Machine Vision and Applications
https://doi.org/10.1007/s00138-019-01036-6

ORIGINAL PAPER

Wide baseline pose estimation from video with a density-based uncertainty model

Nicola Pellicanò[1] · Emanuel Aldea[1] · Sylvie Le Hégarat-Mascle[1]

Received: 26 January 2018 / Revised: 26 April 2019 / Accepted: 23 May 2019

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 toward a high-quality solution. The core idea of our algorithm is to build across the image space a nonstationary mapping of the local pose estimation uncertainty, based on the spatial distribution of interest points. Subsequently, the mapping guides the selection of new observations from the video stream in order to prioritize the coverage of areas of high uncertainty. With an additional step in the initial stage, the proposed algorithm may also be used for refining an existing pose estimation based on the video data; this mode allows for performing a data-driven self-calibration task for stereo rigs for which accuracy is critical, such as onboard medical or vehicular systems. We validate our method on three different datasets which cover typical scenarios in pose estimation. The results show a fast and robust convergence of the solution, with a significant improvement, compared to single image-based alternatives, of the RMSE of ground-truth matches, and of the maximum absolute error.

Keywords Pose estimation · Wide baseline · Camera calibration · Guided matching

1 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 and photogrammetry, with novel contributions and surveys appearing regularly [2–4,10,21,47,50,53]. Recently, the increased focus on safety and surveillance applications has underlined the importance of smart camera networks (the reader may refer to [36,54] for a more detailed 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, it 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 on applying a robust optimization algorithm in order to determine the unknown pose parameterization [22,35,39,58]. Besides being used in surveillance, these approaches stem from and benefit 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

✉ Nicola Pellicanò
nicola.pellicano@u-psud.fr

Emanuel Aldea
emanuel.aldea@u-psud.fr

Sylvie Le Hégarat-Mascle
sylvie.le-hegarat@u-psud.fr

[1] SATIE, Université Paris-Sud, Université Paris-Saclay, Rue Noetzlin, Gif-sur-Yvette 91190, France

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.

Guided matching methods, which propagate the estimation uncertainty in the image space in order to include progressively lower-quality matches, cannot cope with missing information in significant parts of the scene. In this case, two alternatives are popular. First, the estimation problem may be guided by additional sensors (GPS, IMU, radio fingerprinting devices, etc). Alternately, a prior 3D cartography of the area may be used to register the cameras in the same reference. Our work aims to avoid the additional costs entailed by these strategies by integrating information extracted from the temporal data flow. The only underlying assumptions—a constant relative pose and a synchronized video stream—are generally satisfied by surveillance camera networks, at which our algorithm is aimed. However, a second use case for our algorithm is the accurate self-calibration of stereo rigs performing rigid displacement.

The aim of our work is thus to propose a computationally effective algorithm benefiting from synchronized video streams and performing a robust reduction of the pose uncertainty area in the image space. In more detail, our contributions address the following points:

- we propose a strategy for sampling the video streams for corner matches according to the local uncertainty of the current pose estimation (Sect. 4)
- we introduce an algorithm for evaluating the local uncertainty as a nonlinear mapping of the observation density. Our density estimation is a continuous extension of the density-based spatial clustering which was used in our initial work [44] (Sect. 5)
- we propose an algorithm variant which refines an existing pose, and which is more suitable for stereo systems requiring frequent extrinsic calibrations (Sect. 6).

Section 7 details the methodology we propose for building an uniformly distributed ground-truth set for matching. This allows us to perform comprehensive evaluations of candidate poses in real, large-scale, outdoor environments. Section 8 presents the evaluation results, and we conclude in Sect. 9.

The full implementations of the pose estimation algorithm and of the variant for pose refinement are accessible online.[1]

2 Related work

Since the pose estimation requires a set of correct matches, the outlier rejection is a prerequisite step which is usually performed using a RANSAC-based approach [35,58]. A large number of matching observations 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.

Guided matching In order to address these problems, guided matching strategies aim to expand the well-constrained area by encouraging a progressive inclusion of new matches [43]. 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 toward an inadequate solution. More elaborate strategies may relax the quality of matches in addition to guiding the search spatially [61], but this favors the inclusion of incorrect correspondences. Correct matches tend to form clusters with specific motions, and previous works proposed explicit geometrical checks for guaranteeing a consistent transformation of the inlier point set [20,27,61], based on local planarity or local contour invariance. More recently, data-driven strategies for selecting consistent observations have been proposed; for example, in [71] the authors rely on a one-class SVM to select a reliable candidate inlier set, and in [28] a motion model based on bilateral functions is used. However, all these approaches which rely on higher-level perceptual information in order to validate the inlier set coherent motion are not effective in complex urban environments with scarce candidates, abrupt and frequent depth variations of the scene and inconsistent edge detections due to significant viewpoint changes (see, for example, Fig. 4).

An interesting correlation between the pose estimation errors and the number of matches, albeit empirically validated, has been discussed in [30]. This justifies all the more the fact that below a certain level of conveniently distributed inlier information, guided matching will not be able to recover a globally fit solution.

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

Externally guided pose estimation The impact of the challenges raised when facing wide baseline calibration may be mitigated by the use of independent sources of information. One promising avenue is the use of a prior pose hypothesis relying on GPS devices, which provide the approximate locations, coupled with IMUs, which provide the orientations. M-estimators are well adapted for guiding the pose search based on prior information [19], and for real-time applications RANSAC-based strategies are also widely used, i.e.,[17,25].

A second strategy which has gained popularity recently relies on the additional creation of a cartography of the surveyed environment using SLAM [1,18,48]. While this technique is the only way to register cameras with non-overlapping fields of view (using visual information), it can also help in wide baseline scenarios as the pose estimation is reduced to two localization tasks within the cartography.

The externally guided techniques overcome the difficulties of the purely vision-based pose estimation, at a cost. For prior pose hypotheses, the cameras must be fitted with additional devices, and also the systems must be accurately calibrated offline in order to align the sensor and camera reference systems. When using a cartography, the mapping procedure may be cumbersome and is valid as long as the scene does not change significantly. In addition, any dynamic parts of the scene contribute only to the outlier observations, and also access to the scene for mapping is not always possible due to various types of restrictions. Finally, externally guided procedures can not be appended once the dataset has already been acquired—the ideal solution would just rely on the actual video data.

Leveraging temporal information 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 [52], is to extend image-based to video-based registration by temporal accumulation of matches. An alternative strategy identifies corresponding trajectories of salient objects [5] 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 in turn may remove sparse correct matches and deteriorate the solution.

Moreover, in [5], 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 [52], 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. 4), the estimation will not be globally correct. Moreover, the method is unfeasible, in terms of memory requirements, when applied to high-resolution images.

Recent efforts aimed at pose estimation from video use motion barcodes of lines [23]. The authors sample points on the image borders and connect any pair of them in order to build a set of candidate epipolar lines. Then, lines are matched by their motion barcodes, computed from background subtraction, and a RANSAC estimation is performed given the line matches. Besides the need to explore a large search space, the method may fail when people move in a straight line in the scene, due to the extraction of a quasi-degenerate pencil of candidates. Moreover, when applied to real datasets as PETS 2009 [14], the method in [23] as well as other algorithms is benchmarked against the provided ground-truth calibration. However, such ground truth may itself present (as we will discuss in Sect. 8.2 for PETS 2009) local errors resulting in a performance bias of the evaluation.

3 Overview of the proposed approach

We consider a pair of calibrated, synchronized cameras, with overlapping fields of view.

In our approach, we exploit the richness of information provided by an existing video sequence, in contrast to relying on a single image pair. In fact, we have noticed that in such wide baseline scenarios with large-scale regions of interest, 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, rendering the image-based estimation algorithms unreliable (this point will be later illustrated clearly by Fig. 5 in the results Section).

On the contrary, our method starts from an image-to-image initial estimation, and refines it by acquiring new information in the following 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 matches which satisfy specific quality-related metrics (such as the 2NN heuristic proposed in [31]), as well as to enforce 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, a match with a point of the candidate set is extracted only if it is by far the most distinctive among the others. Thus, a filtering procedure is applied, by taking into account only the quality of the candidate matches.

In contrast to this approach, in our match selection stage we extract matches which are distinctive inside the band region, by accounting only for candidate matches lying 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 toward a robust solution.

Moreover, differently from a standard guided matching approach, we not only use 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 a faster convergence of the solution, encouraging the matching in parts deficient in inliers while discouraging the inclusion of conflicting matches in areas rich in information.

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 across all the common fields of view, in order to test as fairly and comprehensively as possible the quality of the solution.

Our method, which allows to automatically recover the relative pose between two cameras in an iterative way in the time dimension, has shown during our 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.

The main functionalities of our algorithm are presented in Fig. 1 and will be detailed in the following sections.

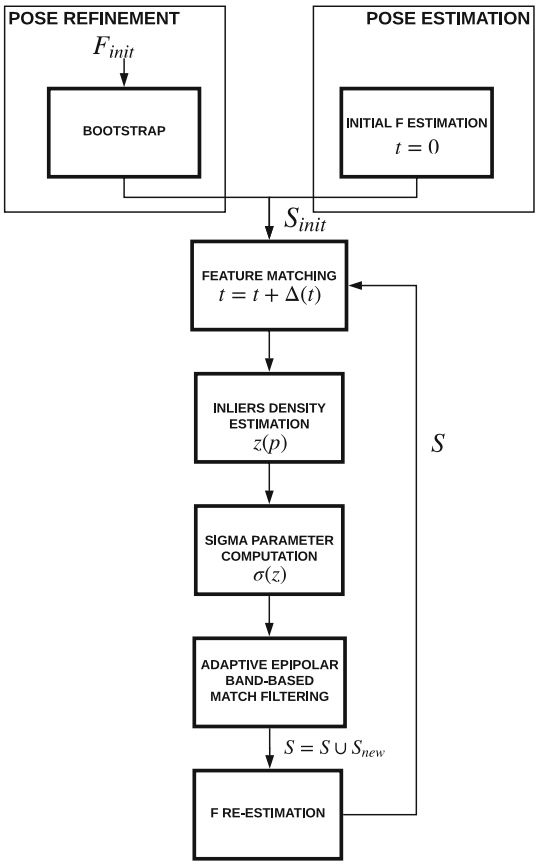

Fig. 1 Overview of our algorithm, which may be executed either for a generic pose estimation (Sect. 4) or for the refinement of an existing prior pose (Sect. 6)

4 Integrating temporal information from synchronized video streams

This section introduces the mechanism that integrates at each time step the current epipolar *local* uncertainty into the matching process.

4.1 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 a significant displacement of dynamic objects, 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 other hand, setting a too high $\triangle_t$ would just cause a slower convergence in time.

4.2 Matching strategy

Given the two frames at the current iteration, the objective is extract a new set of matches S_{new} that will add new information to the current set of inliers S, which represents the output of the previous iteration. The SIFT descriptor [31] is employed in the feature extraction and matching stages. We 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 descriptor's distance score.

Consider the presence of repetitive structures, such as elements on building facades or multiple people with body parts being very similar looking at small scales. Thus, it is common for a point in the first image to relate strongly to multiple points in the second image. Of course such matches would not pass the 2NN heuristic proposed in [31], because descriptor distances would 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 $m = \left[p'_1, p'_2, \ldots, p'_k\right]$, 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 is defined as $\hat{l} = Fp/ \parallel Fp \parallel$. The epipolar band is an envelope around the epipolar line which depends on the epiline covariance [60,75]:

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

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 [22]:

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

where κ^2 is chosen by solving $F_2^{-1}(\kappa^2) = \lambda$, with λ 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 filtered vector $m_{Band} = \left[\widetilde{p}'_1, \widetilde{p}'_2, \ldots, \widetilde{p}'_{k'}\right]$, where $k' \leq k$. In order to retain only high-quality matches, the following constraint must hold:

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

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. Put differently, the inversion of the filtering and heuristic stages has an impact only on the choice of the second best match for score comparison, while it encourages the same matching quality as the standard approach.

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. This consists in 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.

4.3 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 Sect. 8 its use with the ORSA [39] framework. The resulting F is then refined using the Levenberg–Marquardt algorithm, and the 9×9 parameter covariance matrix is evaluated as in [75].

5 Choice of the parameter σ

In the following section, we detail how we relate locally the presence of inliers to the uncertainty of the epipolar constraint, and how the latter is updated as the temporal observations accumulate. We exploit the parameter σ in Eq. (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 underlying 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 consists in varying smoothly the value of σ as a function of the inlier density, which reflects how well constrained locally the solution was at the previous iteration. When σ is small, the first term of Eq. (1) is predominant, and the shape of the epipolar band will likely follow a hyperbola; when σ is high, the second term of Eq. (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.

5.1 The binary density model

In our preliminary work [44], we defined the notion of well-constrained regions by using a fundamental concept introduced in the field of data clustering with noisy data [13]. In [13], a point q is considered a *core point* if, given two parameters ϵ and $MinPts$, $|\ N_\epsilon(q)\ | \geq MinPts$, where $N_\epsilon(q)$ is the set of points at a distance lower than ϵ from q. The following definition of a *directly density-reachable* point p, given ϵ and $MinPts$, has been exploited

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

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 at least one core point $q, q \in S$

Such condition provides a binary check whether the local area of interest is well constrained or not, and it has been used in order to set a low sigma σ_L if it is satisfied, or a high sigma σ_H otherwise. However, as a step function, such decision rule lacks continuity at different density levels, treating regions at medium densities as badly constrained as empty regions.

5.2 A continuous density-uncertainty dependency

In our formulation, we propose to define σ as a continuous sigmoid function which spans between σ_H and σ_L (Fig. 2). While σ_L can be always set to $\sigma_L = 1$, as for the classic guided matching refinement methods, σ_H is a free parameter, which depends on the reliability of the initial solution, reflected by the scarcity of matches. Let us define as η the target density at which we have an α degree of confidence in the solution:

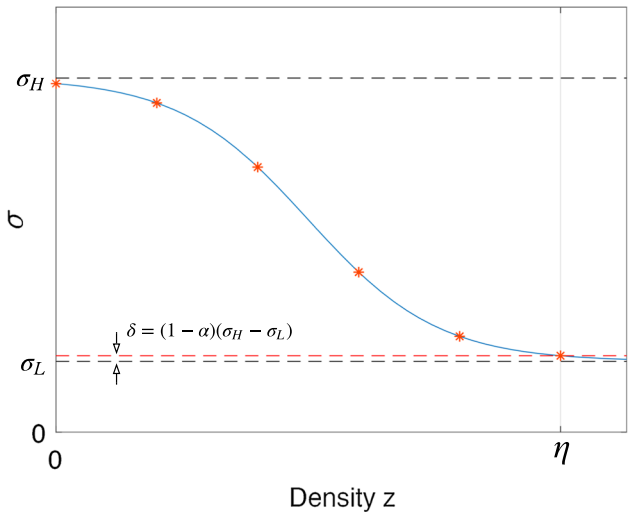

Fig. 2 Sigmoid function which models in our algorithm the impact of the local observation density on the local uncertainty. The stars along the function represent the sampling locations which would be used by a histogram kernel density estimator with $n = 5$

$$\sigma(\eta) = \alpha\sigma_L + (1-\alpha)\sigma_H \tag{5}$$

The use of α is due to the fact that the sigmoid reaches the bounding uncertainties σ_H and σ_L at $-\infty$ and $+\infty$. Thus, the degree of confidence α allows us to define the disparity $\delta = (1-\alpha)(\sigma_H - \sigma_L)$ which is present at 0 and η densities between the sigmoid function and the target uncertainties σ_H and σ_L, respectively (see Fig. 2). The choice of α is not critical, and in all our experiments we set $\alpha = 0.99$.

We can then express the sigmoid as a function of the density z in the following way:

$$\sigma(z) = \sigma_L + \frac{\sigma_H - \sigma_L}{1 + \mathrm{e}^{-b(z-\eta/2)}} \tag{6}$$

where the implicit steepness b has the form:

$$b = \frac{2}{\eta}\log\left(\frac{1-\alpha}{\alpha}\right) \tag{7}$$

We propose to evaluate the density z at each point $\mathbf{p}$ of the image using a Kernel Density Estimation (KDE) in the two-dimensional space:

$$z(p) = \frac{1}{h^2}\sum_{i=1}^{N} K\left(\frac{\mathbf{p} - \mathbf{p}_i}{h}\right) \tag{8}$$

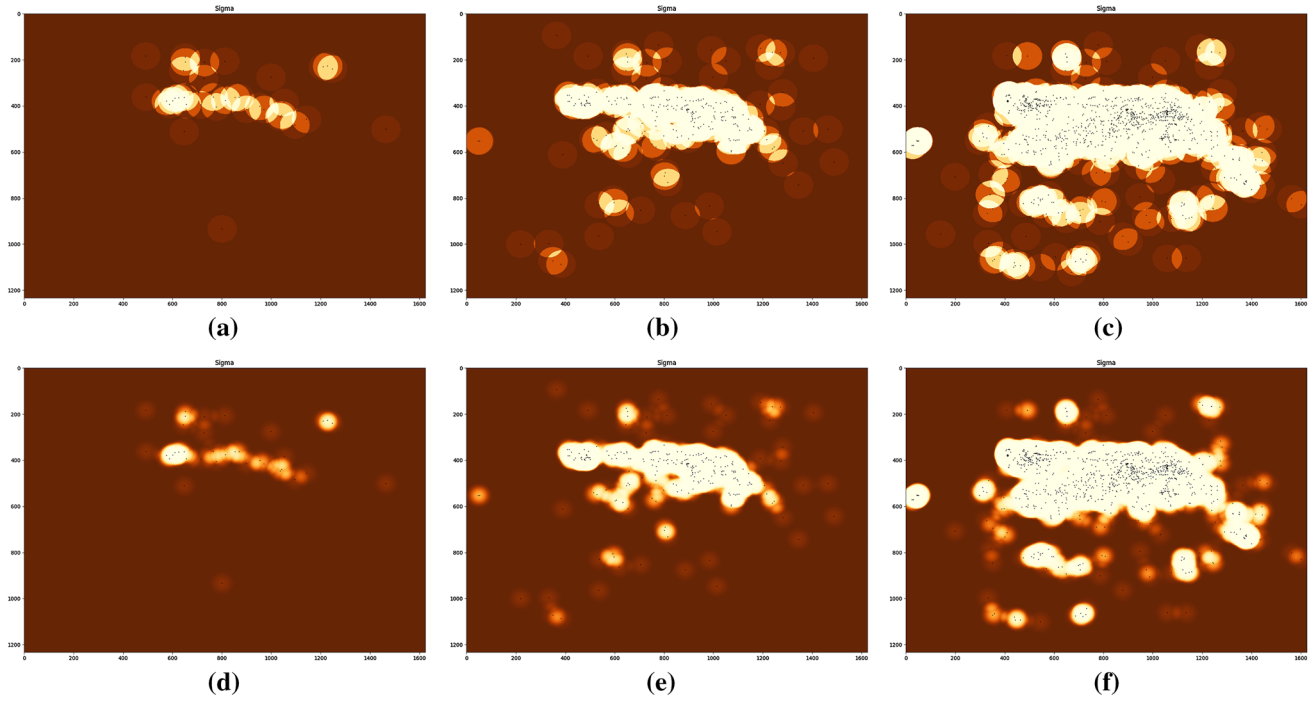

Fig. 3 Sigmoid $\sigma(z)$ evaluation in the image space, with histogram kernel: **a** Iteration 0, **b** Iteration 5, **c** Iteration 30. The lighter the color, the lower σ value. As the method converges to a robust solution, the well-constrained region grows in size. Smoother $\sigma(z)$ estimation can be performed with an Epanechnikov's kernel (**d**–**f**), but the higher computational does not correspond to substantial improvement in the result. The images refer to the *Regents Park* dataset, with camera 2 as the reference (Fig. 4a)

Note that in Eq. 8, differently from the classical KDE formulation, we do not normalize the density by the total number N of inliers. This is justified by the fact that N varies at each iteration of the algorithm, and thus, this would require a continuous rescaling of the target density parameter η, without any change in the σ estimation output.

The choice of the kernel is not critical for our application, and a simple function as the histogram kernel:

$$K_H(\mathbf{u}) = \frac{1}{\pi} \mathbb{1}_{\|\mathbf{u}\| \leq 1} \tag{9}$$

has shown good performance in our experiments, while more complex kernels as Epanechnikov:

$$K_e(\mathbf{u}) = \frac{2}{\pi} \left(1 - \|\mathbf{u}\|^2\right) \mathbb{1}_{\|\mathbf{u}\| \leq 1} \tag{10}$$

do not introduce a significant advantage, while being more computationally costly (the kernels are normalized for the 2D scenario occurring in our case). Figure 3 shows the gradual expansion of the well-constrained areas in the image space when using the histogram kernel (Fig. 3a–c) and the Epanechnikov kernel (Fig. 3d–f).

The target density η is a user-defined quantity depending on the ideal interest point density for a specific type of scene. However, one may reason rather in terms of the expected number of corners n at a relevant spatial scale, while the actual numerical value of η involves a specific KDE function as well as the local relative corner layout. In our framework, we propose the following interpretation of the target density η with respect to the expected number of points via a given kernel K. The η target density may be represented as the density evaluated with n points at distance $h/2$ from the target:

$$\eta = \frac{n}{h^2} K(\mathbf{v}) \tag{11}$$

with $\mathbf{v}$ being any vector such that $\|\mathbf{v}\| = 0.5$. This reasonable assumption allows us to relate the target density to the target number of points via the kernel. A critical parameter for the density estimation task is the bandwidth h, which identifies the radius of interest around a point. As we will show in the results, while the estimation task is sensible to the bandwidth, the actual error of the algorithm after convergence remains stable even for large variations of h.

6 Refining an existing pose estimation

In this section, we consider an adaptation of our algorithm which allows for data-driven refinement of an existing pose. Indeed, numerous existing datasets provide extrinsic calibra-

tions, acquired with different techniques and characterized by various degrees of accuracy.

The main interest of the refinement procedure is that, as video data is analyzed, our algorithm may be used in order to refine the original estimation, which may lack precision in some specific areas of the image space. Moreover, pose refinement may be needed when the camera positions might have changed slightly prior to an acquisition due to mechanical factors or due to internal behavior (e.g., pan-tilt-zoom cameras), but a reasonable prior pose is known. In robotic vision, the pose refinement is often applied to stereo rigs, but our setting is not suitable for continuous refinement in which the pose is time dependent (in this case Kalman filtering is the method of choice [9,21,42]). Our algorithm is suited for the accurate update of a stereo rig pose which is fixed but possibly different slightly from a reference value. Existing algorithms such as [29] rely on bucketing heuristics in order to enforce spatial uniformity of the observations, while devices which refine the stereo pose upon initialization such as the ZED camera from Stereolabs [59] run proprietary code.

The refinement procedure is similar to the estimation presented in Sect. 4, except the requirement of a *bootstrap period* at the beginning of the refinement process. The bootstrap period consists in building an initial set of matches by performing the acquisition and band filtering for several frames (setting inlier density $z(p) = 0$ for the entire period), by using the initial pose F_{init}. The period ends when a target number of matches is reached; we set this number to be proportional to the number m of raw matches acquired from the first frame pair of the sequence (we heuristically set this number to be $5m$, independently from the dataset). Please note that the bootstrap is different from a blind accumulation because it exploits via the band filtering the F_{init} that we intend to refine. The initial set of matches S_{init} will provide an approximate representation of the initial solution. The use of the bootstrap procedure follows from the fact that the convergence properties of our approach are related to the growing percentage of matches which "vote" for a specific solution; thus, the bootstrap period encourages a smooth convergence from F_{init} during the initial steps of the refinement. Viewed from another angle, this means that without any other information related to F_{init}, the bootstrap creates the support set which is needed in order to compute σ adaptively across the image space.

One may argue that in the context of pose refinement, a constant $\sigma = \sigma_L$ would suffice. However, it is still advisable to use a variable σ parameter since the error introduced by the prior (e.g., the error on the tilting angle of a motorized surveillance camera) may be large enough in order to be impossible to sample correct observations; at the same time, the convergence should benefit from the adaptive σ in order to "follow" the pose variation as fast as possible.

1. Manual extraction of ground truth matches S_{gt}.
2. Robust RANSAC ($th = 1$) estimation of F matrix from the S_{gt}.
3. **if** inliers percentage $\geq \alpha$: stop.
4. Manual S_{gt} matches location refinement, from F matrix.
5. Go to step 2.

Outline 1: Ground truth extraction strategy

7 Ground-truth extraction

In order to perform a rigorous evaluation of the algorithm performance for real-world scenes of relevant size, we propose the construction of a manual ground truth which allows to characterize the quality of the solution by performing a local analysis across the whole scene. The main motivator for such ground-truth extraction comes from the observation that defining the error only at a global level may hide local high-error regions, which may be harmful when using the estimation for tasks such as detection, tracking or depth estimation.

Our methodology for building this accurate ground-truth data is the following (the main steps are provided in Outline 1). We define an uniform grid of buckets which provides a partition of the reference image, and we extract matches manually and uniformly inside the buckets belonging to the overlapping field of view. In order to enforce an uniform distribution of ground-truth points, to each bucket we assign a number M of matches, which is weighted by the portion of the bucket which belongs to the common field of view. Such extraction is essential in order to evaluate estimation errors even in regions where an automated process (followed by a manual validation) would not be able to identify meaningful and not degenerate interest points.

At the end of the uniform match extraction step, the measurement noise is too high due to human impreciseness, and occasional gross annotation errors may also occur. Thus, the procedure is followed by the robust estimation of a fundamental matrix from the current matches set, which is then used to refine the position of the generating matches, i.e., the human annotator is shown the annotations presenting high residuals in order to adjust them if necessary. The process is repeated iteratively, until we obtain a set of matches with half-pixel precision, which is at the same time large enough in order to guarantee a comprehensive evaluation of a candidate pose.

The error metrics we employ are the RMSE and the Max symmetric geometric error [22] 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 our estimation process, all the presented results are evaluated over 300 executions of each test.

Fig. 4 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

8 Results

We demonstrate the performance of our algorithm on three different datasets: *Regent's Park*, *PETS 2009*[14] and a laparoscope in-vivo procedure video provided by the Hamlyn Centre, Imperial College London [40]. The proposed algorithm is evaluated in the following contexts:

1. Section 8.1: We evaluate the pose estimation performance in a large scale, realistic scenario (outdoor scene, no volunteers).
2. Section 8.2: We evaluate the pose estimation performance on a less realistic but widely used dataset (volunteers following predetermined directions), as well as the pose refinement initialized with the ground truth provided with the dataset.
3. Section 8.2: We evaluate the pose refinement performance on a widely used, high-quality medical dataset in order to illustrate the versatile character of our algorithm.

Regarding the main parameters, we set for all the tests $\sigma_L = 1$, which is a common choice in guided matching covariance propagation methods [43]. The scale of σ_H depends on the matcher ease to associate features from the views, which is mainly reflected by the inlier set size, and by the inlier percentage of a robust estimation for a single frame (i.e., a small inlier set suggests an unstable estimation, and the value of σ_H should be set high enough in order to allow for a wider exploration). At the same time, small variations of the σ_H value have a negligible influence on the convergence behavior and on the final error. We set $\sigma_H = 5$ for all tests (with the exception of PETS 2009, see Sect. 8.2). 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 use as robust fundamental matrix estimator the ORSA [39] a-contrario framework, which exhibits good robustness without the need to set a sensitive threshold. Please note however that, while the robust method chosen has an influence on the final RMSE achieved, it has no effect on the actual convergence behavior of our approach; thus, other methods based on the popular RANSAC [7,51] may also be employed.

8.1 Pose estimation: Regent's Park dataset

The first part of our experiments is focused on estimating the relative pose in a realistic urban setup exhibiting typical challenges for this context.

8.1.1 Experimental setup

We test our method on synchronized sequences recorded at Regent's Park Mosque, London. The camera network consists of three cameras installed on the roof (see Fig. 4), labeled from 1 to 3. The analysis region is the rectangular-shaped inner courtyard (the *sahn*), surrounded traditionally by arcades and other repetitive structures on all sides. The video streams capture 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. The stream is sampled each 3 seconds (i.e., $\Delta_t = 24$ frames).

8.1.2 Experimental results

We start by highlighting in Fig. 5 the estimation errors obtained independently on single pairs of images extracted from the streams of cameras 1 and 2, with the ORSA estimator. For difficult scenes, the quality of the estimation is highly dependent on how the instantaneous configuration of the dynamic objects in the scene constrains the fundamental matrix, with large areas which may be left uncovered. In this specific case, the best achievable estimation has a maximum error of almost 4 pixels, which leaves room for a consistent improvement. Yet, the main underlying issue is that a single frame-based estimation would provide a result of arbitrary quality. We evaluated at this stage the method in [61], which

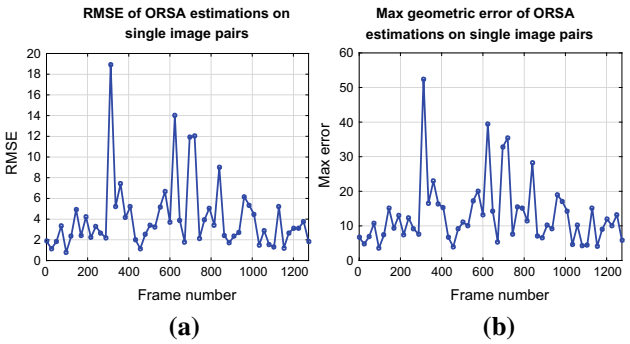

Fig. 5 RMSE and Max geometric error by applying on each frame pair independently ORSA. Large variations in the result demonstrate the unreliability of estimation with still images in such setup. Streams from cameras 1 and 2 are used

aims to extract matches iteratively from an image pair by enforcing spatial uniformity. This method fails to converge toward an acceptable solution (i.e., RMSE=245 for the first frame which was used for evaluation in Fig. 8) as it does not cope with such a wide baseline correlated to a strong depth variation of the scene.

Section 3 underlined 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. 6 show a typical unbalanced inlier configuration which promotes high errors locally, and underline the importance of using a video sequence in the case of wide baseline cameras and large-scale scenarios.

Figure 6 shows the inlier matches which are maintained after running an estimation of the fundamental matrix between frames at $t = 74$ of cameras 2 and 3. 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 as reliable for guiding the geometry estimation during the subsequent iteration. Then, Fig. 7a shows the spatial distribution of the symmetric geometric error on the left image. For each bucket of the image, we highlight the average error of the estimation with respect to the matches drawn from

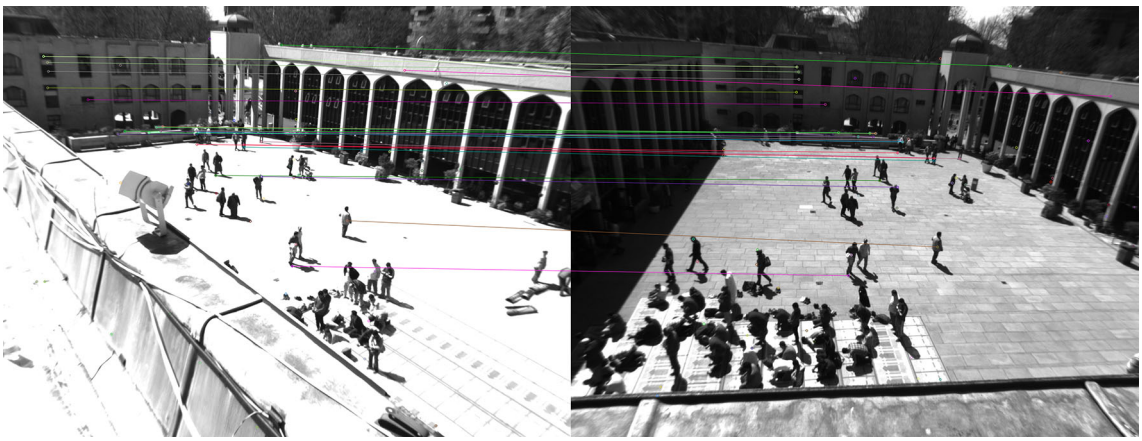

Fig. 6 Sample pair of frames ($t = 74$) exhibiting an unbalanced inlier coverage (it is advisable to zoom in the electronic version for inspecting the inlier matches)

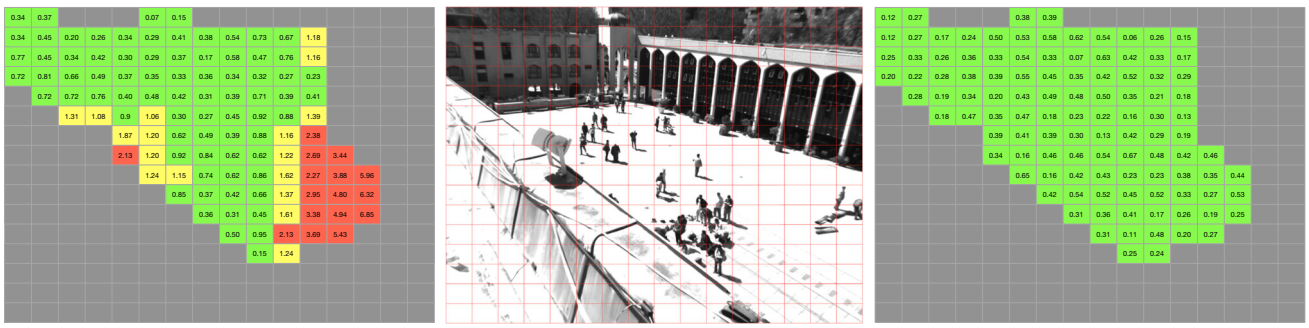

Fig. 7 Resulting spatial distribution of the symmetric geometric error with respect to a dense manually annotated ground truth. Errors less than 1 pixel are highlighted in green, between 1 and 2 pixels in yellow, and more than 2 pixels in red. **a** Average errors per bucket using the single image frame. **b** Reference frame subdivided in buckets. **c** Average errors per bucket using the proposed method

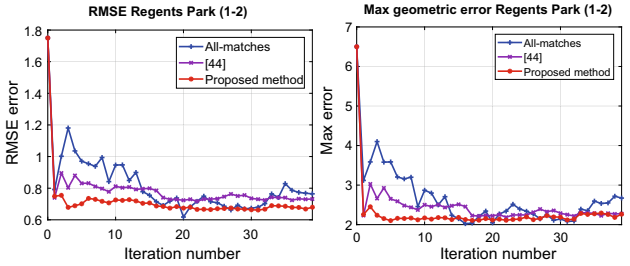

Fig. 8 RMSE and Max geometric error by applying the *All-matches* strategy, the method in [44] and our algorithm on 1–2 camera pair of *Regents Park* dataset. Our selection is more reliable, and we are able to improve the initial estimation significantly and robustly, with a lower RMSE and less oscillations than [44]

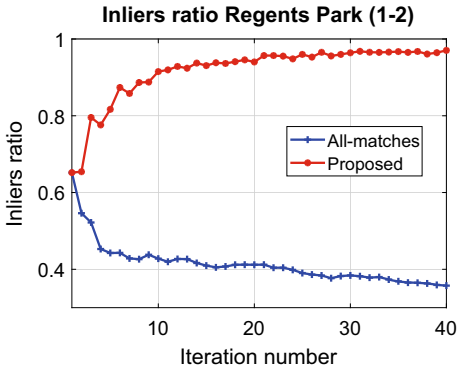

Fig. 9 The inliers ratio at each iteration for the *All-matches* and for our approach

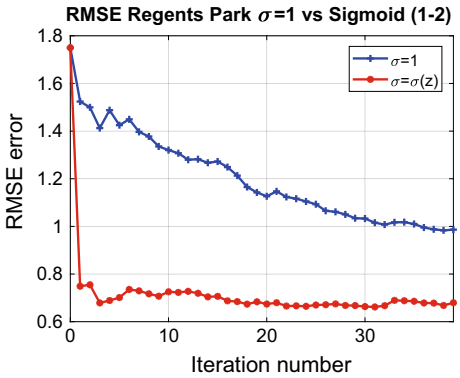

Fig. 10 RMSE by applying our method on the 1–2 camera pair by using a fixed $\sigma = \sigma_L = 1$ value, and by using the adaptive sigmoid shaped σ introduced in Sect. 5

the ground-truth points at that location. While approaching the area lacking inliers, we note the presence of high errors, which makes the single image pair approach unadapted for fitting the entire image space. While the overall RMSE=1.8 which is obtained from this estimation does not fully underline this major limitation, the Max geometric error of 7.53 reflects more accurately the local problems of the solution. This example also explains the significant variation, among different frames from the same video, in the quality of the estimation which depends significantly on how the dynamic elements are disposed spatially. Finally, Fig. 7c shows an example of the error distribution resulting from the proposed approach. The image shows a significant decrease in the error in areas which were challenging for single image pairs methods, but also a reduction in the error on a global scale. The overall RMSE for this example is 0.46, while the Maximum geometric error is 1.23.

Next, 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 this strategy *All-matches*). Figure 8 shows the RMSE and Max geometric errors at different iterations of the algorithms. Our method is able to reduce the RMSE from 1.75 to 0.66, and to consistently decrease the Max error from 6.5 to 2.2 pixels. We note the robustness of our strategy, with the error following a monotonic decreasing trend after a few iterations. Conversely, *All-matches* presents large oscillations in time, which implies that getting more points from the video stream will not improve definitely the batch estimation result, introducing thus a frame window-size choice problem. Our method also shows a smoother and faster convergence than our previous work [44] which sets the adaptive σ parameter by using a binary decision threshold on inlier clustering (final RMSE 0.75 compared to 0.66 for the current algorithm).

The explanation of the behavior of the *All-matches* approach comes from the analysis of the inlier ratios estimated at each iteration (Fig. 9). From the *All-matches* curve, we note that the inlier percentage obtained by accumulating matches drops monotonically. 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 rejection procedure depending on the current inlier configuration. Subsequently, the inlier ratio follows the opposite trend, since being increasingly confident in the current solution, and using lower σ values will improve the probability of including only inliers as new matches. Such trend explains the robust convergence of our approach.

Figure 10 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 $\sigma = \sigma_L = 1$, as in [43], we cannot add new information which is able to correct gross local errors in the estimation, leading to a much slower convergence which is never able to achieve performance, in terms of error, comparable to our strategy.

An important trait of an iterative pose estimation algorithm is its behavior in case of an adverse initialization. In Fig. 11, we show the RMSE and Max geometric error evolu-

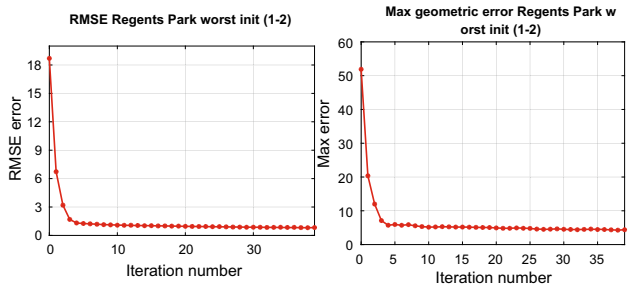

Fig. 11 RMSE and Max geometric error by applying our algorithm on the worst possible initialization of the 1–2 camera pair sequence (*Regents Park* dataset). Our estimation is cabable of successfully converging independently of the initialization chosen

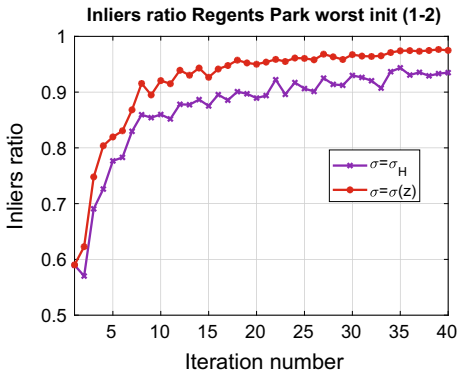

Fig. 12 The inliers ratio at each iteration on the worst initialization of the camera pair 1–2 sequence (*Regents Park* dataset) by using a fixed $\sigma = \sigma_H = 5$ or our adaptive sigmoid shaped σ introduced in Sect. 5

tion for the 1–2 pair when the most unfavorable initialization is selected (frame 312 in Fig. 5). The algorithm is still able to recover and to decrease the RMSE from 18.7 to 0.78 and the Max error from 52 to 4.1 pixels. This result demonstrates that the algorithm is able to converge to a stable, low-error solution, regardless of the starting point.

Then, we compare our adaptive σ solution with the use of a fixed $\sigma = \sigma_H$ for the band filtering step. Such an approach is more aggressive in the way it tries to add as many matches as possible by relaxing more the epipolar constraint. Although this strategy is able to achieve low errors occasionally, it does not trust the current solution locally more or less depending on the observations; this results in lower inlier ratios and a worse convergence stability. In Fig. 12, we compare the inlier ratios when we use the sigmoid function or the $\sigma = \sigma_H$, and it is clear how the use of the sigmoid is able to promote a stronger, smoother increase, especially noticeable at the last iterations. Table 1 summarizes how the sigmoid approach is capable of achieving low errors which are comparable with an aggressive solution, guaranteeing at the same time an inlier ratio up to 0.98.

Table 1 also shows the effect of the choice of the cross point density η in the performance of the algorithm. The

Table 1 RMSE, Max geometric error and inliers ratio on the worst initialization of camera pair 1–2 (*Regents Park* dataset) with different choices of the σ function and of the cross point density η ($n = 5$ is fixed for each selection of h)

σ	RMSE	Max error	Inliers ratio
σ_H	0.778	4.195	0.950
$\sigma(z), h = 30$	0.780	4.086	0.964
$\sigma(z), h = 60$	0.785	4.100	0.974
$\sigma(z), h = 100$	0.799	4.395	0.983

By using the sigmoid, the algorithm is capable of achieving comparable errors as an aggressive $\sigma = \sigma_H$ solution, at an higher inlier percentage

Table 2 RMSE, Max geometric error and inliers ratio on the worst initialization of camera pair 1–2 (*Regents Park* dataset) at constant cross point density η and different choices of the bandwidth h

$\eta = const$	RMSE	Max error	Inliers ratio
$n = 1, h = 2.836$	0.838	4.389	0.974
$n = 5, h = 60$	0.785	4.100	0.974
$n = 14, h = 100.39$	0.775	4.012	0.973
$n = 56, h = 200.79$	0.779	4.135	0.976

parameter η is expressed as the density of a desired number n of points in a h bandwidth. At a constant value of n, the higher the bandwidth h, the lower will be η. A lower η means that one gets confident sooner about the solution. This behavior is explained by the numbers in Table 1. Higher values of η/lower values of h show the smallest errors, while lower values of η/higher values of h present the best inlier ratios. Therefore, the η parameter represents how aggressive the algorithm is in terms of adding new points. However, as it may be noticed from the same table, different choices of η do not have an important impact on the convergence and on the overall goodness of the final solution, which is a desirable property when consistent results with effortless parameter tuning are needed.

Table 2 shows the effects of the choice of the bandwidth parameter h, when η is kept constant. Varying the bandwidth entails different choices of the n parameter, which, being an integer number of points, tunes the resolution at which the sigmoid function is sampled. A specific value of n involves, when using a histogram kernel, sampling the same sigmoid curve $(n + 1)$ times in the $[0, \eta]$ density interval, so higher the bandwidth, higher will be the sampling resolution. The first row of Table 2 corresponds to a binary selection of the σ value, equivalent to the one introduced in [44]. A significant error reduction is obtained by moving away from the binary representation of the inlier density. The table shows that increasing the resolution of the sigmoid has a benefit on the error levels while maintaining stable the inlier ratio. However, the error does not decrease monotonically as we increase h, because at the same time the density estima-

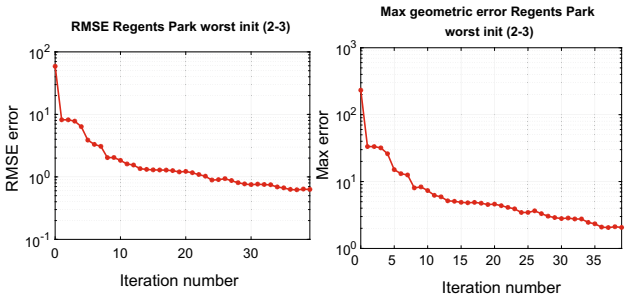

Fig. 13 RMSE and Max geometric error (in semilog scale) obtained by applying our method for the 2–3 camera pair (*Regents Park* dataset), with the worst possible initialization

Fig. 14 Sample frames from PETS 2009 dataset. **a** Camera 1, **b** Camera 3

tion loses its locality, providing inaccurate estimates of the boundaries between well-constrained and badly constrained regions. Overall, as in the case of the choice of η, the selection of the bandwidth h, while being critical in pure density estimation tasks [56], does not affect the convergence of the algorithm. By setting h in a reasonable range, on the basis of the image size, one gets the lowest estimation errors.

Finally, we show the estimation results for the camera pair 2–3 , using as starting point the worst possible initialization of the entire stream. Figure 13 shows again consistent results in terms of both RMSE and of Max error (curves are plotted in semilog scale for easier understanding). We are able to decrease the overall RMSE from 58.9 to 0.6 while reducing the Max error on the whole image space from 232.4 to 2 pixels.

8.2 Pose estimation versus pose refinement: PETS 2009

8.2.1 Experimental setup

PETS 2009[14] is a well-known and widely used dataset [6,11,57,70] which provides multi-sensor sequences of moving pedestrians for tracking [32,38,63,68,74], density estimation and counting [8,15,62], and event recognition [16,69]. The authors provide a full calibration of the system, which was performed using the Tsai calibration method [65]. From the calibration data, the ground-truth pose estimation may be represented in the form of a fundamental matrix F_{GT}. The image resolution is 768 $\times$ 576, and the videos are recorded at 7 fps. We consider for experiments the *City Center 12:34* sequence, which contains a moderate number of freely moving pedestrians (Fig. 14).

There are two main limitations of the provided geometry. First, the pose estimation is more accurate in the central part of the image which was covered comprehensively by the calibration procedure. This fact encourages the use of a limited area of interest for analysis which is more restrictive than the actual common field of view [37,46,66]. Secondly, the calibration allows for multiple camera data fusion at object level (mid-level) or trajectory level (high level). However, and also owing to synchronization issues, the calibration is not accurate enough in order to allow pixel/voxel level (low level) data fusion algorithms [12,24,45,55] to perform reliably due to significant pedestrian displacements [14,66].

Since synchronization errors are critical for pose estimation, we have manually inspected a subset of the sequence in order to evaluate the temporal displacement at each timestep based on the pedestrian precise limb arrangements. Figure 15 presents these displacements for the first 100 timesteps, and the values confirm that most frame pairs exhibit a slight lag, which is occasionally significant. We chose to run the proposed pose estimation algorithm on the raw data in order to evaluate the robustness to persistent desynchronization.

Finally, some additional factors worth noting and leading to a difficult pose estimation problem are the slight errors related to radial distortion which are noticeable on the borders, the photometric differences among the distinct types of camera sensors and the significant scale variations.

We apply the same procedure as presented in Sect. 7, by manually selecting and then refining matches only on accurately synchronized frames. To the extent of our knowledge, this is the first time for PETS 2009 that the accuracy of the provided ground truth is also evaluated quantitatively (the standard approach being the validation against the provided ground truth, i.e.,[23]).

8.2.2 Experimental results

In Fig. 16, we show the errors when performing a robust estimation with the ORSA algorithm on a single image pair. For most frames, we get extremely high RMSE values, which reflect how challenging the calibration procedure is in such scenario. For the PETS dataset, a $\sigma_H = 200$ has been used, an order of magnitude higher than in the *Regents Park* dataset case. The choice of such high σ_H comes directly from the observation of the number of inlier matches retained by the single pair estimation. At frame 0, for example, only nine inliers are maintained in the estimation, and this number is clearly insufficient in

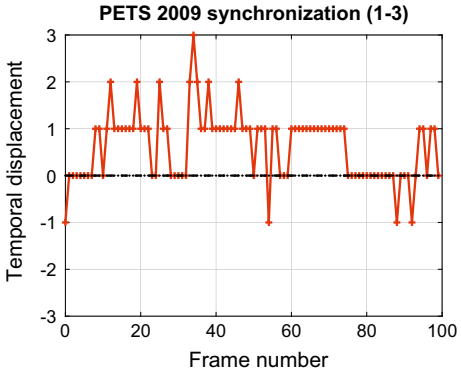

Fig. 15 Temporal displacement of the first 100 frames from view 3 with respect to the ones of view 1 of the *City Center 12:34* sequence (PETS 2009 dataset)

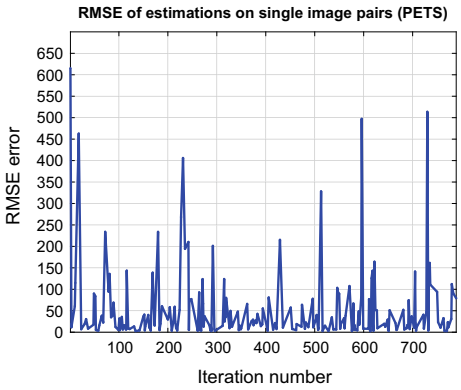

Fig. 16 RMSE by applying ORSA in each frame pair of the *City Center 12:34* (PETS 2009) independently. Streams from cameras 1 and 3 are used

Table 3 RMSE on F_{GT} and different initialization times t_0 of our algorithm on the PETS 2009 dataset

t_0	Init RMSE	RMSE
F_{GT}	–	2.58
$t_0 = 0$	621.88	3.32
$t_0 = 99$	6.05	3.02

The final error is always comparable with the ground truth one, while the initial RMSE does not affect the convergence of the solution to a close final error

order to represent a robust support set for the inferred pose.

Full FOV analysis Table 3 shows the RMSE of the ground truth provided pose F_{GT}, compared with that of our algorithm, at different initialization times, for the entire area which is visible from the two cameras. Comparing the solution directly with F_{GT}, without using the manual ground truth, would have hidden away the actual F_{GT} imprecision. The RMSE values obtained by running our algorithm directly on the video sequence are less than 1 pixel off compared to the errors of F_{GT} estimated using the Tsai

Table 4 RMSE and Max error on the F_{GT}, our algorithm at different initializations times t_0, and the provided pose refinement on the region of interest R0 of PETS 2009 dataset

t_0	Init RMSE	RMSE	Init max	Max	Inl. ratio
F_{GT}	–	1.14	–	2.65	–
$t_0 = 0$	612.45	1.14	2336.94	3.83	0.93
$t_0 = 99$	4.79	1.05	15.05	2.98	0.93
$F_{refined}$	–	0.83	–	2.08	0.97

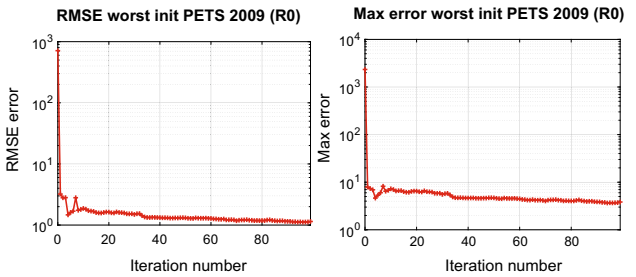

Fig. 17 RMSE and Max geometric error (in semilog scale) obtained by applying our method on region R0 (PETS 2009), with the worst possible initialization ($t_0 = 0$)

calibration. Moreover, for two different initialization times characterized by a low RMSE (6.05 pixels) and by the worst observable configuration (620.2 pixels), we note the minimal impact on the final convergence result. Regarding the Max error, the F_{GT} presents a 11.54 pixel error, while our method reaches Max error of 16.36 (starting from 3444.14) for $t_0 = 0$, and of 15.34 (starting from 17.97) for $t_0 = 99$.

AOI analysis First of all, the localization of the highest errors in the bottom left area of camera 1 suggests that border errors are less reliable for the analysis due to the impact of the image undistortion. More importantly, our method, while being able to decrease significantly the Max error, presents a higher Max final error than F_{GT} due to the fact that on the image borders no pedestrian action occurs (for the manual annotations, we used moving pedestrians from other sequences of the dataset in order to cover border areas). Thus, the lack of observations limits the algorithm to refining locally the solution. For the two reasons above, we consider a region of interest R0 on camera 1, which is defined as the moving pedestrian envelope and which allows us to provide an unbiased comparison in the actual analysis area used for the detection and tracking tasks (see Fig. 18 for the spatial extent of R0). Such area consists in all the walkway region, including also for completeness the area which is strongly cluttered by the tree in camera 3.

Table 4 shows the errors for the F_{GT} and our algorithm (at different initialization times) in the R0 region. Even when starting from an almost random initialization ($t_0 = 0$), our method is able to achieve the same RMSE as the F_{GT} (even

Fig. 18 Resulting spatial distribution of the symmetric geometric error with respect to a dense manually annotated ground truth (PETS 2009), in the region of interest R0 (colored buckets). Errors less than 1 pixel are highlighted in green, between 1 and 2 pixels in yellow, and more than 2 pixels in red. **a** Average errors per bucket using the provided F_{GT}. **b** Reference frame subdivided in buckets. **c** Average errors per bucket after executing the proposed refinement

slightly lower in the case of $t_0 = 99$). The Max error for the two solution is close to the F_{GT} one, showing that our method is able to provide a good-quality solution in the area of interest without relying on any calibration device, as in the F_{GT} case. Fig. 17 shows the error variation in time (both RMSE and Max) when we start from the worst possible initialization ($t_0 = 0$). The characterization of the algorithm behavior in such case is critical due to the use of a large value for σ_H, which, being more permissive, may introduce instabilities in the results. However, due to the use of the sigmoid, the algorithm is capable after a few steps to follow a smooth convergence, due to the gradual increase in confidence in the output solution at higher inliers densities.

Pose refinement Finally, we show the results obtained when refining an existing pose, which is F_{GT} in our case. The interest of pose refinement is that the estimation of F_{GT} has been carried out with helper objects, which may not cover the entire image space exhaustively. Starting from F_{GT}, we aim to refine the pose in the tracking region of interest R0, by including the rich visual information that is provided by the actual data.

Table 4 shows the RMSE and Max error of F_{GT} compared with $F_{refined}$, obtained by refining the provided pose on the entire *City Center 12:34* sequence. The $F_{refined}$ achieves a consistent improvement in both RMSE and Max error. In Fig. 18, it is possible to inspect the average errors for each bucket in R0. The $F_{refined}$ is able to reduce the estimation errors across almost all the discretized image space, and to reach an average error per bucket below 1 pixel, except only two buckets at 1.1 pixels.

8.3 Pose refinement: Hamlyn Centre laparoscopic/endoscopic video dataset

8.3.1 Experimental setup

The dataset [40] consists of multiple monocular and stereo medical video sequences which are widely used for validating a variety of applications such as Shape-from-Shading [67], surface reconstruction [26,34], deformable surface tracking [49,72,73] and SLAM [33,41,64]. For all sequences, the dataset maintainers provide high-quality intrinsic and extrinsic calibration information, estimated in the laboratory using a checkerboard helper object. For our experiments, we consider stereo data provided by a moving laparoscope visualizing an abdominal porcine wall (Dataset6). The image size is 640×480, and the video is recorded at 30 fps. We choose a sampling value $\Delta_t = 15$.

8.3.2 Experimental results

For the medical dataset, our objective is to refine the pose which was provided for the stereo rig, given that for stereo navigation or dense reconstruction algorithms any stereo calibration error weighs on the 3D estimations, since the stereo pose is assumed to be fixed.

The creation of a manually annotated ground truth for validating the pose is unfeasible in practice on these types of data due to the absence of highly salient small structures which are needed by a human subject. Thus, we demonstrate the interest of our refinement step using the live recorded data by showing some qualitative results on eight manually matched structures. The $\sigma_H = 5$ remains unchanged with respect to the *Regents Park* dataset tests. Figure 19 demonstrates the improvements in the proposed refined matrix on the test point selected in the image space. The red epipolar

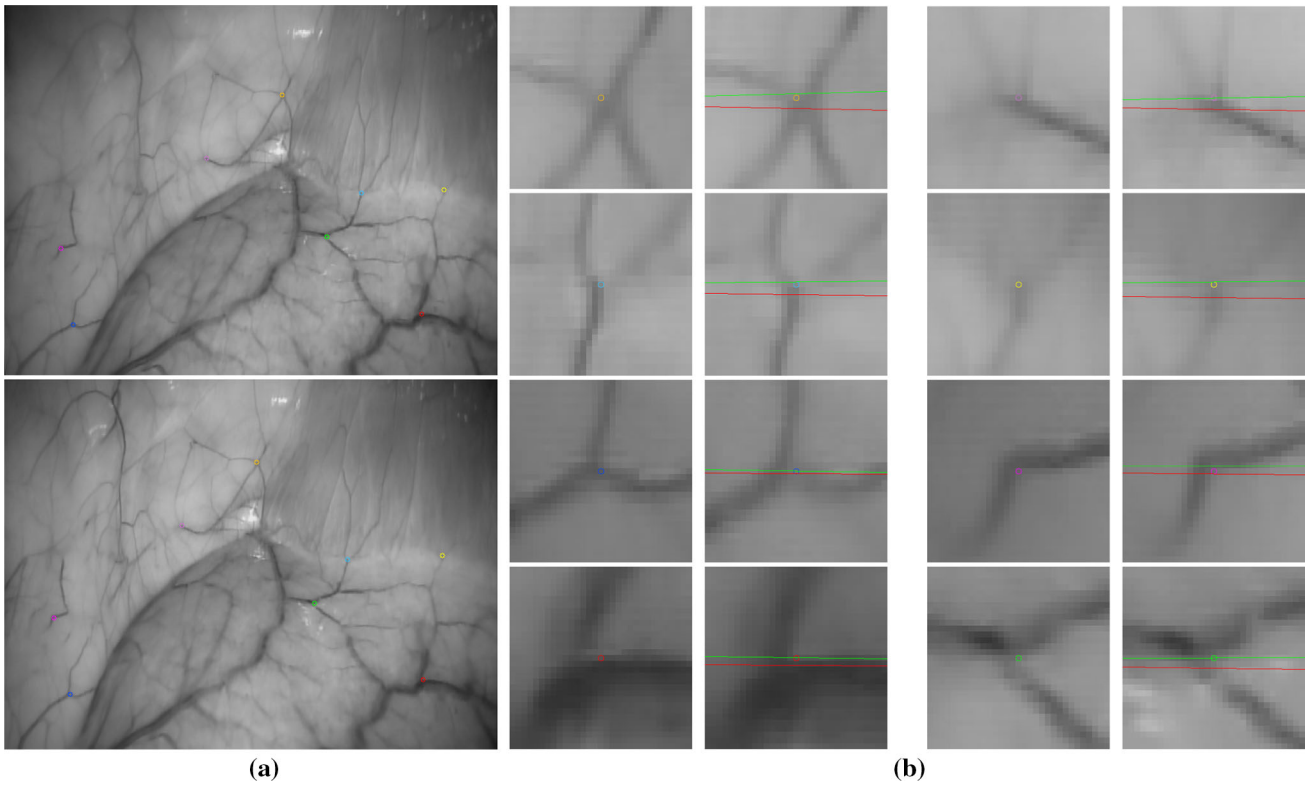

Fig. 19 Qualitative results obtained from the refinement of the provided pose of *Hamlyn Centre Laparoscopic/Endoscopic Video* dataset. **a** Stereo pair, with eight manually selected control points highlighted in different colors. **b** Zoomed views of the local patches around the control points, with two epipolar lines being drawn each time: the one from the provided F_{GT} (red) and the one from our refinement (green). A small, but noticeable displacement is present for F_{GT}; the proposed refinement is successful in removing it

line is drawn from the F_{GT} matrix provided by the dataset maintainers. While F_{GT} shows good performance in the left part of the space, it presents higher errors (up to 3 pixels on the test points) in some border regions of the image, especially in the right and top parts. The green epipolar line is drawn from the $F_{refined}$ matrix, which decreases the errors in the critical areas, while maintaining good performance in the parts which are already well covered (our solution achieves less than 0.5 pixels error in the test points).

Such refinement step has no additional cost in terms of data acquisition (the already available raw data can be used), and is capable of providing a better-quality calibration which is essential when applied to, e.g., 3D projection and reconstruction tasks.

8.4 Final overview

We summarize in Table 5 the performance of the proposed algorithm and of the existing approaches previously considered in this section. As baseline, RANSAC and ORSA are run on individual frame pairs and exhibit as expected a very unstable performance (columns 2 and 3). In these two cases, the RMSE, the Max geometric error and the inliers ratio are computed on the union of all the matches validated on the individual image pairs.

The methods which integrate temporally observations (columns 4–9) that we benchmark are: the RANSAC-based posed estimator with temporal accumulation of matches, the ORSA-based posed estimator with temporal accumulation of matches, the density-based accumulation of [44] with conservative accumulation ($\sigma = 1$), the density-based accumulation of [44] with permissive accumulation ($\sigma = 5$), the RANSAC-based proposed method, and the ORSA-based proposed method. We evaluate the performance in an identical manner to the individual pair based estimators, except that we allow the methods to convergence for 10 frames before considering the algorithm outputs. As it may be seen in the detailed convergence analysis plots, the number of frames allowed for convergence does not have a significant influence as long it is not extremely small (i.e., less than three frames).

In terms of robustness to outliers, the results show that ORSA outlier rejection outperforms systematically the standard RANSAC-based outlier rejection scheme in every scenario (columns 2 vs. 3, 4 vs. 5 and 8 vs. 9). In terms of overall strategy, the baseline accumulation of all the matches

Table 5 Overall comparison for the Regents Park data (camera pair 1–2)

Regents park performance	Frame estimators		Video estimators					
	RS	OR	RS-all	OR-all	[44] $\sigma = 1$ OR	[44] $\sigma = 5$ OR	Ours RS	Ours OR
RMSE	4.8	1.75	2.2	0.78	0.99	0.75	1.17	**0.66**
MaxE	15.4	6.5	6.9	2.8	3.5	**2.2**	3.9	**2.2**
Inl.Rat.	0.45	0.65	0.31	0.36	**0.98**	0.95	0.95	**0.98**

The meaning of each performance measure is recalled in the text (Sect. 8.4). RS and OR correspond to algorithms relying for outlier rejection on RANSAC and ORSA, respectively

accepted by the outlier rejection scheme (RS-all and OR-all) is clearly inferior to the methods which take into account the density of the observations, the proposed method with ORSA exhibiting the best performance for all three main indicators considered here. Note that the convergence rate of the methods does not emerge from Table 5, but in this aspect as well the proposed method exhibits stability after a fast convergence period (see for example Fig. 13).

In terms of reproducibility, the open-source code provided includes the proposed algorithm along with implementations of the other ones considered above. The PETS 2009 dataset is freely available, and although the Regents Park dataset cannot be freely distributed, a sample subset is provided with our source code, and the entire dataset may be shared on an individual basis.

9 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 a fast convergence toward a high-quality solution, which is being highlighted by the manual ground truth we created for two difficult scenes. In our experiments, we show that this video accumulation strategy converges robustly to globally effective pose estimations, irrespective of the scene configuration during initialization. We have also proposed an extension able to perform data-driven pose refinement based on a prior pose initialization, and which is aimed at stereo systems requiring frequent high-quality extrinsic re-calibrations. During experiments, our self-calibration procedure was able to improve consistently the prior pose with no overhead in terms of data acquisition procedures.

In our future work, we are interested in integrating our pose estimation procedure to a multiple camera alignment algorithm, as well as in exploiting relative positioning cues from additional sensors in the computation of the pose uncertainty.

Acknowledgements The authors gratefully acknowledge the support of Regent's Park Mosque for providing access to the site during data collection, and of 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. Ataer-Cansizoglu, E., Taguchi, Y., Ramalingam, S., Miki, Y.: Calibration of non-overlapping cameras using an external slam system. In: 2nd International Conference on 3D Vision (3DV), vol. 1, pp. 509–516. IEEE (2014)
2. Ayaz, S.M., Kim, M.Y., Park, J.: Survey on zoom-lens calibration methods and techniques. Mach. Vis. Appl. **28**(8), 803–818 (2017)
3. Boutros, N., Shortis, M.R., Harvey, E.S.: A comparison of calibration methods and system configurations of underwater stereo-video systems for applications in marine ecology. Limnol. Oceanogr. Methods **13**(5), 224–236 (2015)
4. Brückner, M., Bajramovic, F., Denzler, J.: Intrinsic and extrinsic active self-calibration of multi-camera systems. Mach. Vis. Appl. **25**(2), 389–403 (2014)
5. Caspi, Y., Simakov, D., Irani, M.: Feature-based sequence-to-sequence matching. Int. J. Comp. Vis. **68**(1), 53–64 (2006)
6. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. **117**(6), 633–659 (2013)
7. Chum, O., Matas, J.: Matching with prosac-progressive sample consensus. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 220–226. IEEE (2005)
8. Conte, D., Foggia, P., Percannella, G., Vento, M.: Counting moving persons in crowded scenes. Mach. Vis. Appl. **24**(5), 1029–1042 (2013)
9. Dang, T., Hoffmann, C., Stiller, C.: Continuous stereo self-calibration by camera parameter tracking. IEEE Trans. Image Process. **18**(7), 1536–1550 (2009)
10. Devarajan, D., Radke, R.J., Chung, H.: Distributed metric calibration of ad hoc camera networks. ACM Trans. Sensor Netw. (TOSN) **2**(3), 380–403 (2006)
11. Dubuisson, S., Gonzales, C.: A survey of datasets for visual tracking. Mach. Vis. Appl. **27**(1), 23–52 (2016)
12. Eshel, R., Moses, Y.: Tracking in a dense crowd using multiple cameras. Int. J. Comput. Vis. **88**(1), 129–143 (2010). https://doi.org/10.1007/s11263-009-0307-0
13. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd **96**, 226–231 (1996)
14. Ferryman, J., Shahrokni, A.: Pets2009: Dataset and challenge. In: 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009, pp. 1–6. IEEE (2009)

15. Foroughi, H., Ray, N., Zhang, H.: Robust people counting using sparse representation and random projection. Pattern Recognit. **48**(10), 3038–3052 (2015)
16. Fradi, H., Luvison, B., Pham, Q.C.: Crowd behavior analysis using local mid-level visual descriptors. IEEE Trans. Circuits Syst. Video Technol. **27**(3), 589–602 (2017). https://doi.org/10.1109/TCSVT.2016.2615443
17. Fraundorfer, F., Tanskanen, P., Pollefeys, M.: A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. Comput. Vis.-ECCV **2010**, 269–282 (2010)
18. Gemeiner, P., Micusik, B., Pflugfelder, R.: Calibration Methodology for Distant Surveillance Cameras, pp. 162–173. Springer, Cham (2015)
19. Goldman, Y., Rivlin, E., Shimshoni, I.: Robust epipolar geometry estimation using noisy pose priors. Image Vis. Comput. **67**, 16–28 (2017)
20. Guo, X., Cao, X.: Triangle-constraint for finding more good features. In: International Conference on Pattern Recognition (ICPR), pp. 1393–1396 (2010)
21. Hansen, P., Alismail, H., Rander, P., Browning, B.: Online continuous stereo extrinsic parameter estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1059–1066. IEEE (2012)
22. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
23. Kasten, Y., Ben-Artzi, G., Peleg, S., Werman, M.: Fundamental matrices from moving objects using line motion barcodes. In: European Conference on Computer Vision, pp. 220–228. Springer (2016)
24. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 505–519 (2009)
25. Kneip, L., Chli, M., Siegwart, R.Y.: Robust real-time visual odometry with a single camera and an IMU. In: Proceedings of the British Machine Vision Conference 2011. British Machine Vision Association (2011)
26. Lin, B., Johnson, A., Qian, X., Sanchez, J., Sun, Y.: Simultaneous tracking, 3d reconstruction and deforming point detection for stereoscope guided surgery. In: Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions, pp. 35–44. Springer (2013)
27. Lin, W.Y., Cheong, L.F., Tan, P., Dong, G., Liu, S.: Simultaneous camera pose and correspondence estimation with motion coherence. Int. J. Comput. Vis. **96**(2), 145–161 (2012)
28. Lin, W.Y., Liu, S., Jiang, N., Do, M.N., Tan, P., Lu, J.: Repmatch: Robust feature matching and pose for reconstructing modern cities. In: European Conference on Computer Vision, pp. 562–579. Springer (2016)
29. Ling, Y., Shen, S.: High-precision online markerless stereo extrinsic calibration. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 1771–1778. IEEE (2016)
30. Liu, Z., Monasse, P., Marlet, R.: Match selection and refinement for highly accurate two-view structure from motion. In: European Conference on Computer Vision, pp. 818–833. Springer (2014)
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comp. Vis. **60**(2), 91–110 (2004)
32. Madrigal, F., Hayet, J.B., Rivera, M.: Motion priors for multiple target visual tracking. Mach. Vis. Appl. **26**(2–3), 141–160 (2015)
33. Mahmoud, N., Hostettler, A., Collins, T., Soler, L., Doignon, C., Montiel, J.M.M.: SLAM based quasi dense reconstruction for minimally invasive surgery scenes. ICRA 2017 workshop C4 Surgical Robots: Compliant, Continuum, Cognitive, and Collaborative (2017)
34. Maier-Hein, L., Groch, A., Bartoli, A., Bodenstedt, S., Boissonnat, G., Chang, P.L., Clancy, N., Elson, D.S., Haase, S., Heim, E., et al.: Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. IEEE Trans. Med. Imaging **33**(10), 1913–1930 (2014)
35. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR'07, pp. 1–8. IEEE (2007)
36. Mavrinac, A., Chen, X.: Modeling coverage in camera networks: a survey. Int. J. Comput. Vis. **101**(1), 205–226 (2013)
37. Mehmood, M.O., Ambellouis, S., Achard, C.: Ghost pruning for people localization in overlapping multicamera systems. In: International Conference on Computer Vision Theory and Applications (VISAPP), 2014, vol. 2, pp. 632–639. IEEE (2014)
38. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**(1), 58–72 (2014)
39. Moisan, L., Stival, B.: A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. Int. J. Comp. Vis. **57**(3), 201–218 (2004)
40. Mountney, P., Stoyanov, D., Yang, G.Z.: Three-dimensional tissue deformation recovery and tracking. IEEE Signal Process. Mag. **27**(4), 14–24 (2010)
41. Mountney, P., Yang, G.Z.: Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009, EMBC 2009, pp. 1184–1187. IEEE (2009)
42. Mueller, G.R., Wuensche, H.J.: Continuous extrinsic online calibration for stereo cameras. In: Intelligent Vehicles Symposium (IV), 2016 IEEE, pp. 966–971. IEEE (2016)
43. Ochoa, B., Belongie, S.: Covariance propagation for guided matching. In: Workshop on Statistical Methods in Multi-Image and Video Processing (2006)
44. Pellicano, N., Aldea, E., Le Hégarat-Mascle, S.: Robust wide baseline pose estimation from video. In: 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3820–3825. IEEE (2016)
45. Pellicanò, N., Aldea, E., Le Hégarat-Mascle, S.: Geometry-based multiple camera head detection in dense crowds. In: Proceedings of the 28th British Machine Vision Conference (BMVC)—5th Activity Monitoring by Multiple Distributed Sensing Workshop (2017)
46. Peng, P., Tian, Y., Wang, Y., Li, J., Huang, T.: Robust multiple cameras pedestrian detection with multi-view bayesian network. Pattern Recognit. **48**(5), 1760–1772 (2015)
47. Pollefeys, M., Koch, R., Van Gool, L.: Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. Int. J. Comput. Vis. **32**(1), 7–25 (1999)
48. Pollok, T., Monari, E.: A visual slam-based approach for calibration of distributed camera networks. In: 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23–26, 2016, pp. 429–437 (2016). https://doi.org/10.1109/AVSS.2016.7738081
49. Puig, L., Daniilidis, K.: Monocular 3d tracking of deformable surfaces. In: IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 580–586. IEEE (2016)
50. Radke, R.J.: A survey of distributed computer vision algorithms. Handbook of Ambient Intelligence and Smart Environments pp. 35–55 (2010)
51. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: Usac: a universal framework for random sample consensus. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 2022–2038 (2013)
52. Ravichandran, A., Vidal, R.: Video registration using dynamic textures. Patt. Anal. Mach. Intell. **33**(1), 158–171 (2011)

53. Remondino, F., Fraser, C.: Digital camera calibration methods: considerations and comparisons. Int. Arch. Photogr. Rem. Sens. Spat. Inf. Sci. **36**(5), 266–272 (2006)
54. SanMiguel, J.C., Micheloni, C., Shoop, K., Foresti, G.L., Cavallaro, A.: Self-reconfigurable smart camera networks. IEEE Comput. **47**(5), 67–73 (2014)
55. Sekii, T.: Robust, real-time 3d tracking of multiple objects with similar appearances. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4275–4283 (2016)
56. Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B (Methodol.) **53**(3), 683–690 (1991)
57. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442–1468 (2014)
58. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. Int. J. Comp. Vis. **80**(2), 189–210 (2008)
59. STEREOLABS: ZED Stereo Camera (2018). https://www.stereolabs.com/
60. Sur, F., Noury, N., Berger, M.O.: Computing the uncertainty of the 8 point algorithm for fundamental matrix estimation. In: 19th British Machine Vision Conference-BMVC 2008, p. 10 (2008)
61. Tan, X., Sun, C., Sirault, X., Furbank, R., Pham, T.D.: Feature matching in stereo images encouraging uniform spatial distribution. Pattern Recognit. **48**(8), 2530–2542 (2015)
62. Tang, N.C., Lin, Y.Y., Weng, M.F., Liao, H.Y.M.: Cross-camera knowledge transfer for multiview people counting. IEEE Trans. Image Process. **24**(1), 80–93 (2015)
63. Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., Schiele, B.: Learning people detectors for tracking in crowded scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1049–1056 (2013)
64. Totz, J., Mountney, P., Stoyanov, D., Yang, G.Z.: Dense surface reconstruction for enhanced navigation in mis. Med. Image Comput. Comput.-Assist. Interv.-MICCAI **2011**, 89–96 (2011)
65. Tsai, R.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J. Robot. Autom. **3**(4), 323–344 (1987)
66. Utasi, Á., Benedek, C.: A bayesian approach on people localization in multicamera systems. IEEE Trans. Circuits Syst. Video Technol. **23**(1), 105–115 (2013)
67. Visentini-Scarzanella, M., Stoyanov, D., Yang, G.Z.: Metric depth recovery from monocular images using shape-from-shading and specularities. In: 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 25–28. IEEE (2012)
68. Wang, B., Wang, G., Chan, K.L., Wang, L.: Tracklet association by online target-specific metric learning and coherent dynamics estimation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(3), 589–602 (2017)
69. Wu, S., Wong, H.S., Yu, Z.: A bayesian model for crowd escape behavior detection. IEEE Trans. Circuits Syst. Video Technol. **24**(1), 85–98 (2014)
70. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1834–1848 (2015)
71. Xiao, C.B., Feng, D.Z., Yuan, M.D.: An efficient fundamental matrix estimation method for wide baseline images. Pattern Analysis and Applications pp. 1–10 (2016)
72. Ye, M., Giannarou, S., Meining, A., Yang, G.Z.: Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. Med. Image Anal. **30**, 144–157 (2016)
73. Ye, M., Giannarou, S., Patel, N., Teare, J., Yang, G.Z.: Pathological site retargeting under tissue deformation using geometrical association and tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 67–74. Springer (2013)
74. Zamir, A.R., Dehghan, A., Shah, M.: Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In: Computer Vision–ECCV 2012, pp. 343–356. Springer (2012)
75. Zhang, Z.: Determining the epipolar geometry and its uncertainty: a review. Int. J. Comp. Vis. **27**(2), 161–195 (1998)

Nicola Pellicanò is an applied research scientist at Samsung AI Labs, in Paris. He received his BS and MS degrees in Computer Engineering from University of Pavia (Italy) in 2013 and 2015, respectively, and his PhD in computer vision and image processing from University Paris-Sud (France) in 2018. His research interests include video scene analysis and understanding and data fusion.

Emanuel Aldea is associate professor at the Paris-Sud University. He received his BS and MS degrees in Computer Science from Ecole Polytechnique in 2005 and from Paris 6 University in 2006 respectively, and his PhD in image processing from Télécom ParisTech in 2009. His current research interests include image processing and robotic vision for autonomous systems.

Sylvie Le Hégarat-Mascle received the Ph.D. degree in signal and image from Telecom ParisTech, in 1996, and the HDR degree from Versailles University (France) in 2006. She is currently a Professor of image processing at Paris-Sud University, Orsay (France). Her research activities focus on statistical pattern recognition (gestalt and structure detection), image analysis (classification, change detection), and data fusion using belief function theory. Application interests are for remote sensing and video scene analysis and understanding.