

Estimation of common groundplane based on co-motion statistics

Zoltan Szlavik¹, Laszlo Havasi², Tamas Sziranyi¹

¹ Analogical and Neural Computing Laboratory, Computer and Automation Research Institute of Hungarian Academy of Sciences, P.O. Box 63,
H-1518 Budapest, Hungary
{szlavik, sziranyi}@sztaki.hu

² Peter Pazmany Catholic University, Piarista köz 1.,
H-1052 Budapest, Hungary
{havasi}@digitus.itk.ppke.hu

Abstract. The paper presents a method for groundplane estimation from image-pairs even if unstructured environment and motion. In a typical outdoor multi-camera system the observed objects might be very different due to noise coming from lighting conditions and camera positions. Static features such as color, shape, and contours cannot be used for image matching in these cases. In the paper a method is proposed for matching partially overlapping images captured by video cameras. Using co-motion statistics, which is followed by outlier detection and a nonlinear optimization, does the matching. The described robust algorithm finds point correspondences in two images without searching for any structures and without tracking any continuous motion. Real-life outdoor experiments demonstrate the feasibility of this approach.

1 Introduction

Multi-camera based observation of human or traffic activities is becoming of increasing interest for many applications like cases of semi-mobile traffic control using automatic calibration or tracking humans in a surveillance system. In a typical outdoor scenario, multiple objects, such as people and cars, move independently on a common ground plane. Transforming the activity captured by distributed individual video cameras from local image coordinates to a common frame then sets the stage for global analysis and tracking of the activity in the scene.

Matching different images of a single scene could be difficult, because of occlusion, aspect changes and lighting changes that occur from different views. Over the years numerous algorithms for image and video matching have been proposed. Still-image matching algorithms can be classified into two categories. In “*template matching*” the algorithms attempt to correlate the gray levels of image patches, assuming that they are similar [3][7]. This approach appears to be valid for image pairs with small difference; however it may be wrong at occlusion boundaries and within featureless regions. In “*feature matching*” the algorithms first extract salient primitives from images (edges or contours) and match them in two or more views [1][4][5][6]. An

image can then be described by a graph with primitives as nodes and geometric relations defining the links. The registration then becomes the mapping of the two graphs: subgraph isomorphism. They may fail if the chosen primitives cannot be reliably detected. The views of the scene from the various cameras might be very different, so we cannot base the decision solely on the color or shape of objects in the scene.

In a multi-camera observation system the video sequences recorded by cameras can be used for estimating matching correspondences between different views. Video sequences contain much more information than the spatial scene structure of any individual frame, as also capturing information about scene dynamic. The scene dynamic is an inherent property of the scene; it is common to all video sequences recorded from the same location, even when taken from different cameras from different positions at different zooms. In [9][10] approaches were presented, which align tracks of the observed objects. In these cases the capability of robust object tracking is assumed and this is the weak point of the methods. It must be assumed that the speed doesn't change more than a predefined value and the objects in the scene are moving continuously.

In our experiment we use standard digital cameras with wide angle. So, the field of view is large, consequently, the size of features is small; the images are blurred and noisy. The common field of view of two neighboring cameras is less than 30%. We have tested several correlation-based toolboxes for matching, but they gave poor results. In case of several randomly moving objects on the screen, the conventional 3D registration of cameras usually needs some a priori object definition or human interaction to help the registration.

The approach we propose in the paper is a serious extension of previously published sequence-based image matching methods [9][10] for non-structured estimation. It aims to use statistics of concurrent motions – the so called co-motion statistics – instead of trajectories of moving objects to find matching points in image pairs. The input of the system is video sequences from fixed cameras at unknown positions, orientations and zooms. After matching of images the system aligns the multiple camera views into a single frame making possible to track all moving objects across different views, producing the 3D positions and orientations of cameras up to scale. In our approach no a-priori information is needed, and the method also works well in images of randomly scrambled motion, where other methods fail because of the missing fixed structures.

2 Common Groundplane Estimation

The main steps of our algorithm:

1. Motion detection; record point coordinates where motion is detected;
2. Update local and remote statistical maps (the notion of statistical maps is defined in *Section 2.2*);
3. Extract candidate point pairs from statistical maps;
4. Outlier rejection;

5. Fine tune of point correspondences by minimizing the reprojection error between sets of candidate point pairs;
6. Alignment of the two views.

The major assumption is the time synchronization between the cameras. When it exists, the motion information can be transformed into motion-statistics. Later we will show that by using further processing this assumption can be avoided.

2.1 Motion detection

Our application field has several special requirements to motion extraction. The videos of open-air traffic were created with normal digital cameras with wide angle so the images are blurred and noisy. The size of moving objects has a great variety; there are small blobs (walking people) and huge blobs (trams or buses), too. The background cannot be extracted perfectly, because we do not want to assume any a-priori knowledge about the scene.

In the first step we define pixels, which are considered in the statistical calculus. The motion blobs are extracted by using simple running-average background subtraction with large β to delete the irrelevant parts by using the reference image I_{k-1} :

$$I_k(x, y) = \beta I_k(x, y) + (1 - \beta) I_{k-1}(x, y), \quad 0 < \beta < 1 \quad (1)$$

This method is fast and very sensible with low threshold value. Some disadvantage comes from the cases that often detect noises and background flashings. In the pre-processing algorithm the detected motion blobs are dilated while these are reaching the local maximums of edge maps; we found local maximums of the edge map by using similar algorithm to that of proposed by Canny [12]. This approach seems a usable solution to detect the significant moving objects in the scene. In our method we do not need precise motion detection and object extraction, because of the later statistical processing these minor errors are irrelevant. The binarized image with the detected objects is the motion map, which is used for updating statistical maps.

2.2 Co-motion statistics

For finding point correspondences between two images in case of wide baseline stereo and video sequences we have decided to analyze the dynamics of the scene. To do it co-motion statistics (statistics of concurrent motions) were introduced. In case of single video sequence a motion statistical map for a given pixel can be recorded as follows: when motion is detected in a pixel, the coordinates are recorded of all pixels where motion is also detected at that moment. In the motion statistical map the values of the pixels at the recorded coordinates are updated. After all, this statistical map is normalized to have global maximum equal to 1.

In case of stereo video sequences to each point in the images, two motion-statistic maps are assigned: a local and a remote. Local map means the motion-statistical map in the image from the pixel is selected, the remote motion-statistical map is refer to the motions in the other image. After the motion detected on the local side, for the

points defined by the local motion map the local statistical map updated by the local motion map. For each point where motion is detected on the local side, the local motion map of the remote side updates the corresponding remote statistical map. Examples of co-motion statistics are given in Fig. 1.

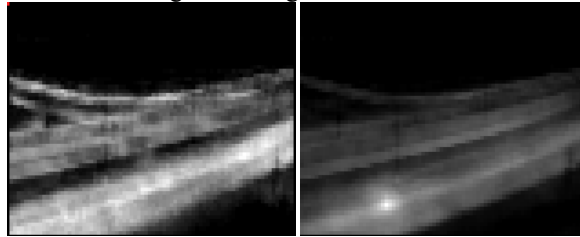


Fig. 1. Remote statistical maps for different cases are in the pictures. In the left one for the point, which is not in the cameras' common field of view; in the right one for the point from cameras' common field of view.

2.3 Outlier rejection

As candidate matches we choose global maximums on local and remote statistical images.

For the rejection of outliers from the set of point correspondences we applied the principle of "good neighbors" and analyze the errors'. The principle of "good neighbors" says that if we have a good match, then we will have many other good matches in some neighbor of it. Consider a candidate match (m_1, m_2) where m_1 is a point in the first image and m_2 is a point in the second image. Let $N(m_1)$ and $N(m_2)$ be the neighbors of m_1 and m_2 . If (m_1, m_2) is a good match, we will expect to see many other matches (n_1, n_2) , where if $n_1 \in N(m_1)$ then $n_2 \in N(m_2)$. So, candidate pairs for which less other candidate pairs could be found in their neighborhood were eliminated.

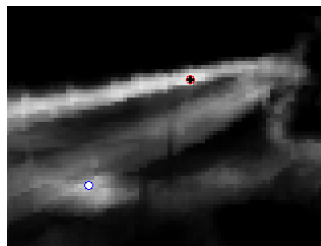


Fig. 2. The global maximum is the black cross, while should be the white dot.

The reduced set of point-correspondences also has erroneous matches due to the errors caused by recording of co-motion statistics:

1. From global statistical map, we know image regions where much more moving objects are detected than in other places. If we have a point in the first image where the correspondent scene location is not in the field of view of

the second camera, then the correspondent maximum in remote statistical image will be in a wrong place, in a point, where value of the motion statistics in the global statistic is high, see Fig. 2.

2. Because of the size of the moving objects, the global maximums could be shifted and it will be somewhere in the neighborhood of the desired correspondent point. These “shifting” results in cases where different points from local statistical images are “mapped” onto the same point in remote statistical images.

To solve the first problem we need to eliminate points from the set of candidate matches if the global maximum on remote statistical images is a pixel where the value of the motion statistics is greater than some predefined parameter. To get over on second problem we also need to eliminate points from the set of candidate matches if the global maximum on remote statistical image is a pixel, which also present in another candidate pair.

2.4 Fine-tuning of point correspondences

The above described outlier rejection algorithm results in point correspondences, but these results must be fine-tuned for the alignment of two views.

For the alignment of two images a transformation is estimated from the extracted point correspondences. The results of the transformation can be seen in Fig. 3.

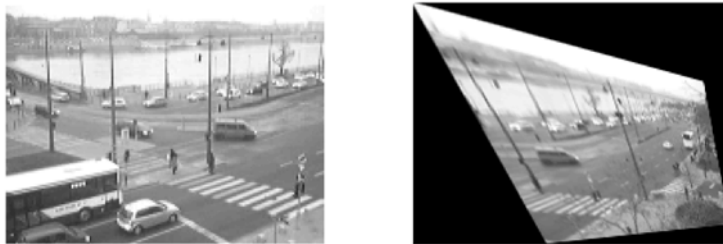


Fig. 3. Left: the view of the left camera; Right: transformed view of the right camera

It can be seen that the resulted transformation is not the desired one: the continuous edges are broken if a composite view is generated from the transformed images. The point coordinates can contain errors; they can be shifted by some pixels, due to the nature of co-motion statistics recording. Even if we have 1 pixel error in point coordinates the fine alignment of the images cannot be done. This simple outlier rejection algorithm must be followed by a robust optimization to fine tune point correspondences and obtain subpixel accuracy.

An iterative technique is used to refine both the point placements and the transformation. The method used is the Levenberg-Marquardt iteration [13] to minimize the sum-of-square difference between the obtained coordinates and the transformed values. The entries of the transformation matrix as well as the coordinates of points in right camera’s image are treated as variable parameters in the optimization, but the

point coordinates of the left camera's image are kept constant. The initial condition for this iteration is the entries of the transformation matrix and point's coordinates estimated by using the above-described outlier rejection algorithm.

3 Time Synchronization

Until now, we have assumed that the cameras' clocks are synchronized. For time synchronization many algorithms have been developed, e.g. the Berkeley algorithm. In our case, if the cameras are not synchronized then the generated co-motion statistics should no longer refer to concurrent motions detected in the two stereo sequences. So, when we apply our algorithm for outlier rejection, we do not get a "large" set of point correspondences, but more point correspondences can be extracted in the case of synchronized sequences.

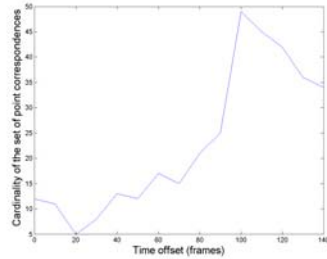


Fig. 4. Cardinality of the set of point correspondences for different time offset values. The maximum is at 100 frames, which means that the offset between two sequences is 100 frames.

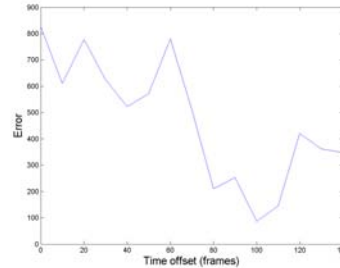


Fig. 5. The change of the error rate for different offset values; the minimum is at 100 frames as the maximum in Fig.4.

Since this observation is obvious and true in practice, we calculate point correspondences for different time offset values then perform a one-dimensional search for the largest set of point correspondences to synchronize the sequences, see Fig. 4.

It can be seen in Fig. 4 that even in the case of unsynchronized sequences the algorithm produces point correspondences. But if we analyze the sum-of-square differences score (the reprojection error in this case), see Fig. 5, we found that the global minimum is at offset value of 100 frames, as the maximum in the Fig. 4 for the cardinalities of sets of point correspondences. This means that the global optimum is at offset value of 100 frames, in all other cases the obtained point correspondences mean that the algorithm finds a local optimum. The above results also show some robustness of the method against the synchronization error.

5 Results

For testing purpose we have chosen scenes containing traffic motion and structured outlines of street scenario. However, we did not exploit this structural information, it only helps to verify the obtained results. The above-described approach was tested on videos captured by two cameras, having partially overlapping views, at Gellert (GELLERT videos) and Ferenciek squares (FERENCIEK videos) in Budapest. The GELLERT videos are captured at resolution 160×120 , at same zoom level and with the same cameras while the FERENCIEK videos are captured at resolution 320×240 , at different zoom levels and with different cameras. The common field of view of the two cameras in both cases is about 30%. The proposed outlier rejection algorithm rejects most (98%) of the candidate point pairs. For the GELLERT videos it results in 49 point-correspondences and in 23 for FERENCIEK videos, which are still enough to estimate common groundplanes.

The computation time of the whole statistical procedure was about 10 minutes for 10 minutes of video presented in the figures. For longer sequences and higher resolution we apply a two-step procedure: the generated statistical maps are of resolution 80×60 , then, based on them, the fine-tuning of point-correspondences was done at the video's native resolution.

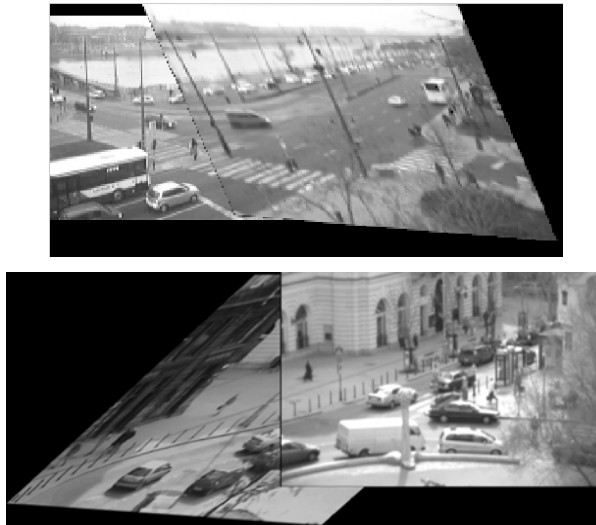


Fig. 6. The constructed composite views are in the pictures. The upper one is generated for the GELLERT videos, the lower one is for the FERENCIEK videos. Based on the extracted point-correspondences the epipolar geometry easily can be estimated by using standard algorithms [11]. The self-calibration of cameras can be done.

6 Conclusions

The paper has shown that for free-placed outdoor cameras the common groundplane can be estimated without human interaction in case of arbitrary scenes. In our approach no a-priori information is needed, and the method also works well in images of randomly scrambled motion, where other methods fail because of the missing fixed structures.

In our approach we introduced co-motion statistics to find matching points in image pairs. First our method records motion statistics and then chooses global maximums as candidate matches. This step is followed by an elimination of outliers from the set of candidate matches and an optimization based on the minimization of the reprojection error between images, to fine-tune the locations of candidate pairs.

In the next step we apply our method for 3D registration of different views of cameras capturing different projections of volumetric motion.

References

1. O. D. Faugeras, Q.-T. Luong, S.J. Maybank: Camera self-calibration: Theory and experiments, ECCV '92, Lecture Notes in Computer Science, Vol. 588 (1992) 321-334
2. R. Hartley: Estimation of relative camera positions for uncalibrated cameras, Proc. of ECCV'92, Lecture Notes in Computer Science, Vol. 588 (1992)
3. D. H. Ballard, C. M. Brown: Computer Vision, Prentice-Hall, Englewood Cliffs NJ (1982)
4. S. T. Barnard, W. B. Thompson: Disparity analysis of images, IEEE Trans. PAMI, Vol. 2(4) (1980) 333-340
5. J. K. Cheng, T. S. Huang: Image registration by matching relational structures, Pattern Recognition, Vol. 17(1) (1984) 149-159
6. J. Weng, N. Ahuja, T. S. Huang: Matching two perspective views, IEEE Trans. PAMI, Vol. 14(8) (1992) 806-825
7. Z. Zhang, R. Deriche, O. Faugeras, Q.-T. Luong: A robust technique for matching two uncalibrated images through the recovery of the unknown Epipolar Geometry, Artificial Intelligence Journal, Vol.78 (1995) 87-119
8. H. C. Longuet-Higgins: A computer algorithm for reconstructing a scene from two projections, Nature, Vol. 293 (1981)
9. L. Lee, R. Romano, G. Stein: Monitoring activities from multiple video streams: establishing a common coordinate frame, IEEE Trans. PAMI, Vol. 22(8) (2000)
10. Y. Caspi, D. Simakov, and M. Irani: Feature-based sequence-to-sequence matching (2002)
11. R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press (2003)
12. J. Canny, A computational approach to edge detection, IEEE Trans. PAMI, Vol. 8(6), (1986) 679-698
13. Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge (1986)