Robust Wide Baseline Pose Estimation from Video

Nicola Pellicanò, Emanuel Aldea, Sylvie Le Hégarat-Mascle

Robust Wide Baseline Pose Estimation from Video

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

***Abstract*—Robust wide baseline pose estimation is an essential step in the deployment of smart camera networks. In this work, we highlight some current limitations of conventional strategies for relative pose estimation in difficult urban scenes. Then we propose a solution which relies on an adaptive search of corresponding interest points in synchronized video streams which allows us to converge robustly towards a high-quality solution. The experiments are performed using a manually annotated ground truth of a large scale scene exhibiting significant depth and perspective variation, uniform areas, repetitive patterns and homogeneous dynamic elements. The results show a fast and robust convergence of the solution, and a significant improvement, compared to single image based alternatives, of the RMSE of ground truth matches, and of the maximum absolute error.**

I. Introduction

The calibration of a camera network with minimal requirements of human intervention (use of calibration objects, guidance of the pose estimation process) has long represented a major field of research in computer vision, with reviews of novel contributions appearing regularly. Recently, the increased focus on safety and surveillance applications has underlined the importance of smart camera networks (the reader may refer to [1] for a concise but complete taxonomy of the major challenges raised by smart cameras). The self calibration part is critical for monitoring projects, for multiple reasons. In order to be able to project image elements from one camera to another in the case of cameras with overlapping fields of view, a relative pose estimation is mandatory and may either help locate an existing element of interest in a different view, or if the calibration is accurate enough, may help identify elements of interest from raw data (i.e. disambiguate using the second view a person who is strongly occluded in the initial view).

Irrespective of the number of cameras deployed, the pose estimation between a pair of cameras is the foundation of any camera network calibration. Existing relative pose estimation algorithms are, for the vast majority, based on matching interest points among the two views and then applying a robust optimization algorithm in order to determine the unknown pose parametrization [2]–[4]. These approaches are used successfully in various domains ranging from aerial imaging to Structure from Motion (SfM) for virtual reality. However, for large scale camera networks in urban environments, some specific scene characteristics complicate or dismiss altogether the use of existing approaches. Due to physical positioning constraints, wide baselines with significant perspective change may be imposed. Even when ignoring positioning constraints, it is beneficial to cope robustly with significant pose variations in order to minimize the number of cameras required for covering a specific area. Another problem is raised by the actual image content; for outdoor surveillance, the scenes are often homogeneous (open spaces) for the most part, or featuring repetitive patterns (human shapes, building facades), and this hampers the use of fully automatic calibration algorithms. Finally, calibration solutions which require significant human intervention, by using calibration objects for example, are time and resource consuming, and in certain situations they are impracticable due to the size of the scene or due to access constraints.

II. Related work

Since the pose estimation requires a set of correct matches, the rejection of outliers is a prerequisite step which is usually performed using a RANSAC-based approach [4], [5]. A large number of matches with a significant ratio of inliers is a positive indicator for, but does not implicitly guarantee, a high-quality pose estimation, as the distribution of matches over the image space is also involved. Wide baseline setups in urban areas exhibit at the same time a low number of matches, a low ratio of inliers as well as a skewed distribution due to large uniform zones (ground, roofs, facades etc). As a result, an uneven distribution leads to a pose estimation which is correct only in covered areas, although the solution is consistent with the observations.

In order to address these problems, guided matching strategies aim to expand the well-constrained area by encouraging a progressive inclusion of new matches [6]; however, in difficult scenes the potential elements to include are sparse and distant, and guided matching may easily include outliers and drive the pose estimation towards an inadequate solution. More elaborate strategies may relax the quality of matches in addition to guiding the search spatially [7], but this favors the inclusion of incorrect correspondences. Also, additional geometrical checks for guaranteeing a local consistency [7], [8] are not effective in scenes which are non-planar and with significant depth variation.

Since one of the fundamental challenges when facing wide baseline calibration is the scarceness of matches, the exploitation of the video stream seems a promising solution (the temporal synchronization of the cameras being convenient, but not a strict requirement). A naive approach, as pointed out by [9], is to extend image-based registration to video-based by temporal accumulation of matches, while other approaches

identify corresponding trajectories of salient objects [10] in order to populate the match set. Despite the richness of video information, the exploitation of video sequences does not address implicitly all the problems previously raised. Although the number of total matches does increase, in scenes with homogeneous dynamic objects such as crowded areas the inlier ratio may actually decrease. Another limitation of straightforward video accumulation is that new matches are clustered around moving objects, and the pose estimation may get constrained locally very strongly, which correspondingly may remove sparse correct matches and deteriorate the solution.

Moreover, in [10], each candidate estimation is performed on a set of matches extracted from a single trajectory (or a pair of them). The authors request non-trivial trajectories to be present, which are trajectories able to cover a large enough part of the image space, and which do not belong to a degenerate configuration (planar trajectory). However, in large scale scenes a representative set of non-trivial trajectories which span most of the image space is often not available; each trajectory is likely to cover a small fraction of the total area, and to be degenerate, when the dynamics of the scene are mostly produced by people walking on the ground plane.

In [9] the authors estimate the geometric constraint by accumulating matches from a fixed number of dynamic texture image pairs. A limitation of this approach (and of the trajectory-based one), is that only dynamic parts of the scene are considered. If a scene contains large static parts (e.g. buildings, see Fig. 2) the estimation will not be globally correct. Moreover, the method is unfeasible, in terms of memory requirements, when applied to high resolution images.

The aim of our work is to propose a computationally effective algorithm for robust pose estimation in difficult scenes, which benefits from synchronized video streams. Our contributions address the following points:

- for a large scale urban scene, we propose a methodology for building an uniformly distributed ground truth set, along with the estimation of a fundamental matrix defining a general transformation between the views;
- based on the previous ground truth data, we highlight that for this type of scene the pose estimation provided by the current state of the art algorithms is highly unstable, with errors which vary strongly across the image space;
- we show that naive temporal accumulation of matches degrades the match inlier ratio, and that convergence towards a high-quality solution is not guaranteed or slow with existing robust estimation techniques;
- we propose a robust temporal accumulation strategy with a fast convergence towards a high-quality solution.

III. Overview and general considerations

We assume a pair of calibrated synchronized cameras, with overlapping fields of view. The SIFT descriptor [11] is employed in the feature extraction and matching stages.

In our approach we exploit the richness of information provided by a video sequence, in contrast with using a single image pair. In fact, we notice that in such wide baseline scenarios with large scale interest regions, it is common that at any given moment only some image locations provide correspondences, increasing the risk of obtaining locally optimal epipolar geometry estimations. As a result, the quality of an estimation based on feature matching may differ a lot for different time instants, making its use unreliable (Fig. 3).

On the contrary, our method starts from an image-to-image initial estimation, and refines it by acquiring new information in the successive frames. At each iteration, the epipolar constraint estimated at the previous step is used to guide the acquisition of new matches between the current frames, through the use of an epipolar band. This new set of matches is combined with the set of inliers identified at the previous step, and a new robust estimation is performed on the new set.

A common practice for match selection is to extract globally distinctive SIFT matches, which pass the 2NN heuristic proposed in [11], as well as a symmetry check which validates pairs only with the best match candidate for both left and right feature points. Given a feature point in the first frame, and a set of candidate features in the second frame, the 2NN heuristic is satisfied *iff* the SIFT distance ratio between the match with the best score and the one with the second best score is lower than a certain threshold. In other words, the test is passed only if that match is by far the most distinctive among the others.

In contrast to this approach, in our match selection stage we extract SIFT matches which are distinctive inside the band region, by applying a modified version of the 2NN heuristic which accounts only for candidate matches in the restricted search space. This procedure is very effective in providing a much larger number of good quality matches, which is critical both because in a wide baseline scene globally distinctive high quality matches are scarce, and because the algorithm is capable to converge faster towards a robust solution.

Moreover, differently from a standard guided matching approach, we do not use only the uncertainty of the estimation of the fundamental matrix to compute the band size, but we adjust the band based on the inlier distribution in the image. This approach has two advantages: it guarantees faster convergence of the solution, encouraging the matching in parts deficient in inliers, while discouraging strong inlier clustering in a localized area of the image, which could bias the estimation.

The illustration of all the proposed steps is supported by a ground truth that we have manually created from the testing scenes. The ground truth consists in manual matches uniformly extracted from all the common field of view, in order to test the quality of the solution across all the analysis area.

Our method allows to automatically recover the relative pose between two cameras in an iterative way. It has shown, in the proposed experiments, to reach a quasi-monotonic decreasing of the geometric error with respect to the number of iterations, while strongly improving the robustness of the estimation, even with different choices of the robust estimator employed.

IV. Exploiting temporal information from synchronized cameras

A. Temporal sampling

An important parameter of our process is the stream sampling period $\triangle_t$. Since we want to exploit the dynamic behavior of the objects in the scene, $\triangle_t$ should be large enough in order to allow dynamic objects moving significantly, and to avoid new information being mostly redundant. This constraint is in opposition with a tracking-based approach which needs small inter-frame difference in order to work properly. On the contrary, setting a too high $\triangle_t$ would cause a slower convergence in time.

B. Matching strategy

Given the two frames at the current iteration, we make use of SIFT to extract an initial set of candidate feature matches M_{init}. Each element of the M_{init} set consists of an array m of the best k candidate matches involving a specific point p in the first frame. The array is ordered in ascending order on the basis of the SIFT distance score.

We do not apply the 2NN heuristic directly on the array m. For example, consider the presence of repetitive structures, e.g. the ones in building facades. A point in the first image could match strongly with multiple points in the second one. Of course such matches would not pass the 2NN heuristic because SIFT distances will be very similar. However, if we first restrict the search space using an epipolar band, provided by the approximate fundamental matrix F computed at the previous iteration, we could find that there is only one possible match which is coherent with the geometry. In such case, that match should be considered a valid candidate because it is distinctive within the area of interest.

For this reason we invert the order of filtering stages which is typical of guided matching approaches: instead of getting global distinctive matches and then checking them against the epipolar bands, we first perform the band filtering and then we isolate the distinctive matches. Given $m = [p'_1, p'_2, ..., p'_k]$, we can compute the epipolar bands in both views for each pair (p, p'_i), as a function of the uncertainty of the estimation and of the point location. The normalized epipolar line in the second image can be defined as $\hat{l} = Fp/ \parallel Fp \parallel$. The epipolar band is an envelope around the epipolar line which depends on the epiline covariance [12][13]:

$$\Sigma_l = J_F \Sigma_F J_F^T + \sigma^2 J_p J_p^T . \tag{1}$$

We assume that the point p is independent from F, since it has not been used in the estimation procedure. The first term encodes the uncertainty of the nine F parameters, while the second one encodes the uncertainty of the position of point p in the image. The standard deviation σ represents the isotropic uncertainty in both image directions.

The conic which gives the mathematical representation of the epipolar band can be retrieved as [2]:

$$C = \hat{l}\hat{l}^T - k^2 \Sigma_l, \tag{2}$$

where k^2 is chosen by solving $F_2^{-1}(k^2) = \alpha$, with α the confidence level parameter, commonly set to 95%, and F_2 the cumulative χ_2^2 distribution.

If p or p'_i are not contained in one of the corresponding epipolar bands, then p'_i is removed from m. We call the new vector $m_{Band} = [\widetilde{p}'_1, \widetilde{p}'_2, ..., \widetilde{p}'_{k'}]$, where $k' \leq k$. In order to retain only high quality matches, the following constraint must hold:

$$\widetilde{p}'_1 = p'_1, \tag{3}$$

which means that if the match with best score is not contained in the epipolar band, we discard the entire current set of candidate matches, and continue. This constraint avoids the inclusion in the final set of matches with a poor absolute score.

We are now able to perform the 2NN heuristic on m_{Band}:

$$\frac{d(p, \widetilde{p}'_1)}{d(p, \widetilde{p}'_2)} < \tau, \tag{4}$$

where d is the SIFT distance measure, and τ is a threshold usually set in the range 0.6-0.8.

Together with the test in (4), we perform also a symmetry check in order to improve considerably the quality of the matching process. It consists into applying the same procedure in the opposite sense, from the second to the first frame. If $\widetilde{p}'_1$ is the best match for p, and p is the best match for $\widetilde{p}'_1$, the symmetry check is respected. If both tests are passed, then the match $(p, \widetilde{p}'_1)$ is added to the set S_{new}, which contains all the matches discovered at the current iteration.

C. Choice of the parameter σ

We exploit the parameter σ in (1) in order to be able to deal with large errors in the epipolar constraint. If the epipolar line is correct, the σ value represents the error in the matching process which leads to a small deviation from the epipolar line. On the other hand, when the epipolar line is shifted because of an estimation error in some part of the image, σ can represent the error due to the bad localization of the line.

The idea is that in areas of the image which lack inliers, there is a high risk that the current estimation is biased with respect to the optimal one. Our approach is the following: in well-constrained areas of the image, we set a low σ_L value representing errors in the matching procedure; we impose a much higher σ_H value in regions which are not well covered by inliers. When σ is small, the first term of (1) is predominant, and the shape of the epipolar band will likely follow a hyperbola; when σ is high, the second term of (1) dominates the first, and the epipolar band will be likely enclosed by two straight lines. Possible outliers included in the process are taken into account by using a robust estimation technique at every iteration.

We define the critical notion of well-constrained areas by using a fundamental concept introduced in the field of data clustering with noisy data [14]. In [14], a point q is considered a *core point* if, given two parameters ϵ and $MinPts$, $\mid N_\epsilon(q) \mid \geq MinPts$, where $N_\epsilon(q)$ is the set of points at a distance lower than ϵ from q. The authors of [14] provide also

the definition of a *directly density-reachable* point p, given ϵ and $MinPts$:

1) $p \in N_\epsilon(q)$
2) q is a core point

A *density reachable-point* is a neighbor of some core point q. **Condition 1:** Given the inlier set S, a new point p belongs to a clustered region if one of the two conditions holds:

1) p is a core point of the set $S \cup p$
2) p is *directly density-reachable* by any point q, $q \in S$

Condition 1 provides a simplified check to state whether a point is contained or not inside a region of the image in which the current estimation is well-constrained due to a strong inlier presence. Finally we are ready to choose the parameter σ as:

$$\sigma = \begin{cases} \sigma_L & p \,\text{satisfies Condition 1} \\ \sigma_H & \text{otherwise} \end{cases} \tag{5}$$

D. Fundamental matrix re-estimation

Once the matching stage has been completed, the set S_{new} containing the new matches may be added to the inlier set S obtained from the previous estimation. All these matches can be used as input of a robust estimation algorithm, in order to obtain F for the current iteration.

Our approach is independent from the specific algorithm employed at this stage, and we will demonstrate in section V its use with both the popular RANSAC [15] and with ORSA [3] frameworks. The resulting F is then refined using the Levenberg-Marquardt algorithm, and the 9×9 parameter covariance matrix is evaluated as in [12]. Fig. 1 provides an outline of the proposed algorithm.

Input: $Stream0, Stream1, k, \triangle_t, max_t, \epsilon, MinPts, \sigma_H, \sigma_L$
Output: F

$M_{init} \leftarrow SIFTMatches(Stream0[0], Stream1[0], k)$
$M \leftarrow filterMatches(M_{init})$ {2NN, symmetry}
$F, inlierSet, Cov \leftarrow estimateAndRefineF(M)$
$S \leftarrow inlierSet$
$t \leftarrow \triangle_t$
while $t \leq max_t$ **do**
  $C \leftarrow detectCorePoints(S, \epsilon, MinPts)$
  $M_{init} \leftarrow SIFTMatches(Stream0[t], Stream1[t], k)$
  $Snew \leftarrow \emptyset$
  for all m in M_{init} **do**
    {m array of candidate correspondences of length k}
    if $isInCluster(m, C, \epsilon, MinPts)$ **then**
      $\sigma \leftarrow \sigma_L$
    else
      $\sigma \leftarrow \sigma_H$
    end if
    $mBand \leftarrow epipolarBandFiltering(m, F, Cov, \sigma)$
    if m_{Band} passes $2NNheuristic$ and $SymmetryCheck$ **then**
      $S_{new} \leftarrow S_{new} \cup m_{Band}[0]$ {add best score match}
    end if
  end for
  $S \leftarrow S \cup S_{new}$
  $F, inlierSet, Cov \leftarrow estimateAndRefineF(S)$
  $S \leftarrow inlierSet$
  $t \leftarrow t + \triangle_t$
end while

Fig. 1. Outline of the proposed approach.

V. Results

A. Experimental setup and ground truth construction

We test our method on sequences recorded at Regent's Park Mosque, London. The camera network consists of three cameras installed on the roof (see Fig. 2). The analysis region is a square surrounded by buildings which present repetitive structures. We record video streams of the square, capturing the dynamic behavior of people who are free to move in the area. The grayscale video is recorded at 8 fps, with a 1624×1234 resolution. In order to demonstrate the independence of the method from the dynamic context, we calibrate different pairs at different times.

In order to perform a rigorous evaluation of the performance we have built a ground truth which allows us to understand the quality of the solution across the whole scene. By defining an uniform grid (buckets), we extract matches manually and uniformly inside the overlapping field of view. The extraction procedure is followed by the estimation of a fundamental matrix, which is used to refine the position of the matches. The process is repeated iteratively in order to obtain a set of matches with half-pixel precision. Such ground truth extraction is essential to evaluate errors in the estimation even in regions where an automated process would not consider interest points.

The measurements metrics we employ are the RMSE and the Max symmetric geometric error [2] on the ground truth. The use of the Max Error is the strictest possible metric, and is essential for revealing localized errors, which would be mitigated by RMSE. Due to the stochastic nature of the process, results are evaluated over 300 executions of each test.

B. Experimental results

Regarding the parameters, the most significant inputs are Δ_t, for which we set a value of 24 frames, $\sigma_L = 1$, which is a common choice in guided matching covariance propagation [6], and finally we set $\sigma_H = 5$. The k parameter has shown to have little influence on the final results if chosen in a range of 2-5 (results with $k = 3$ are presented). We choose, as robust fundamental matrix estimator, the ORSA[3] a-contrario framework, which has proven to guarantee great robustness without the need of setting a threshold. In order to demonstrate the independence of our method from the estimator, we also present an estimation result using the RANSAC [15] method implemented in the USAC [16] framework.

Fig. 3 shows the errors in the estimations that are obtained by using a single pair of images extracted from the streams of cameras 1 and 2, and applying the ORSA estimator. The quality of the estimation is highly dependent on the specific pair considered and on the position of dynamic objects in the scene, which can leave large areas without any interest point. In any

(a) (b) (c)

Fig. 2. Sample frames acquired from the three cameras. (a) Camera 1, (b) Camera 2, (c) Camera 3. Two large featureless regions can be seen on the bottom-right and top-left of the square.

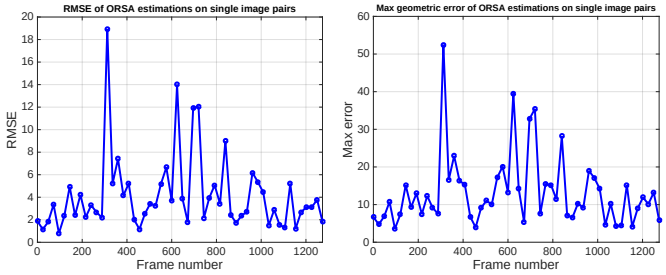

Fig. 3. RMSE and Max geometric error by applying ORSA on each frame pair independently. Large variations in the result demonstrate the unreliability of estimation with still images in such setup. Streams from cameras 1 and 2 are used.

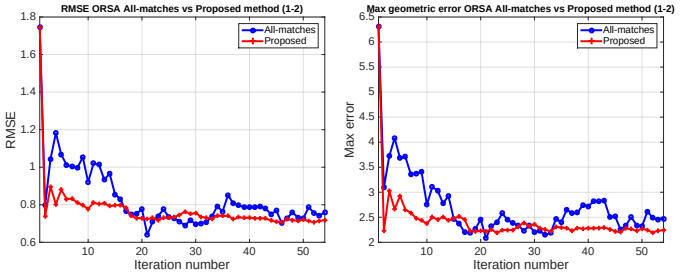

Fig. 4. RMSE and Max geometric error by applying the *All-matches* strategy and our method on 1-2 camera pair, with ORSA. Our selection is more reliable, and we are able to improve the initial estimation significantly and robustly.

case, the best estimation achievable has a maximum error of almost 4 pixels, leaving room for a consistent improvement. The method in [7] fails to converge towards an acceptable solution (RMSE=245) as it does not cope with wide baselines and strong depth variation.

Then we show our estimation results for cameras 1-2, presenting them against the results obtained by performing robust estimation on a set of matches accumulated naively from frame pairs (we call it *All-matches*). Fig. 4 shows the RMSE and Max geometric errors at the different iterations of the algorithm. Our method is able to reduce the RMSE from 1.75 to 0.75, and to consistently decrease the Max error from 6.5 to 2.2 pixels. We highlight the robustness of our strategy, with the error following a monotonic decreasing trend after a few iterations. This does not happen in the *All-matches* case, which presents large oscillations in time, which implies that getting more points from the video stream will not certainly improve the batch estimation result, introducing thus a frame window size choice problem.

The explanation for this behavior comes from the analysis of the inlier ratios estimated at each iteration (Fig. 5), which we are able to present due to the manual ground truth. The *All-matches* curve relates to the inlier percentage curve (checked against the ground truth F) that is obtained by accumulating matches, which drops independently from the time interval chosen. Thus the benefit of adding new points is negated by a lowering ratio of good matches, which implies the existence of a trade-off. On the other hand, our approach is based on a strict

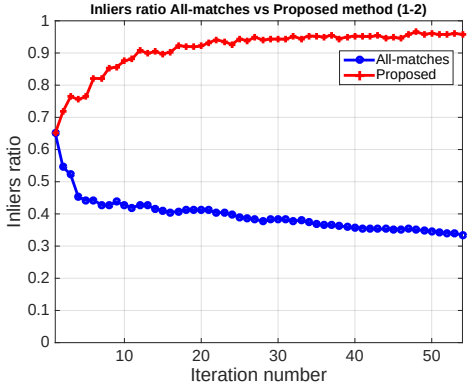

Fig. 5. The inliers ratio at each iteration for the *All-matches* and our approach.

rejection procedure, so the inlier ratio follows the opposite trend, since the constraining of the solution will improve the probability of including only inliers as new matches.

In Fig. 6 we show the RMSE for the same camera pair, but using the RANSAC framework to estimate the fundamental matrix. The behavior of our algorithm remains the same, irrespective of the estimator, and of the initial values of the RMSE. In the RANSAC case the drawbacks of the a straightforward match accumulations are even more evident.

Fig. 7 demonstrates the benefits of adapting the σ parameter of the covariance of the epipolar band to the actual spatial distribution of inlier matches in the image. It follows that by setting a $\sigma = 1$, as in [6], we can not add new information

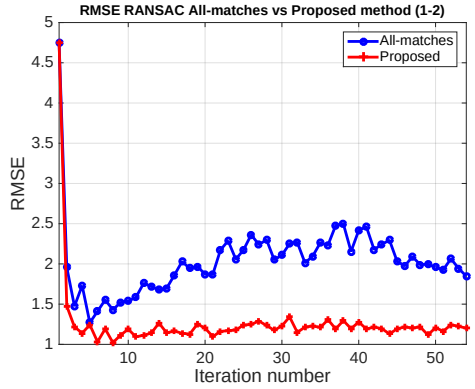

Fig. 6. RMSE by applying the *All-matches* strategy and our method on 1-2 camera pair, by exploiting a RANSAC-based estimation.

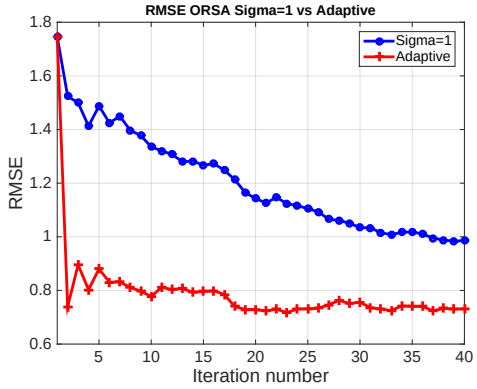

Fig. 7. RMSE by applying our method on the 1-2 camera pair by using a fixed σ value, and by using the adaptive σ introduced in Section IV-C.

which is able to correct gross local errors in the estimation, leading to a much slower convergence which is never able to reach, in terms of error, the results obtained by our strategy.

Finally, we show the estimation results for the camera pair 2-3 , using a video stream captured at a completely different time instant, in order to deal with different dynamics of the scene. Fig. 8 shows again consistent results both in terms of RMSE and of Max error. We are able to decrease the RMSE from 1.9 to 0.5, while reducing the Max error on the whole image space from 11 to 1.5 pixels, with a substantial decrease of initial errors in just 4 iterations.[1]

VI. Conclusions

This paper proposed a new approach for solving difficult relative pose estimation problems based on a guided selection of new matches from video. We select new matches in order to constrain the estimation robustly, by adapting the search process with respect to the local inlier distribution. This results in an fast convergence towards a high-quality solution, which is being highlighted by the manual ground-truth we produced for a difficult scene. In our experiments, we show that this video accumulation strategy clearly outperforms current pose estimation solutions. Directions for future research include

[1]Implementation at https://github.com/MOHICANS-project/fundvid

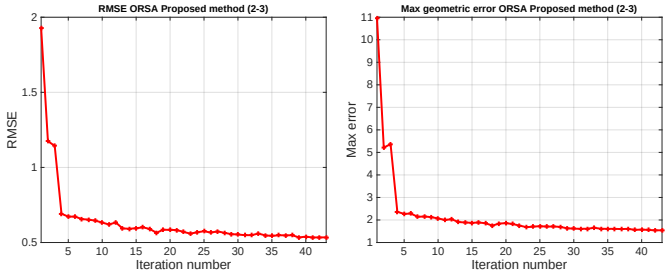

Fig. 8. RMSE and Max geometric error obtained by applying our method for the 2-3 camera pair, with the ORSA estimator.

acquiring and distributing for the academic community a multiple camera dataset in an urban environment, applying our pose estimation within a security application and improving the accuracy of tasks such as detection, counting or tracking.

Acknowledgment

The authors gratefully acknowledge the support from Regent's Park Mosque for providing access to the site during data collection, and from K. Kiyani. This work was partly funded by ANR grant ANR-15-CE39-0005 and by QNRF grant NPRP-09-768-1-114.

References

[1] J. C. SanMiguel, C. Micheloni, K. Shoop, G. L. Foresti, and A. Cavallaro, "Self-reconfigurable smart camera networks," *IEEE Computer*, vol. 47, no. 5, pp. 67–73, 2014.

[2] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[3] L. Moisan and B. Stival, "A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix," *Int. J. Comp. Vis.*, vol. 57, no. 3, pp. 201–218, 2004.

[4] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *Int. J. Comp. Vis.*, vol. 80, no. 2, pp. 189–210, 2008.

[5] D. Martinec and T. Pajdla, "Robust rotation and translation estimation in multiview reconstruction," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[6] B. Ochoa and S. Belongie, "Covariance propagation for guided matching," in *Workshop on Statistical Methods in Multi-Image and Video Processing*, 2006.

[7] X. Tan, C. Sun, X. Sirault, R. Furbank, and T. D. Pham, "Feature matching in stereo images encouraging uniform spatial distribution," *Pattern Recognition*, vol. 48, no. 8, pp. 2530–2542, 2015.

[8] X. Guo and X. Cao, "Triangle-constraint for finding more good features," in *Pattern Recognition (ICPR), Int. Conf. on*, 2010, pp. 1393–1396.

[9] A. Ravichandran and R. Vidal, "Video registration using dynamic textures," *Patt. Anal. Mach. Intell.*, vol. 33, no. 1, pp. 158–171, 2011.

[10] Y. Caspi, D. Simakov, and M. Irani, "Feature-based sequence-to-sequence matching," *Int. J. Comp. Vis.*, vol. 68, no. 1, pp. 53–64, 2006.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *Int. J. Comp. Vis.*, vol. 27, no. 2, pp. 161–195, 1998.

[13] F. Sur, N. Noury, and M.-O. Berger, "Computing the uncertainty of the 8 point algorithm for fundamental matrix estimation," in *19th British Machine Vision Conference-BMVC 2008*, 2008, p. 10.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J. Frahm, "Usac: A universal framework for random sample consensus," *Patt. Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, 2013.