COLLABVP: STRUCTURE-ENHANCED VP DETECTOR
Abdelkarim Elassam, Gilles Simon, Marie-Odile Berger

To cite this version:
Abdelkarim Elassam, Gilles Simon, Marie-Odile Berger. COLLABVP: STRUCTURE-ENHANCED VP DETECTOR. IEEE International Conference on Image Processing (ICIP 2023), IEEE, Oct 2023, kuala Lumpur, Malaysia. hal-04192288

HAL Id: hal-04192288
https://hal.science/hal-04192288
Submitted on 31 Aug 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
COLLABVP: STRUCTURE-ENHANCED VP DETECTOR

Abdelkarim Elassam  Gilles Simon  Marie-Odile Berger
INRIA, Université de Lorraine, LORIA, FR-54600

ABSTRACT

We introduce in this paper a novel structure-enhanced VP detector, CollabVP, which uses a multi-task learning framework to estimate multiple horizontal vanishing points (VPs) from a single RGB image. CollabVP exploits contextual information and scene structures through masks of vertical structures to accurately estimate the horizon line and VP positions. The proposed approach is not limited to Manhattan worlds and can detect any number of VPs.

Index Terms— Vanishing points detection, Horizon line estimation, Scene structure segmentation, Multi-task learning

1. INTRODUCTION AND RELATED WORKS

With a pinhole camera, a vanishing point (VP) is a point in the image where parallel lines in 3D space appear to converge. Traditional VP detection methods start by extracting line segments from the image [1], then try to group them by using an algorithm such as RANSAC [2], J-linkage [3], the Hough transform [4] or more recently CONSAC, a learning-based robust estimator [5]. Recent papers [6, 7] dealing with man-made environments have shown that detecting first the horizon line, with a CNN approach in Zhai’s paper [6], allows constraining the search of horizontal VPs on this line and greatly improve the accuracy of the VPs detected with line segments. In order that any set of lines that meet accidentally on the HL can generate a false positive, a robust method based on the a-contrario methodology [8] was proposed by Simon in [7], which leads to far fewer false positive and duplicate VPs than the optimization method proposed in [6].

Following the trend of learning-based methods, several CNN-based approaches for VP detection have been proposed, often with constraints on the number of detected VPs or on the environment. In [9, 10], only a dominant VP is detected inside a grid map of the image. Other works [3, 11, 12, 13] operate under the assumption of a Manhattan World, assuming three mutually orthogonal VPs, among which is the zenith. For example, [13] introduced TLC, a transformer-based line segment classifier designed to classify a line segment into one of the three Manhattan directions. A less strict hypothesis is the assumption of an Atlanta world where the zenith and an unknown number of horizontal VPs are assumed. Several works [6, 14, 15], including ours, rely on this assumption.

Other methods operate on the bounded Gaussian sphere instead of the unbounded image plane as in Kluger et al. [16] where VP detection is formulated as a multi-label classification task on the sphere. Zhou et al. [14] presented NeurVPS, a CNN with geometry-inspired convolutional operators for detecting VPs. Using an image and a candidate point on the unit sphere as input, their network predicts the probability of the point being near a ground-truth (GT) VP. To exploit the geometric properties of VPs as the intersection of parallel lines, the authors developed an operator named conic convolution, which explicitly enforces feature extractions and aggregations along the structural lines. However, this approach relies heavily on the initial random sampling on the sphere. Recently, Liu et al. introduced VaPiD [15], a more efficient version of NeurVPS that uses learned optimizers and a computation-sharing scheme to process VP anchors efficiently. VaPiD [15] performs better than all previous state-of-the-art VP detection approaches. However, this network and NeurVPS have been only evaluated against precision but not against recall on the Holicity dataset [17], which prevents evaluating the number of VPs found among those expected.

This paper proposes Collab-VP, a horizon-first VP detection method that exploits contextual information and structural VPs through a multi-task CNN. As in [14, 15], we consider a CNN supervised by couples $<\text{image}, \text{VP}>$. We additionally exploit the fact that man-made environments contain structural VPs which correspond to the intersection of vertical planes—whose orientations can be roughly inferred with learning-based techniques—with the horizontal plane. We thus propose a flexible two-branch framework (see Fig. 1) that outputs the horizon line, the distribution of VPs along the HL and the masks of discretized orientations of vertical structures. We finally propose a robust method for VP extraction inspired from [7] that jointly processes these outputs.

2. METHOD

2.1. Network architecture

The two-branch CollabVP architecture is shown in Fig. 1. The DirectVP branch uses the ResNet50 architecture [18]
2.2. Loss functions

To train the DirectVP branch, we use the cross-entropy loss of the distributions of the HL and VP parameters: 

\[ L_{\text{parameter}} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i), \]

where \( y_i \) and \( \hat{y}_i \) are discretized in \( N = 100 \) bins for each parameter. Mask segmentation is also treated as a classification problem. For each pixel \( i \), the cross-entropy loss is calculated between the predicted and GT masks and the final loss can be written as 

\[ L_{\text{Mask}} = -\frac{1}{M} \sum_{i=1}^{M} y_i \log(\hat{y}_i), \]

where \( M \) is the number of pixels in the input image. The overall losses are defined as:

\[ L_{\text{DirectVP}} = L_{\text{Slope}} + L_{\text{Offset}} + L_{\text{VP}} \]

\[ L_{\text{CollabVP}} = L_{\text{DirectVP}} + L_{\text{Mask}} \}

2.3. VP detection and collaboration strategy

The HL is computed from the slope and offset corresponding to the best scores of the related output distributions. VPs are detected as local maxima of the third distribution provided by DirectVP (Fig. 1). The probabilistic a contrario (AC) framework [8] is used for this purpose, which has proven relevant to reliably detect modes of a histogram based on a parameter-free multi-scale approach. It is robust to irrelevant peaks and can detect peaks made of several consecutive bins. However, it can happen that VPs obtain low scores, judged as not meaningful by the AC method, mainly when they are associated with small or very low-textured structures. These scores are, in fact, judged not meaningful because other VPs obtain much higher ones. But isolated from the others, i.e. considered within an orientation interval provided by MaskVP, such VPs can be recovered.

Therefore, our collaboration strategy adds VPs in intervals found by MaskVP only if no VP was detected by the AC method in these intervals. In that case, the value with the highest score inside the mask is retained as VP (Fig. 1). This procedure is illustrated in Fig. 2. The input image is shown in Fig. 2a, with the three VPs finally detected in bold stars. The output scores of DirectVP are shown in Fig. 2b, with crosses showing the peaks detected by the AC method: two VPs were found. By contrast, MaskVP predicts three VP intervals (2c): two of them (blue and orange) contain a previously detected VP, but the third one (maroon) allows the detection of an additional VP whose score was too low to be detected by the AC method. This interval corresponds to the

---

Fig. 1. Overview of CollabVP (better seen when zooming).
small facade in the middle of the left building (2c). In comparison, Fig. 2d shows the VP distribution obtained with DirectVP trained alone (TA). It can be seen that the third VP is not detected, whereas a small peak is obtained when DirectVP is trained together with MaskVP (fig.2.b).

3. EXPERIMENTAL RESULTS

3.1. Datasets and evaluation criteria

To conduct our experiments, we used the recently published Holicity dataset [17] both for training and for testing, and we also evaluated our model on the YU dataset [19], which provides GT VPs but lacks normal maps. The Holicity Dataset consists of 54354 real-world images covering downtown London. It provides labels for different 3D structures in urban environments, such as surface normal maps and VPs. We adopted the split the authors proposed, containing 45032 training samples and 2504 validation samples. We applied several data augmentation techniques to increase the training dataset’s variety and quality. Firstly, we rotated the original images to different angles that follow a normal distribution with $\sigma = 10^\circ$ to simulate variations in the camera’s roll. We also utilized homography transformations to introduce changes in the camera’s pitch. Using these techniques, we generated more diverse training samples, resulting in better performance and improved generalization to new, unseen images. The York Urban dataset (YU) [19] consists of 102 indoor and outdoor scene images, each with provided camera parameters and GT VPs that satisfy the Manhattan world assumption.

Our method was evaluated using the mean and median angle errors. Also, the angle accuracy (AA) defined in [14] is used to quantify the quality of VP detections. We compute the angle difference between each predicted VP and the closest GT VP. The $AA^\theta$ value is defined as the area under the angle accuracy curve between $[0, \theta]$ divided by $\theta$. We calculate $AA^\theta$ using various precision levels $\theta$ as in [14, 15]. A second metric called AA-R (in reference to the recall term) is introduced to quantify the ratio of GT VPs detected with the different methods. We compute the angle difference between each GT VP and the closest predicted VP. The AA-R $\theta$ indicates the percentage of GT VPs detected at a certain precision level $\theta$ and shows the method’s ability to detect multiple VPs.

3.2. Results on Holicity

Our method was compared on Holicity to state-of-the-art line-based methods such as [7] and to recent learning-based methods such as NeurVPs [14] and VaPiD [15], which are the closest to our work since working on non-Manhattan scenes. Since the source code of VaPiD [15] is unavailable, we conducted a comparison on Holicity by referring to the tables provided in [15] and utilizing the same angular parameters (1, 2, and 10 degrees). The results presented in Tab. 1 show that CollabVP outperforms other baselines regarding the precision of detected VPs, as indicated by the AA and AA-R values. CollabVP achieves an $AA^1$ of 30.9% ($AA^2$ of 42.8% and $AA^{10}$ of 79.5%), surpassing the previous best performance of VaPiD by a relative improvement $(AA_{new} - AA_{old})/(1 - AA_{old})$ of 11.3% (14.5%, 16.6%). Although the precision of NeurVPs [14] and VaPiD [15] methods on Holicity is reported, they do not provide the number of detected GT VPs (their method detects a fixed number of VPs). In contrast, we report the AA-R metrics: in the test set, there are 4973 GT VPs, and CollabVP detects AA-R$^{10} = 86.1\%$ of them with a precision of less than $10^\circ$ (AA-R$^1 = 48.3\%$, AA-R$^2 = 65.1\%$). In comparison, Simon et al.’s method [7] only detects 80.5% of them with such precision (AA-R$^1 = 41.3\%$, AA-R$^2 = 56.6\%$). This demonstrates the advantage of extracting structures with similar normal vectors, as it enables the detection of a more comprehensive set of structural VPs. It is worth noting that although our method’s mean error is slightly higher than VaPiD’s, it detects additional points that are correct but not present in the ground truth. These additional points can significantly impact the mean error metric, as the angle difference between the predicted VP and the closest GT VP will be large, inflating this average.

3.3. Results on York Urban

We also compared CollabVP to several state-of-the-art methods, namely J-Linkage [3], Simon et al. [7], Li et al. [20], Wu et al. [21], Lu et al. [22], CONSAC [5], and NeurVPs
Table 1. Comparisons with baseline methods on HoliCity.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC - HL</th>
<th>AA3</th>
<th>AA5</th>
<th>AA10</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simon et al. [7]</td>
<td>72.9</td>
<td>22.0</td>
<td>35.0</td>
<td>62.1</td>
<td>19.1</td>
<td>1.91</td>
</tr>
<tr>
<td>NeurVPS [14]</td>
<td>-</td>
<td>18.2</td>
<td>31.7</td>
<td>62.1</td>
<td>8.32</td>
<td>1.78</td>
</tr>
<tr>
<td>VaPiD [15]</td>
<td>-</td>
<td>22.1</td>
<td>39.6</td>
<td>75.4</td>
<td>3.00</td>
<td>1.19</td>
</tr>
<tr>
<td>CollabVP</td>
<td>93.1</td>
<td>30.9</td>
<td>48.4</td>
<td>79.5</td>
<td>4.23</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 2. Comparisons with baseline methods on YU [19].

<table>
<thead>
<tr>
<th>Method</th>
<th>AA3</th>
<th>AA5</th>
<th>AA10</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-linkage [3]</td>
<td>40.2</td>
<td>50.5</td>
<td>64.1</td>
</tr>
<tr>
<td>Simon et al. [7]</td>
<td>40.1</td>
<td>58.2</td>
<td>77.5</td>
</tr>
<tr>
<td>Wu et al. [21]</td>
<td>44.3</td>
<td>61.4</td>
<td>77.4</td>
</tr>
<tr>
<td>Li et al. [20]</td>
<td>51.1</td>
<td>66.1</td>
<td>80.5</td>
</tr>
<tr>
<td>Lu et al. [22]</td>
<td>58.0</td>
<td>73.2</td>
<td>86.2</td>
</tr>
<tr>
<td>CONSAC [5]</td>
<td>62.1</td>
<td>73.7</td>
<td>84.1</td>
</tr>
<tr>
<td>NeurVPS [14]</td>
<td>39.9</td>
<td>50.3</td>
<td>65.0</td>
</tr>
<tr>
<td>Tong et al. [13]</td>
<td>65.5</td>
<td>77.1</td>
<td>87.4</td>
</tr>
<tr>
<td>CollabVP - trained on Holicity [17]</td>
<td>57.6</td>
<td>73.2</td>
<td>83.3</td>
</tr>
<tr>
<td>CollabVP - trained on NYU [23] + Holicity [17]</td>
<td>66.8</td>
<td>77.8</td>
<td>86.6</td>
</tr>
</tbody>
</table>

3.4. Origin of the detected vanishing points

Our experiments on the Holicity dataset reveal that out of the total detected VPs, 2573 were detected by both AC and masks, achieving a precision of 82.0% for 5 degrees. When relying on AC alone, 1039 VPs were detected, with an accuracy of 79.2% for 5 degrees. Similarly, when using masks alone, 1089 VPs were detected with an accuracy of 72.2% for 5 degrees. These findings emphasize that points detected using only masks are meaningful, even if their score in the

3.5. A sample result

Finally, Fig. 3a shows sample VPs obtained in an indoor scene, compared with CONSAC [5] (black crosses +) and Simon et al.’s [7] (black cross x) results. CONSAC [5] and CollabVP outperform the line-based method of Simon et al. [7] regarding prediction accuracy. Simon et al.’s performance is affected by brightness and shadows, which are not handled well by their method despite the scenes containing long lines. Notably, two GT VPs were identified, but Simon et al. [7] only detected one, while CONSAC [5] detected both. In contrast, CollabVP detected all VPs with better precision and accurately segmented the associated vertical structures (Fig. 3b).

4. CONCLUSION

CollabVP accurately estimates the HL and VP positions by leveraging contextual information and scene structures, outperforming traditional line-based and learning-based methods. Our approach does not impose constraints on the number or orthogonality of VPs, yet outperforms methods that do. This provides a more robust and flexible solution for VP detection in various environments. Although the generated masks are by-products, they could benefit robotics tasks such as semantic SLAM. We plan to explore this direction in our future work.
5. REFERENCES


