Bandit-Based Online Candidate Selection for Adjustable Autonomy

Boris Sofman, J. Andrew Bagnell, and Anthony Stentz

Abstract In many robot navigation scenarios, the robot is able to choose between some number of operating modes. One such scenario is when a robot must decide how to trade-off online between autonomous and human tele-operation control. When little prior knowledge about the performance of each operator is known, the robot must learn online to model their abilities and be able to take advantage of the strengths of each. We present a bandit-based online candidate selection algorithm that operates in this adjustable autonomy setting and makes choices to optimize overall navigational performance. We justify this technique through such a scenario on logged data and demonstrate how the same technique can be used to optimize the use of high-resolution overhead data when its availability is limited[1].

1 Introduction

Autonomous UGVs have advanced to a point where they are competent and reliable a large portion of the time. However, even the most robust autonomous robotic systems will struggle with certain situations. Fortunately, in some domains it is reasonable to assume that a human operator may be available for periods of time to provide remote tele-operation support. Full tele-operation is prohibitively expensive for many applications due to the degree of required human attention and communications bandwidth, so a policy must determine under which conditions the robot or the human are to take control.

It is important for such a system to be well-suited for online use. Not only is it difficult to asses in advance how well the autonomy system will perform in novel environments, but human operator performance can also vary depending on factors

Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213
e-mail: {bsofman,dbagnell,axs}@ri.cmu.edu

[1] Many of the images in this paper are best viewed in color

Fig. 1 For our experiments, we used logs from the Crusher unmanned ground vehicle. The robot operates in complex, natural environments where the goal is to navigate across large distances with the aid of onboard and overhead sensor data. Further information about the system can be found in [1]. The overhead processing capabilities used to plan prior routes and generate features for our experiments are described in [2].

such as bandwidth limitations, operator handicaps such as limited skill or familiarity with the interface, fatigue and weather conditions. In such situations, a learning system can observe the performance of the autonomous vehicle in particular situations and compare that to performance under remote human-control in similar situations. When the vehicle encounters similar situations in the future, it can then invoke whichever expert demonstrated better performance: the remote human or autonomous vehicle. Such a capability would enable a single operator to assist many UGVs, ensuring peak performance for the entire team with minimal human involvement.

We pursue this problem using an on-line, reinforcement learning approach and demonstrate its performance on logged data from the rugged, all-terrain UGV shown in Fig. 1. The candidate selection system's goal is to learn to interpret features from available overhead sensor data in order to make candidate choices that maximize its overall long-term performance. This inevitably becomes a trade-off between exploring candidates' performances in situations that will allow it to learn more about the world and taking advantage of their learned models to maximize current performance.

We also show how this technique can be used to deal with scenarios where limited high-resolution overhead data is available to aid the robot in navigating through an environment. Higher resolution overhead data can be used to produce more accurate traversal cost estimates that the UGV can use for better prior path calculations but often requires expensive and time-consuming aerial surveying and a large amount of bandwidth if remotely supplied to the vehicle. In scenarios where the availability of such data is limited, our algorithm allows the robot to learn to allocate it to areas where its impact will be maximized.

The next section presents background on adjustable autonomy techniques and some example applications. Section 3 presents our online candidate selection algorithm, followed by experimental results in Section 4 and concluding remarks in Section 5.

2 Related Work

We deal with the scenario where a human can contribute limited attention to improve a mobile robot's performance. In this scenario, a robotic system operates somewhere on the spectrum between full autonomy, where there is no human involvement, and full tele-operation, where the human is in complete control at all times. Scenarios where the degree and methods of human interactions with robots within a system can be varied dynamically in order to optimize performance are often referred to as ones of *sliding autonomy* or *adjustable autonomy* [3, 4]. While most mobile robot systems tend to lie on one of the two extremes of this spectrum, effectively balancing autonomy with limited human involvement can lead to significant improvements in safety, efficiency and overall cost.

In some scenarios where the human is the primary operator, the autonomy system is intended to aid by request or when it detects a dangerous situation [5, 6, 7, 8]. Similar approaches have been applied to automating repetitive tasks in surgery to decrease surgeon fatigue [9].

In scenarios where the autonomy system is the default operator, the system must determine whether and when to transfer control to a human [10, 11, 12]. Some have suggested relinquishing control when there is an expectation of high benefit [13, 14] or the degree of uncertainty is high [15].

Goodrich and Schultz have written an extensive survey article on the field of Human-Robot Interaction exploring many additional approaches and applications [16].

The key difference in our approach from the above-mentioned approaches is that our system operates with no prior performance model or pre-determined operator transition rules. In many scenarios where prior performance information is unavailable, the ability to learn the capabilities of each potential candidate online allows systems to better adapt to more diverse and challenging environments.

3 Approach

3.1 Contextual Multi-Armed Bandit Setting

The candidate selection problem involves choosing an operator for each encountered situation from a set of candidate systems, in our case the autonomy system and the human tele-operator, whose performance we assume comes from some unknown distribution. It is therefore intuitive to frame this problem as an instance of the commonly studied *multi-armed bandit* problem [17, 18, 19].

In the k-armed bandit setting, at each time step t the world chooses k losses (or rewards), $l_t^1, \ldots, l_t^k$, and the player makes a choice of an arm $i \in \{1, \ldots, k\}$ without

knowledge of the hidden losses. The player then observes only the loss l_t^i corresponding to the chosen arm. Since the loss distributions are unknown, there is an inevitable conflict between minimizing the immediate loss and gathering information that will be useful for long-term performance. This is often referred to as the *exploration-exploitation trade-off* since we must choose between *exploring* our unknown loss distributions and *exploiting* the arm we currently believe to be best.

We deal with a more suitable variation of the *bandit* setting called the *contextual bandits* setting where at each time step t the player also observes some contextual information x_t which can be used to determine which arm to pull [20]. We compute these features from commonly available overhead imagery and DTED 3 elevation data[2] for the given environment as described in [2].

As is common with bandit problems, our goal is to minimize regret, the difference between the performance of the algorithm and that of the optimal algorithm in hindsight:

$$R = \sum_{t=1}^{T} (l_t - l_t^*) \tag{1}$$

where l_t^* is the loss incurred in round t by the optimal strategy.

3.2 Exploration-Exploitation Trade-off

We choose to deal with the exploration-exploitation trade-off through the use of confidence bounds. With a model that is able to supply confidence bounds, the widths of the confidence bounds reflect the uncertainty of the algorithm's knowledge. By choosing the candidate with the highest upper confidence bound estimate for the specified features x_t at each time step, the algorithm elegantly trades off between exploration and exploitation. When uncertainty is high, choosing that candidate will provide information that will quickly reduce uncertainty in that region of the feature space. As we gain knowledge about each candidate, confidence bounds will shrink and the candidates' expected performances will begin to dominate the selection process. This approach was well-justified for the bandits setting and is shown to have small regret [21].

[2] The Digital Terrain Elevation Data (DTED) level of an overhead elevation data set specifies its density of coverage where a higher level specifies denser coverage.

3.3 Formalization

We frame online candidate selection problems as follows. At each time step t, we get some contextual features x_t for our environment and try to minimize our incurred loss by choosing from one of k candidates for that time step[3].

After each selection, the algorithm observes the noisy estimate of l_t^i for only the chosen candidate i. We model the distribution for l_t^i as a Gaussian whose mean is a linear function of the contextual features x_t modeled by vector μ^i:

$$E(l_t^i|\mu^i, x_t) = \mu^i x_t \tag{2}$$

We assume the estimates have Gaussian noise and are therefore distributed:

$$\tilde{l}_t^i \sim Normal(l_t^i, \sigma^2) \tag{3}$$

We model this distribution online using a Bayesian linear regression model as described in [22]. Given a new data point $\tilde{l}_t^i$ estimating the true variable l_t^i, our goal is to compute a new estimate of the variable $l_{t+1}^{1...k}$, assuming we have already seen data $D = \{\{x\}_{1...n}, \{\tilde{l}^i\}_{\tau \subseteq \{1...t-1\}}\}$. We can compute this by integrating over μ^i:

$$p(l_{t+1}^i|\tilde{l}_t^i, x_t, D) = \int d\mu^i \; p(l_{t+1}^i|\mu^i, \tilde{l}_t^i, x_t) p(\mu^i|\tilde{l}_t^i, x_t, D)$$

We can compute the required distribution over μ^i as:

$$p(\mu^i|\tilde{l}_t^i, x_t, D) \propto p(\mu^i|D) \int dl_t^i p(\tilde{l}_t^i|l_t^i) p(l_t^i|\mu^i, x_t)$$

In our linear-Gaussian model, this can be understood as revising the posterior distribution from $p(\mu^i|D)$ in light of a Gaussian likelihood that takes into account noise.

Our computation of the posterior distribution $p(\mu^i|\tilde{l}_t^i, x_t, D)$ is as follows. We first initialize our distribution to the prior distribution $p(\mu^i)$. Then, for every training example t, we multiply our distribution by $p(\tilde{l}_t^i|\mu^i, x_t)$. Since the prior distribution and $p(\tilde{l}_t^i|\mu^i, x_t)$ are normal, the posterior distribution is also normal.

This not only allows us to efficiently perform online updates of our model but also provides a variance estimate for each prediction that we treat as our confidence bound. We therefore track k Bayesian linear regression instances in parallel, one for each candidate, and at each time step choose the candidate with the highest upper confidence bound prediction for that scenario.

[3] In the case of choosing between a human and the autonomy system, $k = 2$. We discuss this problem in the more general case as it could also be applied to any candidate selection setting such as choosing between multiple autonomy systems, multiple human operators or multiple overhead data sources as shown later.

4 Experimental Results

4.1 Adjustable Autonomy

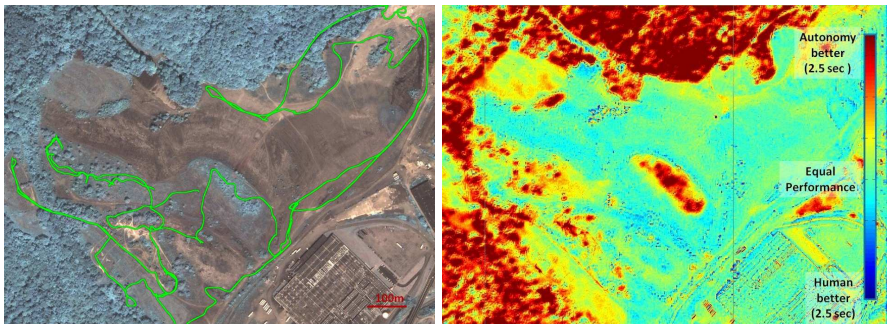

Fig. 2 Aerial image of test site with course driven using each operating mode (left) and the estimated differences in traversal time in seconds per meter for this site using the final models learned by the online candidate selection algorithm (right). The algorithm found that human performance tended to excel in open areas where the human was better able to interpret sparse obstacles and drive more aggressively and at perimeters of heavy obstacles when the human's situational awareness allowed him to better handle environmental complexity.

While we do not have the system infrastructure to be able to trade-off online between tele-operation and autonomous vehicle control, we simulated such an online scenario by using a pair of logged traversals of the same long course in western Pennsylvania by each candidate: a human tele-operator using a high-bandwidth camera system and the autonomy system. All locations where the path of the human driver and the autonomous driver were in sufficient proximity were used as test points for the system where the loss was measured by the period of time it took to enter and exit a 3 meter radius window around that location. Raw overhead features were convolved with a Gaussian kernel in order to blur the data, in effect introducing an influence from surrounding areas into each location. This allowed the system to learn a more realistic model since the rate of progress at a given location is heavily influenced by factors from the surrounding area. Each time a candidate was chosen, the traversal time for only the specified candidate was revealed to the algorithm.

The course and estimated relative performance of each candidate using a trained model appear in Fig. 2. Quantitative results comparing our algorithm's performance to various alternatives appear in Fig. 3 and Table 1. The optimal, worst-case and random candidate algorithms used for comparison throughout this paper are those that always choose the best-performing, worst-performing and at random candidate at each time step respectively.

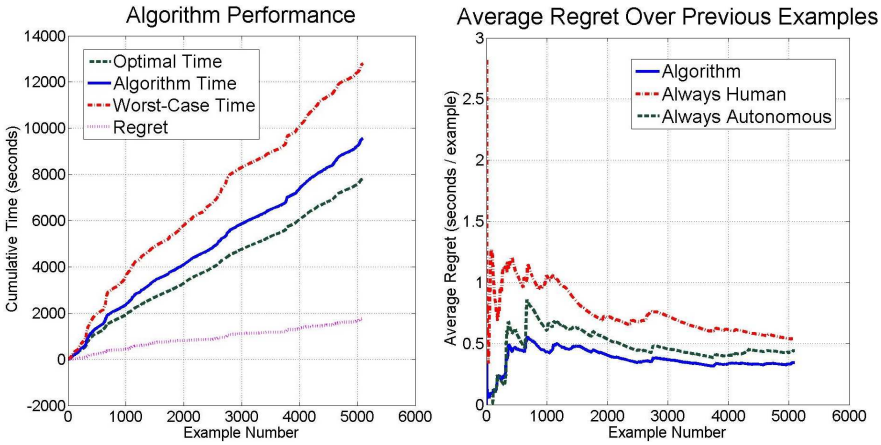

Fig. 3 Online operator selection performance: cumulative navigation time for our algorithm and various alternatives (left) and the average regret of our algorithm over previous examples compared to alternatives (right). Lower times represent better performances.

Table 1 Online Operator Selection Performance

Algorithm	Cumulative Time (seconds)[a]	Percent Improvement over Always-Human
Online Algorithm	9551.7	9.41
Optimal	7809.3	25.94
Worst-Case	12791.0	-21.31
Always-Human	10544.4	0.00
Always-Autonomy	10055.9	4.63
Random Driver	10307.4	2.95

[a] Note that since 3 meter regions at example locations often overlapped with each other, these cumulative traversal times are greater than the total navigation time.

4.2 Online Overhead Data Selection

We also show how this algorithm can be applied to scenarios where the vehicle must decide online how to best utilize the availability of various-density overhead data for upcoming navigation. We simulated this scenario by analyzing sets of multi-waypoint logged runs from a field test at Fort Carson in Colorado on various courses using DTED levels 3, 4 and 5 overhead data. The candidates for each waypoint in this case were the single choice of density of overhead data to be used for computing a prior path for that path segment. The candidate selection system therefore had the task of learning a mapping from the average of feature values (computed from overhead imagery and DTED 3 elevation data) within the segment's bounding box to the average traversal speed for the vehicle over that segment of the path using each candidate type of data. While DTED 5 data almost always resulted in the best performance, we simulated a scenario where high-density data is available for only

a fraction of all segments: a maximum of 20% availability for DTED 5 and 30% availability for DTED 4.

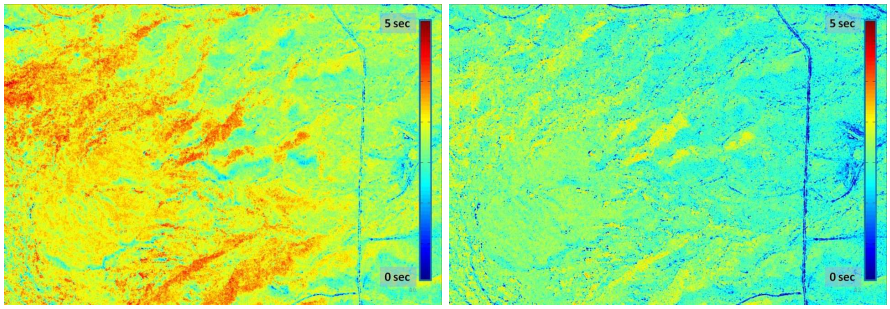

Fig. 4 Estimated traversal time in seconds per meter (in effect, brighter colors identify for difficult or hazardous locations) is shown for sample terrain used for overhead data selection experiments with DTED 3 (left) and 5 (right) data. As expected, DTED 5 data shows large improvements in navigation speed for difficult terrain (left portion of images) but does not provide nearly as much benefit on roads and open fields (right portion of images).

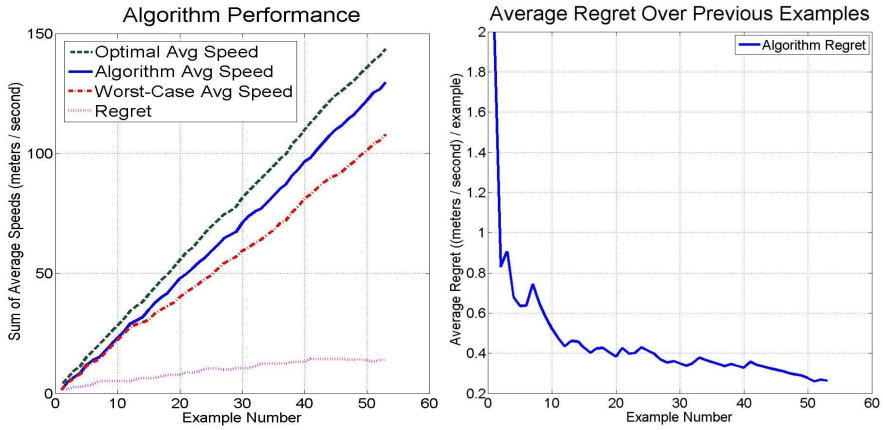

Fig. 5 Overhead data selection performance: sum of average navigation speed over each path segment for our algorithm and various alternatives (left) and the average regret of our algorithm over previous segments (right).

At each step we used a linear program to optimize the allocations of remaining data availability using the predicted performance on all remaining segments from the current learned models for each candidate. Selections at each step were based on the initial step of this locally computed optimal allocation. To avoid having to solve an integer programming problem, we chose the candidate with the highest allocation at the first step.

Table 2 Online Overhead Data Selection Performance

Algorithm	Average Speed (meters / second)	Percent Improvement over Random
Online Algorithm	2.45	5.60
Optimal	2.71	16.81
Worst-Case	2.04	-12.07
Random Data Source	2.32	0.00

The estimated rates of progress predicted by the trained models using DTED 3 and DTED 5 data sources appear in Fig. 4. Quantitative results for this scenario appear in Fig. 5 and Table 2. Our algorithm shows a clear improvement over naive or random approaches for both scenarios with quickly-converging regret properties.

5 Conclusion

We have presented an online algorithm for dealing with scenarios where the robot must learn to trade-off between multiple operating modes. The proposed approach relies on a bandit-based framework and uses confidence bounds to deal with exploration-exploitation trade-offs. The algorithm was demonstrated on two scenarios relevant to the mobile robotics domain and showed improved performance over several alternatives. We hope that such techniques will increase the potential real-world applications of mobile robots by allowing them to adapt in real-time to changing environments and better allocate available resources.

Acknowledgements

This work was partially sponsored by DARPA under contract Unmanned Ground Combat Vehicle - PerceptOR Integration (contract number MDA972-01-9-0005) and by the U.S. Army Research Laboratory under contract Robotics Collaborative Technology Alliance (contract number DAAD19-01-2-0012). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Boris Sofman is partially supported by a Sandia National Laboratories Excellence in Engineering Fellowship. The author gratefully acknowledges Sandia's Campus Executive Laboratory Directed Research and Development Program (LDRD) for this support.

References

1. A. Stentz, J. Bares, T. Pilarski, and D. Stager, "The crusher system for autonomous navigation," in *AUVSIs Unmanned Systems North America*, August 2007.
2. D. Silver, B. Sofman, N. Vandapel, J. A. Bagnell, and A. Stentz, "Experimental analysis of overhead data processing to support long range navigation," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, October 2006, pp. 2443 – 2450.
3. M. B. Dias, B. Kannan, B. Browning, E. Jones, B. Argall, M. F. Dias, M. B. Zinck, M. M. Veloso, and A. Stentz, "Sliding autonomy for peer-to-peer human-robot teams," in *10th International Conference on Intelligent Autonomous Systems 2008*, July 2008.
4. P. Scerri, D. V. Pynadath, and M. Tambe, "Towards adjustable autonomy for the real world," *Journal of Artificial Intelligence Research*, vol. 17, p. 2002, 2002.
5. R. Grace, V. Byrne, D. Bierman, J.-M. Legrand, D. Gricourt, B. Davis, J. Staszewski, and B. Carnahan, "A drowsy driver detection system for heavy vehicles," in *Proceedings of the 17th Digital Avionics Systems Conference*, vol. 2, 2001, pp. I36/1 – I36/8.
6. A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 4, no. 3, pp. 143–153, 2003.
7. R. Bishop, "Intelligent vehicle applications worldwide," *IEEE Intelligent Systems*, vol. 15, no. 1, pp. 78–81, 2000.
8. E. Krotkov, R. Simmons, F. Cozman, and S. Koenig, "Safeguarded teleoperation for lunar rovers: From human factors to field trials," in *In Proc. IEEE Planetary Rover Technology and Systems Workshop*, 1996.
9. A. Krupa, M. de Mathelin, C. Doignon, J. Gangloff, G. Morel, L. Soler, and J. Marescaux, "Development of semi-autonomous control modes in laparoscopic surgery using automatic visual servoing," *Lecture Notes in Computer Science*, vol. 2208, pp. 1306–??, 2001.
10. F. W. Heger and S. Singh, "Sliding autonomy for complex coordinated multi-robot tasks: Analysis & experiments," in *Robotics: Science and Systems*, G. S. Sukhatme, S. Schaal, W. Burgard, and D. Fox, Eds. The MIT Press, 2006.
11. T. W. Fong, C. Thorpe, and C. Baur, "Multi-robot remote driving with collaborative control," *IEEE Transactions on Industrial Electronics*, 2003.
12. A. Stentz, C. Dima, C. Wellington, H. Herman, and D. Stager, "A system for semi-autonomous tractor operations," *Auton. Robots*, vol. 13, no. 1, pp. 87–104, 2002.
13. E. Horvitz, A. Jacobs, and D. Hovel, "Attention-sensitive alerting," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, K. B. Laskey and H. Prade, Eds. S.F., Cal.: Morgan Kaufmann Publishers, Jul. 30–Aug. 1 1999, pp. 305–313.
14. H. Hexmoor, "A cognitive model of situated autonomy," *Lecture Notes in Computer Science*, vol. 2112, pp. 325–??, 2001.
15. J. P. Gunderson and W. N. Martin, "Effects of uncertainty on variable autonomy in maintenance robots," in *In Workshop on Autonomy Control Software*, 1999, pp. 26–34.
16. M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.
17. H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc*, vol. 58, no. 5, pp. 527–535, 1952.
18. T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics(Print)*, vol. 6, no. 1, pp. 4–22, 1985.
19. P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, *47*, vol. 2, no. 3, pp. 235–256, 2002.
20. C. Wang, S. Kulkarni, and H. Poor, "Bandit problems with side observations," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 338–355, 2005.
21. P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *The Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2003.
22. B. Sofman, E. L. Ratliff, J. A. Bagnell, J. Cole, N. Vandapel, and A. Stentz, "Improving robot navigation through self-supervised online learning," *Journal of Field Robotics*, vol. 23, no. 1, December 2006.