Multiple partial solutions for the point-to-point correspondence problem in three views

Miguel Carrasco
Escuela de Informática y Telecomunicaciones,
Universidad Diego Portales, Vergara 432. Santiago, CHILE.
Email: miguel.carrasco@udp.cl

Abstract—**This paper presents a method of point-to-point correspondence analysis based on a combination of two techniques: (1) correspondence of multiple points through similarity of invariant features in three views by using a standard feature method, and (2) a combination of multiple partial solutions through the trifocal geometry. This method allows the determination of point-to-point geometric correspondence by means of the intersection of multiple partial solutions that are weighted through the MLESAC algorithm. The main advantage of our method is the extension of the algorithms based on the correspondence of invariant descriptors, generalizing the problem of correspondence to a geometric model in multiple views. For all the images analyzed, we showed that the point-to-point correspondence can be generated through a multiple geometric relation between three views. An important characteristic of our method is that can be used in sequences of images that have a low signal-to-noise ratio.**

I. Introduction

Correspondence analysis consists basically in determining a set of points in an image such that they are identified as the same in other images of the same scene. Determining these correspondences computationally is not a simple task. We must consider that corresponding points can undergo various transformations depending on the points of view from which they have been captured. This is due to the geometric and/or photometric transformations caused by the motion of the object as well as by the movement of the camera with respect to the object. To increase the problem's complexity, it is possible for other points to have a texture and/or color similar to that of the point whose correspondence we want to determine, increasing the complexity of the task of discriminating among possible corresponding pairs and triplets.

Different approaches for solving the correspondence matching have been developed over the last 30 years. Some of them are, for example, methods based on the analysis of invariant descriptors [1], estimation of affine transformations, homographies and estimation of perspective transformations [2], epipolar geometry analysis [3], and methods based on optical flow [4]. In general, all these methods differ in the type of motion of the objects contained in a video sequence, or in the simplest case through correspondence between two images. If the scene is static and there is no continuous change of the camera's position, the problem is reduced mainly to the analysis of the epipolar geometry for two images through stereo vision. Unfortunately, none of the above methods filter out wrong matchings in multiple views. To avoid this problem, here we propose a method to determine the point-to-point correspondence in sequences of three images through multiple partial solutions. Since we use a geometric model that is independent of the objects, it is possible to determine the position of corresponding points in those views in which the point may be occluded. In the following sections we detail our methodology.

II. Proposed methodology

The first task to solve is to find a set correspondences between two images; here we used the method described in [6]. The main idea was to model the error through the robust estimation of the MLESAC algorithm considering a multiple epipolar lines analysis. In this research we propose a similar analysis but considering three corresponding views.

A. Tracking in three views

The analysis in three views allows modeling all the geometric relations generated in the 3D space, regardless of the structure contained in each image [5], [7]. One of the great advantages of geometric modeling in three views, and particularly of the estimation of the *Trifocal tensor* [5], is that depends only on the motion between the views and on the internal parameters of the cameras. Additionally, it can be completely defined by the projection matrices of which it is composed. Therefore, our analysis is based on how to estimate the error of the projection matrices from the set of correspondences in three views.

Formally, the trifocal tensor[1] $\mathbf{T} = (T_t^{rs})$ is a $3\times3\times3$ matrix that codes the relative motion between the $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3$ views. As already mentioned, one of its most relevant properties is that from the estimation of the tensor we can determine the position of a point $\mathbf{s}$ in the $\mathbf{I}_3$ plane using the positions of the correspondences $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ of the first and second views, respectively. The re-projection is defined in terms of the $\mathbf{r} = [x_1, y_1, 1]^\top$, $\mathbf{m} = [x_2, y_2, 1]^\top$ positions in homogeneous coordinates and of the tensor $\mathbf{T}$, derived from the first two trilinearities of Shashua [8]. In particular, we use the re-projection by means of the point-line-point method proposed by [5, pp.373]. For that purpose, let $\hat{\mathbf{s}}$ be the re-projection of the trifocal tensor in the third view, defined as $\hat{\mathbf{s}} = [x_3, y_3, 1]^\top$.

[1] See Hartley and Zisserman [5] for details on the computation of the trifocal tensors.

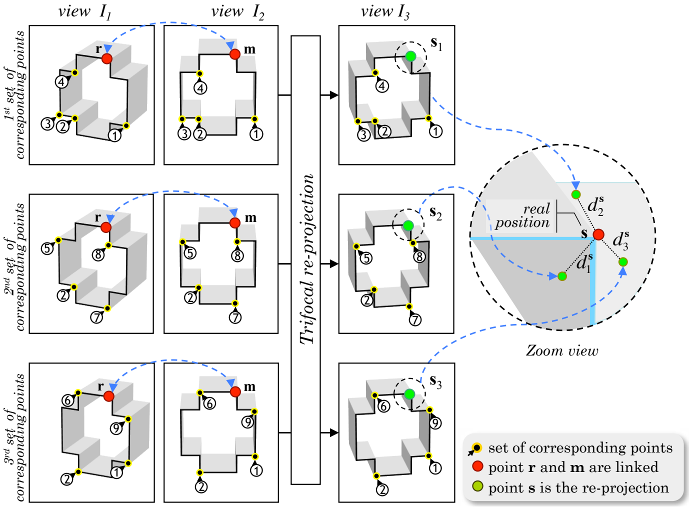

Fig. 1. Reprojection process of hypothetical points using a third view through the trifocal tensors. In the example, the projection of point $\hat{\mathbf{s}}_1$ was determined with the correspondences {1,2,3,4}, that of point $\hat{\mathbf{s}}_2$ with correspondences {2,5,7,8}, and that of point $\hat{\mathbf{s}}_3$ with correspondences {1,2,6,9}.

Unfortunately, the estimation of the projected point $\hat{\mathbf{s}}$ is subject to an error that can be generated for two reasons: (1) The intrinsic error in the estimation of the tensors due to an incorrect choice of the set of correspondences, and (2) the correspondence error between the $\mathbf{r}$ and $\mathbf{m}$ pairs. Even in the ideal case, when the tensors are relatively stable in uncalibrated sequences, there is always an error between the hypothetical correspondence $\mathbf{s}$ and the reprojected point $\hat{\mathbf{s}}$. For simplicity, we assume that the distance between these points is the Euclidian distance $d^{\mathbf{s}}$ of point $\mathbf{s}$, defined as $d^{\mathbf{s}} = \|\hat{\mathbf{s}} - \mathbf{s}\|$. Each subset generates a re-projection in the $\hat{\mathbf{s}}_1$, $\hat{\mathbf{s}}_2$, $\hat{\mathbf{s}}_3$ positions, corresponding to the re-projection of the tensor in the third view, as shown in Fig.1.

Extending the above example, let $i \in [1, \ldots, k]$ be the number of correspondences used. In this way, from each i subset it is possible to estimate the trifocal tensor $\mathbf{T}_i$. Each tensor is unique and independent of the previous one, provided the selection of the subsets is different. Assuming independence between i sets, let $\hat{\mathbf{s}}_i$ be the re-projection of the tensor $\mathbf{T}_i$ generated from the re-projection of the pair of points in correspondence $\mathbf{r}$ and $\mathbf{m}$ (or $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$).

B. Error compensation by MLESAC algorithm

MLESAC is a robust estimation algorithm to establish point correspondences in multiple views [9], generalizing the RANSAC estimator [10]. In our proposal, MLESAC is an intermediate step in the error estimating process that allows for weighting of individual errors. One of the main advantages of estimating the error in this way is that the inliers, or correct correspondences, have a high weight, in contrast with the RANSAC algorithm, in which only the outliers are considered in the cost function. MLESAC is designed considering that the error L_i is a mixture of Gaussian and uniform distribution, where $d_i^{\mathbf{s}}$ is the error of the estimation of the trifocal tensor, for all $i \in [1, \ldots, k]$ subset such that

$$L_i = \left(\gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp \left(- \frac{(d_i)^2}{2\sigma^2} \right) + (1-\gamma) \frac{1}{\nu} \right) \qquad (1)$$

where γ is a mixing parameter, ν is an a priori diameter of the search window used to handle false matches, and σ is the standard deviation of the error in each coordinate. Parameters γ and ν are not known, but they can be estimated by means of the EM [11] algorithm (for more details see [9]). Thus, the EM algorithm estimates the parameters and the probability that a putative selection will be an inlier or an outlier. In this way, the objective function is to minimize the log-likelihood of the error, which in our case is the distance $d_i^{\mathbf{s}}$ between a subset of trifocal re-projection (Fig.1). Normally three iterations are needed for the algorithm to converge. Recall that MLESAC uses the random selection of random solutions. In this way the estimation of the log-likelihood of the i-th hypothesis of each partial solution allows us to weight correctly the real distance d_i. In order to perform this task, we re-scale the values contained in the log-likelihood vector L for all $i \in [1, \ldots, k]$, as:

$$S = |max(L) - L_i| + 1, \qquad (2)$$

where S is a vector that give more relevance to the lower values of the log-likelihood vector L. For instance, when L_i is a maximum, the result is one. Conversely, when L_i is a minimum, the result is a maximum. Partial log-likelihood (L_i) values are used in this estimate, so that the $d_i^{\mathbf{s}}$ distance is weighted according to:

$$\tilde{d_i^{\mathbf{s}}} = d_i^{\mathbf{s}} \left(\frac{S_i}{\sum_{i=1}^{k} (S_i)} \right), \qquad (3)$$

where $\tilde{d_i^{\mathbf{s}}}$ is a weighted distance that considers the error associated of each i trifocal tensor. This procedure allows weighting and reestimating the distance of each tensor according to the log-likelihood of the projection error with respect to the set of hypothetical points in the third view. The estimation of the error allows weighting correctly the distance $d_i^{\mathbf{s}}$, increasing or decreasing it according to the size of its error. Therefore, to determine a correspondence, we determine the distance with respect to the set Θ. Finally, to identify the correspondence of pairs $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ in the third view, the following relation must be satisfied:

$$\{\mathbf{r} \leftrightarrow \mathbf{m} \leftrightarrow *\} = \begin{cases} \mathbf{s_i} & \text{if } \tilde{d_i^{\mathbf{s}}} < \epsilon \\ \nexists & \text{if } \tilde{d_i^{\mathbf{s}}} \geq \epsilon \end{cases} \qquad (4)$$

where $\{*\} \in \Theta = \{\mathbf{s_1}, \mathbf{s_2}, \mathbf{s_3}\}$, and ϵ is a distance measured in pixels. The final result allows the determination of which points are corresponding and which, depending on a threshold level, must be discarded. A complete description of the proposed methodology is showed in Algorithm 1.

Algorithm 1 : *Trifocal Geometric Correspondence* (TRIGC) algorithm in three views.

```
Require: Set of matchings candidates in three views.
Ensure: Set of wrong matchings filtered out in three views.
1: Determine $n$ corresponding points in three views. These triplets are known or estimated in a process that can be off-line, or automatic by means of the analysis of correspondences; for example, by SIFT [1] or SURF [12].
2: Determine pairs of point-to-point correspondence in the first and second view (e.g. BIGC [6]).
3: Determine $i$ trifocal tensors $\mathbf{T}_i$, where $i \in [1, \ldots, k]$. Each $i$-subset is composed by multiple corresponding points depending of the algorithm used to estimate the trifocal tensor.
4: Determine the re-projection of the trifocal tensor $\mathbf{T}_i$ for each pair corresponding to step 2.
5: Determine the error associated with each trifocal tensor with the MLESAC algorithm and re-estimate the distance $\tilde{d}_i^{\mathbf{s}}$ between the hypothetical correspondence and the projected position.
6: Assign the correspondence with point $\mathbf{s}$ provided that the $\tilde{d}_i^{\mathbf{m}} < \epsilon$ restriction is fulfilled for every pair $\{\mathbf{r} \mapsto \mathbf{m}\}$.
```

III. Experimental Results

This section presents the experimental results generated with sequences of uncalibrated images in three views. A set of 120 images of bottle necks with manufacturing faults generated in [13] (Fig. 2) is used. In all the experiments we have considered two standard indicators [14]: $r = \frac{\text{TP}}{\text{TP+FN}}$ (recall) and $p = \frac{\text{TP}}{\text{TP+FP}}$ (precision). TP is the number of *true positives* or correctly classified correspondences. FN is the number of *false negatives* or real correspondences not detected by our algorithm. FP is the number of *false positives* or correspondences classified incorrectly. These two indicators can be joined in a single measure F-score $= \frac{2 \cdot p \cdot r}{p + r}$ [14]. Ideally, one can expect that $r = 100\%, p = 100\%$, and F-score $= 1$.

Next, we evaluated the influence of parameter $i \in [1, \ldots, k]$ when the number of solutions of the proposed method is varied. In the same way we evaluated the influence of the Euclidian distance ϵ. We recall that both parameters can be modified in combination. For that we separated the analysis varying each of them independently. Recall that the variations of parameter i increase the number of trifocal tensors for three views. Also, parameter ϵ determines the Euclidian distance between the re-projection of the trifocal tensor and the hypothetical correspondence (see Fig. 1). As we already established, the determination of a new solution of the geometric problem in three views implies the re-projection of a new geometric solution, restricting the search space for a correspondence. In the latter case we must consider that the re-projection requires correspondence in the first two views to determine the reprojection in the third view, making it necessary to have three corresponding points in three images. In all the experiments we considered the average performance of the set of images. In the following subsections we will detail these aspects.

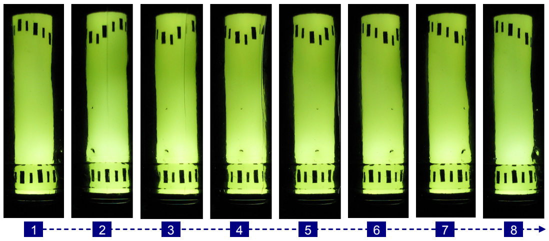

Fig. 2. Sequence of images of bottle necks for the tracking process as a quality control method.

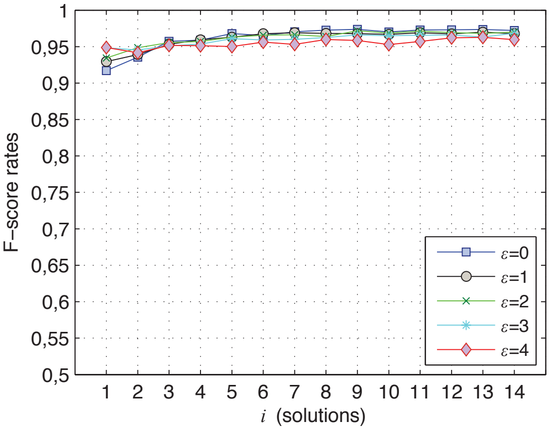

Fig. 3. Influence of parameter i as the maximum tolerance distance in pixels ϵ is varied

Sequences of bottle images: The set contains 120 sequences of images of bottle necks with faults or regions with defects generated in [13] (Fig. 2). Each sequence is composed of three images with an angle of rotation $\alpha = 15$. From the captured images we have extracted subimages of 1000 × 250 pixels. The base correspondence was determined by means of markers outside the object that comply with the object's motion. In this case the objective of the point-to-point correspondence was to determine the trajectory of multiple defects in the sequence that must be detected to determine the quality of the bottle in a multiple view inspection process.

Evaluation in relation to the number of partial solutions: We got the best performance when a discretized distance $\epsilon = 0$ was used (Fig. 3). These results indicate that at $\epsilon = 0$ we get a trifocal correspondence with a performance F-score$= 0.97$. It is interesting to note that as parameter ϵ is increased, the performance of the method starts dropping.

The maximum performance of the system using the optimum ϵ value for each variation of parameter i is shown in Fig. 4. It is seen that from five combinations ($i = 5$), the best distance remains at zero pixels. The results of this graph agree with those presented previously, because with $i = 9$ we get the best performance.

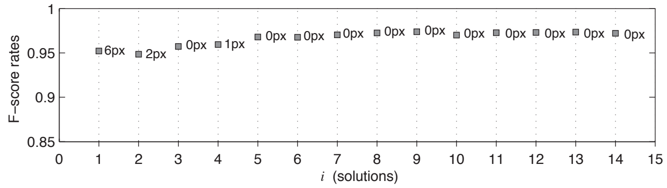

Fig. 4. Best performance of correspondence varying according to the number of solutions. The best ϵ value has been chosen in each performance curve.

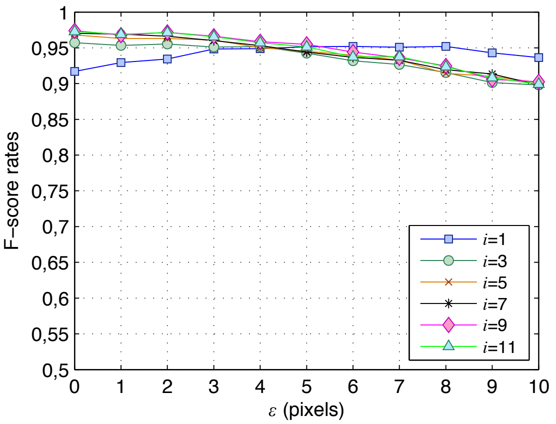

Fig. 5. Influence of distance (ϵ) on the detection of correspondences for different numbers of solutions (k). The larger the number of solutions, the shorter the distance ϵ.

Evaluation according to re-projection distance: In the previous evaluation we varied parameter i to determine its influence on the performance of the system of correspondences. In this case we evaluated the influence of the distance ϵ keeping parameter i fixed. Because of the large number of curves generated, we graphed only odd numbers of parameter i (Fig. 5). Below we detail the results. Clearly there is an important difference when using a trifocal tensor versus multiple trifocal tensors. For example, when using a single trifocal tensor, the maximum performance is obtained at a distance of 6 pixels ($\epsilon = 6$), with an F-score= 0.95. In contrast, when using nine combinations we get a performance F-score= 0.97 at a discretized distance $\epsilon = 0$. The latter result shows the effectiveness of using the multiple intermediate solutions combination by the proposed method.

IV. Conclusions

In this paper we have developed two important contributions. First, we presented a method that uses multiple geometric solutions in three views to determine point-to-point correspondence and filter out wrong matchings. Second, for each geometric model we have determined the real distance with respect to corresponding point by means of the MLESAC estimator, in that way weighting the error associated with each intermediate solution. The main novelty of our proposal is the geometric methodology for solving the problem of the estimation of point-to-point correspondence, regardless of the angles of the points of view of the objects contained in the images and of the geometric transformations present in them. We call this algorithm as *Trifocal Geometric Correspondence* (TRIGC). It is important to note that the point can be occluded in the following views, but its position remains valid because our method is based on a geometric model that defines the scene. We also show that the use of multiple random solutions makes it possible to improve the performance of the correspondence. Although our method starts from the basis that there is a set of points in previous correspondence necessary to determine the trifocal tensors, it is designed to maximize the correspondences in specific regions of each image and not necessarily in a specific point that is not relevant to that method. Finally, for the images analyzed, we showed that the point-to-point correspondence can be generated through a multiple geometric relation between three views and it can be used in sequences of images that have a low signal-to-noise ratio. In those cases invariant algorithms will not achieve a good performance due to the appearance of many false alarms.

Acknowledgment

This work was supported by the National Commission of Science and Technology (CONICYT, Chile). Fondecyt grant no. 11100098.

References

[1] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[2] A. Fitzgibbon, Robust registration of 2d and 3d point sets, Image and Vision Computing 21 (13–14) (2003) 1145–1153.

[3] R. Vidal, Y. Ma, S. Soatto, S. Sastry, Two-view multibody structure from motion, International Journal of Computer Vision 68 (1) (2006) 7–25.

[4] J. L. Barron, D. J. Fleet, S. S. Beauchemin, Performance of optical flow techniques, International Journal of Computer Vision 12 (1) (1994) 43–77.

[5] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, Cambridge, UK, 2000.

[6] M. Carrasco, D. Mery, Bifocal matching using multiple geometrical solutions, in: Proceedings of the 5th Pacific Rim conference on Advances in Image and Video Technology - Volume Part II (PSIVT), no. 7087, Springer, 2011, pp. 192–203.

[7] O. Faugeras, Q.-T. Luong, T. Papadopoulo, The geometry of multiple images: The laws that govern the formation of multiple images of a scene and some of their applications, The MIT Press, Cambridge MA, London, 2001.

[8] A. Shashua, Algebraic functions for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8) (1995) 779–789.

[9] B. J. Tordoff, D. W. Murray, Guided-mlesac: faster image transform estimation by using matching priors, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10) (2005) 1523–1535. doi:10.1109/TPAMI.2005.199.

[10] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[11] A. P. Dempster, N. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B 39 (1977) 1–38.

[12] H. Bay, A. Ess, T. Tuytelaars, L. Gool, Surf: Speeded up robust features, Computer Vision and Image Understanding (CVIU) 110 (3) (2008) 346–359.

[13] M. Carrasco, L. Pizarro, D. Mery, Visual inspection of glass bottlenecks by multiple-view analysis, International Journal of Computer Integrated Manufacturing 23 (11) (2010) 925.

[14] D. Olson, David L.; Delen, Advanced Data Mining Techniques, Springer, 2008.