Utilizing Q-Learning to allow a radar to choose its transmit frequency, adapting to its environment

Leon O. Wabeke* and Willem A.J. Nel†
DPSS, CSIR, Pretoria
South Africa
*lwabeke@csir.co.za
†wajnel@csir.co.za

***Abstract*—Recent research show that utilization of knowledge of the environment can allow a radar system to adapt its processing to improve its performance. Furthermore, a radar system that utilize both a-priori and measured knowledge in an adaptive close loop manner could seem to be cognitive of its environment, able to adapt to changes to optimize performance. Reinforced learning could play a vital role as part of such a closed-loop cognitive radar system. The Q-Learning algorithm is hypothesized to be useful for this cognitive radar domain. This paper investigates the problem of adaptively choosing the radar transmit frequency through application of Q-Learning on measured radar data. A comparison is made against other frequency selection algorithms and its shown that Q-Learning manages to learn a good strategy to adaptively select radar transmit frequency, mostly outperforming the other methods tested in the scenario investigated here.**

I. Introduction

This paper concerns itself with the application of radar to perform persistent, ubiquitous surveillance of a littoral region from a land based radar, with particular focus on the use of cognitive processing principles to make decisions about the radar waveform.

In missions like anti-abalone poaching or sea rescue the targets of interest can be very small e.g. a single person drifting in the water. This requires very high radar sensitivity and can be exasperated by missions coinciding with bad weather conditions, making environmental clutter an even bigger problem.

Recently the term Cognitive radar has been introduced [1] to describe the concept that the radar needs to learn from all information it receives from the environment and adjust both its processing and waveforms based on this knowledge. In the context of a persistent ubiquitous surveillance system, the persistent nature of the system allows it to experience a large variety of environmental conditions to gain "experience". A cognitive system that harness and learns from the "experience" has the potential to improve its performance over time as measured against a particular objective. The ubiquitous nature of the system can also benefit small target detection and tracking. One such benefit could be easier track maintenance through areas of high clutter once detected.

II. Traditional Radar adaptation

Certain a priori knowledge about the environment and mission of the radar is used at design time to define the hardware, algorithms and performance specifications needed by the system.

Other forms of information can also be useful to the radar system. In particular the use of geographical information systems (GIS) has been shown to be of value in a number of the radar receiver processing steps: for example in terms of determining expected clutter statistics [2] [3] and in terms of adjusting tracking association rules [4]. Radar systems employing such techniques are often termed knowledge-aided radar.

Beyond this a priori knowledge Constant False Alarm Rate (CFAR) detectors [5] have been used to adapt to the average noise/interference level. But in many radar systems very little further direct adaptation to the environment is performed.

Adaption to the environment can either be done by changing the way the radar processes received signals termed, *reactive* in the context of this paper, or can also change the radar transmission to influence the sensing process, termed *active*.

Traditional techniques were only concerned with the detection and tracking of targets in Gaussian noise, including the pulse compressor with matched filter [6] or mis-match filtering [7], many variants of CFAR detectors [5] and Kalman filters [8]. To deal with the practical case of non-Gaussian noise, pre-whitening is proposed [9], to flatten the interference spectrum. Alternatively for interference from specific clutter scatterers, an iterative adaptive matched filtering process can be used to adapt to the reflected signal and the clutter from the nearby range cells [10]. These algorithms adapt how the received signal is interpreted, but does nothing to control the radar waveform, making them reactive.

Active adaptation techniques have also been proposed. In [11] matching of the transmitted waveform to target categories to maximise the probability of classification is proposed. Matched illumination techniques to maximise target detectability has also been proposed [12].

In radar target classification both the transmit waveform and receive processing are typically heavily dependent on the exact target type. Various techniques exist, each optimized to deal with a specific class of target. Such techniques include High Range Resolution (HRR) [13], Inverse Synthetic Aperture Radar (ISAR), micro-Doppler [14] [15] and many others. Even though these techniques have been shown to be successful for their particular target class, there remains a requirement for

978-1-4244-6458-6/10/$26.00 ©2010 IEEE

an adaptive decision process to choose the correct technique based on the target under consideration.

As the radar learns about its environment, it can also attempt to utilize that knowledge for resource scheduling. Recent research in this area has focused on beam-steering in phased-array radar [16] necessitated by the flexibility and maneuverability that such an antenna provides. Resource scheduling for mechanically steered systems hasn't received a lot of attention, although some work has been done for multi-function search and tracking radars on constant angular speed positioners [17].

III. Cognitive Radar adaptation

Inspired by the complex behavior observed in bat echo-location, the idea of Cognitive radar has been introduced [1] with the aim of designing adaptive radar systems that learn from all information it receives from the environment (both targets of interest and background effects to be suppressed) and then adjusting its transmission waveforms based on this knowledge to optimize its mission performance. This concept is further expanded in [18] where it is proposed to use a Bayesian filter implemented as a Cubature Kalman filter to estimate information from the channel and Q-Learning to select the transmitted waveforms from a library of available waveforms.

Q-Learning is a computationally efficient form of a re-inforced learning algorithm for dynamic programming in a Markovian environment [19]. Dynamic programming is a much older approach to determining optimal decision making policies for sequential optimization, which has helped form the foundation of such algorithms like the Viterbi decoder [8].

Thus the approach of using a Cubature Kalman filter in combination with a Q-Learning algorithm to form a closed-loop cognitive learning system seems promising. Since the application of Q-Learning in radar is relatively new, there are questions that still need to be addressed regarding the performance, stability, optimal state-mapping and the extent of waveform libraries required to ensure sufficient freedom for the system to learn and adapt to its environment.

The use of Q-Learning in a radar tracking problem is considered in [20] [21]. The state space representation proposed is the target location(s) in the range-Doppler map. An example is shown in [20] for a 4 state simulation. In operational systems, the radar range-Doppler or range-bearing map is frequently over 100000 elements. The authors couldn't locate any reports in the literature of how the state space have been/ would be scaled to practical sizes in an operational radar. In [21] this scaling problem is bypassed by using a modified Q-Learning algorithm.

This paper considers the practical application of Q-Learning on measured radar data with an alternate approach to the state-space definition.

A. Q-Learning in more detail

As mentioned Q-Learning is a reinforced learning algorithm to learn the optimal strategy in a Markovian environment.

The Markov process consists of a set of states at each time instant with a set of actions available from each state. For the process to be considered Markovian, the transition to the next state may only depend on the current state and current action and potentially a memory-less random variable. Associated with each state-transition pair is a reward distribution. This reward need not be fixed, but can depend on a particular probability distribution. In dynamic programming in general there can potentially be a unique reward associated with the state at the final time instant, but the type of processes considered to be Markovian are stationary, of infinite duration, with no special end-time. Each time instant is similar to the next, in that the states, actions and transition and reward probabilities do not change.

In such an environment, it is useful to be able to have a policy with which to choose the action in each state in such a way as to optimize the sum of the likely reward over time. In an infinite duration environment, the sum of the rewards will generally be infinite, thus to make this optimization meaningful an average reward rate normalized per time instant is typically used. An alternative way to achieve a bounded reward function that is used in the Q-Learning environment, is to discount future rewards, i.e. the further into the future a reward will occur, the less its current value is. Thus

$$C_t(R_{t+n}, n) = \gamma^n R_{t+n} \tag{1}$$

where C_t is the current discounted value at instant t of the future (expected) reward R_{t+n} earned at instant $t+n$ and γ is the discount factor, with $0 \leq \gamma < 1$.

The sum of the expected discounted future rewards at time instant t and using action A_t at time instant t and a specific policy thereafter is thus

$$\begin{aligned} Q_{t,S_t,A_t} &= \sum_{n=0}^{\infty} C_t(R_{t+n}, n) \\ &= R_t + \gamma \sum_{n=1}^{\infty} C_t(R_{t+n}, n) \\ &= R_t + \gamma Q_{t+1,S_{t+1},A_{t+1}} \end{aligned} \tag{2}$$

where S_{t+1} is the expected next state by following the policy and $Q_{t+1,S_{t+1},A_{t+1}}$ is the expected reward obtainable from that future state. But since the system is stationary, $Q_t = Q_{t+n}$ and the subscript t for Q can be dropped, resulting in the simpler notation

$$Q_{S_t,A_t} = R_t + \gamma Q_{S_{t+1},A_{t+1}} \tag{3}$$

Low values of γ would correspond to a short term planning policy, where short-term rewards are valued more than longer term future rewards. In the extreme $\gamma = 0$, would be a myopic/greedy policy that only looks at the immediate reward, without considering future resulting states. Values of γ close to 1, would result in decision strategies that correspond to a long term planning horizon.

The optimal policy at a state S_t would correspond to choosing the action corresponding to the maximum Q value

in a particular state, thus

$$V_S = \max_A Q_{S_t,A_t} \tag{4}$$

where V_S is the expected discounted future reward for choosing the optimum action A in state S and following the optimum policy thereafter.

The form of Q-Learning utilized in this paper, is based on a direct implementation of 3 and 4. A matrix Q is maintained, consisting of an approximation of the $Q_{S,A}$ values. This form has been shown to converge to the real $Q_{S,A}$ values under specific conditions [19]. At each time instant the action is chosen based on this Q matrix as if it was an accurate approximation. After the reward for that specific state-transition has been received, the element of Q corresponding to this transition is updated as follows:

$$Q^*_{S,A} = (1-\alpha)Q_{S,A} + \alpha(R + \gamma V_{S_1}) \tag{5}$$

where $Q^*_{S,A}$ is the new approximation of the element corresponding to the state S (old state) and the action A that was used,

Although it is not explicitly required in the convergence proof of [19], other descriptions of implementations of Q-Learning in the literature for example [22], use an exploration factor. This defines a probability with which a random action is chosen instead of the optimal action based on the policy. This helps prevent early convergence to a local maximum, since too early convergence will prevent other state-action transitions from being explored.

The convergence proof [19] also doesn't put requirements on the initial value of Q. It is just assumed that an initial set of values is given. In practice, underestimating the values of Q leads to local maximums, since as soon as a state-action transition is explored, it becomes a better path, with higher rewards and no further state-action pairs are explored. Using initial values of Q that overestimate the expected rewards, seem to work better: As soon as a state-action is evaluated, the resulting reward seems "disappointing" and next time a different state-action pair is rather explored, until all expectations become more realistic. This forces exploration of all state-action pairs initially during this orientation phase.

But this exploration comes at a price: the orientation phase, to try all the combinations of state-action pairs, will mostly result in bad choices. The real learning will be delayed by this time. This highlights the curse of dimensionality: A large number of states and/or actions will result in an even bigger set of combinations, taking a long time to learn.

In cases where there are correlations between neighboring states and/or actions, variations of Q-Learning could be used to leverage this correlation and simultaneously update multiple of the elements of the Q matrix based on this correlation, cutting down dimensionality. The highly ambiguous nature of the phase of clutter returns mostly decorrelates adjacent frequencies [23], thus such techniques have not been utilized in the experiment considered in this paper.

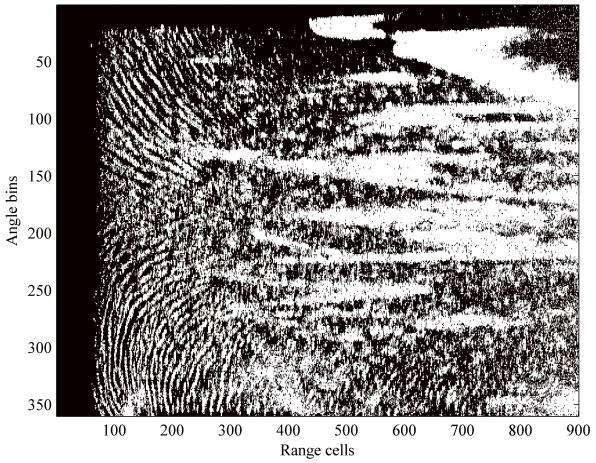

Fig. 1. Randomly chosen example of a single frequency clutter map showing cells above the threshold

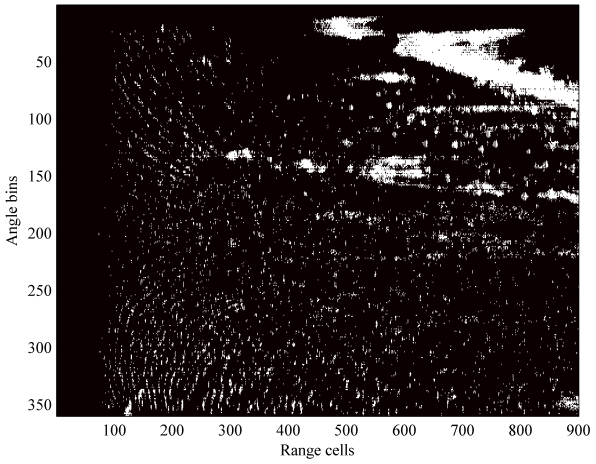

Fig. 2. Clutter map showing lowest clutter value versus frequency in each cell

IV. Experiments

In order to explore Q-Learning in a radar context, the search mode of a surveillance radar is considered. The challenge presented to the Q-Learning system was to learn about the frequency response of the clutter interference and to control the radar's transmit frequency in order to provide good target detection performance.

A. The dataset

Recorded data from an X-band pencil beam tracking radar was used for this simulation. The radar was deployed on the coast near Simon's Town, overlooking the False Bay area. The presence of a mountain behind the radar gave it an effective 180° area that could be observed. A scan was done over this sector at a 0° elevation angle. A 10 kHz PRF was utilized, giving an unambiguous range of 15 km. This unambiguous range meant that some of the mountains across the bay was

in the second and third ambiguity and folded back to produce land clutter spots within the ocean area. During this scan the radar interleaved transmit waveforms covering a part of the X-band frequency band. The dataset used, consisted of 49 different frequencies in this band.

The radar's range resolution is 15m. Due to transmitter guard times and processing setup delays, 900 range samples where obtained for each pulse. The data was binned into 0.5° azimuth bins, thus giving 360 azimuth bins over the sector.

A constant STC level was used when the dataset was recorded, thus the data exhibits a drop off in amplitude as range increases. A 44 sample non-linear frequency modulated pulse was used and matched filtering (pulse compression) was applied. The data was not explicitly calibrated over frequency, except for the inherent radar design which aims for constant transmit power versus frequency.

Thus the complete data can be represented as a 3 dimensional matrix of 49 frequencies by 360 angular bins by 900 range bins.

For the purposes of the experiment, the clutter energy was scaled by a R^4-law to represent equivalent clutter RCS. An arbitrary threshold of RCS level was chosen such that about half of the samples are below this threshold. A target would be assumed to be detectable if the clutter level in its range-azimuth cell was below the threshold RCS, whereas it would assume to stay undetected if the clutter level is above this RCS threshold. A typical CFAR will look at the statistics of surrounding cells to estimate the threshold, but in order not to get bogged down in CFAR technique optimization, this simplified approach was adopted for the simulation.

Given this threshold 57.78%, of the range-azimuth cells were below the threshold. For each frequency individually, this varied between 54% and 63%. An example of a clutter map for a randomly chosen frequency is shown in Figure 1.

Taking the lowest clutter level across all 49 frequencies in each cell gives 92.38% of the cells that are below the threshold. Thus in 7.62% of the range-azimuth cells a target of the particular RCS level would stay masked by clutter energy across all the frequencies available to the radar system. This is also the upper bound for the simulation's performance, assuming that the simulated target positions are uniformly distributed across the range-azimuth map. The resulting clutter map is shown in Figure 2.

These figures show how frequency diversity can help suppress most of the close-in wave structure of the sea clutter and in particular the second and third time around land clutter regions.

B. The experimental setup

For this experiment the radar had to employ a strategy to choose its transmit frequency for each scan. At each scan 10 targets' locations where randomly chosen within the radar return. For these locations, the returned clutter energy was compared against the detection threshold that had been chosen to decide which of these targets would have been detectable against the competing clutter. Those locations where a detection would have been possible, because the clutter return was below the threshold, were removed from the next scan. If a target hadn't been detected for a certain amount of scans, it was assumed to be undetectable and removed from the list of targets. The radar's reward was calculated based on the amount of new detected targets each scan. This process was repeated for 100000 scans, giving 1 million targets in total that could be detected.

The total score of the system was calculated as the direct sum of the rewards (no discounting) (or the total number of targets detected over the entire period).

Four different algorithms were compared to choose the transmit frequency:

- Randomly choosing a frequency
- Sequentially sweeping over all the frequencies
- Sequentially jumping to every 21th frequency starting from a random frequency index, resulting in a 7 frequency sequence
- Q-Learning was used to learn a strategy to select the frequency

For the Q-Learning algorithm, the state space was chosen to be the current transmission frequency. The actions were then the next frequency to transmit (and thus also the next state). In this state-space definition there is no uncertainty about the current state or what the next state would be after an action has been chosen. This state space doesn't allow the radar to explicitly adapt itself to the targets in its environment (tracking mode), but does allow sufficient flexibility for a search mode.

Using such a deterministic state-space would after convergence result in a strategy that corresponds to a sequence of frequencies being used, i.e. from f_A, the action is to use f_B, resulting in state f_B as the next state, from which f_C is chosen, etc. until from some state the optimum action is to switch to f_A, at which point the sequence repeats.

The choice of discount factor that is used, has a direct impact on the number of frequencies being utilized by the trained Q-Learning system: Low discount factors (myopic) approaches tended to converge to a single frequency strategy, i.e. always switch to the "cleanest" frequency. Higher discount factors resulted in longer sequences of frequencies being utilized. A discount factor of 0.5 was chosen, since it seemed to typically result in a sequence of between 3 and 8 frequencies being used, which gave good detection performance results.

A constant exploration factor of 0.01 was used. The learning rate was 0.9995^N, where N is the scan number.

The Q-matrix was initialized with an overestimate of the Q-values, forcing exploration to all state-transitions initially. This over-estimate was calculated as

$$Q_{init} = \frac{E(R)}{1-\gamma} + \epsilon \tag{6}$$

where $E(R)$ is the expected reward per turn for a "perfect" strategy that never missed a target detection and ϵ is a small random value which ensured that the initial transitions for each simulation is randomized. With 10 targets being generated per turn, the best optimistic reward would be 10 and therefor the

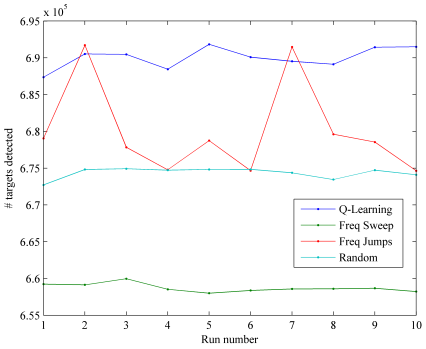

Fig. 3. Results for simulation with targets at locations with lifetimes of 2 scans.

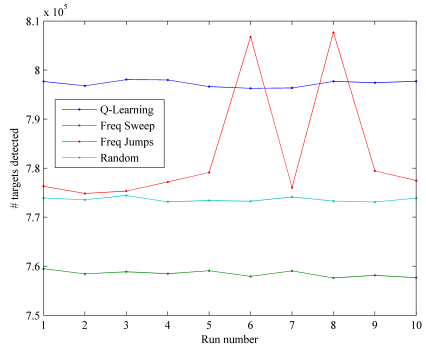

Fig. 4. Results for simulation with targets at locations with lifetimes of 5 scans.

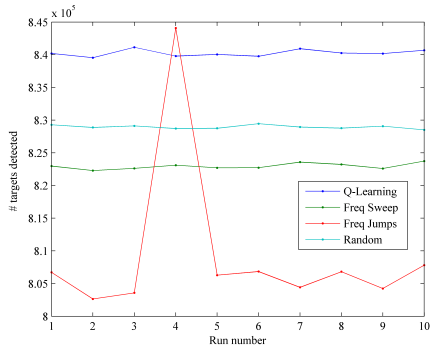

Fig. 5. Results for simulation with targets at locations with lifetimes of 10 scans.

Q-values was initialized with random number between 20 and 21.

C. Experiment with random target locations

An experiment was performed where the 10 targets per scan, was placed at random locations on the range-azimuth map. They stayed in that position for a number of turns. For different trials this duration was change between 2, 5, 10 and 50 turns.

Longer target lifetimes allow more frequency diversity to be utilized. Some of the results are shown in Figures 3 - 6.

These results can be summarized as: The random strategy is a benchmark against which the other strategies can be

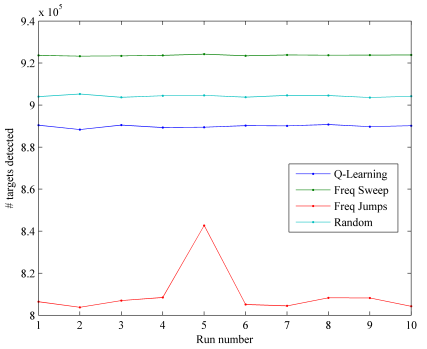

Fig. 6. Results for simulation with targets at locations with lifetimes of 50 scans.

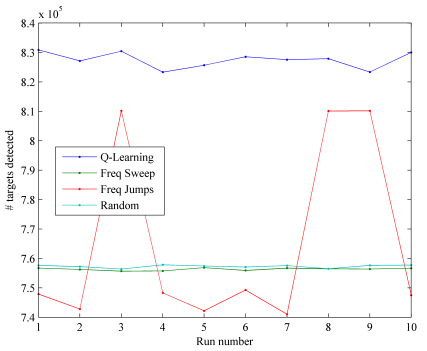

Fig. 7. Results for simulation with incoming targets with lifetimes of 10 scans.

compared. The sequential frequency approach, scores poorly for short lifetimes, but achieves optimum performance at maximum lifetimes, since it then used all available frequencies. The frequency jump approach, quickly leverages diversity, but for long lifetimes, it falls behind, since it is limited to 7 frequencies and for the shortest lifetimes, it is also not ideal, since it doesn't always use the best single frequencies. The Q-Learning algorithm does seem to consistently outperform the other algorithms, except for the longest lifetimes, where it suffers from trying to use only a subset of the frequencies.

D. Experiment with incoming targets

In a long term surveillance scenario, very few targets are likely to just appear randomly at any location in the radar coverage area. Most targets will be detected shortly after they enter the coverage area, either from maximum range or from behind a shadow region (e.g. mountain). Thus manually designing an algorithm to optimize clutter across the whole coverage area, might actually not give optimal performance in a surveillance role, if the radar can specifically adjust its tracking mechanisms based on the target's now known (estimated) location. Thus a subsequent experiment was performed, where the target behavior was adjusted: The targets where now assumed to spawn at the maximum range cell (still at a random bearing) and then move a range cell closer each scan. In this case, the reward was only achieved if they could be detected

within 10 scans (i.e. before they left the last 10 range bins). The detection criteria were kept the same, although for this scenario the radar detector would probably leverage Doppler information to further separate the target from clutter. This effect wasn't included in the simulation.

Results for this scenario are shown in Figure 7, where it can be seen how well the Q-Learning strategy outperforms the other strategies.

Thus without any direct changes to the radar frequency selection algorithm (Q-Learning is still used unchanged), the radar now better adapts to its environment than any of the compared algorithms.

V. Conclusion

Published results seem to support the idea that a cognitive radar can learn from its environment and modify its transmitted waveforms and processing to adapt to the environment. The cognitive control and performance increase such an approach can obtain still needs to be clarified through future research.

As an initial attempt to answer some of these questions, the use of a Q-Learning algorithm was explored here in the search mode of a surveillance radar. This is done using prerecorded clutter data. The radar is allowed to adjusts its transmit frequency in an attempt to minimize the clutter interference and thus maximize the target detection probability.

Q-Learning with a state space of only the current transmit frequency seems sufficient for the radar to learn to use sequences of frequencies that complement each other for the purposes of improving detection.

The algorithm was compared against other manually techniques. Utilizing Q-Learning to decide upon a policy for frequency sequences consistently outperform the manually chosen methods, except in the extreme case where utilizing most or all of the frequencies together give a better result.

With targets more "realistically" entering the surveillance scene the Q-Learning algorithm seems to learn to better exploit this slightly reduced problem space and outperforms the manual techniques by a significant margin. These results are encouraging and illustrate the potential power of utilizing reinforced learning algorithms to build cognitive radars.

Future work could look at more complex state spaces, in particular allowing the system to adapt its waveform for different angular sectors separately. However initial investigations shown that state space growth will hamper the practical application of Q-Learning alone. It is believed a fuzzy state assignment will be needed. Lessons from other application domains will be considered, before deciding on a good approach for the radar domain.

References

[1] S. Haykin, "Cognitive radar: a way of the future," *Signal Processing Magazine, IEEE*, vol. 23, no. 1, pp. 30 – 40, January 2006.

[2] M. C. Wicks, M. Rangaswamy, R. S. Adve, and T. B. Hale, *Knowledge Based Radar Detection, Tracking and Classification*. Gini, F. and Rangaswamy, M. Ed. Rosewood Drive, Danvers, MA: Wiley-Interscience, 2008, ch. Space-Time Adaptive Processing for Airborne Radar: A Knowledge-Based Perspective, pp. 76–102.

[3] C. T. Capraro, G. T. Capraro, A. De Maio, A. Farina, and M. C. Wicks, *Knowledge Based Radar Detection, Tracking and Classification*. Gini, F. and Rangaswamy, M. Ed. Rosewood Drive, Danvers, MA: Wiley-Interscience, 2008, ch. CFAR Knowledge-Aided Radar Detection and its Demonstration Using Measured Airborne Data, pp. 103–128.

[4] A. Benavoli, L. Chisci, A. Farina, S. Immediata, and L. Timmoneri, *Knowledge Based Radar Detection, Tracking and Classification*. Gini, F. and Rangaswamy, M. Ed. Rosewood Drive, Danvers, MA: Wiley-Interscience, 2008, ch. Knowledge-Based Radar Tracking, pp. 167–196.

[5] G. Minkler and J. Minkler, *CFAR*. Baltimore, MD: Magellan Book Company, 1990.

[6] M. Cohen, "An overview of high range resolution radar techniques," in *Telesystems Conference, 1991. Proceedings. Vol.1., NTC '91., National*, March 1991, pp. 107 –115.

[7] J. E. Cilliers and J. C. Smit, "Pulse compression sidelobe reduction by minimization of L_p-norms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, pp. 1238–1247, July 2007.

[8] D. P. Bertsekas, *Dynamic Programming Deterministic and Stochastic models*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1987.

[9] D. E. Bowyer, P. K. Rajasekaran, and W. W. Gebhart, "Adaptive clutter filtering using autoregressive spectral estimation." *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-15, no. 4, pp. 538–546, July 1979.

[10] S. D. Blunt and K. Gerlach, "Adaptive pulse compression via MMSE estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 2, pp. 572–584, April 2006.

[11] N. A. Goodman, P. R. Venkata, and M. A. Neifeld, "Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 1, pp. 105–113, June 2007.

[12] D. Garren, A. Odom, M. Osborn, J. Goldstein, S. Pillai, and J. Guerci, "Full-polarization matched-illumination for target detection and identification," *IEEE Trans. Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 824–837, July 2002.

[13] Y. D. Shirman, *Computer simulation of aerial target radar scattering, recognition, detection, and tracking*. 685 Canton Street, Norwood, MA: Artech House, Inc., 2002.

[14] V. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-doppler effect in radar: phenomenon, model, and simulation study," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 42, no. 1, pp. 2 – 21, January 2006.

[15] A. Cilliers and W. Nel, "Helicopter parameter extraction using joint time-frequency and tomographic techniques," in *Proc. 2008 Int. Conf. on Radar*, no. art. no. 4653993. Adelaide, Australia: Radar 2008, September 2008, pp. 598–603.

[16] S. L. C. Miranda, C. Baker, K. Woodbridge, and H. D. Griffiths, *Knowledge Based Radar Detection, Tracking and Classification*. Gini, F. and Rangaswamy, M. Ed. Rosewood Drive, Danvers, MA: Wiley-Interscience, 2008, ch. Multifunction Radar Resource Management, pp. 225–264.

[17] K. Veeramachaneni and L. A. Osadciw, "Multiple sectors, multi function, multi radar dwell time management using particle swarm optimization (M3RTM)," vol. 2006, 2006, pp. 425–431.

[18] S. Haykin, "Cognition is the key to the next generation of radar systems," in *2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, DSP/SPE 2009, Proceedings*, Marco Island, FL, January 2009, pp. 463–467.

[19] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[20] B. Wang, J. Wang, X. Song, and F. LIU, "Q-learning-based adaptive waveform selection in cognitive radar," *Int. J. Communications, Network and System Sciences*, vol. 2, no. 7, pp. 669–674, Oct 2009.

[21] F. Liu and F. Wang, "An optimal adp algorithm for waveform selection in cognitive radar systems," in *Progress in Electromagnetics Research Symposium Proceedings, China*, March 2009, pp. 728–730.

[22] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[23] P. L. Herselman, C. J. Baker, and H. J. De Wind, "Analysis of X-band calibrated sea clutter and small boat reflectivity at medium-to-low grazing angles," *International Journal of Navigation and Observation*, vol. 2008, pp. 1–14, 2008.