A Machine Learning Approach to the Visual Perception of Forest Trails

Alessandro Giusti[1], Jerome Guzzi[1], Dan Ciresan[1], Fang-Lin He[1], Juan Pablo Rodríguez[1]
Flavio Fontana[2], Matthias Faessler[2], Christian Forster[2]
Juergen Schmidhuber[1], Gianni Di Caro[1], Davide Scaramuzza[2], Luca Gambardella[1]

***Abstract*— We study the problem of perceiving forest or mountain trails from a single monocular image acquired from the viewpoint of a robot traveling on the trail itself. Previous literature focused on trail segmentation, and used low-level features such as image saliency or appearance contrast; we propose a different approach based on a Deep Neural Network used as a supervised image classifier. By operating on the whole image at once, our system outputs the main direction of the trail compared to the viewing direction. Qualitative and quantitative results computed on a large real-world dataset show that our approach outperforms alternatives, and yields an accuracy comparable to the accuracy of humans which are tested on the same image classification task. Preliminary results on using this information for quadrotor control in unseen trails are reported.**

SUPPLEMENTARY MATERIAL

A narrated video summary of this paper is available at `http://youtu.be/xOr_zrAIR-c`.

Additional data and figures are available at `http://bit.ly/perceivingtrails`.

I. INTRODUCTION

Autonomously following a man-made trail (such as those normally traversed by hikers or mountain-bikers) is a challenging and mostly unsolved task for robotics. Solving such problem is important for many applications, including wilderness mapping [1] and search and rescue; moreover, following a trail would be the most efficient and safest way for a ground robot to travel medium and long distances in a forested environment: by their own nature, trails avoid excessive slopes and impassable ground (e.g. due to excessive vegatation or wetlands). Many robot types, including wheeled, tracked and legged vehicles [2], are capable of locomotion along real-world trails. Moreover, Micro Aerial Vehicles (possibly collision-resilient [3]) flying under the tree canopy [4], [5] are a compelling and realistic option made possible by recent technological advances.

In order to successfully follow a forest trail, a robot has to somehow *perceive* where the trail is heading, then *move* in order to stay on the trail and progress along it. In this paper we focus on the former problem; we consider as input a monocular image (i.e. we do not use depth information which may be acquired by ad-hoc sensors, structure from motion, or stereo vision) acquired from a viewpoint lying on the trail and an approximately horizontal direction of view (i.e. not top-down, as in remote sensing or aerial mapping).

Perceiving real-world trails in these conditions is an extremely difficult and interesting pattern recognition problem (see Figure 1), which is often challenging even for humans (e.g., losing a trail is a common experience among casual hikers). Computer Vision and Robotics literature mainly focused on paved road [6] and forest/desert road perception [7]. The latter is a significantly more difficult problem than the former, because unpaved roads are normally much less structured than paved ones: their appearance is very variable and often no well-defined boundaries can be seen. Compared to roads, the perception of trails poses an even harder challenge, because their surface appearance can change very frequently, their shape and width is not as constrained, and they often seamlessly blend with the surrounding area (e.g. grass).

Several previous works [8], [9] dealing with trail perception solve a *segmentation* problem: i.e., determines which areas of the input image correspond to the image of the trail. In order to solve this task, one needs to explicitly define *which visual features characterize a trail*. Rasmussen et al. [8] rely on *appearance contrast*, whereas Santana et al. adopt image conspicuity [9]; both features are conceptually similar of image saliency [10]. For every pixel of an image, saliency quantifies how much such pixel visually "stands out" from the rest; for example, the pixels belonging to a small colored object on an uniform background will be characterized by an high saliency with respect to the saliency computed for the background pixels. If we assume that the trail image exhibits some sort of marked visual difference with respect to its surroundings, then saliency will be high for trail pixels, and low elsewhere. This information, which by itself is expected to be very noisy, is aggregated [11] with a number of simple (symmetric, triangular shape [8]) or complex (spatial-temporal integration based on virtual ants [9]) geometry priors in order to infer the trail position and direction in the image, thus producing a rough segmentation of the trail.

We follow a different approach and cast the trail perception problem as an *image classification* task:we estimate the approximate direction the trail with respect to the direction of view, by adopting a supervised machine learning approach based on Deep Neural Networks (DNNs), a state-of-the-art *deep learning* technique which operates directly on the

[1] Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Lugano, Switzerland

[2] Robotics and Perception Group (RPG), University of Zurich, Switzerland

This research was supported by the Swiss National Science Foundation (SNSF) through: the National Centre of Competence in Research (NCCR) Robotics (www.nccr-robotics.ch); and the Supervised Deep / Recurrent Nets grant (project code 140399)

Fig. 1: Three images from our dataset. Given an image, we aim to determine the approximate direction the trail is heading with respect to the viewing direction (respectively, left, right and straight ahead).

image's raw pixel values (Section III-B).DNNs have recently emerged as a powerful tool for various computer vision tasks (e.g. object classification [12], [13], biomedical image segmentation [14]), often outperforming other techniques. One of the advantages of DNNs over common alternatives for supervised image classification is *generality*: in fact, features are learned directly from data, and do not have to be chosen or designed by the algorithm developers for the specific problem on which they are applied.

Machine learning techniques have been used for a long time [15] in order to map visual inputs to actions. When the goal is obstacle avoidance, several works [16], [17], [18] obtained good results with simple biologically-inspired controllers based on optical flow features. More recently, deep learning techniques have also been adopted by Sermanet, Hadsell et al. [19], [20] for autonomous navigation of a robot in various nonstructured environments; in these works, the terrain visible in front of the robot is classified for traversability, which provides high-level information for obstacle-free path planning. Ross et al. [4] used imitation learning [21], [22] to steer a quadrotor in order to avoid trees in an outdoor environment; the controller is previously trained by manually piloting the robot for a short time. In our case, the visual perception task is harder (Fig. 1) since real-world trails have much more appearance variability than trees at close range. This requires a more powerful classifier to be trained with a significantly larger training dataset, which would be impractical to acquire by manually piloting a robot: such dataset is acquired offline by means of a simple but efficient expedient introduced in Section III-A.

We formalize the problem in Section II. Our **main contribution** is a trail perception technique based on Deep Neural Networks (Section III-B) which bypasses the challenging problem of determining the characteristic features of a trail. The system is trained on a large dataset efficiently acquired on real-world hiking trails (Section III-A). We quantitatively compare our approach with alternatives on an unseen testing set in Section IV, and report preliminary results on the problem of controlling a robot to follow a previously-unseen trail, relying only on the outputs of our system (some videos of the system in action are reported in supplementary material [23]).

II. Problem Formulation

Consider a generic scene with a single trail in a wilderness setting. Our input is an image acquired by a camera situated above the trail. In the following, we assume the viewpoint height is similar to the height of a person (approximately 1.7m), because it is high enough to provide a good view over the surrounding ground, but still a realistic sensor position for medium-sized all-terrain ground robots; moreover, we can expect that such height is mostly free of obstacles on trails in forested areas, and as such, a reasonable choice for micro aerial robots. Let $\vec{v}$ be the direction of the camera's optical axis; we assume that $\vec{v}$ lies on the horizontal plane. Furthermore, let $\vec{t}$ be the dominant direction of the trail: we define $\vec{t}$ as the (horizontal) direction towards which a hiker would start walking if standing at the position of the robot, with the goal of remaining on the trail.

Let α be the signed angle between $\vec{v}$ and $\vec{t}$: we consider three classes, which correspond to three different actions that the (human or robotic) carrier of the camera should implement in order to remain on the trail, assuming that the camera is looking at the direction of motion.

Turn Left (TL) if $-90° < \alpha < -\beta$; i.e., the trail is heading towards the left part of the image.

Go Straight (GS) if $-\beta \leq \alpha < +\beta$; i.e., the trail is heading straight ahead, at least in the close range.

Turn Right (TR) if $+\beta \leq \alpha < +90°$; i.e., the trail is heading towards the right part of the image.

Given the input image, our goal is to classify it in one of the three classes. In the following, we consider $\beta = 15°$.

Note that, in case the absolute value of α is large, the trail may entirely lie outside of the camera view (i.e., the robot is looking in a perpendicular direction with respect to the trail, which is not visible anywhere in the image). In that case, the image only allows us to infer that the true class is *not* GS.

III. Visual Perception of Forest Trails

We solve the problem as a supervised machine learning task, which is extremely challenging because of the wide appearance variability of the trail and its surroundings: perceptions are heavily affected by lighting conditions, vegetation types, altitude, local topography, and many other factors. We deal with such challenges by gathering a large and representative labeled dataset, covering a large variety of trails and a long distance on each.

Fig. 2: *Left:* stylized top view of the acquisition setup; *Right:* our hiker during an acquisition, equipped with the three head-mounted cameras.

A. Dataset

To acquire such a dataset, we equip a hiker with three head-mounted cameras: one pointing 30° to the left, one pointing straight ahead, and one pointing 30° to the right. The hiker then swiftly walks a long trail, by taking care of always looking straight along its direction of motion. The dataset is composed by the images acquired by the three cameras.

Each image is labeled, i.e. it is associated to its ground truth class. Because of the very definition of our classes, all images acquired by the central camera are of class GS: in fact, they were acquired while the hiker was walking along the trail, and looking straight ahead (i.e., $\alpha \approx 0°$) in the direction of motion. Conversely, the right looking camera acquires instances for the TL class, with $\alpha \approx 30°$; and the left-looking camera acquires instances of the TR class ($\alpha \approx -30°$).

The dataset [1] is currently composed by 8 hours of 1920x1080 30FPS video acquired using 3 GoPro Hero3 Silver cameras in the configuration outlined above, and covers approximately 7 kilometres of hiking trails acquired at altitudes ranging from 300m to 1200m, different times of day and weather. Exposure, dynamic range and white balance are automatically controlled by the cameras. To avoid long exposure times which would yield to motion-blur, all sequences are acquired during daytime, excluding twilight. Many different trail types and surroundings are represented, ranging from sloped narrow alpine paths to wider forest roads. Acquisitions are normally uninterrupted unless for technical reasons or to avoid long sections on paved roads; this ensures that the dataset is representative not only of ideal, "clean" trails but also of frequent challenging or ambigous spots often observed in the real world. Synchronized GPS and compass information has been recorded for most sequences, but is unused at the moment. Supplementary material [23] reports a random sample of images from the dataset, as compared to the datasets used by Santana et al. [9].

The dataset has been split in disjoint training (17119 frames) and testing (7355 frames) sets. The split was defined by carefully avoiding that the same trail section appears in both the training and testing set. The three classes are evenly represented within each set.

[1]The whole dataset is available as supplementary material [23]

B. Deep Neural Networks for Trail perception

We use a DNN as a black-box image classifier, and adopt a network architecture which has been shown to perform well when applied to a large amount of image classification problems [13]; in particular, we consider a matrix of $3 \times 101 \times 101$ neurons as the input layer, followed by a number of hidden layers (Figure 3) and three output neurons.

The input image is first anisotropically resized to a size of 101×101 pixels; the resulting $3 \times 101 \times 101$ RGB values are directly mapped to the neurons in the input layer. For a given input, the DNN outputs three values, representing the probability that the input is of class TL, GS, TR, respectively.

Training a net: The 17119 training frames are used as training data. The training set is augmented by synthesizing left/right mirrored versions of each training image. In particular, a mirrored training image of class TR (TL) yields a new training sample for class TL (TR); a mirrored GS training sample yields another GS training sample. Additionally, mild affine distorsions ($\pm 10\%$ translation, $\pm 15°$ rotation, $\pm 10\%$ scaling) are applied to training images to further increase the number of samples. The DNN is trained using backpropagation for 90 epochs, which requires about 3 days on a workstation equipped with an Nvidia GTX 760 GPU.

DNN architecture: A DNN is a feed-forward connectionist model built out of successive pairs of convolutional and max-pooling layers, followed by several fully connected layers (the architecture adopted in our system is illustrated in Figure 3). Input pixel intensities are passed through this complex, hierarchical feature extractor. The fully connected layers at the end of the network act as a general classifier. The free parameters (weights), initialized with small random numbers, are jointly optimized using stochastic gradient descent to minimize the misclassification error over the training set.

Convolutional layers [24] perform 2D convolutions of their input maps with a rectangular filter. When the previous layer contains more than one map, the results of the corresponding convolutions are summed and transformed by a nonlinear activation function. Higher activations will occur where the filter better matches the content of the map, which

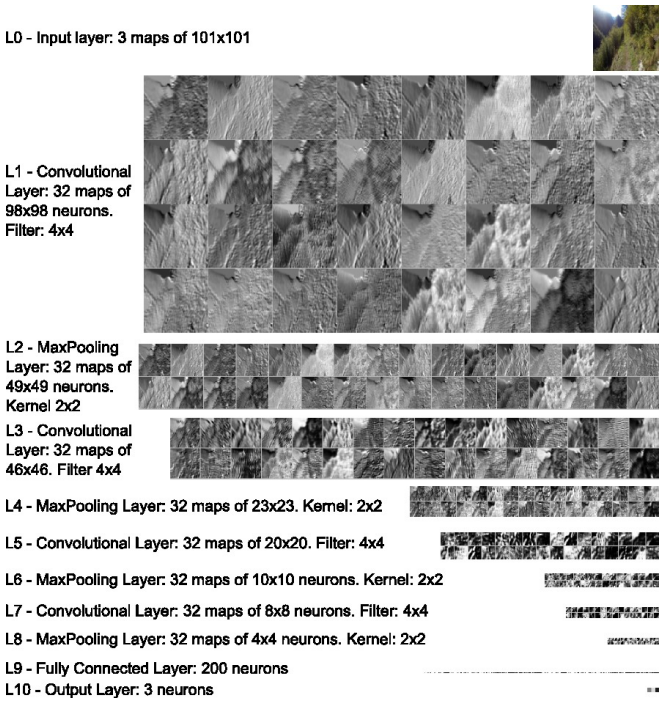

Fig. 3: Architecture for the DNN used in our system, and representation of the maps in each layer

can be interpreted as a search for a particular feature. The output of the **max-pooling (MP) layers** [25] is formed by the maximum activations over non-overlapping square regions. MP layers decrease the map size, thus reducing the network complexity. MP layers are fixed, non-trainable layers selecting the winning neurons. Typical DNNs are much wider than previous CNN, with many more connections, weights and non-linearities. A GPU implementation is used in order to to speed up training. The **output layer** is a fully connected layer with one neuron per class (i.e. TL, GS and TR), activated by a softmax function. Each output neuron's activation can be interpreted as the probability of the input image belonging to that class.

IV. Experimental Results

Performance metrics: We use the testing set defined in Section III-A (7355 images) in order to compute performance metrics for different classification techniques.

For the three-class classification problem defined in Section II, we compute the absolute accuracy (i.e. fraction of correctly classified images) and the confusion matrix.

We additionally consider a derived two-class classification problem, on which additional, more robust performance measures can be defined. In the two-class problem, one has to decide whether an input image is of class GS or not (i.e., whether a trail is visible straight ahead, or not). The image is classified as GS if and only if $P(\text{GS}) > T$. We report the accuracy of the binary classifier for $T = 0.5$, and the corresponding precision, recall, and the area under the ROC curve (the last is a robust metric and does not depend on the choice of T).

Comparisons: We test two versions of our proposed technique and four alternatives.

DNN, where $P(\text{TL})$, $P(\text{GS})$ and $P(\text{TR})$ are directly computed by applying the DNN model to the input frame.

Simple Saliency-based Model , where we compute saliency maps of the input frame using Itti's model [26], as in Santana et al. [9]. Such map is computed on the image hue only, which preliminary experiments shown to be the configuration where saliency is most correlated to trail location. The saliency map is discretized to 16×9 blocks, and the average saliency for each block yields a 144-dimensional feature vector. A SVM model with an RBF kernel is learned from the training set to map such feature vector to $P(\text{TL})$, $P(\text{GS})$ and $P(\text{TR})$.

Santana et al., whose algorithm is applied to the frames extracted from our videos (50 iterations per frame) and its output trail soft segmentation is sampled at each of the testing frames. In order to map a class to a segmentation, we follow the quantitative evaluation in [9]: a single representative point for the trail is computed as the centroid of the largest connected component in the binarized trail probability map; the threshold is computed as $0.85 \cdot M$, where M is the maximum value of the probability map. Then, we classify an image as TR (respectively, TL) if the x coordinate of such point is larger than $(0.5 + k) \cdot W$ (respectively, smaller than $(0.5 - k) \cdot W$), where W is the image width; else, the image is classified as SC. k is chosen in order to optimize the accuracy of the resulting classifier.

Two human observers, each of which is asked to classify 200 randomly sampled images from the testing set in one of the three classes.

TABLE I: Results for the three-class problem.

	DNN	Saliency	[9]	Human1	Human2
Accuracy	85.2%	52.3%	36.5%	86.5%	82.0%

TABLE II: Results for the two-class problem.

	DNN	Saliency	[9]	Human1	Human2
Accuracy	95.0%	73.6%	57.9%	91.0%	88.0%
Precision	95.3%	60.9%	39.8%	79.7%	84.0%
Recall	88.7%	46.6%	64.6%	95.1%	81.6%
AUC	98.7%	75.9%	–	–	–

Tables I and II report quantitative results for the three- and two-class problems, respectively. We observe that DNN methods perform comparably to humans, meaning that they manage to extract much of the information available in the inputs. The Simple Saliency Model and [9] perform poorly on this data, which is expected as image saliency is not correlated to the trail location in most of our dataset.

(a) GS frames with highest $P(\text{GS})$, i.e. frames where the path is easily found as being straight ahead

(b) not-GS images with highest $P(\text{GS})$, i.e. failure cases where the path is not straight ahead but was detected as such

Fig. 4: Success and failure cases. More examples are reported in supplementary material [23].

Failure Cases and Qualitative Results

Figure 4 reports success and failure cases on the testing set. We observe that instances which are easy for our system are also trivial for human observers, whereas instances that our system failed (third and fourth row) are in fact difficult and ambigous cases.

We also implemented an additional qualitative test using videos acquired from a Samsung Galaxy SIII cellphone on a sloped forest trail in a different region than those in which our dataset was acquired. These videos have a much lower FOV than the dataset videos (about $60°$ vs $120°$), are highly compressed, and frequently exhibit under/over exposed areas due to the limited dynamic range of the sensor. The viewing direction rotates frequently in such a way that the trail is not always in the center of the frame, and is often not visible at all. Figure 5 reports three representative frames for such video, along with the outputs of our system; this and other videos, annotated with the outputs of our system, are available as supplementary material [23].

Discussion: Control of a Robot for Autonomous Trail Following

This paper focused on the problem of perceiving the trail, and proposed a robust solution which provides a rough indication on the direction the trail is heading with respect to the direction of view. This information alone is unlikely to be descriptive enough to reliably steer a robot in order to follow a trail. Instead, it may act as an additional input to a robot control framework equipped with state estimation, obstacle detection and avoidance, and high-level path planning. Nonetheless, it is interesting to study how well a robot navigates in the real world when using exclusively with our vision system. In order to investigate this, we implemented a simple reactive controller which translates our system's outputs to control signals as follows:

yaw (i.e. steering) is proportional to $P(\text{TR}) - P(\text{TL})$; a positive value steers the robot to the right, and a negative value steers the robot to the left;

speed is proportional to $P(\text{GS})$.

We performed preliminary tests of such controller on two platforms: a Parrot ARDrone controlled by a laptop; and a standalone quadrotor equipped with a forward-looking MatrixVision mvBlueFox global shutter camera, with an onboard Odroid-U3 system for image processing and control, also implementing a semi-direct monocular visual odometry pipeline [27]. The Odroid processor runs both vision systems symultaneously at more than 15 FPS. The main problem we observed during our tests was the much lower image quality acquired by the quadrotors' cameras as compared to the gopro images in the training dataset; this yielded a lower performance of the classifier compared to the testing datasets. This was expecially apparent in situations with strong sky-ground contrast, as the dynamic range of the mvBlueFox camera can not capture well-exposed footage. An additional problem we observed is that the quadrotor is unable to negotiate narrow trails, since the classifier does not promptly react as the quadrotor's lateral drift if the yaw is roughly correct. On wider trails and in even lighting conditions, the robot was able to successfully follow the trail for a few hundreds of meters (see supplementary videos).

V. Conclusions

We trained a Deep Neural Network for visually perceiving the direction of an hiking trail from a single image. Trained on a large real-world dataset and tested on a disjoint set, the system performs better comparably to humans. By operating on the raw RGB frames, we bypass the need to define characteristic features of trails, which is a very difficult task given the huge variability of their appearance. Preliminary field tests showed promising results.

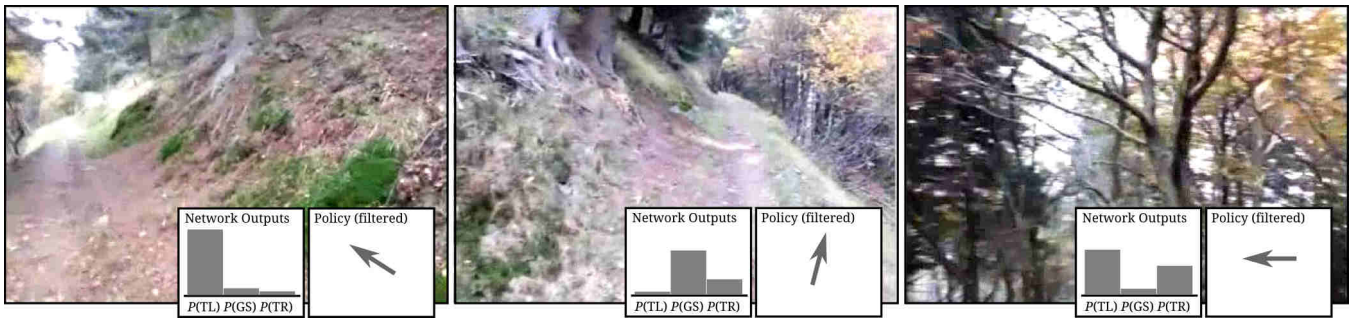

Fig. 5: Three representative frames form the cellphone video robustness test. For each frame, we report the raw network outputs for that frame (bar graph), and the motion policy (arrow) derived from such outputs averaged over the previous 10 frames (see text). The rightmost frame is acquired when looking sideways, so the trail is not visible; the DNN is then rightfully confused among TR and TL, but returns a very small value for $P(\text{GS})$.

.

Fig. 6: Images from preliminary field testing. Rightmost figure shows quadrotor with Odroid-U3 featuring full on-board processing.

.

REFERENCES

[1] Google, "Maps trekker." `www.google.com/maps/about/partners/streetview/trekker`.

[2] M. Hutter *et al.*, *StarlETH: A compliant quadrupedal robot for fast, efficient, and versatile locomotion*. 2012.

[3] A. Briod, P. Kornatowski, J.-C. Zufferey, and D. Floreano, "A collision-resilient flying robot," *Journal of Field Robotics*, vol. 31, no. 4, 2014.

[4] S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert, "Learning monocular reactive uav control in cluttered natural environments," in *Proc. ICRA*, 2013.

[5] A. J. Barry, A. Majumdar, and R. Tedrake, "Safety verification of reactive controllers for uav flight in cluttered environments using barrier certificates," in *Proc. ICRA*, 2012.

[6] A. B. Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: a survey," *Machine Vision and Applications*, pp. 1–19, 2012.

[7] S. Thrun *et al.*, "Stanley: The robot that won the darpa grand challenge," *Journal of field Robotics*, vol. 23, no. 9, 2006.

[8] C. Rasmussen, Y. Lu, and M. Kocamaz, "Appearance contrast for fast, robust trail-following," in *Proc. IROS*, pp. 3505–3512, 2009.

[9] P. Santana, L. Correia, R. Mendonça, N. Alves, and J. Barata, "Tracking natural trails with swarm-based visual saliency," *Journal of Field Robotics*, vol. 30, no. 1, pp. 64–86, 2013.

[10] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Transactions on PAMI*, vol. 33, no. 11, pp. 2131–2146, 2011.

[11] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," in *Proc. ECCV*, pp. 581–594, 2006.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, pp. 1106–1114, 2012.

[13] D. C. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," in *Proc. CVPR*, pp. 3642–3649, 2012.

[14] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images," in *Proc. NIPS*, pp. 2852–2860, 2012.

[15] D. A. Pomerleau, "Neural network perception for mobile robot guidance," 1993.

[16] J. Conroy, G. Gremillion, B. Ranganathan, and J. S. Humbert, "Implementation of wide-field integration of optic flow for autonomous quadrotor navigation," *Autonomous robots*, vol. 27, no. 3, 2009.

[17] M. V. Srinivasan, "Visual control of navigation in insects and its relevance for robotics," *Current opinion in neurobiology*, vol. 21, no. 4, pp. 535–543, 2011.

[18] A. Beyeler, J.-C. Zufferey, and D. Floreano, "Vision-based control of near-obstacle flight," *Autonomous robots*, vol. 27, no. 3, 2009.

[19] P. Sermanet *et al.*, "A multirange architecture for collision-free off-road robot navigation," *Journal of Field Robotics*, vol. 26, no. 1, 2009.

[20] R. Hadsell *et al.*, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.

[21] J. Kober and J. Peters, "Learning motor primitives for robotics," in *Proc. ICRA*, pp. 2112–2118, 2009.

[22] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 627–635, 2011.

[23] "Supplementary material." `http://bit.ly/perceivingtrails`, 2015.

[24] Y. LeCun, L. Bottou, G. Orr, and K. Muller, "Efficient BackProp," in *Neural Networks: Tricks of the trade* (G. Orr and M. K., eds.), Springer, 1998.

[25] D. Scherer, A. Müller, and S. Behnke, "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition," in *Proc. International Conference on Artificial Neural Networks*, 2010.

[26] L. Itti, C. Koch, E. Niebur, *et al.*, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[27] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. ICRA*, 2014.