Presented as late breaking results contribution
21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Combination of reinforcement learning and bio-inspired odor source searches under a time-variant gas mapping ⋆

Cesar Hernandez Reyes * **Kei Okajima** ** **Shunsuke Shigaki** ***
Daisuke Kurabayashi * **Kazushi Sanada** **

* *Dept. of Systems and Control Engineering, Tokyo Institute of Technology, Tokyo, Japan (e-mail: hernandez.cesar@irs.sc.e.titech.ac.jp, dkura@irs.sc.e.titech.ac.jp)*
** *Division of Systems Research, Yokohama National University, Kanagawa, Japan (e-mail: okajima-kei-hx@ynu.jp, sanada-kazushi-sn@ynu.jp)*
*** *Department of Systems Innovation, Osaka University, Osaka, Japan(e-mail: shigaki@arl.sys.es.osaka-u.ac.jp)*

Abstract: Searching for odor sources such as hazardous objects or gas leaks is desirable in a robot. In the literature, many approaches to odor source search are bio-inspired or probabilistic. However, the performance of either method can decrease if exploring and exploiting olfactory information is unbalanced. In this paper we investigated whether a balance can be achieved by a hybrid strategy composed of a bio-inspired search and infotaxis, which is an RL-based method of the literature. We tested infotaxis and the hybrid algorithm under a time-variant virtual odor plume. We obtained this plume from readings of a gas sensor array and a wind tunnel. From this we found that the hybrid algorithm showed better search performance and less deviation from the plume centerline. Therefore we believe that combining probabilistic and bio-inspired policies might be useful to balance exploration and exploitation and efficiently perform olfactory searches.

Keywords: Guidance, navigation and control; intelligent robotics; robot ethology

1. INTRODUCTION

Searching for olfactory sources is a challenging problem that requires fast decision-making to track chemical particles that are frequently diluted in unstructured or turbulent wind flows. It is also a desired capability for an autonomous robot since it would enable it to search for dangerous gas leaks at industrial plants or in disaster-struck areas as well as finding other hazardous materials such as explosives.

In the literature, many olfactory search algorithms are bio-inspired, which means they are extracted from observations of the average behavioral response of insects such as the silkmoth *Bombyx mori* (Kanzaki et al. (1992)). Unfortunately, these have shown to need conditions that are very similar to the habitat of the studied animal to perform well. Other studies have developed probabilistic algorithms that use particle filters (Li et al., 2011), Bayesian inference, and Reinforcement Learning (RL). A widely-cited example of an RL olfactory search agent is the algorithm known as "infotaxis" (Vergassola et al. (2007)), which navigates towards an odor source by minimizing the entropy of a *belief*, which is a probabilistic representation of the location of the source. Although infotaxis has shown good results in simulations, Rodríguez et al. (2017) reported that if the parameters of the belief function are not optimally set, the balance between exploration and exploitation of rewards (decrease in information entropy) could shift towards either, hence resulting in a decreased performance. Interestingly, recent works have performed probabilistic analyses of the olfactory behavior of insects. For example, Pang et al. (2018), found that fruit flies turn less towards the wind flow as they accumulate odor detections; also, Shigaki et al. (2018) found that silkmoths are less likely to move forward as they experience more detections.

In this paper we investigate whether combining an RL and a bio-inspired algorithm could balance exploration and exploitation of olfactory information and effectively perform odor source searches. For this, we performed simulations of infotaxis under a virtual odor plume obtained from readings of a gas sensor array and a wind tunnel. From these simulations we found that the largest expected reward shifted from forward to lateral movements when the information entropy of the belief and the cumulative number of odor hits reached the same value. From this tendency we designed a hybrid policy that switched from infotaxis to a bio-inspired policy after the crossing point between entropy and odor hits was met.

⋆ This work was supported in part by the JSPS KAKENHI under Grants JP19H04930, JP19K14943 and 19H02104; the first author also acknowledges financial support from Instituto de Innovación y Transferencia de Tecnología (I2T2) and Consejo Nacional de Ciencia y Tecnología.

Presented as late breaking results contribution
21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

From simulating the hybrid algorithm, we found that it showed a higher success rate than using only infotaxis. We believe that this suggests that animals might combine cognitive behaviors based on memory or predictions and reactive ones based on instincts to balance exploration and exploitation, hence resulting in a high odor source search performance.

2. MATERIALS AND METHODS

2.1 Generation of instantaneous gas distribution maps

To generate the virtual odor plume we used in our simulations, we built a sensor array inside a 550 × 200 mm wind tunnel as seen in figure 1. We chose these dimensions to conduct experiments with live silkmoths in a future study. We placed 27 gas sensors (MiCS 5524, Adafruit, USA) in the tunnel in a 9 × 3 matrix as seen in figure 1a. We chose the MiCS5524 sensor because of its small size and availability. Then, we pulsated an ethanol source into the wind tunnel at 1 Hz using a solenoid valve (open: 0.2 s; closed: 0.8 s) and a flow meter (1.0 L/min). We measured the wind speed at the gas source as 0.68 m/s. We released ethanol into the tunnel and recorded the response of the gas sensors during 50 s.

With these recordings and the Kernel DM + V algorithm (Lilienthal et al., 2009) we estimated the gas distribution map at each time step. The Kernel DM + V algorithm employs a uni-variate Gaussian kernel $\mathcal{N}$ to weight a measurement r_i obtained at a given location $\mathbf{x}_i$ to model the gas distribution as a 2D lattice with k cells. Although this algorithm provides predictive mean and variance maps, we focused on the latter because Lilienthal et al. (2009) reported that the variance map showed better results at estimating the location of a gas source. The calculation of the variance map is as follows: first, the importance weights $\Omega^{(k)}$ and the weighted reading $R^{(k)}$ maps are obtained:

$$\begin{aligned}\Omega^{(k)} &= \sum_{i=1}^{n} \mathcal{N}\left(\left|\mathbf{x}_i - \mathbf{x}^{(k)}\right|, \sigma\right) \\ R^{(k)} &= \sum_{i=1}^{n} \mathcal{N}\left(\left|\mathbf{x}_i - \mathbf{x}^{(k)}\right|, \sigma\right) \cdot r_i\end{aligned} \tag{1}$$

Here, $\left|\mathbf{x}_i - \mathbf{x}^{(k)}\right|$ denotes the distance between the measurement point and the position of a given cell k of the predicted map, the parameter σ is the width of the Gaussian kernel; we set this value to 25 mm. To normalize the values of $\Omega^{(k)}$, a confidence map $\alpha^{(k)}$ is obtained as follows:

$$\alpha^{(k)} = 1 - e^{-\left(\Omega^{(k)}\right)^2/\sigma_\Omega^2} \tag{2}$$

Where σ_Ω is the scaling parameter. We set this value to $1/\left(2\pi\sigma^2\right)$ as recommended by Lilienthal et al. (2009). Next, the predictive mean of the gas distribution can be obtained as:

$$r^{(k)} = \alpha^{(k)} \frac{R^{(k)}}{\Omega^{(k)}} + \left\{1 - \alpha^{(k)}\right\} r_0 \tag{3}$$

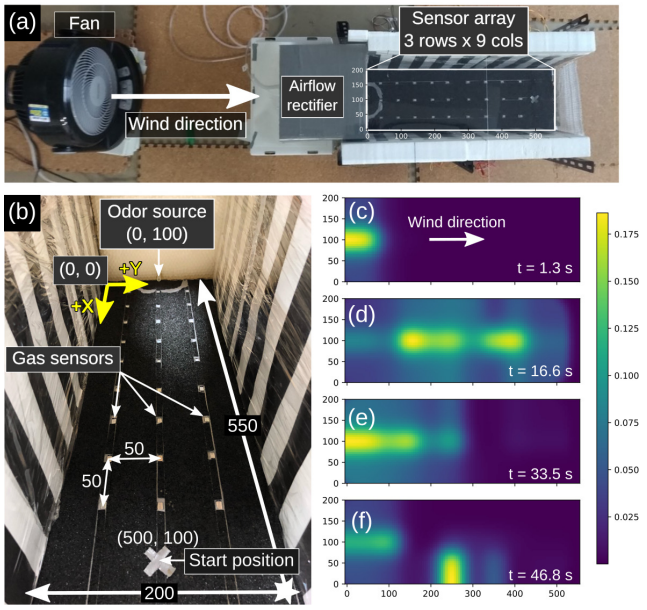

Fig. 1. (a) and (b) Wind tunnel and gas sensor array, (c–f) snapshots of the predictive variance maps obtained with the Kernel DM+V method. Dimensions are in millimeters

Here, r_0 is the mean of the values of all gas sensors in the wind tunnel. Afterward, the weighted variance $V^{(k)}$ can be obtained as:

$$V^{(k)} = \sum_{i=1}^{n} \mathcal{N}\left(\left|\mathbf{x}_i - \mathbf{x}^{(k)}\right|, \sigma\right) \left(r_i - r^{(k(i))}\right)^2 \tag{4}$$

Where $r^{(k(i))}$ is the predictive mean of the nearest cell to the measurement point x_i. Finally, the predictive variance map $v^{(k)}$ can be obtained by:

$$v^{(k)} = \alpha^{(k)} \frac{V^{(k)}}{\Omega^{(k)}} + \left\{1 - \alpha^{(k)}\right\} v_0 \tag{5}$$

We obtained the predictive variance map for each time step of the data recorded from the sensor array. Some snapshots are shown in figure 1 (b–e). Since these maps essentially represent the distribution of ethanol in the wind tunnel, we used these maps as the input odor plume of a simulated olfactory search agent as described in the next section.

2.2 Infotaxis: reinforcement learning-based olfactory search

Infotaxis was proposed by Vergassola et al. (2007) to perform olfactory searches on turbulent plume environments. In this algorithm, a point-mass agent located at $\mathbf{r}$ searches an odor source located at $\mathbf{r}_s$ throughout a 2D workspace (similar to a *Gridworld* (Sutton and Barto, 2018)) denoted as $\mathcal{W}$ based on the probability $P_t(\mathbf{r}_s)$ of the location of the source; which is also called a *belief*.

The goal of the agent is to minimize the information entropy $S_t = -\int_{\mathcal{W}} P_t(\mathbf{r}) \log\left(P_t(\mathbf{r})\right) \mathrm{d}\mathbf{r}$ of the belief. At each time step the agent calculates how much entropy would be decreased by moving from its current location $\mathbf{r}$ at time t to a next location $\mathbf{r}'$ (front, back, left, right, stay still) at time $t + \Delta t$ as defined in equation 6.

Presented as late breaking results contribution
21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

$$\mathbf{E}[\Delta S_t(\mathbf{r} \mapsto \mathbf{r}')] = P_t(\mathbf{r}')[-S_t] + [1 - P_t(\mathbf{r}')][\rho_0(\mathbf{r}')\Delta S_0 + \rho_1(\mathbf{r}')\Delta S_1] \quad (6)$$

Where $\rho_k(\mathbf{r}') = h(\mathbf{r}')^k e^{-h(\mathbf{r}')}/k!$ is the Poisson probability of detecting k odor hits during the time Δt given that $h(\mathbf{r}') = \Delta t \int P_t(\mathbf{r}_s) R(\mathbf{r}'|\mathbf{r}_s) d\mathbf{r}_s$, and $R(\mathbf{r}'|\mathbf{r}_s)$ is the rate of encounters with odor particles for an agent located at $\mathbf{r}'$. In this paper empirically set the values of the parameters for the rate of encounters $R(\mathbf{r}'|\mathbf{r}_s)$ as follows: *particle diffusivity* $D = 0.06$, *release rate* $\mathcal{R} = 1$, *particle lifetime* $\tau = 1500$, *agent size* $a = 0.01$ (m). For details on the derivations of infotaxis equations we refer the reader to the original paper by Vergassola et al. (2007). At each time step, the agent moves to the location $\mathbf{r}'$ that decreases entropy the most, in other words, selecting the action with the largest expected reward as it is common in RL.

2.3 Design of an hybrid algorithm

To identify the tendencies towards exploration or exploitation of an infotaxis agent, we conducted 500 simulation runs of infotaxis in a virtual environment with the same dimensions as our wind tunnel (550 × 200 mm). In these simulations, the odor plume was represented by the Kernel DM+V variance maps $v^{(k)}$ as described in section 2.1. The agent was represented as a 10 mm round particle to emulate the size of a silkmoth (Kanzaki et al., 1992). When the agent passed a location where the value of $v^{(k)}$ was higher than the average value along the whole field (denoted as v_0), we determined that the agent experienced an odor hit (see algorithm 1). From these simulations we used equation 6 to measure the expected entropy decrease $\mathbf{E}[\Delta S_t(\mathbf{r} \mapsto \mathbf{r}')]$; these measurements are shown in figure 2a. We also recorded the entropy S and the cumulative number of hits H experienced by the agent as seen in figure 2b. Furthermore, we found that when the values of the entropy S and the number of hits H intersect, the expected entropy decrease from moving forward reaches its minimum. In other words, after the time step when $S = H$, the agent expects larger rewards by moving sideways (exploration) rather than forward (exploitation). This intersection point is indicated by a red dashed line in figure 2. This is similar to the findings of Pang et al. (2018), where odor hits happening later in a sequence triggered weaker upwind turns, and Shigaki et al. (2018), where the likelihood of a silkmoth to move forward decreases inversely against the cumulative number of odor hits. Based on this, we propose a hybrid policy in which after the entropy-hits intersection, the agent switches its navigation policy from infotaxis to a bio-inspired algorithm consisting of moving forward for 0.5 s after an odor hit and alternating between left and right movements during 0.7 s after a new hit is experienced. These timing values were inspired by the programmed behavior of the male silkmoth (Kanzaki et al., 1992).

3. RESULTS

We evaluated the performance of infotaxis and the hybrid algorithm by observing the trajectories generated by each algorithm as well as their success rate, search time, and traveled distance. A successful run is one where the agent reaches a radius of 0.05 m around the source position under a time limit of 50 s. We executed 500 simulation runs for each algorithm. The performance metrics are shown in table 1. It should be noted that we only considered the successful trials for the calculation of search time and traveled distance. The trajectories generated by each

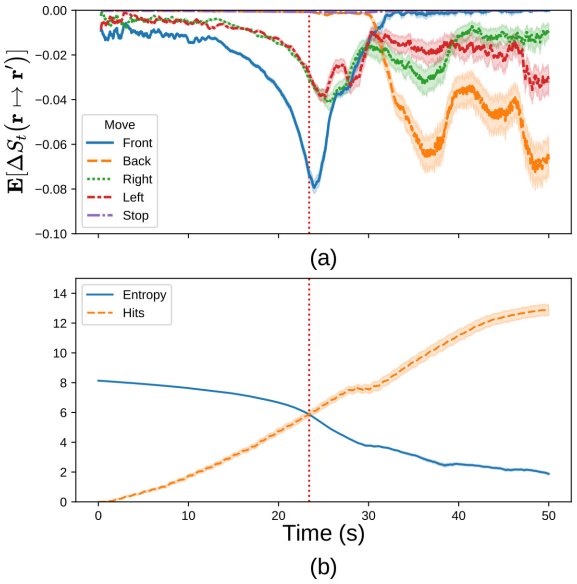

Fig. 2. (a) Expected entropy decrease for each action of an infotaxis agent; (b) relationship between entropy and the cumulative number of odor hits, the red dashed line shows the time when these values are equal. For all curves the average and standard deviation are indicated by solid and shaded colors, respectively (n=500 simulations)

```
OdorHit, HybridPolicy = False;
Entropy, NumOfHits, TimeSinceLastHit = 0;
while t < TimeLimit do
    Sample odor plume v^(k) at time t;
    if v^(k)(r_i) > v_0 then
        OdorHit = True;
        NumOfHits += 1;
    else
        OdorHit = False;
    end
    Update entropy;
    if Entropy − NumOfHits < 0.01 then
        HybridPolicy = True;
    end
    if HybridPolicy then
        if OdorHit then
            Move forward;
            TimeSinceLastHit = 0;
        else
            if TimeSinceLastHit > 0.5 s then
                Randomly move left or right;
                Invert direction every 0.7 s until next OdorHit;
            end
        end
    else
        Choose next move with infotaxis policy;
    end
end
```

Algorithm 1: Proposed odor source search algorithm with an hybrid policy

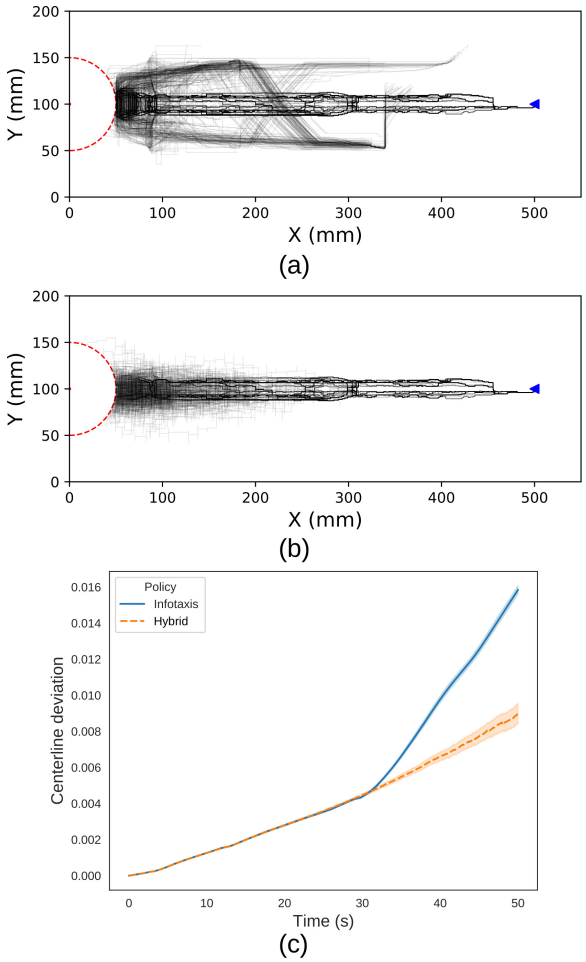

Fig. 3. Trajectories generated by (a) infotaxis and (b) the hybrid policy. (c) Centerline deviation of each policy; average and standard deviation are indicated by solid and shaded colors, respectively (n=500 simulations)

Table 1. Performance metrics for infotaxis and hybrid algorithm simulations

	Infotaxis	Hybrid
Success rate (%)	49.8	69.0
Search time (s)	29.70 ± 1.34	38.13 ± 5.70
Traveled distance	1.04 ± 0.03	1.21 ± 0.11

algorithm are shown in figures 3a and 3b. To visualize the exploration-exploitation balance in terms of the geometry of the search area, we propose the concept of centerline deviation which we define as $C_{dev} = \sum_{t=0}^{N} |y(t) - y_s|$ where $y(t)$ and y_s are the y-axis coordinate of the agent and the source, respectively. This value indicates how much the agent moved away from the center of the plume, and as it is shown in figure 3c, the agents using infotaxis abruptly started moving away from the center of the plume after around 30 s while the ones using a hybrid algorithm maintained a mostly linear tendency.

4. DISCUSSION

In this paper, we investigated whether combining infotaxis, which is a probabilistic and RL based method, and a bio-inspired algorithm could balance the exploration and exploitation of information rewards to effectively perform olfactory searches. We achieved this by analyzing the expected rewards of infotaxis and found that these are larger for lateral rather than frontal movements after the cumulative number of odor hits and the information entropy of the agent's belief reach the same value. Based on this, we designed a hybrid algorithm that starts a search with infotaxis and then shifts to a bio-inspired strategy partially based on the programmed behavior of the silkmoth. By conducting simulations under a time-variant virtual odor plume obtained from readings of a gas sensor array and a wind tunnel, we found that the hybrid algorithm had a better search performance than pure infotaxis. Additionally, the hybrid algorithm had less deviation from the plume centerline.

We believe that our findings suggest that efficient odor source search could be performed by initiating a search with an RL agent such as infotaxis and shifting to a bio-inspired agent according to the relationship between accumulated odor hits and the information entropy of the belief of the odor source location. We also believe that these results might indicate that insects modulate their olfactory searches by the frequency or amount of odor detections. In future works, we will conduct experiments with silkmoths and analyze their behavior as an RL agent and investigate whether a similar shift from exploration to exploitation is exhibited by them.

REFERENCES

Kanzaki, R., Sugi, N., and Shibuya, T. (1992). Self-generated zigzag turning of bombyx mori males during pheromone-mediated upwind walking (physology). *Zoological science*, 9(3), 515–527.

Li, J.G., Meng, Q.H., Wang, Y., and Zeng, M. (2011). Odor source localization using a mobile robot in outdoor airflow environments with a particle filter algorithm. *Autonomous Robots*, 30(3), 281–292.

Lilienthal, A.J., Reggente, M., Trincavelli, M., Blanco, J.L., and Gonzalez, J. (2009). A statistical approach to gas distribution modelling with mobile robots-the kernel dm+ v algorithm. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 570–576. IEEE.

Pang, R., van Breugel, F., Dickinson, M., Riffell, J.A., and Fairhall, A. (2018). History dependence in insect flight decisions during odor tracking. *PLoS computational biology*, 14(2), e1005969.

Rodríguez, J.D., Gómez-Ullate, D., and Mejía-Monasterio, C. (2017). On the performance of blind-infotaxis under inaccurate modeling of the environment. *The European Physical Journal Special Topics*, 226(10), 2407–2420.

Shigaki, S., Sakurai, T., Ando, N., Kurabayashi, D., and Kanzaki, R. (2018). Time-varying moth-inspired algorithm for chemical plume tracing in turbulent environment. *IEEE Robotics and Automation Letters*, 3(1), 76–83.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vergassola, M., Villermaux, E., and Shraiman, B.I. (2007). 'infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126), 406.