msec, which leads to an increase in the prediction signal of 1% for each 100 msec. The reward prediction error signal is at the time of the stimulus equal to
the change in the reward prediction. Since dopamine responses decrease proportionally to the learned duration of the interval between the stimulus and the reward, dopamine neuron activity was used to estimate the value of the discount factor (Suri & Schultz, 1999). At the time of the reward, the reward prediction error is zero because the change in the prediction signal cancels out the reward signal.

The prediction error signal of the TD model is strikingly similar to activities of midbrain dopamine neurons (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998; Suri & Schultz, 1999). The prediction error signal is phasically increased by unpredicted reward and by the earliest reward-predicting stimulus, and it is negative when a predicted reward is omitted (Figure 2, left). This signal closely resembles dopamine responses (Figure 2, right). The depression in dopamine activity below baseline levels at the time of the predicted but omitted reward reflects a central timing mechanism because no stimulus is present at the time of the omitted reward.

The reward prediction error signal of the TD model by Suri and Schultz (1999) reproduces dopamine neuron activity in the situations: (1) upon presentation of unpredicted rewards, (2) before, during, and after learning that a stimulus precedes a reward, (3) when two stimuli precede a reward with fixed time intervals, (4) when the interval between the two stimuli are varied, (5) in the case of unexpectedly omitted reward, (6) delayed reward, (7) reward earlier than expected (Hollerman & Schultz, 1998), (8) in the case of unexpectedly omitted reward-predictive stimulus, (9) in the case of a novel, physically salient stimulus that has never been associated with reward (see allocation of attention, below), (10) and for the blocking paradigm (Waelti,
Dickinson, & Schultz W., 2001). To reach this close correspondence, three constants of the TD model were tuned to characteristics of dopamine neuron activity (learning rate, decay of eligibility trace, and temporal discount factor), some weights were initialized with positive values to achieve (9), and some ad hoc changes of the TD algorithm were introduced to reproduce (7).

In many computational frameworks, it is not possible to implement active information gathering, as they process the image as a whole. This is for example the case for Eigenface methods, which use singular value decomposition, cluster analysis methods, or Bayesian methods.

Frameworks based on Hidden Markov Models (HMMs) process the image regions sequentially, but the theory is not sufficiently advanced to incorporate actions. Applied to face recognition, these methods identify for each person the coefficients of a HMM. The image is typically sliced in a series of sub-images and the observation vector for the HMM corresponds to one image slice (Samaria, 1994). It would be advantageous if the observation slices would not have to be fixed, but would contain the same face components for all images. However, a proper solution to this problem would require the definition of actions as a part of Markovian Decision Process (MDP) (Minut, 2000; Henderson, 2000). Unfortunately, the underlying states of this MDP are not directly observable, as each state corresponds to an image region and two image regions may look identical. This leads to a Partially Observable Markovian Decision Process for which finding the optimal actions is intractable even for problems with as few as 30 states (Littman, 1996). The proposed framework addresses this issue by defining the sequence of action outcomes on the current image as the current state. This also has the advantage that the same model parameters can be used for all images, which drastically reduces the number of required states. Furthermore, the problem becomes a MDP that can be solved with TD learning (Sutton and Barto, 1998).

In the proposed framework, the sequence of image processing steps is optimized using TD learning. Image processing steps can include the computation of correlations with templates for parts of the object of interest, as well as rotation, color-adjustments and scaling of templates. In the case of face detection, such templates may consist of typical images of a nose, an ear, or an eye. Such templates could also be fabricated using local feature detection, independent component analysis, or similar algorithms that find localized templates. The TD algorithm uses example images to optimize the sequence of image processing steps. The optimized sequence of image processing steps may consist of: 1) search eye by correlation with eye template, 2) turn eye template by 10 degrees, 3) search eye by correlation with eye template, 4) search nose by correlation with nose template, and so on (Figure 3).

![Figure 2.](image.png)
Figure 2. Prediction error signal of the TD model (left) similar to dopamine neuron activity (right) (figure adapted from Suri & Schultz, 1998; discount factor $\gamma = 0.98$). If a neutral stimulus A is paired with reward, prediction error signal and dopamine activity respond to the reward (line 1) (activities reconstructed from Ljungberg, Apicella, & Schultz, 1992; Mirenowicz & Schultz, 1994). After repeated pairings, the prediction error signal and dopamine activity are already increased by stimulus A and on baseline levels at the time of the reward (line 2). After training with an additional stimulus B, which precedes stimulus A, prediction error signal and dopamine activity are increased by stimulus B and neither affected by stimulus A nor by the reward (line 3). If the stimulus A is conditioned to a reward but is occasionally presented without reward, the prediction error signal and dopamine activity are decreased below baseline levels at the predicted time of reward (line 4). (Activities lines 2-4 reconstructed from Schultz, Apicella, & Ljungberg, 1993).

2. ACTIVE IMAGE RECOGNITION WITH TD LEARNING

For most, or perhaps all, image recognition tasks, the performance of human observers is unmatched by computational approaches. Studies of human eye movements suggest that humans actively gather information: Humans are able to use some clues to form expectations and verify such expectations by looking for more specific markers. In this process, humans are likely to use a model of the suspected object and may adjust this model according to observed clues. Such adjustments may include turning, scaling, gray scale changes, and view angle changes. Current image recognition systems are not able to actively gather information.

For most, or perhaps all, image recognition tasks, the performance of human observers is unmatched by computational approaches. Studies of human eye movements suggest that humans actively gather information: Humans are able to use some clues to form expectations and verify such expectations by looking for more specific markers. In this process, humans are likely to use a model of the suspected object and may adjust this model according to observed clues. Such adjustments may include turning, scaling, gray scale changes, and view angle changes. Current image recognition systems are not able to actively gather information.

For most, or perhaps all, image recognition tasks, the performance of human observers is unmatched by computational approaches. Studies of human eye movements suggest that humans actively gather information: Humans are able to use some clues to form expectations and verify such expectations by looking for more specific markers. In this process, humans are likely to use a model of the suspected object and may adjust this model according to observed clues. Such adjustments may include turning, scaling, gray scale changes, and view angle changes. Current image recognition systems are not able to actively gather information.
Figure 3. A subset of potential image processing sequences for face detection. The proposed TD framework will use Temporal Difference learning to optimize the sequence of actions for reliable image recognition. Actions include choices such as “look for eye”, “turn eye template,” and “face detected.” The outcome of each action (blue highlighted) will be evaluated to determine the next state (shown as circle). Temporal Difference learning (TD learning) will use training images to learn to estimate the value of each state (bright = low value, dark = high value), and thus to maximize the chance to correctly classify the image. Inconsistent information, such as “eye found” and “nose not found” reduces the value of a state, whereas consistent information, such as “left eye found” and “right eye found” increases the value of the state. TD learning uses the values to select the optimal image processing actions.

There are several popular TD learning methods that may be considered for the current application, such as Actor-Critic architecture, SARSA, and Q-learning. Although all TD learning algorithms replicate dopamine neuron activity, the biological implementation of TD learning may be best described with an Actor-Critic architecture (Suri and Schultz, 1998; Suri and Schultz, 1999; Suri, Bargas and Arbib, 2001; Suri, 2002). However, in contrast to biology, a non-linear Actor network may not be necessary for the current application. Furthermore, the training of Actor-Critic architectures requires the simultaneous learning in the Actor and in the Critic, which requires setting the additional Actor learning rate to a reasonable value. I thus favor Q-learning, which is one of the most popular TD learning algorithms. Q-learning uses only one learning mechanism, which makes it easier to train than an Actor-Critic architecture. In contrast to other TD learning frameworks, it learns a value for each state-action pair instead of learning a value of each state. In the current framework, no function approximators will be used to learn the value function; as for the current application I am not aware of rules how to design a function approximator with good generalization properties. Instead of using a function approximator, each state-action pair will be represented by a binary vector with only one non-zero entry that corresponds to this state-action pair. The dimension of the state space will thus be equal to the number of states. This sparse coding guarantees that the Q-learning algorithm will find the optimal action sequence (Sutton and Barto, 1998).

Acknowledgements
I thank Igor Ternovskiy (IOS) and Alexander Trunov (IOS) for comments on earlier versions of this manuscript.

4. REFERENCES


