Salient Object Detection Using UAVs

Mo Shan*, Feng Lin*, and Ben M. Chen§
*Temasek Laboratories, National University of Singapore
§Department of Electrical and Computer Engineering, National University of Singapore

Abstract

A salient object detection approach is proposed in this paper, tailored to the aerial images collected by Unmanned Aerial Vehicles (UAVs). In particular, the aerial images are classified. The selected image is segmented using superpixels, and then a weak saliency map is constructed based on image priors. Positive and negative samples are selected in accordance with the weak saliency map for training a boosted classifier. Consequently, the classifier is used to produce a strong saliency map. The weak and strong saliency maps are integrated to locate the candidate objects, and false alarms are pruned after post-processing. Experiments on aerial images collected above meadow and roof demonstrate the effectiveness of the proposed approach.

Keywords: Saliency detection, UAVs

1 Introduction

UAVs have already been widely employed for a range of tasks, including inspection and surveillance. Furthermore, the advent of visual inertial odometry presented recently in [1, 2], which utilizes on-board camera and IMU in a tightly coupled manner, reduces their dependence on GPS dramatically and enables the UAVs to operate in clustered environment. For instance, researchers in the SFLY project have demonstrated that the UAVs are capable of performing localization and mapping in GPS-denied environment [3].

This paper studies salient object detection in aerial images. Specifically, illegal dumping detection is selected as a case-study. It refers to the unauthorized waste disposal of garbage, appliances or furniture upon public or private properties. A patrol team is usually required to deter such offensive activities. Meanwhile, the automatic illegal dumping detection from the images captured by UAVs is very critical to lower the burden of manual labor. Since the amount of images obtained from the on-board camera is significant, it is troublesome to scan through every image. Consequently, detection algorithms are necessary to select those images that may contain dumped objects. Normally, the waste is left at the places where it stands out from the environment, such as a plastic bag on the meadow, and this makes the dumped object to be salient. Therefore salient object detection becomes handy for detecting illegal dumping automatically. This problem may seem trivial at first glance, because the dumped waste is usually quite distinguishable from its environment. However, other man-made objects may exist in the scene as well, such as basketball court on a meadow, or windows on the roof. The existence of these outliers makes it more challenging to detect the garbage.

Saliency detection could be divided into three categories: bottom-up, top-down, and hybrid approaches [4]. The first category includes [5, 6, 7, 8, 9], and considers primitive information, such as the intensity contrast, color histogram, and global distribution of colors, whereas the second category contains [10, 11], and concerns about the application of prior knowledge for the specific task at hand. An example of bottom-up saliency detection is to detect a red ball against a while background, while top-down approach is used when the cyclists are searching for the bicycle lane [12].

In this paper, a task driven salient object detection algorithm is proposed. The images are pre-processed by scene classification. For the images that are likely to contain debris, a weak saliency map is generated from the input image based on prior knowledge, and then both positive and negative training samples are selected to produce a strong saliency map. Those two saliency maps are combined and thresholded to identify salient objects. Post-processing is carried out to prune outliers. The proposed approach is implemented in MATLAB based on the open-source code of [13] [†]. Aerial images collected by our UAV are used for the experiments.

The remainder of the paper is organized as follows: Section II presents a brief literature review; the methodology is described in Section III; Section IV contains the experiments for salient object detection; Section V consists of the conclusion.

2 Related works

2.1 Bottom-up approaches

One of the early bottom-up models is presented in [5]. It is based on a biologically-plausible architecture related to the feature integration theory. The primitive visual features considered include intensity contrast, color contrast, and local orientation, all of which are computed at multiple scales. To integrate the feature maps with different modalities, a normalization operator is designed to promote the unique maps, mimicking the cortical lateral inhibition mechanisms. The feature maps are combined into conspicuity maps, which are normalized and summed to produce the saliency map. While

[†] http://202.118.75.4/lu/publications.html

this framework could be tailored to different tasks via dedicated feature maps, it may not be able to detect objects salient for feature types not implemented.

Another biologically-plausible bottom-up approach is proposed in [6]. The structure of graph algorithms is exploited to compute saliency efficiently. Specifically, Markov chains are defined for the images, and the equilibrium distribution is used as activation. Because each node is independent, this process can be computed in a parallel way. In the normalization phase, a graph is constructed to concentrate activation into key locations. This approach outperforms [5] by 14% on human fixation prediction, and it could be extended to multiresolutions for improved performance. Nevertheless, this work favors the central-bias and thus may not be suitable to detect salient objects in the image periphery.

The recent work [9] relies on a convolutional neural network (CNN) trained from ImageNet dataset to perform feature extraction. The features are extracted from three rectangular windows enclosing the target region, its surrounding regions, and the whole image, in order to evaluate the visual contrast. Fully connected layers are trained from these multiscale CNN features to infer the saliency score. The saliency maps are refined to maintain spatial coherence and fused to obtain an aggregated map. The drawbacks of this approach include its requirement of training datasets and long processing time if GPU is not used.

2.2 *Top-down approaches*

As for the top-down approach, [10] focuses on salient object detection by incorporating the high level concept. The problem is modeled by a condition random field (CRF) to combine multiscale contrast, center-surround histogram, and color spatial distribution as local, regional as well as global salient features. Temporal cues are also exploited for dealing with sequential images. Being only able to detect a single salient object is one of the remaining issues for this approach.

An alternative top-down model is described in [11]. The image is decomposed into multiscale segmentations. A random forest regressor is learnt to map the regional descriptors representing the contrast, property and backgroundness to a saliency score. An aggregated saliency map is obtained from fusing the saliency maps of different segmentation levels. The key differences of this approach is that it computes a contrast vector instead of a value, and it combines the features to generate the integrated saliency map, other than combining the saliency maps generated by varied features. This method requires the collection of training samples with groundtruth labels as well.

2.3 *Hybrid approach*

In addition to the two categories mentioned above, there are hybrid approaches as well. In [13], a novel salient object detection approach is proposed. Image priors and multiple features are exploited to generate positive and negative training samples for bootstrap learning. Since the training samples are selected using the bottom-up model, the off-line training and groundtruth labeling are alleviated. This is critical as there is a lack of datasets for illegal dumping detection. Our work is similar to [13] with the following contributions:

1. Saliency map is used to detect illegal dumping using aerial images captured by a UAV.
2. A simple yet efficient pre-processing algorithm is used for scene classification.
3. The color and size priors are used to take into account the features of the scene and the objects to be detected.
4. Steps for post-processing the saliency map to prune the outliers are proposed.

3 Methodology

Referring to the overview in Fig.1, the proposed approach consists of scene classification as pre-processing, generating a saliency map, which is thresholded to identify the candidates for the dumped waste, and post-processing to prune the false alarms.

3.1 *Pre-processing*

UAVs are often needed to survey a building to detect illegal dumping, and it is common that the building is surrounded by meadow and tree. Since the debris is most likely to exist on the meadow and roof, the aerial images are classified into three categories, namely meadow, roof, and tree images. The salient object detection is only carried out in the meadow and roof images. In this way, not only the number of false alarms is reduced, but also the computational time is saved by discarding the tree images.

In scene classification, green regions are detected first, and then blob detection of these regions is performed to count the cluster number. If there is few clusters, then the scene is probably the meadow. In contrast, the existence of more clusters indicate that the green regions are not connected, and hence the image is likely a tree image. If no green region is found, the probability of the image depicting the roof is high.

The green region detection is based on [14]. The input RGB image is transformed to grayscale image by

$$I = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \qquad (1)$$

where R, G, B are the red, green, and blue channels of the input, I is the grayscale image. Then I is subtracted from G to obtain the green part of the image, followed by applying the median filter to suppress the noise. The resulting image is converted to a binary one by using the threshold of $thresh_{gr}$. Afterwards, blob detection is performed in the binary image, and the green regions whose area is smaller than $thresh_{ga}$ are discarded as noise. Suppose the number of the remaining blobs is k, then the scene is classified as meadow if $0 < k \leq thresh_k$, or tree if $k > thresh_k$. If $k = 0$, then the scene depicts roof. In the experiments, the parameters are set to $thresh_{gr} = 0.07$, $thresh_{ga} = 1000$, and $thresh_k = 3$.

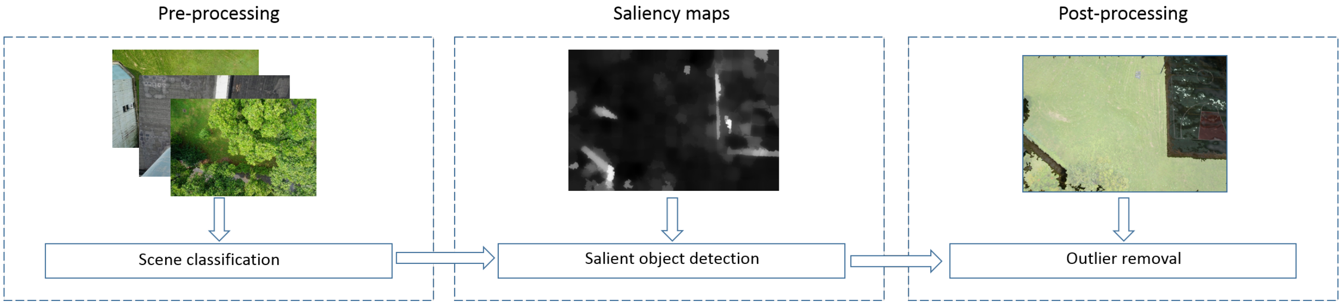

Figure 1: Overview of the proposed approach for salient object detection.

3.2 *Saliency detection*

A weak saliency map is constructed using color and size priors. As for the color prior, two cases are tested: meadow and roof, since it is common to find debris on the meadow, and it is also important to detect debris on the roof using UAVs because it is difficult for people due to out of view.

The color prior of a pixel p for the meadow is defined as

$$S_c(p) = \sqrt{(h(p) - H_m)^2 + (s(p) - S_m)^2 + (v(p) - V_m)^2} \tag{2}$$

where h, s, v are the Hue, Saturation and Value channels of the RGB image. H_m, S_m, V_m are the typical values for the meadow in the HSV colorspace, and $H_m = 0.2$, $S_m = 1$, $V_m = 0.5$. This color prior essentially computes the distance of the pixel to the green color. In this way, non-green objects are assigned with higher weights, and thus they are more likely to be chosen as positive training samples.

Similarly, the color prior for the roof is $S_c(p) = s(p)$, which is the value of the Saturation channel. Based on the observation that the Saturation for roof is low, the higher the Saturation, the more likely that it is not part of the roof, indicating the pixel belongs to objects such as debris.

The superpixels proposed in [15] are computed in the image to exploit the size prior, defined as $S_s(p) = 1 - A$, where A is the normalized region of the superpixel that pixel p belongs to. The size prior penalizes large area, because these regions often correspond to buildings, and are unlikely to be dumped objects.

Besides color and size priors, another useful criterion for determining salient objects is the difference from the image border, assuming that the majority of the border region contains only background. To compute the distance of each superpixel to those close to the boundary, the RGB, CIELab, and Local Binary Pattern (LBP) features are used. Therefore, the per-pixel saliency could be obtained from

$$S_{weak}(p) = S_c(p) \times S_s(p) \times d(p) \tag{3}$$

where $S_c(p), S_s(p), d(p)$ are the color prior, size prior, and the superpixel difference to the image border respectively. This weak saliency map is smoothed by Graph Cut.

Next, the weak saliency map is used to generate the training samples for the strong saliency map. The superpixels whose average saliency values are below the lower threshold is selected as negative training sample, whereas those with average saliency values exceeding the higher threshold are chosen as positive training samples.

Since how to integrate different features for training is unclear, Multiple Kernel Boosting (MKB) presented in [16] is employed. Several weak classifiers, which are Support Vector Machines (SVMs) with linear, polynomial, RBF, and sigmoid kernels, are combined to form a strong classifier. The details of the boosting process is described in [13]. The aggregated classifier is then used to generate a pixel-wise strong saliency map S_{strong}. The map is smoothed by Graph Cut and guided filter.

To address the multiscale issue, superpixels with four different granularities are generated, producing four pairs of weak and strong saliency maps. Averaging these maps gives the multiresolution weak and strong saliency maps.

To summarize, a weak saliency map is generated based on priors, and then a strong saliency map is constructed using the training samples. The former detects fine details due to its bottom-up model, while the latter focuses on global shapes. As the two maps are complementary, they are integrated to produce the final saliency map as

$$S_{aggregated} = \frac{S_{weak} + S_{strong}}{2} \tag{4}$$

The final saliency map $S_{aggregated}$ is thresholded by $thresh_{sa}$ to identify objects that possess a high saliency value.

3.3 *Post-processing*

The post-processing algorithm is described in Algorithm 1. The first step is to detect homogenous regions. This is necessary to exclude the false alarms existed on non-homogenous regions, such as the basketball court besides the meadow, or the sidewall on the roof.

The meadow is considered first. Green region detection as described in the pre-processing step is conducted. For the roof, the image is converted from RGB colorspace

Algorithm 1 Post-processing algorithm

```
Detect homogenous region
if The scene is meadow then
    Detect green region
    $thresh_{gr} \leftarrow 0.05$
end if
if The scene is roof then
    Threshold the HSV image
    $thresh_{sl} \leftarrow 0, thresh_{su} \leftarrow 0.1$
end if
Retain the salient objects on the homogenous region
for Each salient object candidate do
    if Rectangularity $< thresh_{rt}$ then
        Remove outlier
    end if
    if (Area $< thresh_{al}$ or Area $> thresh_{au}$) then
        Remove outlier
    end if
    if (ECD $< thresh_{dl}$ or ECD $> thresh_{du}$) then
        Remove outlier
    end if
end for
```

to HSV colorspace, and thresholded using pre-defined lower and upper thresholds for the Saturation channel, where $thresh_{sl} = 0$, $thresh_{su} = 0.1$, since the roof part has very low saturation. The holes in these binary maps are filled to produce masks for homogeneous regions.

Even though the salient object could be detected on the homogenous regions, there may still exist some outliers such as the boundaries of the meadow or the building. To remove these outliers, the rectangularity of the detected object could be taken into account, since the garbage usually consists of boxes or appliances that are rectangular. The rectangularity could be computed as the ratio of pixels belonging to the object to the total pixels in its bounding box. The objects whose rectangularity is lower than the threshold $thresh_{rt}$ are discarded.

Besides rectangularity, the size is also taken into account to prune the objects that are either too small, such as the shadow, or too large, such as the entire wall. The parameters for size include the area and the equivalent circular diameter (ECD), computed by $\sqrt{4\pi \times area}$, which specifies the diameter of a circle with the same area as the detected object. Their lower and upper thresholds are $thresh_{al}$, $thresh_{au}$, and $thresh_{dl}$, $thresh_{du}$ respectively.

4 Experiment

4.1 Image collection

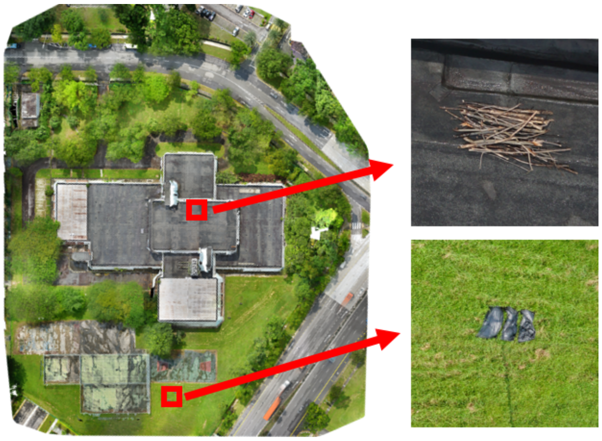

Figure 2: Stitched map of the test site, generated by Pix4Dmapper. The locations of plastic bag and tree branches are marked in red.

The images used in the experiments are captured when the UAV surveys a building surrounded by a meadow, as shown in Fig. 2 [‡]. Black plastic bags and tree branches are placed on the meadow and the roof to simulate illegal dumping. The UAV operates at about 35m, and the height of the building is 18m. In other words, the height in the aerial images for garbage detection on meadow is about 35m, while the height for garbage detection on roof is about 17m.

The camera carried on-board is Sony A6000, with a focal length of 16mm. The original resolution of the image is 6000 × 4000, and to reduce the computational time, it is downsampled to 640 × 427.

4.2 Scene classification

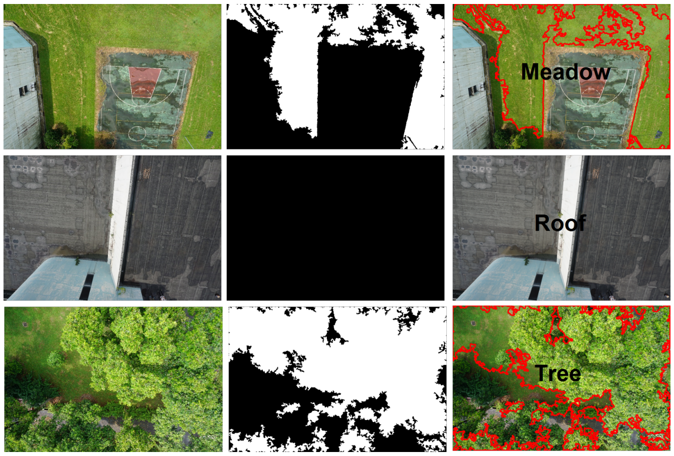

Figure 3: Scene classification results. From left to right: original image, binary image after green region detection, classification result with blobs marked in red.

The images are classified into different categories prior to salient object detection. From Fig. 3, it is evident that the

[‡] These images used for the experiments are available at https://github.com/shanmo/IMAV2016-Dataset

green regions can be effectively detected in the images collect by the UAV. Moreover, the number of blobs in meadow image is smaller than that in the tree image, where the green regions tend to be discontinuous. Furthermore, there is no green region in the roof image. With the help of the scene classification, illegal dumping detection will only be performed in the meadow and roof images to save computational time.

4.3 Garbage detection on the meadow

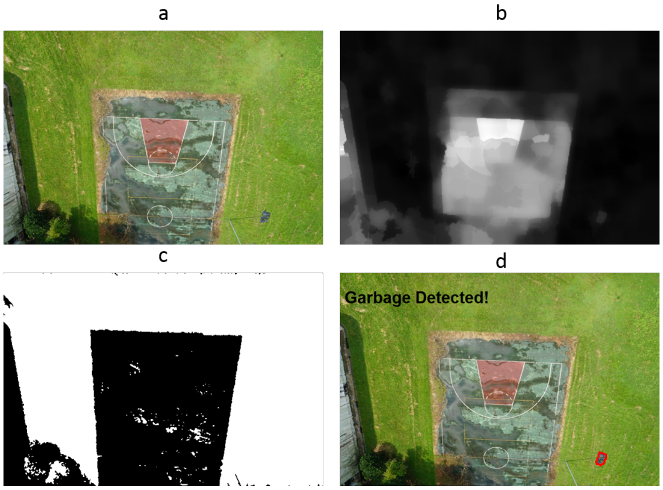

Figure 4: Garbage detection result on the meadow. a: original image. b: saliency map. c: mask image of the meadow region. d: result of detected garbage marked in red. Best viewed in color.

The experimental results for the illegal dumping detection on a meadow will be presented in this section, where some black plastic bags are placed on the meadow as the dumped waste. The rectangularity threshold $thresh_{rt}$ is set to 0.5. The size thresholds are $thresh_{al} = 30$, $thresh_{au} = 600$, and $thresh_{dl} = 20$, $thresh_{du} = 40$ respectively. The saliency threshold is $thresh_{sa} = 0.1$.

As shown in Fig. 4, there is a plastic bag on the meadow, which could be clearly observed in the saliency map. Moreover, the mask image of the meadow effectively covers the entire meadow region, such that the salient objects on the basketball court can be removed as outliers. After postprocessing, the boundaries of the meadow could be further removed, and the resulting saliency map only contains the plastic bag.

For comparison, using the same original image from Fig. 4, the saliency map from other methods are displayed in Fig. 5. It shows the saliency map generated from the approaches proposed in [6, 9] and the original version of [13] respectively. Since the dumped waste is not in the center, the methods that emphasize central-bias, for instance [6] and [13], may not work well.

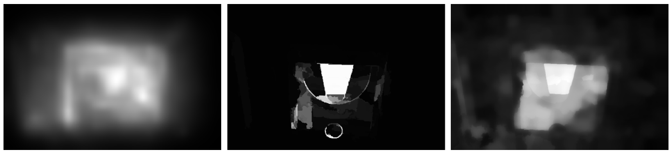

Figure 5: Comparison of saliency maps. From left to right: saliency maps generated by [6], [9], [13].

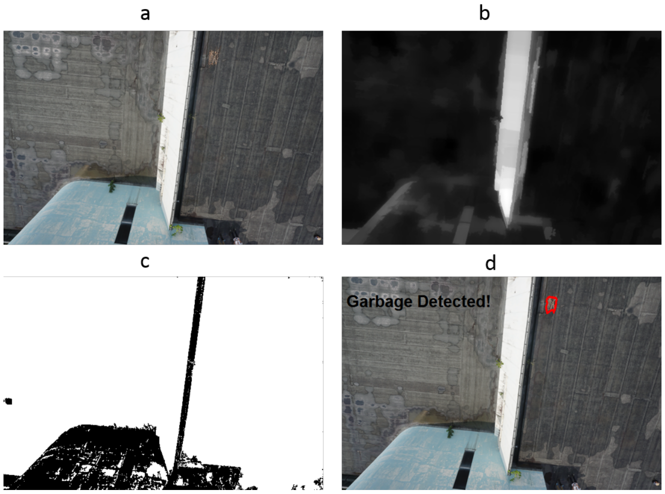

Figure 6: Garbage detection result on the roof. a: original image. b: saliency map. c: mask image of the roof region. d: result of detected garbage marked in red. Best viewed in color.

4.4 Garbage detection on the roof

This section presents the results of the proposed approach for garbage detection on roof, and the dumped waste are tree branches. The rectangularity threshold $thresh_{rt}$ is set to 0.6, which is higher than that of the previous experiment, because the water pond on the roof whose boundaries have irregular shapes produce many outliers. The size thresholds are $thresh_{al} = 300$, $thresh_{au} = 600$, and $thresh_{dl} = 20$, $thresh_{du} = 30$. The saliency threshold is also set to $thresh_{sa} = 0.1$ as used in the meadow case.

It could be observed in Fig. 6 that although the tree branches are not as salient as the plastic bag shown in Fig. 4, it could still be detected.

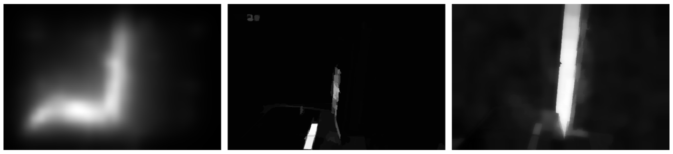

Figure 7: Comparison of saliency maps. From left to right: saliency maps generated by [6], [9], [13].

To compare the results of different saliency detection algorithms, the same image from Fig. 6 is used, and the saliency maps are shown in Fig. 7. Similar to what is observed in Fig. 5, the algorithms assuming that the object is close to the image center fail to detect the tree branches.

4.5 Discussion

Some may contend that instead of using saliency, HSV thresholds could be applied to the meadow or roof regions to detect the plastic bags or tree branches, as they possess unique colors such as black or brown. Admittedly, color thresholding may work well as the meadow or roof regions are quite homogeneous. However, the type of garbage is not limited to only two kinds, and thus saliency is preferable over color because of its capabilities to detect all kinds of dumped waste.

Though the proposed approach is extensible to detect varied kinds of garbage, there are also certain limitations associated with its extensibility. For instance, sometimes the lighting pole on the meadow is detected as dumped waste, because it stands out from the scene.

Another drawback of our work is its limited dataset size, which contains few images for each category, and thus the number of images that have debris in the scene is very small. Hence it is difficult to evaluate the proposed approach quantitively. To obtain statistically meaningful results, a larger dataset is needed. In addition, the small dataset also hinders the usage of CNNs for scene classification and garbage detection as the training samples are quite limited. One possible way is to apply transfer learning to resolve this issue.

5 Conclusion

In this paper, a new saliency detection algorithm is proposed, which consists of pre-processing, saliency detection and post-processing. During the first stage, the image is classified based on the number of green blobs. For the second stage, color and size priors are designed to obtain a weak saliency map. This map is used to generate the training samples for a boosted classifier. The classifier then produces a strong saliency map, which is fused with the weak saliency map. In the third stage, the outliers that are located on the non-homogenous regions, or inconsistent with the size requirements are pruned. Experimental results show that the proposed approach could perform illegal dumping detection on aerial images captured above the meadow and roof. Consequently, the proposed approach could be exported to other salient object detection applications as well.

Acknowledgment

The authors would like to thank the members of NUS UAV Research Group for their kind support, and Prof Lu Huchuan for the generous sharing of their code.

References

[1] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 15–22. IEEE, 2014.

[2] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

[3] Davide Scaramuzza, Michael C Achtelik, Lefteris Doitsidis, Felice Friedrich, Elias Kosmatopoulos, Alessio Martinelli, Markus W Achtelik, Maria Chli, Savvas A Chatzichristofis, Laurent Kneip, et al. Vision-controlled micro flying robots: from system design to autonomous navigation and mapping in gps-denied environments. *Robotics & Automation Magazine, IEEE*, 21(3):26–40, 2014.

[4] Kate Duncan and Santonu Sarkar. Saliency in images and video: a brief survey. *Computer Vision, IET*, 6(6):514–523, 2012.

[5] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[6] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[7] Ming Cheng, Niloy J Mitra, Xumin Huang, Philip HS Torr, and Song Hu. Global contrast based salient region detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(3):569–582, 2015.

[8] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[9] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5455–5463, 2015.

[10] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):353–367, 2011.

[11] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[12] Simone Frintrop, Erich Rome, and Henrik I Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6, 2010.

[13] Na Tong, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Salient object detection via bootstrap learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1884–1892, 2015.

[14] Arindam Bose. How to detect and track red, green and blue objects in live video, 2013–2014.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[16] Fan Yang, Huchuan Lu, and Yen-Wei Chen. Human tracking by multiple kernel boosting with locality affinity constraints. In *Computer Vision–ACCV 2010*, pages 39–50. Springer, 2010.