Robust Wide Baseline Pose Estimation from Video (Supplementary Material)

Nicola Pellicanò, Emanuel Aldea and Sylvie Le Hégarat-Mascle
SATIE - CNRS UMR 8029
Paris-Sud University, Paris-Saclay University, France
{nicola.pellicano, emanuel.aldea, sylvie.le-hegarat}@u-psud.fr

I. Estimation using single image pairs - effect of the inlier distribution

In Section 3 of the paper, we underline the importance of encouraging an uniform inlier distribution, and of accounting for the local inlier coverage in the estimation uncertainty. The two images in Fig. 1 show an example of the impact of the inlier distribution on the spatial distribution of the errors, demonstrating the importance of using a video sequence in the case of wide baseline cameras and large scale scenarios.

Fig. 1 shows the inlier matches which are retained after running an estimation of the fundamental matrix between frames 74 of cameras 2 and 3 (following the notation in the paper). We note the presence of a large region lacking correspondences on the bottom right of camera 2, where no feature matches can be acquired. As a result, that area could not be considered in constraining the estimation of the geometry.

Then, Fig. 2a shows the spatial distribution of the symmetric geometric error on the first image. We subdivide the image in buckets, and each bucket is assigned to a certain number of ground truth matches. In order to ensure the maximum achievable uniformity of the ground truth matches the same number of matches is assigned to each bucket, weighted by the portion of the bucket which belongs to the overlapping field of view (so if N is the number of points extracted from a bucket fully included in the overlapping field of view, N/2 will be the number of points considered for a bucket which has half of its area inside the common FOV). Fig. 2a shows, for each bucket, the average error of the estimation with respect to the matches drawn from the points at that location. Errors under 1px are labeled in green, between 1 and 2px in yellow, and over 2px in red. It is clear that inside the area lacking inliers, there is a presence of high errors, which makes this solution unadapted for fitting the entire image space. The RMSE=1.8 which is obtained from this estimation does not fully underline this major limitation. This shows why the use of the Maximum geometric error, which is 7.53 in this example, is crucial for determining if an estimation is well-defined in all the common field of view or not. This example also explains why there is so much variation of the quality of the estimations between different frames from the same video (Fig. 3 in the paper), which depend significantly on how the dynamic elements are disposed spatially.

II. Estimation using video data - effect of encouraging an uniform inlier selection

Fig. 2b shows an example of error distribution resulting from the proposed approach (refer to Fig. 8 on the paper for the error evolution in this test), for the 2-3 camera pair. The image shows a significant decrease of the error in areas which were challenging for single image pairs methods, but also a reduction of the error on a global scale. The RMSE for this example is 0.47, while the Maximum geometric error is 1.45, caused by a single match which is at the very extremity of the overlapping FOV (inside the bucket with 0.97 average error). The background image is the same as in Fig. 2a just for a faster visual comparison (since our method does not use a single frame in particular).

III. Evaluating the quality of the inlier coverage

These three images in Fig. show the temporal evolution of the areas of the image which are considered well-constrained by the current estimation (at iterations 2,17 and 43 respectively). The orange regions in the images represent the areas where a keypoint that we try to match would be considered inside an area which is already covered by a sufficient number of inlier features. We recall that in this case a low Sigma parameter should be set, because the current estimation is well-defined in that location. Green circles define the boundary of a core point in the image: all core points (highlighted in red) may be extracted at the beginning of the current iteration, so their epsilon-neighborhoods are labeled as "orange" regions. However, the well-constrained region is not only defined by the core point neighborhoods, but also from those locations for which the new point would itself become a core point. The evolution of these regions shows how the solution gets more confident in a continuously larger part of the image space, which is coherent with the moving trend of the dynamic objects. This implies an automatic adaption of the Sigma parameter: the larger the well-constrained area, the less frequent will be the use of a large Sigma parameter in order to deal with gross estimation errors, guiding automatically the convergence to a stable solution.

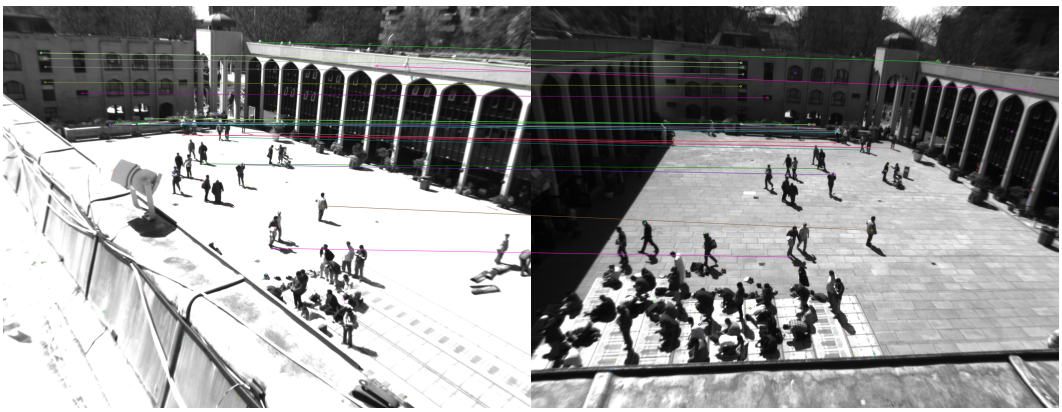
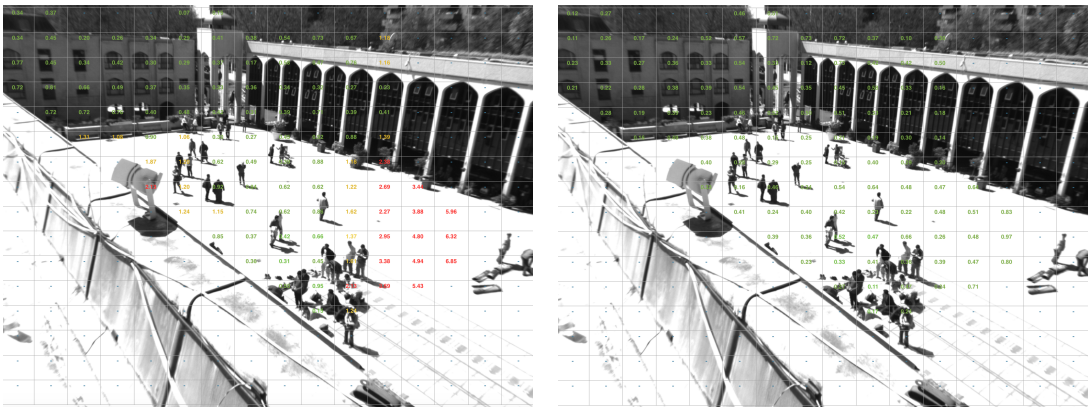
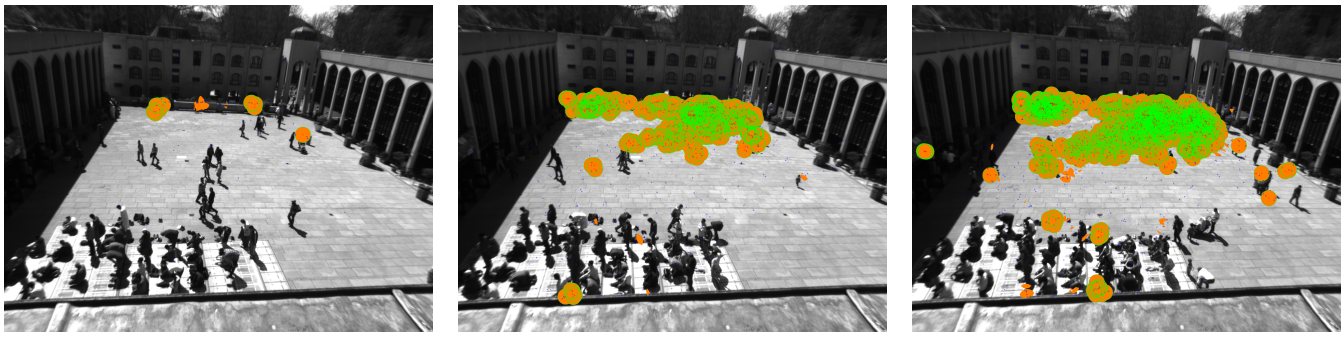


Fig. 1. Sample pair of frames acquired from cameras (it is advisable to zoom for details in the electronic version), and selected matches in this pair.



(a) (b)

Fig. 2. Resulting spatial distribution of the symmetric geometric error with respect to a dense manually annotated ground truth. (a) using a single pair of images. (b) using the proposed method (the same frame as in Fig. 2a is used for ease of reference)



(a) (b) (c)

Fig. 3. Temporal evolution of the areas of the image, highlighted in orange, which are considered well-constrained (it is advisable to zoom for details in the electronic version). (a) Iteration 2. (b) Iteration 17. (c) Iteration 43.