Stereo Localization Using Dual PTZ Cameras

Sanjeev Kumar, Christian Micheloni, and Claudio Piciarelli

Department of Mathematics and Computer Science
University of Udine, Via Della Scienze 206, Udine-33100, Italy
{sanjeev.kumar,christian.micheloni,claudio.piciarelli}@dimi.uniud.it

Abstract. In this paper, we present a cooperative stereo system based on two pant-tilt-zoom (PTZ) cameras that can localize a moving target in a complex environment. Given an approximate target position that can be estimated by a fixed camera with a wide field of view, two PTZ cameras with a large baseline are pointed toward the target in order to estimate precisely its position. The overall method is divided in three parts: offline construction of a look-up-table (LUT) of rectification matrices, use of the LUT in real time for computing the rectification transformations for arbitrary camera positions, and finally 3D target localization. A chain of homographic transformations are used for finding the matching between different pairs of wide baseline stereo images. The proposed stereo localization system has two advantages: improved localization on a partially occluded target and monitoring a large environment using only two PTZ cameras without missing significant information. Finally, through experimental results, we show that the proposed system is able to make required localization of targets with good accuracy.

1 Introduction

Modern video surveillance has been an active area of research. Nowadays, a number of research works are going to develop more intelligent and smart video monitoring systems according to the requirements and applicability [1], [2], [3]. The computation of reliable objects' trajectories by means of localization is really important for different contexts like traffic monitoring, behaviour analysis, suspicious event detection, sensor network configuration, etc. From the low-level to the high level techniques three main steps can be identified: a) detection and localization of interesting objects, b) frame-to-frame tracking of detected objects and c) behaviour recognition. To achieve all these goals visual surveillance systems usually exploit a network of cameras [4]. Existing non-stereo systems often localize objects in the environment by defining homographies between single cameras and a 2D map [4]. Such homographies are based on a ground plane constraint. When the detected object is occluded in such a way that its point of contact with the ground plane is not visible, such an approach introduces relevant localization errors. To overcome such a problem, stereo vision can be taken into account.

Stereo vision has the advantage that it is able to estimate an accurate and detailed 3-D representation of the position of an object with respect to a given

co-ordinate system using its two or more perspective images [5]. Traditional stereo vision research usually uses static cameras for their low cost and relative simpleness in modelling. PTZ camera is a typical and the simplest active camera, whose pose can be fully controlled by pan, tilt and zoom parameters [3]. As PTZ cameras are able to obtain multi-view-angle and multi-resolution information (i.e. both global and local image information), they are used for many real applications specially in video surveillance. The PTZ camera based stereo system is able to cover large environments and, if overlapped fields of view are considered, to reduce the occlusions. However, PTZ cameras based stereo vision is much more challenging when compared to traditional static cameras based stereo vision as the intrinsic and external parameters of each camera can be changed in utility.

Recently, a novel stereo rectification method for dual-PTZ-camera system is presented to greatly increase the efficiency of stereo matching [6]. In this dual-PTZ-camera based stereo method, the problem related to inconsistency of intensities in two camera images is solved by addressing a two-step stereo matching strategy. An interesting approach to solve stereo vision problems by means of rotating cameras has been recently proposed with its analytic formulation [7]. An off-line initialization process is performed to initialize essential matrix using calibration parameters. During on-line operations the rotation angles of the cameras are retrieved and exploited to compute the essential matrix. When the zoom is considered, it would require the calibration for any zoom level of both cameras.

In this paper, we propose a stereo system based on two PTZ cameras from a network of cooperative sensors. The proposed solution is able to accurately localize a moving object in outdoor areas. Once a target is selected by the surveillance system, a pair of PTZ cameras are focused on the target with the required zoom to provide stereo localization of such a target. To solve stereo matching problem in case of dual PTZ camera, an uncalibrated approach that computes the rectification by interpolating the transformations contained in a LUT is proposed. Such a LUT is defined off-line by sampling the pan and tilt ranges of both PTZ cameras and using the same zoom level. The transformations contained in the LUT are computed on image pairs computed with a chain of homographies to solve the wide base-line problem. During on-line operations an interpolation based on neural network is proposed to estimate the rectification transformations for the given orientations of both cameras.

2 Pre-localization Steps

The localization is performed using various rectified pairs of stereo images. Therefore, few steps for real time rectification are needed before performing the task of localization. These steps involve wide baseline stereo matching, construction of LUT and learning of neural network using LUT data.

2.1 Construction of the Look-Up Table

A rectification transformation is a linear one-to-one transformation of the projective plane, which is represented by a 3×3 non-singular matrix. For a pair of stereo images $\mathbf{I}_l$ and $\mathbf{I}_r$, the rectification can be expressed in the following ways

$$\mathbf{J}_l = \mathbf{R}_l * \mathbf{I}_l \qquad \mathbf{J}_r = \mathbf{R}_r * \mathbf{I}_r$$

where $(\mathbf{J}_l, \mathbf{J}_r)$ are the rectified images and $(\mathbf{R}_l, \mathbf{R}_r)$ are the rectification matrices. These rectification transformations can be obtained by minimizing

$$\sum_i [(m_l^i)^T \mathbf{R}_r^T \mathbf{F}_\infty \mathbf{R}_l m_l^i] \tag{1}$$

where (m_l^i, m_r^i) are pairs of matching points between images $\mathbf{I}_l$ and $\mathbf{I}_r$ and $\mathbf{F}_\infty$ is the fundamental matrix for rectified pair of images. Generally, the minimization of (1) is time-consuming and therefore it is not possible to compute the rectifications in real time [8]. Here, an offline LUT containing rectification matrices corresponding to various image pairs captured at predefined pan and tilt angles is constructed. The rectification transformations can then be interpolated in real-time for any arbitrary orientation of both PTZ cameras by using this LUT data. The main steps to construct the LUT are:

1. Sample the different pan and tilt angles $(p_l^i, t_l^i)_{i=1:1:n_1}$ for the whole pan and tilt ranges of left PTZ camera into n_1 equal intervals. Similarly, sample the different pan and tilt angles $(p_r^i, t_r^i)_{i=1;1:n_1}$ for the right camera.
2. Capture $n_1 \times n_1$ different of images $(\mathbf{I}_l^{i,j})_{i=1:1:n_1}^{j=1:1:n_1}$ for left camera. Same time of instance capture their corresponding right stereo images $(\mathbf{I}_r^{i,j})_{i=1:1:n_1}^{j=1:1:n_1}$.
3. Compute the possible $k(> n_1 \times n_1)$ pairs of rectification transformations pairs $(\mathbf{R}_l^k, \mathbf{R}_r^k)$ for the different combination of these stereo images. Here, we use the constraint that the rectification transformations are computed for two images only if they share at least 30% of their field of view. This criterion is considered also during the sampling of pan and tilt angles for both cameras.
4. Store all these pairs of rectification transformations in a LUT in such a way that by choosing a combination of four independent variables (p_l, t_l, p_r, t_r), their corresponding rectification transformations $(\mathbf{R}_l, \mathbf{R}_r)$ can be easily computed. This is done through a neural network described in section 2.3.

The main problem to be addressed in the creation of the LUT is the automatic computation of the rectification transformations. Many works on rectification assume that the baseline (the distance between the two cameras) is small if compared to the distance of the object from the cameras, and thus the two images acquired by the cameras are similar. This allows the detection of the matching points using standard techniques such as SIFT matching [9]. However, in the proposed system this assumption is no longer valid for some combinations of pan-tilt values. The problem of finding matches in wide-baseline configurations is addressed in the next section.

2.2 Point Matching Between Wide Baseline Stereo Images

SIFT matching [9] is a popular tool for extracting pairs of matching points between stereo images. However, this method fails to provide good results in case of wide baseline images. In this work, we have used a method based on a chain of homographic matrices for extracting pairs of matching points in these kinds of image pairs. In the case of wide baseline images, if the object is far enough along the optical axis then it is possible to extract pairs of matching points manually or using an approach proposed in [10]. Let $(\mathbf{I}_l^1, \mathbf{I}_r^1)$ be a pair of images of a 3D scene which is far from the cameras along their optical axis. An initial homography $\mathbf{H}^1$ is generated by using extracted pairs of matching points between $\mathbf{I}_l^1$ and $\mathbf{I}_r^1$ using standard approaches. Let $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ be a pair of images captures from left and right cameras of a scene/object near to cameras along their optical axis. The problem is to autonomously extract the pairs of matching points between the images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$. To solve such a problem, a set of n images is captured for each camera by moving the cameras from the initial position (the one at which $(\mathbf{I}_l^1, \mathbf{I}_r^1)$ are acquired) to the current position. Let these two sets of images be $(\mathbf{I}_l^1, \mathbf{I}_l^2 \ldots \mathbf{I}_l^n)$ and $(\mathbf{I}_r^1, \mathbf{I}_r^2 \ldots \mathbf{I}_r^n)$. Now we use the following steps to solve this matching problem for wide-baseline image pairs:

1. Perform the SIFT matching between image pairs $(\mathbf{I}_l^1, \mathbf{I}_l^2)$, $(\mathbf{I}_l^2, \mathbf{I}_l^3)$, $\ldots$, $(\mathbf{I}_l^{n-1}, \mathbf{I}_l^n)$ and use these sets of pairs of matching points for computing their respective homography matrices $\mathbf{H}_l^{1,2}$, $\mathbf{H}_l^{2,3}$, $\ldots$, $\mathbf{H}_l^{n-1,n}$.
2. Repeat the procedure given in above step on the sequences of images of right camera and compute $\mathbf{H}_r^{1,2}$, $\mathbf{H}_r^{2,3}$, $\ldots$, $\mathbf{H}_r^{n-1,n}$.
3. Compute the homography matrix $\mathbf{H}_l$ and $\mathbf{H}_r$

$$\mathbf{H}_l = \prod_{i=0}^{n-2} \mathbf{H}_l^{n-(i+1),n-i} \qquad and \qquad \mathbf{H}_r = \prod_{i=0}^{n-2} \mathbf{H}_r^{n-(i+1),n-i}$$

4. Compute the homography matrix $\mathbf{H}^n$ for the pairs of matching points between current images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ as

$$\mathbf{H}^n = \mathbf{H}_r * \mathbf{H}^1 * (\mathbf{H}_l)^{-1} \tag{2}$$

Figure 1 gives an intuitive interpretation of the procedure. The final homography matrix $\mathbf{H}^n$ can be computed for any value of n; however, the above procedure can accumulate errors in the final homography due to multiplication of several matrices. In order to minimize this error, we

1. keep the sampling step n as low as possible, with the constraint that we require at least a 30% image overlap for SIFT matching;
2. minimize errors due to bad matches by using a robust estimator for outlier detection and removal: we use the Iterative re-weighted least-square (IRLS) technique for computing the homography matrix from the pairs of matching points between any two images. IRLS provide a robust solution for homography computation when compared to other approaches like standard least square or Singular value decomposition.

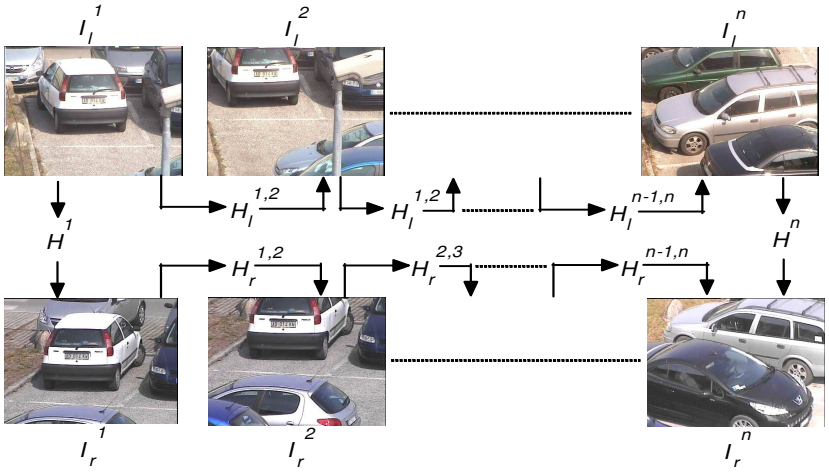

Fig. 1. Wide baseline stereo matching using a chain of homographic matrices

2.3 Neural Network Based Interpolation

Once the LUT is constructed, its content is used for offline training of the neural networks using k combinations of pan and tilt angles $(p_l^i, t_l^i, p_r^i, t_r^i)_{i=1:1:k}$ as input and the elements of their respective rectification transformations $(\mathbf{R}_l^i, \mathbf{R}_r^i)_{i=1:1::k}$ as the network output. A multilayer feed-forward neural network containing one five-nodes hidden layer with backpropagation (BP)learning algorithm has been used in this work. The optimal input-to-hidden nodes weight matrix $\mathbf{W}$ and hidden-to-output nodes weight matrix $\mathbf{V}$ is stored and used for interpolating the rectification transformations for any arbitrary orientations of both cameras in real time.

Note that the LUT is built with constant and equal zoom levels for both cameras. Based on the requirements for the monitoring of selected target, the zoom levels of the two cameras can actually be different, and the LUT data cannot be directly applied. In this case, compensation of unequal zoom settings is needed before stereo matching. This problem can be handled easily by using a focal-ratio-based methodology [11].

3 Localization

The 3D position of the target has to be computed in terms of its coordinates $[\mathbf{x}_w, \mathbf{y}_w, \mathbf{z}_w]$ in a world reference system. Once the pair of stereo images is rectified, the disparity between the matching pairs can be computed only for the pixels belonging to the target. Starting from the pixels in the left camera image, the search for its matching pixels is restricted only on the corresponding epipolar lines in the right camera image. In particular, for each pixel, starting from its x, y position, similarity scores are computed considering a normalized SSD measure that quantifies the difference between the intensity patterns as:

$$C(x,y,d) = \frac{\sum\limits_{(\xi,\eta)} [\mathbf{J}_l(x+\xi, y+\eta) - \mathbf{J}_r(x+d+\xi, y+\eta)]}{\sqrt{\sum\limits_{(\xi,\eta)} \mathbf{J}_l(x+\xi, y+\eta)^2 \sum\limits_{(\xi,\eta)} \mathbf{J}_r(x+\xi, y+\eta)^2}} \tag{3}$$

where $\xi \in [-n, n]$ and $\eta \in [-m, m]$ define a window centred in (x, y), while d is the disparity. The required disparity value is the one that minimizes the SSD error:

$$d_0(x,y) = \min_d C(x,y,|d|) \tag{4}$$

Once the disparity d is computed between the position of the target in the left and right images, the distance of the target $\mathbf{z}_w$ from the camera along optical axis is estimated by

$$\mathbf{z}_s = f_r \frac{B}{d} \tag{5}$$

where f_r is focal length for the rectified pair of images and B denotes the base line distance. Let $(\mathbf{x}_l, \mathbf{y}_l)$ be the position of the target in the left camera image, then its position in the plane orthogonal to the optical axis of camera is given by

$$\mathbf{x}_w = \frac{\mathbf{x}_l \mathbf{z}_w}{f_r} \qquad \mathbf{y}_w = \frac{\mathbf{y}_l \mathbf{z}_w}{f_r}$$

The location of target $(\mathbf{x}_m, \mathbf{Y}_m)$ in a ground plane map is given by

$$[\mathbf{x}_m, \quad \mathbf{y}_m, \quad 1]^T = \mathbf{H}_m^w \, [\mathbf{x}_w, \quad \mathbf{y}_w, \quad 1]^T$$

where $\mathbf{H}_m^w$ is the homography computed offline between the homogeneous coordinates of ground plane position $(\mathbf{x}_w, \mathbf{y}_w)$ of some selected points and their respective position in the map $(\mathbf{x}_m, \mathbf{y}_m)$. The iterative re-weighted least square (IRLS) algorithm allows to robustly estimate such a homography.

4 Experimental Results

Experimental study have been conducted to evaluate the performance of proposed localization algorithm in a parking lot scenario. The experimental results have been obtained from four different pairs of frames by considering different cases, i.e., partially occluded targets and using different zoom levels for both PTZ cameras. Six different pan and tilt angles have been selected in each direction by sampling with a step size of 3.0 degree along pan direction and 4.0 degree along tilt direction for both cameras to cover entire experimental outdoor environment. In this way, a total of 36 images have been captured by each camera. Out of these 36×36 combination of images, only $k = 120$ pairs of images have been selected for network training by considering the fact that at least 30% part of field of view should be common between both images. The rectification transformations have been computed using the matching pairs of feature points from these 120 pairs of images and stored in a LUT.

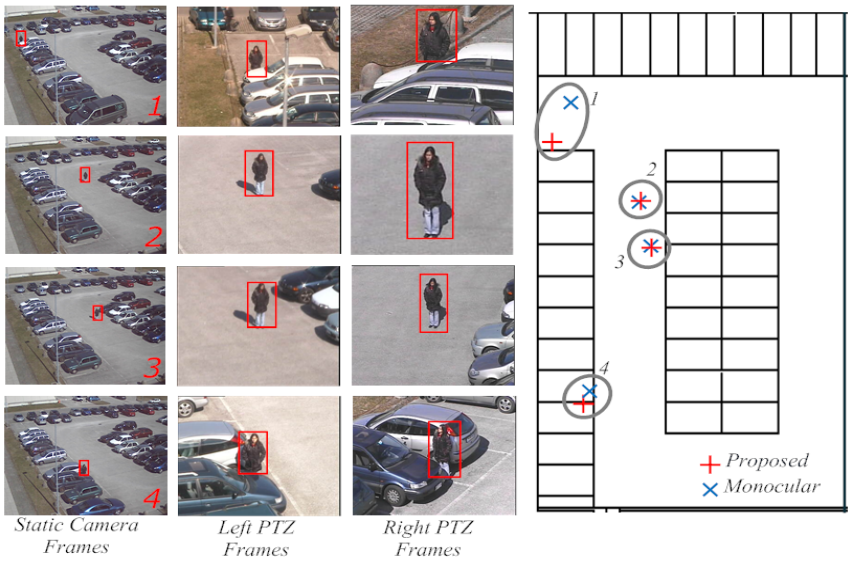

Fig. 2. Localization of a target in various stereo frames

Experiments with real sequences have been carried out in order to test the performance of the proposed localization algorithm. Localization results are shown in Figure 2 for four different pairs of frames captured at different camera settings (pan, tilt and zoom). The selection of these four frames have been performed to check the performance of proposed localization algorithm in different cases such as partially occluded target (frame pairs 1 and 4), images having unequal zoom (frame pairs 1 and 2) and non-occluded (frame pairs 2 and 3). Simultaneously, the localization results are computed based on a monocular camera based technique [4](here we have used a static camera having wide field of view) for making a comparison of the achieved results and for showing the superiority of proposed method on monocular camera based techniques in case of partially occluded targets (see localization for frame pairs 1 and 4). Localization has been made in a 2D ground-plane map (30×40) meters.

Figure 3 represents a surface plot for localization error computed for a target with ground truth obtained from known marks. It can be seen from this plot that the error is increasing if the distance of object from the left camera or the height of occlusion is increasing in case of monocular camera based scheme. In the case of the proposed method, the error is almost constant and does not depend the occlusion's height or object's distance from camera.

5 Conclusions

We have presented an approach for the localization of an object in a given test map using dual PTZ camera based wide baseline stereo system. A neural network is used for finding the rectification transformations in real time using an offline

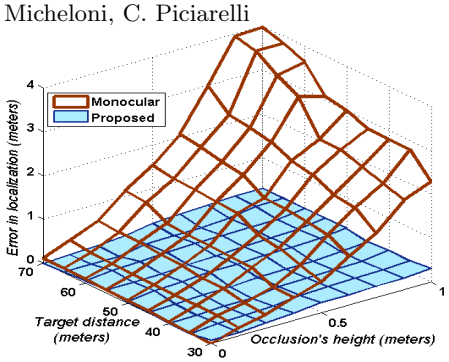

Fig. 3. Error in localization corresponding to the occlusion's height and object's distance.

LUT; and a method has been proposed for extracting pairs of matching points from wide base line stereo images. The required targets have been localized on a given test map using stereo based 3D position. Experimental results have proven that the proposed technique leads to better results than standard monocular camera based localization.

References

1. Abidi, B., Koschan, A., Kang, S., Mitckes, M. Abidi, M.: Automatic target Acquisition and Tracking with Cooperative Static and PTZ Video Cameras,Multisensors Surveillance Systems: The Fusion Perspective, 43–59 (2003).
2. Haritaoglu, S., Harwood, D., Davis, L.: W^4: Real-Time Surveillance of People and their Activities, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 809–830 (2000).
3. Jain, A., Kopell, D., Kakligian, K., Wang, Y.F.: Using Stationary-Dynamic Camera Assemblies for Wide-area Video Surveillance and Selective Attention, in proc. of IEEE Int. Conf. of Computer Vision and Pattern Recogination, 1, 537–544 (2006).
4. Micheloni, C., Foresti, G.L., Snidaro, L.. A network of Co-operative Cameras for Visual Surveillance, IEE-proc. Vis. Image Signal Process., 152(2), 205–212 (2005).
5. Brown, M., Burschka, D., Hager, G.D.: Advances in Computational Stereo, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(8), 993–1008 (2003).
6. Wan, D., Zhaou, J.: Stereo Vision Using Two PTZ Cameras, Computer Vision and Image Understanding, 112(2), 184–194 (2008).
7. Hart, J., Scassellati, B., Zucker, S.W.: Epipolar Geometry for Humanoid Robotic Heads, In proc. of 4^{th} International Cognitive Vision Workshop, 24–36 (2008).
8. Isgro, F., Trucco, E.: On Robust Rectification of Uncalibrated Images, In proc. of 10^{th} International Conference on Image Analysis and Processing, 297–302 (1999).
9. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 2(60), 91–160 (2004).
10. Meltzer, J., Soatto, S.: Edge descriptors for robust wide-baseline correspondence, In proc. of IEEE Int. Conf. on Computer Vision and Pattern Recogination, 1–8, (2008).
11. Kumar, S., Micheloni, C., Foresti, G.L.: Stereo Vision in Cooperative Camera Networks, Smart Cameras, Springer Science+Business Media, Inc.(in press) (2009).