Finding 3D Positions of Distant Targets from a Moving Drone

Julius Pesonen[1,2], Arno Solin[2], Eija Honkavaara[1]

[1]Department of Photogrammetry and Remote Sensing, Finnish Geospatial Research Institute

[2]Department of Computer Science, Aalto University

My homepage

What and why?

How to determine the location of a target object such as a wildfire from a drone looking towards the horizon?

- Segmentation models can be taught to distinguish smoke clouds even from ~10 kilometres in real-time with on-board resources [1]
- Finding the 3D positions on-board the observing drones enables communicating wildfire positions directly
- Most 3D vision solutions are computationally expensive and/or limited to shorter distances

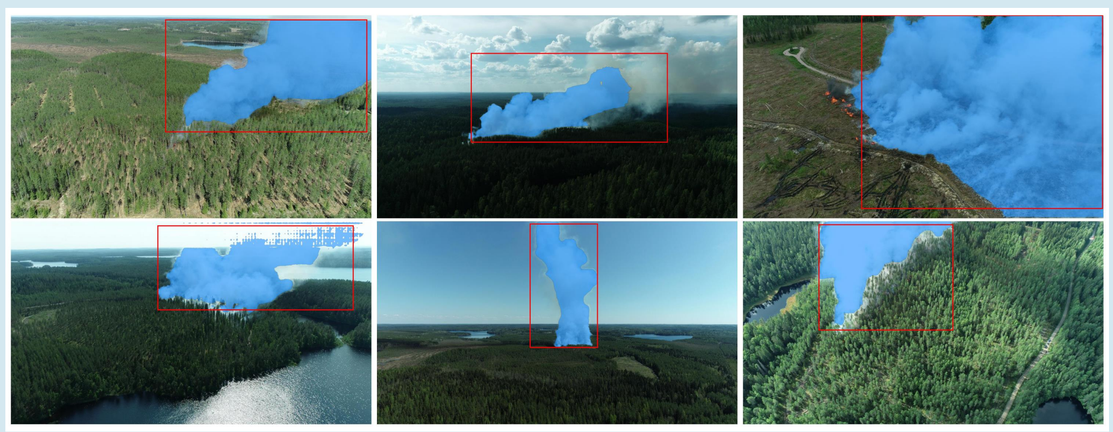

Figure 1. Wildfire smoke segmentation examples from our smoke segmentation dataset [2].

Perspective-based solution – the particle filter

The filter fits a distribution, modelled with random points, to the 3D position of the target object using a sequence of segments and camera poses.

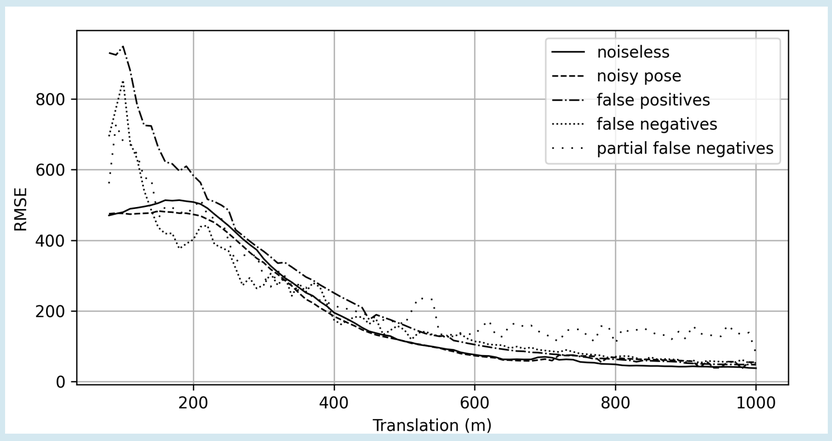

Figure 1. Simulation results of a particle filter positioning a target observed from two kilometres away in the presence of different types of noise.

Pros:

+ Intuitive
+ Provides uncertainty
+ No data required

Cons:

- Requires engineering for new tasks
- Manually defined behaviour
- No estimates without camera movement

Learning-based options

If the target can be found with a particle filter or a human can estimate the distance from a video, why couldn't a neural network do it?

How can we train them *without the right data?*

- Mono depth estimation

- Combining it with the segment would tell how far each segmented pixel is from the camera
- Metric depth estimation models fail with aerial imagery due to lack of available training data
- Obtaining ground truth for such data is difficult

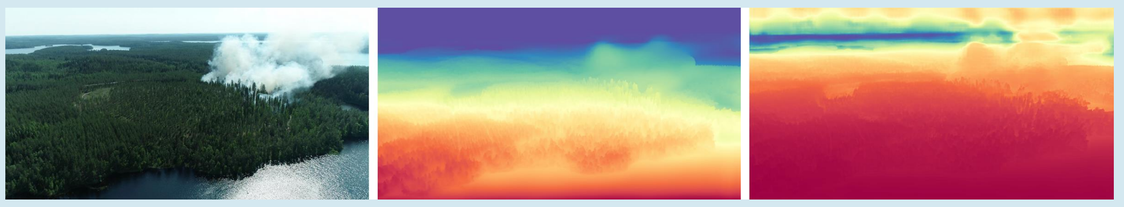

Figure 2. Relative and metric estimates from Depth anything v2 [3]. The metric depth estimate was 3-5 km and required setting the max depth manually.

- End-to-end learning

- Predicting the target position directly from the camera image and pose requires the least software components
- For most tasks there is **no data**

- Diffusion generated simulation videos

- Solves the lack of data with simulations obeying the required physical constraints
- Generative models improve the visual diversity [4]

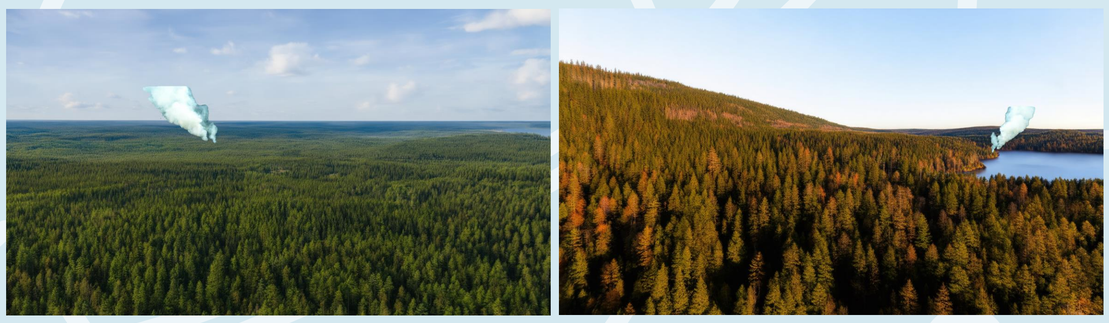

Figure 3. Fusion of Stable Diffusion 3.5 [5] generated backgrounds and real smoke that could be used as a basis for the simulated videos.

References

[1] Pesonen, Julius, et al. "Detecting Wildfires on UAVs with Real-time Segmentation Trained by Larger Teacher Models." *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025.

[2] Pesonen, Julius, et al. "Boreal Forest Fire: UAV-collected wildfire detection and smoke segmentation dataset." *Scientific Data* 12.1 (2025): 1419.

[3] Yang, Lihe, et al. "Depth anything v2." *Advances in Neural Information Processing Systems* 37 (2024): 21875-21911.

[4] Yu, Alan, et al. "Learning visual parkour from generated images." 8th Annual Conference on Robot Learning. 2024.

[5] Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." *Forty-first international conference on machine learning*. 2024.

NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

Aalto University