Learn On The Fly

Yang Cai

Visual Intelligence Studio and Cylab Institute
College of Engineering and School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213, USA
ycai@cmu.edu

Abstract. In this study, we explore the biologically-inspired Learn-On-The-Fly (LOTF) method that actively learns and discovers patterns with improvisation and sensory intelligence, including pheromone trails, structure from motion, sensory fusion, sensory inhibition, and spontaneous alternation. LOTF is related to classic online modeling and adaptive modeling methods. However, the aim is to solve more comprehensive, ill-structured problems such as human activity recognition from a drone video in a disaster recovery environment. It helps to build explainable AI models that enable human-machine teaming with visual representation, visual reasoning, and machine vision. It is anticipated that LOTF would have an impact on Artificial Intelligence, video analytics for searching and tracking survivors' activities for humanitarian assistance and disaster relief (HADR), field augmented reality, and field robotic swarms.

Keywords: AI, machine learning, drone, UAV, video analytics, SLAM, HADR

1. Introduction

We often do things "on the fly" in everyday life. We gain experience without preparation, responding to events as they happen [1]. We often learn new things in that way. For example, children learn to walk, talk, and ride a bike on the fly. Historical examples include Neil Armstrong landing the lunar module on the moon. Apollo 13 crews managed to return to Earth after an explosion. Network administrators responded to the first computer worm created by Robert Morris. More recently, epidemiologists have been fighting the COVID-19 coronavirus outbreak based on live data.

Learn-on-the-fly (LOTF) is a way of *active learning* by improvisation under pressure. It is an *active learning* method to learn quickly in challenging situations of mobility, remote operation, and uncertainty, where other data-centric *passive learning* methods often fail. LOTF is related to classic "online modeling" or "adaptive modeling" methods such as Kalman Filter, Particle Filter, Recursive Time Sequence Models, and System Identification, which adapt to dynamic environments. However, LOTF aims to tackle more robust, complex problems such as human activity recognition from a drone video in a disaster recovery environment, in which small unmanned vehicles are sent out to scale recovery operations.

With LOTF we aim to build explainable AI models that enable human-machine teaming (including visual representation and visual reasoning) where humans and machines interact visually. LOTF can also incorporate lightweight machine learning algorithms such as Bayesian networks, which support low *size*, *weight and power* (SWAP) requirements. In this paper, the author overviews biologically-inspired LOTF

algorithms in non-technical terms, including pheromone trails, structure from motion, sensory fusion, sensory inhibition, and spontaneous alternation. It is anticipated that LOTF will have an impact on artificial intelligence (AI), in particular, video analytics for searching and tracking survivors' activities for humanitarian assistance and disaster relief (HADR), augmented reality, and robotic swarms.

2. Pheromone Trails

It has long been known that social insects such as ants use pheromones to leave information on their trails for foraging food, leaving instructions for efficient routes, for searching, and for making recommendations. Similarly, Amazon's retail website suggests similar products based on the items in a user's online shopping cart. In practice, the term "pheromone" proves useful in describing behaviors such as trail formation in a sequence of spatial and temporal data.

The generalized pheromone update model can help us to discover motion patterns in videos, which transforms invisible patterns of moving objects into visible trails that accumulate or decay over time, much like a scent. Pheromones decay at a certain rate, thereby reducing the risk of repeating the same route. It also helps prevent from reacting to a rapidly changing single event. Here, we generalize pheromone deposits and decay at the pixel level in two-dimension, where a "deposit" function is to add a unit of digital pheromone (in color) each time an object passes that pixel location until the value reaches its maximum. The "decay" function is to remove a unit of pheromone at a certain rate until the existing pheromone at the pixel location reaches zero. Figure 1 shows an example of traffic patterns over time from an aerial video. The heat map shows that the center lane has the heaviest traffic. In a disaster scenario, motion patterns derived from pheromone models could help identify passable routes and survivor movement, helping aid agencies to understand where survivors are headed to allocate supplies accordingly. In the COVID-19 scenario, pheromone trails can reveal traffic patterns in public spaces and can help to assess quarantine situations.

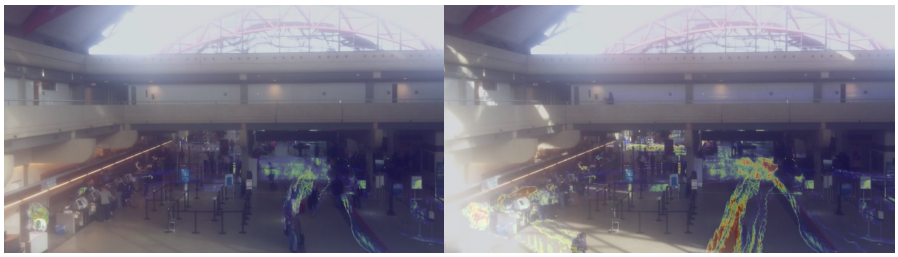

Fig. 1. The digital pheromones show the traffic flow over time from an aerial video

3. Structure from Motion

Motion perception is part of our instinct for survival. It is a vital channel for us to map our world and allocate attention to changes. To extract the motion features, we can use Optical Flow [10] to describe motion, direction, and strength in terms of motion vectors. Optical Flow assumes the brightness distribution on moving objects in a sequence of images is consistent, which is referred to as "brightness constancy." We use the Horn-

Schunck algorithm to minimize the global energy over the image. This algorithm generates a high-density of global optical flow vectors, which is useful for measurement purposes. We then use grid density to define the number of motion vectors in a frame. For example, we can plot a motion vector for every 10 pixels horizontally and vertically respectively.

Dynamic visualization of the field of optical flow is a critical component to reveal the changes of flow patterns over time. This is called a *flow map*. In addition to the flow map, we can visualize the motion vector in the color space of hue, saturation, and value (HSV), wherein hue represents the angle of the vector, and value represents the magnitude of length of the vector. The optical flow vector angle can be naturally mapped to hue in the HSV color system, both in range between 0 and 360 degrees. The magnitude of the vector can be mapped to a value between 0 and 1. Saturation value for this visualization is constant, so we can set it as 1 - the highest value by default. We chose the HSV color space for mapping the two parameters because of its simplicity. Figure 2 shows that the optical flow heat map visualizes the slow-moving utility truck in the wrong direction. This method is based on the assumption that the video is from a stationary camera. The heuristic algorithm for segmentation from a moving camera is in reference [9].

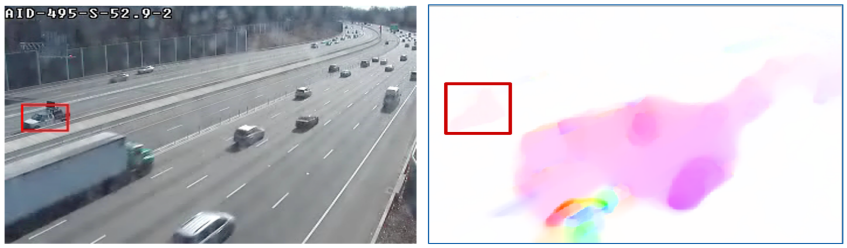

Fig. 2. The optical flow map visualizes the slow-moving utility truck in the wrong direction

Motion creates depth perception that can be used for reconstructing three-dimensional objects, which is beneficial for disaster recovery, for example, assessing the fire damage of Notre-Dame de Paris. Given a 2D video from a drone camera, we use Stereo-photogrammetry [11] to extract the 3D measurements. By analyzing the motion field between frames, the algorithm is designed to find corresponding points shared between frames, allowing for reconstruction of 3D structural coordinates from a single camera. The key assumptions of this method are: the video contains enough high-contrast corner-like feature points, which are used for matching the corresponding structural features; and the geometric transformation caused by the motion is Homographic Transformation [12]. Note that stereo-photogrammetry is computationally intensive. We must down-scale the 4K video to a manageable size in order to achieve a reasonable computation time. Figure 3 shows the results of a 3D reconstructed archeological site in Paspardo in the Italian Alps.

For the last two decades, Structure-from-Motion (SfM) has been evolved into a popular technology for 3D imaging with an affordable single camera, a pair of stereo cameras, or multiple cameras [13]. The RGB camera-based SfM methods commonly need structural features such as Difference of Gaussian (DoG) SIFT features [14-15] or FAST corner features [16] to match the structural features between frames in the video,

and to calculate the homographic transformation matrix accordingly for Simultaneous Localization and Mapping (SLAM) [17]. Similar to stereo-photogrammetry, the matching algorithm requires a minimal number of features in consecutive frames of the video. Unfortunately, in many cases, there are not enough matching features between frames, due to "featureless" smooth walls, blurry images, or rapid movement of the camera. Figure 4 shows results of the SLAM of the floor of an office building with a stereo camera, where the green dots represent the camera's motion path, and the other color dots represent the walls and the floor. The point cloud of the ceiling has been cut away to increase visibility.

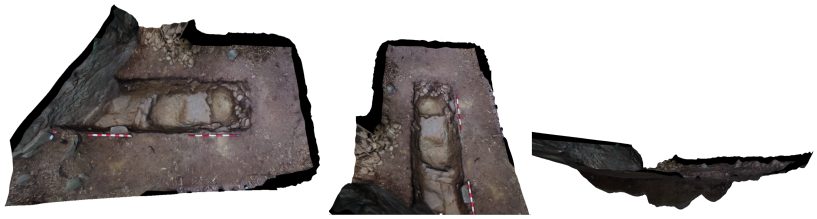

Fig. 3. The archeological site is 3D reconstructed from a drone video with an RGB camera

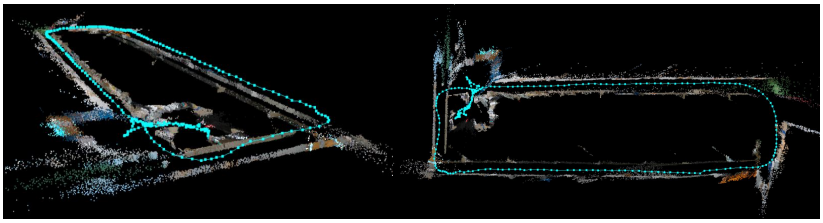

Fig. 4. The SLAM results of a floor plan and path in a building from a stereo camera

4. Sensory Fusion

The LOTF approach with the most potential for disaster recovery applications is sensory fusion. Modern electronic systems such as drones and mobile phones carry many sensors: cameras, microphones, motion sensors, magnetic field sensors, GPS, WiFi, Bluetooth, cellular data, proxy distance sensors, near infrared sensors, and so on. In contrast with prevailing machine learning methods such as Convolutional Neural Networks (CNN), which require massive historical training data, learn-on-the-fly focuses on real-time lateral sensory data fusion to reveal patterns. For example, fusing laser distance sensor data with inertial motion unit (IMU) sensor data can enable activity recognition of firefighters with a Decision Tree [18]. Adding more sensory dimensions increases the confidence of pattern recognition. It also improves human-machine teaming in the field of humanitarian assistance and disaster relief (HADR) tasks. For example, thermal imaging helps to detect humans and vehicles, but it has relatively low resolution compared to the visible channel. Superimposing edges on objects would assist humans and machines to identify and track the objects. Figure 5 shows screenshots of a drone video and the thermal image with edge enhancement from the visible channel.

Fig. 5. The drone image (left) and thermal image (right) of Paspardo with edge enhancement

5. Sensory Inhibition

In contrast to sensory fusion, sensory inhibition prioritizes the sensory channels in order to reduce the computational burden [6]. Sensory inhibition is also referred to as "lateral inhibition [19]," which is common in nature. For example, in neurobiology, lateral inhibition disables the spreading of action potentials from excited neurons to neighboring neurons in the lateral direction in order to create a contrast in stimulation. This happens to visual, tactile, auditory, and olfactory processing as well. For example, we do not taste our own saliva and we do not hear the sound of our jaw moving while eating. Artificial lateral inhibition has been incorporated into vision chips, hearing aids, and optical mice. Typical sensory inhibition is implemented by methods of thresholding, delaying, and adapting. In our case, given multiple sensory channels, we find the channel that has the most contrast with the minimal processing load. Figure 6 shows a depth map and thermal image of two men laying on the floor. The thermal image shows more temperature contrast than the depth map shows distance. Therefore, to detect the human body in this case, thermal imaging would be easier. However, this imaging preference is relative and dynamic. If the person were to stand up, or if the floor temperature were as warm as the human body, then the figure-background contrast relationship would change. Figure 7 shows a color image and a depth map of two men on stairs. The depth map appears to have advantages in human detection, gesture recognition, and spatial relationship estimation when compared to a color image. Adaptation is a form of inhibition.

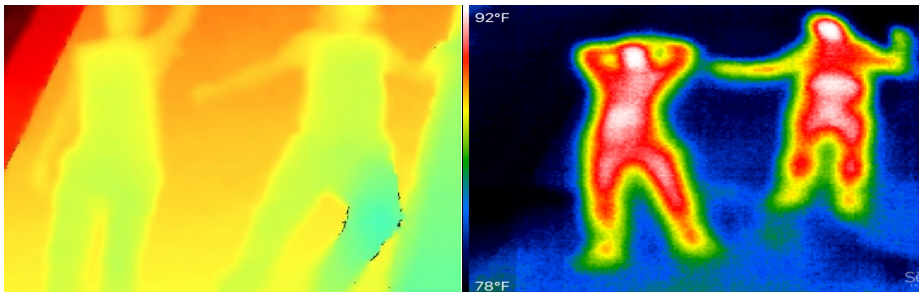

Fig. 6. The depth map (left) and thermal image (right) of men laid on the floor

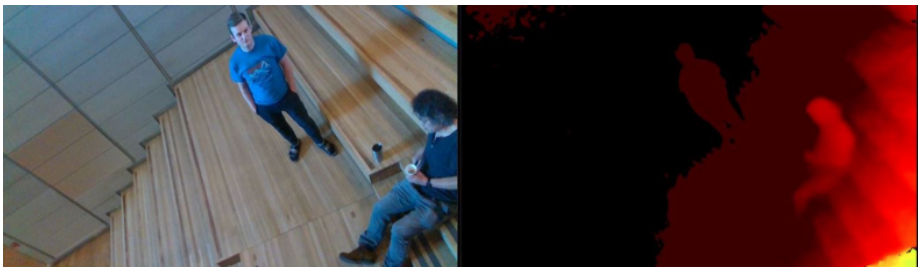

Fig. 7. The aerial color image (left) and depth map (right) of men on stairs

6. Spontaneous Alternation Behavior (SAB)

Creatures in nature commonly learn on-the-fly to adapt to changing environments. One instinctual behavior is randomization in order to search for alternative foraging paths or to avoid collision situations. When an ant gets lost, it will randomly wander until it hits a trail marked with pheromones. This pattern occurs in tests with many different animals. It is called *spontaneous alternation behavior* (SAB) [7]. Spontaneous alternation of paths for an autonomous robot, a search engine, or a problem-solving algorithm can help to explore new areas and avoid deadlock situations. Spontaneous alternation is also a primitive strategy for collision recovery. Collisions can be found in many modern electronic systems in various fields, from autonomous driving vehicles to data communication protocols. There is a variation of the SAB strategy for collision recovery. When a collision occurs, the system spontaneously switches to different sensors or channels, or the system waits for random intervals and reconnects. The "back down" and reconnect process is similar to SAB, which solves the problem of deadlock. SAB is necessary for missions involving the search for and tracking of survivors for humanitarian assistance and disaster relief (HADR) when existing maps of the environment are inaccurate due to changes that occurred during a disaster such as fallen trees or collapsed buildings. This is true especially in cases where communication breaks down, the system collapses or runs into a deadlock, or when deep, extended searches for victims in missing spots is required.

7. Summary

In this study, we explore the biologically-inspired Learn-On-The-Fly (LOTF) method that actively learns and discovers patterns with improvisation and sensory intelligence, including pheromone trails, structure from motion, sensory fusion, sensory inhibition, and spontaneous alternation. LOTF is related to classic "online modeling" or "adaptive modeling" methods. However, it aims to solve more comprehensive, ill-structured problems such as human activity recognition from a drone video in a disaster scenario. LOTF helps to build explainable AI models that enable human-machine teaming, including visual representations and visual reasoning, toward machine vision. It is anticipated that LOTF will have an impact on Artificial Intelligence, video analytics for searching and tracking survivors' activities for humanitarian assistance and disaster relief (HADR), field augmented reality, and field robotic swarms.

LOTF is an evolving approach that moves away from data-centric to sensor-centric, from rigid to adaptive, from unexplainable to explainable, from numeric to intuitive,

and from curve-fitting to semantic reasoning. Our challenges include how we can scale up the system, how we will implement sensory adaptation as inhibition and, finally, how we achieve a balance between the flexibility and efficiency of the algorithms. We intend to address these challenges in a disaster recovery scenario.

Acknowledgement

The author would like to thank Sean Hackett and Florian Alber for data collection and prototyping, Professor Mel Siegel for his discussions and references on sensors and sensing, and Dennis A. Fortner for his organization. This study is in part sponsored by Northrop Grumman Corporation and NIST PSCR / PSIA program. The author is grateful to Mission Systems AI Architects, Neta Ezer and Nick Molino for their detailed commends and editing. The author would like to thank Program Managers Justin King, Erin A. Cherry, Isidoros Doxas, Donald D. Steiner, Paul Conoval, Jason B. Clark Jeb Benson, and Scott Ledgewood for discussions, reviews and advice.

References

1. Wikipedia. On the fly. captured in 2020
2. Richman, C. and Dember, W.N.: Spontaneous alternation behavior in animals: a review. Current Psychological Research & Reviews, Winter 1986-87, vol. 5, no. 4, 358-391 (1986)
3. Hull, C.L.: Principles of behavior. New York: Appleton-Century (1943)
4. Hughes, R.N.: Turn alternation in woodlice. Animal Behavior, 15, 282-286 (1967)
5. DARPA Grand Challenge: https://en.wikipedia.org/wiki/DARPA_Grand_Challenge (2016)
6. von Békésy, G.: Sensory Inhibition. Princeton University Press (1967)
7. Cai, Y.: Instinctive Computing, Springer-London (2016)
8. Wigglesworth,V.B.: Insect Hormones, W.H. Freeman and Company (1970).
9. Cai, Y.: Ambient Diagnostics, CRC Press (2014 and 2019)
10. Horn, B.K.P. and B.G. Schunck, "Determining optical flow." *Artificial Intelligence*, vol 17, pp 185-203, 1981. Manuscript available on MIT server (1981)
11. Photogrammetry: https://en.wikipedia.org/wiki/Photogrammetry (2020)
12. OPENCV, Basic concept of the homography explained with code: https://docs.opencv.org/master/d9/dab/tutorial_homography.html (2020)
13. WikiPedia, Structure from Motion: https://en.wikipedia.org/wiki/Structure_from_motion
14. Rowe, D.: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60(2), 91-110 (2004)
15. SURF: https://en.wikipedia.org/wiki/Speeded_up_robust_features (2020)
16. FAST Corner detection: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_fast/py_fast.html (2020)
17. SLAM, WikiPedia: https://en.wikipedia.org/wiki/Simultaneous_localization_and_mapping (2020)
18. Hackett, S., Cai, Y., Siegel, M.: Activity Recognition from Firefighter's Helmet, Proceedings of CISP-BMEI, Huaqiao, China (2019)
19. Lateral Inhibition, WikiPedia: https://en.wikipedia.org/wiki/Lateral_inhibition (2020)