SDC – Stacked Dilated Convolutions

Extended Abstract

René Schuster
Oliver Wasenmüller
Didier Stricker
DFKI - German Research Center for Artificial Intelligence
firstname.lastname@dfki.de

Christian Unger
BMW Group
christian.unger@bmw.de

Figure 1: Our SDC feature descriptor improves dense pixel-wise matching. From left to right: Disparity map for ELAS [6] on ETH3D [16], optical flow for CPM [10] on Sintel [4], and scene flow (disparity and optical flow components) for SFF [18] on KITTI [14]. Results are shown after consistency check.

ABSTRACT

Dense matching is a fundamental problem in many tasks and applications of Computer Vision. Of utmost importance for robust matching algorithms is a powerful representation of image points. With SDC (Stacked Dilated Convolution), we have presented a universal design element that was successfully used in a deep neural network for dense feature description of images. Using these descriptors, we could improve matching in wide variety of problems and domains.

1 INTRODUCTION

Advanced Driver Assistance Systems (ADAS) or (partially) autonomous systems require accurate and reliable perception of the environment. Two important examples of perception are geometric reconstruction of the surroundings and the prediction of motion.

Scene flow is the joint problem of 3D geometry and 3D motion estimation based on a stereo camera system. The underlying problem is dense (pixel-wise) matching across multiple (at least four) images. As such, scene flow is subjected to all challenges of matching, like image noise, changes in illumination, occlusions, fast motions, and so on. Next to the matching algorithm itself, the performance of scene flow algorithms is defined by the capabilities of the used feature representation to describe single image points. Famous among top-performing algorithms are CENSUS [23], SIFT [13], SIFTFlow [12], or dedicated learned features within an end-to-end deep network [5, 7, 21] to name a few. The list of problems with these descriptors is long. Some lack the ability to generalize, some are not applicable for dense description (only for sparse key points), others are specialized on a single matching problem or domain, most of them reduce the spatial resolution, many have too small receptive fields to cope with difficult image situations.

In this extended abstract we re-present SDC (Stacked Dilated Convolution) [20], a design element and deep neural network that handles all of the mentioned problems which leads to a significant improvement for dense matching across different algorithms and domains (cf. Figure 1).
2 STACKED DILATED CONVOLUTIONS

Motivation. For dense matching, spatially variant features are required that differ sufficiently even across neighboring pixels. Therefore, any form of sub-sampling (pooling, strided convolution, and others) should be avoided to make feature representations not overly smooth. As a consequence, our network operates at full resolution, i.e. the stride for all layers is always 1. Also importantly to note is that by this choice, a dense feature map can easily be predicted with a single forward pass for the entire image. Otherwise if a network contains strided layers, overlapping image patches need to be extracted to produce a dense feature map, or even more involved techniques need to be used [1].

The second important requirement for a robust feature descriptor is a large receptive field. Many image regions suffer from low local entropy (due to lighting, over- or under-exposure, texture-less regions, repetitive patterns, and so on) or are in general difficult or impossible to match (e.g. at occlusions) which makes it difficult to describe them in a recognizable way. Due to this, we argue that information of a large context needs to be considered for the description of single pixels.

Lastly, a universal feature descriptor that is applicable to many diverse domains and matching problems is desirable. This goal is obtained by collecting training data across many different data sets.

Method. The main challenge of our architecture is to obtain a very large receptive field, maintain full resolution, and at the same time keep the network size reasonable. To achieve all this, we did propose SDC [20]. Dilated convolution [22] is an effective strategy to increase the receptive field without increasing the kernel size or the number of parameters. Typically, dilated convolutions are applied in sequence. In contrast, we apply dilated convolution in parallel. Since the dilation rates correspond to sub-scales, with SDC we can make sure that every layer operates (at least partially) on the original scale. Yet, SDC produces a multi-scale filter response.

Table 1: Relative reduction of outliers when using SDC [20] as feature descriptor for different matching algorithms on different data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Stereo</th>
<th>Optical Flow</th>
<th>Scene Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELAS [6]</td>
<td>39.7 %</td>
<td>13.7 %</td>
<td>19.5 %</td>
</tr>
<tr>
<td>SGM [8]</td>
<td>39.8 %</td>
<td>3.6 %</td>
<td>-</td>
</tr>
<tr>
<td>CPM [10]</td>
<td>-</td>
<td>-</td>
<td>-11.3 %</td>
</tr>
<tr>
<td>FF++ [17]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SFF [18]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

One example for such a SDC layer is given in Figure 2. It is important to note that this design is not limited to specific hyper-parameters (parallel convolutions, dilation rates, kernel sizes, etc.). The complete SDC network consists of five such layers in a sequence which all process the input at full resolution. This way, a dense feature map for arbitrarily sized images can be computed in a single forward pass of the network.

For training, we follow a triplet training strategy [9] where a reference patch along with its correct correspondence and a negative match is fed through the network. The loss objective is to reduce the feature distance (in Euclidean space) for the matching pair below the feature distance of the negative correspondence. The thresholded hinge-embedding loss [2] is adopted for this.

Results. Three experiments were conducted to verify the superior performance of SDC features in the original submission. The first two did compare SDC to heuristic descriptors and state-of-the-art descriptor networks in terms of accuracy and robustness. SDC could outperform all methods within the comparison. The third experiment did test SDC in actual matching algorithms. Six data sets for different matching problems (stereo disparity, optical flow, and scene flow) were tested with five different matching algorithms. A summary of the results is given in Table 1. Here, we list the improvement of each algorithm on each data set when replacing the original feature descriptor with SDC features. The improvement is given in relative reduction of outliers (according to the KITTI outlier metric [14]). For all but one combination, SDC brought an improvement. In many cases, the improvement was significant, cutting the rate of outliers by half. For more details, we refer the reader to the original submission [20], the original supplementary material, and the follow-up study on SDC [19]. These documents provide a lot more detailed information and additional experiments to validate the design and performance of SDC.

3 CONCLUSION

The re-presented concept of SDC is a straightforward way to obtain a large receptive field, keep the network size small, and allow the network to operate at full image resolution. The requirements of these properties are motivated by the challenges of general dense matching problems. Extensive experiments did show the outstanding performance of SDC in comparison to state-of-the-art and in practice when applied in matching algorithms. Lastly, we would like to note that the ideas of SDC might be of high interest in other dense prediction problems, which is yet to be shown.
REFERENCES


