Supplementary material for: Deformable Spatial Pyramid Matching for Fast Dense Correspondences

Jaechul Kim[1] Ce Liu[2] Fei Sha[3] Kristen Grauman[1]
Univ. of Texas at Austin[1] Microsoft Research New England[2] Univ. of Southern California[3]
{jaechul,grauman}@cs.utexas.edu celiu@microsoft.com feisha@usc.edu

In this document, we provide supplementary information on (1) mapping data term values across different scales in multi-scale matching, (2) the definition of the localization error metric, and (3) analysis of matching accuracy in terms of pyramid levels. We addressed (1) and (2) in Sec 3.3 and Sec 4 of the main paper, respectively. The analysis in (3) offers further insights on our pyramid model, but does not fit in the main text.

Mapping data terms across different scales: In Sec. 3.3 of the main paper, we defined a mapping between data terms across different scales: the data term $D_i(\mathbf{t}_i, \mathbf{s}_i)$ at scale $\mathbf{s}_i$ maps to $D_i((\mathbf{s}_i - 1)\mathbf{q} + \mathbf{s}_i\mathbf{t}_i, \mathbf{s}_i = 1.0)$ of the reference scale. We derive it as follows:

The data term $D_i(\mathbf{t}_i, \mathbf{s}_i)$ computes a descriptor distance between $d_1(\mathbf{q})$ at a point $\mathbf{q}$ of the first image and $d_2(\mathbf{s}_i(\mathbf{q} + \mathbf{t}_i))$ in the second image (see Eq. 4 in the main paper). Here, the corresponding location of descriptor d_2 for a descriptor d_1 is determined by a translation $\mathbf{t}_i$ followed by a scaling $\mathbf{s}_i$ on the point $\mathbf{q}$.

However, if we suppose those two corresponding locations are associated by a common reference scale ($\mathbf{s}_i = 1.0$), their translation can be represented by a simple coordinate difference between them: $\mathbf{s}_i(\mathbf{q} + \mathbf{t}_i) - \mathbf{q} = (\mathbf{s}_i - 1)\mathbf{q} + \mathbf{s}_i\mathbf{t}_i$. That is, a translation $\mathbf{t}_i$ at a scale $\mathbf{s}_i$ is equivalent to the translation $(\mathbf{s}_i - 1)\mathbf{q} + \mathbf{s}_i\mathbf{t}_i$ at the reference scale. As a result, once we have computed the data term at the reference scale, we can map it to other scales without repeating the computation per scale.

Localization error (LOC-ERR) metric: To define the localization error of corresponding pixel positions (Sec. 4 Evaluation Metrics of the main paper), we first designate each image's pixel coordinate using its ground-truth object bounding box: the pixel coordinate of an object in each image is set such that the top-left of the bounding box becomes the origin and x-and y-coordinate are normalized by width and height of the box respectively. Then, we define the localization error of two matched pixels as: $e = 0.5(|x_1 - x_2| + |y_1 - y_2|)$, where (x_1, y_1) is the pixel coordinate of the first image and (x_2, y_2) is its corresponding location in the second image. We apply this metric to Caltech-101 dataset as it provides bounding box annotations for the foreground objects. Note that LOC-ERR metric is evaluated for the foreground pixels only, as we define bounding box coordinates only for the pixels inside the box.

Pyramid levels	LT-ACC	IOU
level 1	**0.745**	0.442
level 1 + 2	0.745	0.462
level 1 + 2 + 3	0.736	0.477
level 1 + 2 + 3 + pixels	0.732	**0.482**

Table 1. Matching accuracy in Caltech 101 in terms of the number of pyramid levels. The first three results come from grid-layer pyramid, in which pyramid level increases from 1 to 2 to 3, respectively. The last row denotes the result of our original implementation in the main paper, adding a pixel-level layer on top of three levels of grid-layer pyramid.

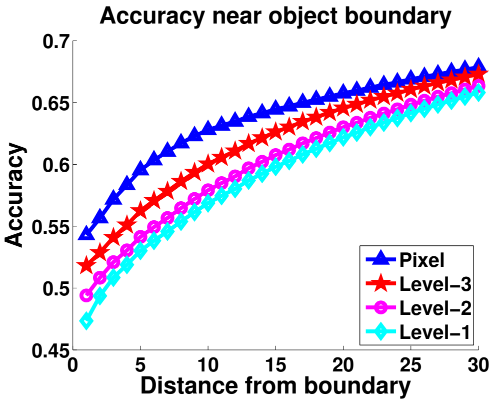

Figure 1. Matching accuracy near the object boundary. We evaluate matching accuracy among different pyramid levels as a function of pixel distance from the object boundary (up to 30 pixels). A pyramid with finer spatial nodes (e.g., Pixel) achieves better accuracy for the pixels near the object boundaries.

Levels of pyramid: In this section, we show how various spatial supports from our pyramid model achieve a balance between robustness to image variations and accurate localization with fine detail. To this end, we compare matching accuracy from different spatial extents by varying the number of pyramid levels.

Table 1 summarizes the results. Each row in the table adds another finer level to the pyramid. The accuracy is then evaluated using the matching given at the finest level in that pyramid, as we did in the main paper. We see that larger spatial nodes from lower pyramid levels provide better LT-ACC, whereas smaller nodes from higher levels offer better IOU. Given that LT-ACC takes all the pixels for evaluation whereas IOU accounts for foreground pixels only, our results point out (1) larger spatial nodes regularize the matching ambiguity from noisy background pixels, reducing the error at the background; (2) smaller nodes enhance the matching quality on the foreground pixels with fine details.

Figure 1 supports our point further, where we evaluate the matching accuracy for the pixels near the object boundaries. We see that as the level of pyramid gets higher, it achieves the larger gain near the object boundaries, demonstrating smaller spatial nodes (e.g., pixels) do better at localizing the finer object structures such as object boundaries.