Dual PTZ Stereo Rectification for Target Localization and Depth Map Computation

Sanjeev Kumar and Christian Micheloni, *Member, IEEE*

***Abstract*—We present an active stereo vision system composed by two pan-tilt-zoom (PTZ) cameras for video surveillance applications. The rectification of stereo images is performed based on the sigmoid interpolation with a set of neural networks. The orientation parameters (pan and tilt values) and the rectification transformations of corresponding images are used as the input-output pairs for network's training. The input data is directly read from cameras, whereas the output data is computed off-line. The trained neural network is used to interpolate rectification transformations in real time for the stereo images captured at arbitrary pan and tilt settings. The correspondence between the stereo images is obtained using a chain of homographies based scheme. Heterogeneity between the intrinsic parameters of the cameras is managed through zoom compensation to improve the quality of stereo rectification. This stereo active vision system is used for two different video surveillance applications: localization of partially occluded targets and construction of multiresolution depth map mosaic for scene understanding.**

***Index Terms*—Disparities, Look-Up-Table (LUT), PTZ Camera, Sigmoid Interpolation, Stereo Vision, Target Localization.**

I. Introduction

The developments of modern video surveillance systems have been attracted a lot of interests [1]–[5]. Recently, the concept of stereo vision has been exploited in surveillance systems to make them more efficient. Stereo vision has the advantage to estimate the 3D position of an object in a given coordinate system from two perspective images [6]. Traditional stereo vision reserach uses static cameras for their low cost and relative simplicity in modeling. A pan-tilt-zoom (PTZ) camera is a typical and the simplest active camera, whose pose can be fully controlled by pan, tilt and zoom parameters. As PTZ cameras are able to obtain multi-angle-views and multiresolution information (i.e. both global and local image information), these are used for wide area monitoring [7]. Thus, a dual PTZ stereo vision system, composed by a pair of PTZ cameras, is able to cover large environment and to reduce the occlusions drawbacks. However, such type of active stereo vision systems are much more challenging when compared to the traditional stereo vision system. PTZs, on purpose (e.g. zoom on face, zoom on license plate etc.), can vary both the intrinsic and the extrinsic parameters thus changing the stereo properties. In this context, it is not an easy task to perform some operations like stereo image rectification in an active stereo vision system.

Recently, a novel image rectification algorithm has been proposed for a dual-PTZ-camera based stereo system [8]. In such a system, the inconsistency of the intensities in two camera images is solved by addressing a two-step stereo matching strategy. Another interesting approach in case of active stereo vision system has been proposed with the analytic formulation in [9]. An off-line initialization process is performed to initialize essential matrix using known calibration parameters. During on-line operations the rotation angles of the cameras are retrieved and exploited to compute the current essential matrix. However, if the zoom is considered, the calibration for any adopted zoom level is required for both cameras. Moreover, the discrepancies [10] in the field-of-view (FOV) and in magnification of the two cameras lead difficulties not only into the stereo rectification but also in the depth estimation.

To solve some of the aforementioned problems, a new image rectification process for an active stereo system composed by two PTZ cameras is here proposed. A Look-Up-Table (LUT) associating the rectification transformation for specif pan and tilt values of the PTZ pair is constructed off-line. The rotation ranges (pan and tilt) of the two cameras are sampled. The rotation parameters are interpolated with respect to given pan and tilt values for computing the required rectification transformations. Neural network based sigmoid interpolation is adopted due to its function approximation property in case of highly nonlinear data. The data from a Look-Up-Table (LUT) are used for the network's training. Such a LUT is constructed off-line by sampling the pan and tilt ranges of the two PTZ cameras for a common zoom level. It contains the pan-tilt combinations and the rotation parameters of the rectification transformations as the independent and the dependent variables, respectively. In case of zoom-in or zoom-out operations in any PTZ camera, a focal ratio based approach is used to compensate the effect of unequal zoom levels [11] between the two cameras.

To show the effectiveness of the proposed approach, two different applications are considered. The first application shows the stereo vision based localization of a partially occluded target on a given ground plane map. Existing non-stereo systems often localize objects in the environment by defining homographies between single cameras and a 2D map [12]. 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. The proposed

This work was partially supported by the Italian Ministry of University and Interreg 4687-SRSNet Project.

Sanjeev Kumar is with Department of Mathematics, IIT Roorkee, India and Christian Micheloni is with the Department of Mathematics and Computer Science, University of Udine, Via Della Scienze 206, Udine 33100, Italy

Email: malikfma@iitr.ernet.in, christian.micheloni@dimi.uniud.it

Corresponding Author: Christian Micheloni

active stereo vision system solves this problem by making the localization based on a stereo camera system instead of a single camera based system. It can be used subsequently in the computation of reliable object's trajectories [13]–[18] which is really important for different contexts like traffic monitoring, behavior analysis [19]–[21], suspicious event detection [22], sensor network configuration [17], [23]–[33], etc. In the second application, we emphasize the importance of zoom settings of PTZ cameras in the scene understanding. In case of static cameras based stereo system, the images are captured with the same resolution. However in case of PTZ cameras, we can consider the two facts: 1) if a region has small depth variations, i.e. almost flat in nature, low resolution images can be used for obtaining the depth map, and 2) when large depth variations occur in a region, high resolution images are required. Based on these two facts, PTZ cameras based stereo vision system provides a multiresolution depth map that can be used for a better scene understanding and required low computational cost in case of a wide-area. In this context, another application of dual PTZ camera based stereo system is to grab images with the above facts in an automatic manner and creation of a multiresolution depth map mosaic of a wide area.

In brief, the main advantages of the proposed PTZ camera based stereo vision system are:

- there is no need to assume a fixed center of projection for the PTZ camera during rotations.
- the 3D localization of the objects works with high accuracy even in case of partially occluded objects.
- both localization and depth-map computations can be achieved with wide baseline stereo systems.
- only limited a-priori information (i.e., information provided by a static camera) is required to compute multiresolution depth maps for a large environment.

In particular, concerning the last advantage, contrarily to [34], there is no need to have the coarse depth map and the precise FOV of the left camera. In addition, instead of using wide baseline feature matching techniques [35], [36], that even though efficient are computationally expensive, an approach based on a chain of 2D homographies is proposed to find corresponding points between wide baseline images in real-time.

II. System Architecture and Correspondence between Stereo Images

A. System Description

The proposed system contains mainly two different units of cameras. The first unit, called static camera unit (SCU), is composed of a generic number of static cameras. These static cameras have wide FOV and cover a large environment with limited overlapping FOV. The second unit contains two different PTZ cameras placed at a wide distance (7 meters) from each other and considered as a dual-PTZ based stereo system. This unit is called active stereo unit (ASU), and maintains a good cooperativeness with SCU. The main functionalities of SCU are object detection [12], behavior understanding and anomalous event detection [22]. Once a region of interest is detected by the SCU, the system delivers the information to the ASU for focusing the two PTZ cameras towards the selected region. The ASU starts the stereo task as soon as the selected target appears in the FOVs of both cameras. The handover of scene information [37] between the different cameras allows a cooperative tracking of the objects within the monitored environment. The sequences acquired by the two PTZ cameras in ASU are transmitted to a central node together with their respective orientation and resolution information. A communication system based on a multi-cast protocol [5] is used for the cooperation within these cameras' network. This communication system is designed in such a way that it requires a low bandwidth. A logic architecture of the proposed system is shown in Fig. 1, where the top layer of the cameras represents the SCU, while the ASU is shown in the second layer.

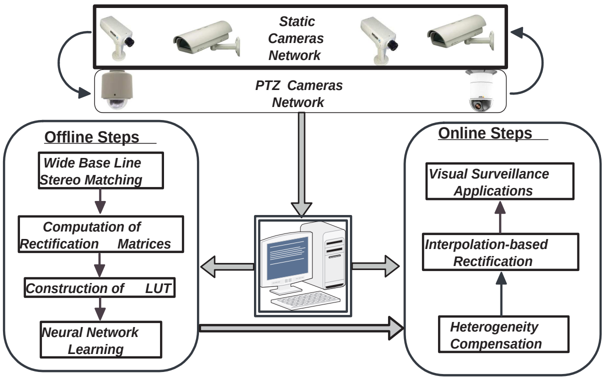
Static Cameras Network
PTZ Cameras Network
Offline Steps
Wide Base Line Stereo Matching
Computation of Rectification Matrices
Construction of LUT
Neural Network Learning
Online Steps
Visual Surveillance Applications
Interpolation-based Rectification
Heterogeneity Compensation

Fig. 1. A virtual design of the proposed stereo system.

The properties of the PTZ camera deployment make the stereo vision problem more difficult when compared to classical stereo systems. In particular, in the proposed system the captured images from the pair of PTZ cameras are heterogeneous in nature (may have different intrinsic parameters). If we perform rectification on these heterogeneous pairs, it will introduce relevant errors (distortion effect) in the rectified images. Thus, it is difficult to compute disparity maps from these erroneously rectified images. Therefore, the effect of these unequal intrinsic parameters must be compensated before rectification. Here, the images are made homogeneous in terms of internal image parameters using the resolution information. The focal lengths are estimated directly from the zoom value. The ratio between the zoom values of two cameras is used to compensate the effect of heterogeneity. Once the frames are homogeneous, the rectification transformations are interpolated using a neural network.

B. Correspondence Between Wide Baseline Stereo Images

SIFT matching [38] is a popular tool for extracting matching points from a pair of stereo images. However, this method is not very accurate in a case when images do not share sufficient common FOV. It can happen when objects are close to both the cameras placed at a wide baseline (See Fig. 2). To sidestep such a problem, we propose a method based on a chain of homographies for establishing the correspondence between the pair of stereo images. Our idea is to initialize the correspondence between the images of two PTZ cameras captured for a far scene and then subsequently use it for other pair of images. To do this, we require the correspondence between different overlapped images captured at different pan and tilt settings in case of each PTZ camera which can be obtained using SIFT.

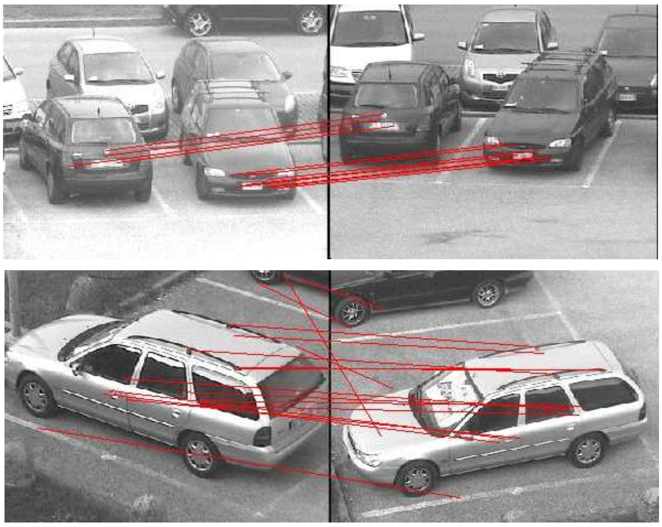

Fig. 2. SIFT matching between wide baseline stereo images of a far (top) and near (bottom) scenes along the optical axis of camera.

Let $(\mathbf{I}_l^1, \mathbf{I}_r^1)$ be a pair of images of a far scene and $\mathbf{H}^1$ be the homography obtained from the SIFT based matching points between these two images. Let $(\mathbf{I}_l^n, \mathbf{I}_r^n)$ be a pair of images of a scene/object near to the cameras along their optical axis. The problem is to autonomously establish the correspondence between the images $(\mathbf{I}_l^n, \mathbf{I}_r^n)$. Such a correspondence can be established by capturing n images between the scenes those are in $\mathbf{I}_l^1$ and $\mathbf{I}_l^n$ from the left PTZ camera. A similar image grabbing process is required for right PTZ camera to capture the images between the scenes those are in $\mathbf{I}_r^1$ and $\mathbf{I}_r^n$. Let these two sets of images be $(\mathbf{I}_l^1, \mathbf{I}_l^2 \ldots \mathbf{I}_l^n)$ and $(\mathbf{I}_r^1, \mathbf{I}_r^2 \ldots \mathbf{I}_r^n)$. The required correspondence between the images $(\mathbf{I}_l^n, \mathbf{I}_r^n)$ in terms of a homography $\mathbf{H}^n$ can be achieved in the following steps:

1) Establish the correspondence between image pairs $(\mathbf{I}_l^1, \mathbf{I}_l^2)$, $(\mathbf{I}_l^2, \mathbf{I}_l^3)$, $\ldots$, $(\mathbf{I}_l^{n-1}, \mathbf{I}_l^n)$ in terms of their respective homographies $\mathbf{H}_l^{1,2}$, $\mathbf{H}_l^{2,3}$, $\ldots$, $\mathbf{H}_l^{n-1,n}$ such that $\mathbf{I}_l^i = \mathbf{H}_l^{i,i+1}\mathbf{I}_l^{i+1}$ for $i = 1, \ldots, n-1$.
2) Repeat the procedure given in the above step to compute $\mathbf{H}_r^{1,2}$, $\mathbf{H}_r^{2,3}$, $\ldots$, $\mathbf{H}_r^{n-1,n}$ for the images captured with the right camera.
3) Compute the homographies $\mathbf{H}_l$ and $\mathbf{H}_r$ as

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

4) Compute the required homography matrix $\mathbf{H}^n$ for the pair the images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ as

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

The homography $\mathbf{H}^n$ can be used to establish correspondence between the images $\mathbf{I}_l^n$ and $\mathbf{I}_r^n$ which is not easy to obtain directly in case of wide baseline stereo systems. Fig. 3 gives an intuitive interpretation of the above described 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 the multiplication of several matrices. In order to minimize this error: 1) we keep the sampling step (i.e. the difference in pan and tilt angles) as low as possible, with a constraint that any pair of images (e.g. $\mathbf{I}_{l/r}^i, \mathbf{I}_{l/r}^{i+1}$) has to share at least 30% of the FOV; 2) outliers from matching points should be removed before applying a robust approach for the homography estimation.

III. Offline Steps

It is necessary to perform an offline initialization for deriving all the information necessary to determine the rectification transformations during online operations. This includes the computation of the rectification transformations for image pairs captured at different pan and tilt sampling from two PTZ cameras. The rotation parameters related to these rectification transformations are stored in the LUT corresponding to the respective pan and tilt values of the PTZ cameras. The LUT data is used for the training of a set of neural networks that are used for sigmoid interpolation of these transformations in real time.

A. Computation of Rectification Transformations and 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 as:

$$\mathbf{J}_l = \mathbf{A}_l \mathbf{I}_l \qquad \mathbf{J}_r = \mathbf{A}_r \mathbf{I}_r$$

where $(\mathbf{J}_l, \mathbf{J}_r)$ are the rectified images and $(\mathbf{A}_l, \mathbf{A}_r)$ are the rectification matrices. In case of uncalibrated cameras based stereo system, a quasi epipolar rectification [39] has been proposed for computing these rectification transformations by minimizing the following function.

$$\sum_i [(\mathbf{m}_l^i)^T \mathbf{A}_r^T \mathbf{F}_\infty \mathbf{A}_l \mathbf{m}_l^i]^2 \tag{3}$$

where $(\mathbf{m}_l, \mathbf{m}_r)$ are pairs of matching points between images $\mathbf{I}_l$ and $\mathbf{I}_r$. $\mathbf{F}_\infty$ is the fundamental matrix for the rectified pair of images. Generally, the minimization of (3) is time-consuming and therefore it is not easy to compute the rectification transformations in real time. Here, we use this scheme [39] for computing rectification transformations offline for the image pairs captured at different pan and tilt sampling. In real time, this information can be used for computing rectification transformations for a given pan and tilt setting by using sigmoid interpolation. Recently, in [40] such an interpolation based method is adopted to make rectification of stereo pairs in real time. An offline LUT containing rectification matrices corresponding to various image pairs captured at predefined

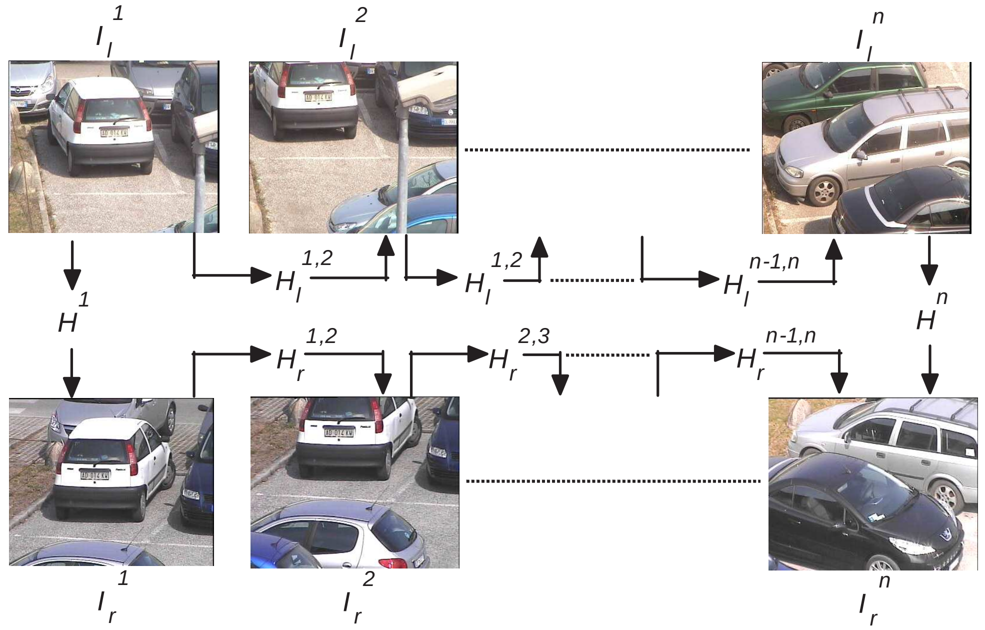

Fig. 3. Wide baseline stereo matching using a chain of homographic matrices.

pan and tilt angles is constructed. Then, the rectification transformations can be interpolated in real-time for any arbitrary orientation of both PTZ cameras by using LUT data. However, the interpolation of eighteen parameters (nine elements for each rectification transformation) is again computationally expensive. Here, our effort is to reduce the number of these interpolated parameters into six instead of eighteen by using some suitable assumptions of camera projection matrix.

The meaning of stereo image rectification is that for any given pair of the original camera projection matrices $\mathbf{P}_l$ and $\mathbf{P}_r$, two new virtual projection matrices $\hat{\mathbf{P}}_l$ and $\hat{\mathbf{P}}_r$ can be obtained to rotate the cameras around their optical centers until the focal planes become coplanar. Therefore, the rectification transformations $\mathbf{A}_l$ and $\mathbf{A}_r$ can be decomposed as

$$\mathbf{A}_l = \hat{\mathbf{P}}_l \mathbf{P}_l^{-1}, \quad \mathbf{A}_r = \hat{\mathbf{P}}_r \mathbf{P}_r^{-1} \tag{4}$$

A camera matrix $\mathbf{P}$ can be decomposed into the intrinsic and extrinsic matrices $\mathbf{P} = \mathbf{KD}$, where $\mathbf{K}$ is the intrinsic matrix and $\mathbf{D} = [\mathbf{R} \;\; \mathbf{t}]$ denotes the extrinsic matrix containing rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. Since there is no translation involved in the rectification process, (4) can be rewritten as

$$\mathbf{A}_l = \hat{\mathbf{K}}_l \bar{\mathbf{R}}_l \mathbf{K}_l^{-1}, \quad \mathbf{A}_r = \hat{\mathbf{K}}_r \bar{\mathbf{R}}_r \mathbf{K}_r^{-1} \tag{5}$$

where $\bar{\mathbf{R}}_l = \hat{\mathbf{R}}_l \mathbf{R}_l^{-1}$ and $\bar{\mathbf{R}}_r = \hat{\mathbf{R}}_r \mathbf{R}_r^{-1}$ are the rotation matrices involved in rectification process. Here, the original intrinsic parameter matrices $(\mathbf{K}_l, \; \mathbf{K}_r)$ and the rotation matrices $(\mathbf{R}_l, \; \mathbf{R}_r)$ are unknown, whereas the new intrinsic matrices $(\hat{\mathbf{K}}_l, \; \hat{\mathbf{K}}_r)$ can be set arbitrarily, provided that the focal lengths and the coordinates of the principal points must be equal. During the rectification process, the unknown intrinsic parameters can be reduced by considering the zero skew, square pixel and principal point in the center of the image assumptions. Then the intrinsic matrices can be written as:

$$\hat{\mathbf{K}}_l = \begin{pmatrix} f_l & 0 & w/2 \\ 0 & f_l & h/2 \\ 0 & 0 & 1 \end{pmatrix} ; \hat{\mathbf{K}}_r = \begin{pmatrix} f_r & 0 & w/2 \\ 0 & f_r & h/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where w and h are the width and the height of the image. The focal lengths f_l and f_r can be computed directly by reading the zoom parameter of the two PTZ cameras. Thus, the problem of computing a pair of rectification transformations is converted into the computation of only two rotation matrices $(\bar{\mathbf{R}}_\mathbf{l}, \bar{\mathbf{R}}_\mathbf{r})$. Hence, for any pan and tilt combination, only three rotation parameters has to be stored in the LUT instead of nine entries of a rectification transformation.

Thus, the main steps to construct the LUT are:

1) The overall monitoring wide-area is divided into a number of subarea in such a way that each subarea is covered in the FOV of each PTZ camera just by changing the pan and tilt angles setting $(p_l^i, t_l^j)_{i=1:1:n_p}^{j=1:1:n_t}$.
2) Capture $n_{tot} = (n_p \times n_t)^2$ pairs of images of all these local subarea with the two PTZ cameras at equal zoom.
3) Compute the possible $k(> n_{tot})$ pairs of rectification transformation $(\mathbf{A}_l^k, \mathbf{A}_r^k)$ for the different combination of stereo images. The used images pairs should have images sharing at least the 30% of their FOV.
4) Decompose rectification transformations as per earlier described scheme and compute their corresponding rotation parameters.
5) Store the rotation parameters in a LUT as dependent variables corresponding to their four independent variables (p_l, t_l, p_r, t_r).

The main problem to be addressed in the creation of the LUT is the establishment of the correspondence between wide baseline stereo images. This has been solved by exploiting the earlier described chain of homographies based approach.

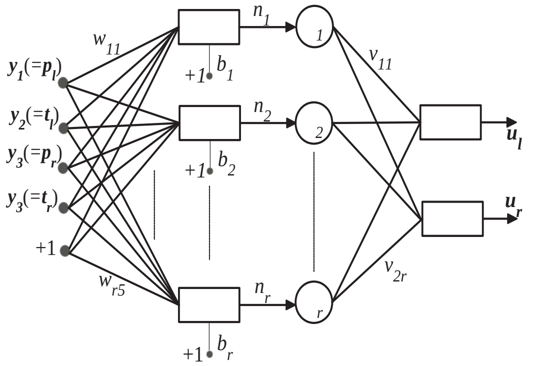

Fig. 4. Architecture of employed neural network

B. Training a neural network using LUT

sigmoid interpolation via a set of neural networks is used for computing the rectification transformations in real time corresponding to any arbitrary orientation of two PTZ cameras. The data stored in LUT is used to train the neural networks. The neural network based interpolation has been chosen due to its strong function approximation property with respect to highly non-linear data. A supervised learning scheme [11] using LUT data has been adopted for the off-line training of the neural networks.

The network considers the pan and tilt angles as input and returns the parameters of the rotation matrices corresponding to the required rectification transformations $(\mathbf{A}_l, \mathbf{A}_r)$ as output. The sets of input and output data are related by a non-linear mapping $\mathbf{U} = f(p_i, t_i)$. For a known set of input-output values, the problem is to find the function $F(\cdot)$ that approximates $f(\cdot)$ over all inputs. That is,

$$\| F(p,t) - f(p,t) \| < \epsilon \qquad \text{for all;} \, (p,t), \tag{7}$$

where ϵ is a small error. The architecture of the proposed neural network is shown in Fig. 4, where two output nodes are corresponding to the angles for left and right rotation matrices. Three different networks are trained for yaw, pitch and roll elements of the rotation matrices. A detailed learning process for the proposed network is given in [11], where back-propagation algorithm is used with gradient information.

C. Zoom to Focal Length Fitting

As aforementioned, the proposed framework is based on a zoom compensation process in case of heterogeneous image-pairs. The effect of such an unequal zoom is compensated by using a focal ratio information which requires the focal lengths corresponding to both images. For a static camera, the focal length can be estimated offline once considering that the image parameters (specifically focal length) will remain constant for the whole process. In case of PTZ cameras, the focal length changes as the zoom level is changed to zoom in/out. Thus, the determination of the accurate focal length associated to any acquired frame is a fundamental even though not an easy task. Moreover, if its computation is not precise enough, the rectification accuracy of the proposed algorithm could be significantly affected. To overcome such a problem, the focal length is computed in two steps: a) offline fitting of focal lengths corresponding to zoom settings and b) online estimation given a particular zoom level.

Concerning the first step, the aim is to find out a mapping between the zoom value and the corresponding focal length. For such a purpose, the whole zoom range is sampled and the focal length is estimated by using a calibration process for every sampled zoom tick. In the case of motorized lenses [41], the relation between a given zoom tick z and corresponding focal length f is

$$f(z) = \frac{a_0}{1 + a_1 z + a_2 z^2 + a_3 z^3 + \ldots + a_n z^n} \tag{8}$$

where the order n and the unknown $a_0, \ldots, a_n$ are camera dependent. For the adopted camera, following the methodology in [41], the estimated optimal value of n is 2. Therefore, a_0, a_1 and a_2 can be estimated by minimizing the following nonlinear function

$$C(a) = \sum_{i=1}^{K} \left[f(z_i) - \frac{a_0}{1 + a_1 z + a_2 z^2} \right]^2 \tag{9}$$

However from (9), the estimation of the focal length is not reliable for small values of zoom, then (9) can be written as

$$C(b) = \sum_{i=1}^{K} \left[p(z_i) - (b_0 + b_1 z + b_2 z^2) \right]^2 \tag{10}$$

where $b_0 = 1/a_0$, $b_1 = a_1/a_0$, $b_2 = a_2/a_0$ and $p(z_i) = 1/f(z_i)$ denotes the lens power. The minimization of (10) is reliable for lower as well as higher zoom settings. The values b_0, b_1 and b_2 corresponding to the minimum value of $C(b)$ are chosen to define the optimal values of a_0, a_1 and a_2. In the real time, the focal length f for any given zoom level z is estimated as

$$f(z) = \frac{a_0}{1 + a_1 z + a_2 z^2} \tag{11}$$

The above method has been tested on various zoom samples and it has been found reliable for estimating the focal length corresponding to a given zoom.

IV. Online steps

During tracking, stereo tasks can be performed by applying a zoom compensation followed by the rectification of the resulting images. This section contains a detailed description of these two steps.

A. Unequal Zoom Compensation

The proposed framework allows to operate with couples of PTZ camera acquiring images with different zoom levels. This introduces a heterogeneity between internal imaging parameters of both cameras. However, equivalent zoom values have been used for the two PTZ cameras during the construction of the LUT containing rectification transformations. Therefore, a compensation is required to deal with this heterogeneity with real time performance. A novel approach based on the focal lengths of the two cameras is used to tackle such heterogeneity. In a perspective projection model, the position of any pixel is always proportional to the focal length for the respective camera. Therefore, if the two images are acquired with different zoom levels, then this heterogeneity can be compensated by shrinking the higher zoom image with a focal ratio information.

Let $\mathbf{I}_l$ and $\mathbf{I}_r$ of size $w \times h$ be the two images captured at different zoom levels z_l and z_r from the dual PTZ cameras. Let the corresponding focal lengths be f_l and f_r obtained from earlier described scheme. The idea behind the process of heterogeneity compensation is achieved by shrinking the image having longest focal length by mean of a focal ratio. The overall compensation algorithm is given in Algorithm 1.

Algorithm 1. Compensation of unequal zoom settings in PTZ stereo

```
Read $(z_l, z_r)$
$Calculate\{f_l, f_r\} = Interpolation(z_l, z_r)$
**if** $f_l = f_r$ **then**
    STOP
**else if** $f_l > f_r$ **then**
    $R = f_l/f_r$
    $I_l' = Shrink(I_l, R)$
    $I_l^h = Zeropad(I_l', Size\{I_r\}) \;\; and \;\; I_r^h = I_r$
**else**
    $I_r' = Shrink(I_r, 1/R)$
    $I_r^h = Zeropad(I_r', Size\{I_l\}) \;\; and \;\; I_l^h = I_l$
**end if**
```

where the function $Shrink(I_l, R)$ represents that the image I_l is shrunk by a factor of R. The function $Zeropad(I_l', Size\{I_r\})$ denotes that the zero padding is performed around image I until its size becomes equal to the size of I_r. The image pair (I_l^h, I_r^h) is homogeneous in terms of the intrinsic image parameters which is necessary to rectify the stereo images correctly.

B. Rectification of Images

Once the zoom compensation is completed, the new pair of images has to be rectified for further stereo processing. This operation can be achieved in the following steps:

1) Interpolate the parameters for generating rotation matrices $\bar{\mathbf{R}}_l^c$ and $\bar{\mathbf{R}}_r^c$ from the trained neural network by giving the pan and tilt angles as input for current pair of frames.
2) Calculate rectification transformations for current frames as

$$\mathbf{A}_l^c = \hat{\mathbf{K}}_l^c \bar{\mathbf{R}}_l^c (\mathbf{K}_l^c)^{-1}, \quad \mathbf{A}_r^c = \hat{\mathbf{K}}_r^c \bar{\mathbf{R}}_r^c (\mathbf{K}_{or}^c)^{-1} \tag{12}$$

where, $(\mathbf{K}_l^c, \mathbf{K}_r^c)$ and $(\hat{\mathbf{K}}_l^c, \hat{\mathbf{K}}_r^c)$ are the pairs of intrinsic matrices in the original and the rectified cameras' geometries.
3) Warp the current pair of frames as a rectified pair of images using $\mathbf{A}_l^c$ and $\mathbf{A}_r^c$.

$$\mathbf{J}_l^c = \mathbf{A}_l^c \mathbf{I}_l^c, \qquad \mathbf{J}_r^c = \mathbf{A}_r^c \mathbf{I}_r^c$$

The above procedure is performed in real time. In this way, rectified pairs of frames can be obtained by using the orientation information of left and right PTZ cameras.

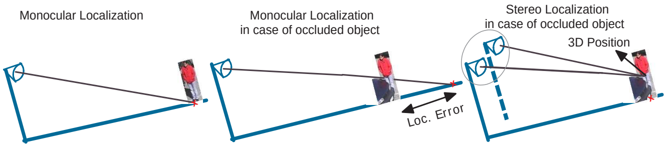

Fig. 5. Target localization. Left to right: monocular camera based approach, monocular camera based approach in case of partially occluded target and proposed stereo camera based approach in case of partially occluded target.

V. Applications

The rectified stereo frames can be used for several applications in the intelligent video surveillance domain. These applications include target recognition/detection, localization on a 2D map, tracking and activity recognition. In this work, we focused on target localization and generation of high resolution depth map images of a selected region. Finally, the high resolution depth maps are used to generate a depth map mosaic of a large environment for scene understanding.

A. Stereo Vision based Target Localization

The localization of the moving targets on a given 2D map, is an essential step for tracking in a complex environment. Usually, the localization task is done by transforming the target's position (usually its lower pixel) in the image into a position onto a 2D map. A priori estimated 2D homography matrix between ground plane of test environment and a given 2D map is used in transformation. The accuracy of the localization depends on the contact position of the target with the ground plane and on the robustness of the homography estimation. When the target is partially occluded, i.e. its contact with the ground plane is not visible, the accuracy of the localization does not meet its requirements (see Fig. 5). In such cases, stereo vision can be used in the localization by taking the advantage of 3D position of a target, i.e. by calculating the ground plane position from stereo 3D data and transforming it onto given 2D map.

For reducing the computational cost for calculating the 3D position (X_w, Y_w, Z_w), we restrict the correspondence search only on the corresponding epipolar lines related to target region in the two rectified images. To match a point of the target in the left image, a sliding window is applied only on the corresponding epipolar line in the right image. The disparity $d(x, y)$ for a pixel (x, y) is computed by minimizing the cost function $C(x, y, d)$ from (13). To do this, windows are compared through the normalized sum of square differences (SSD) measure, which quantifies the difference between the intensity patterns.

$$C(x,y,d) = \frac{\sum\limits_{(\xi,\eta)} [\mathbf{I}_l(x+\xi, y+\eta) - \mathbf{I}_r(x+d+\xi, y+\eta)]}{\sqrt{\sum\limits_{(\xi,\eta)} \mathbf{I}_l(x+\xi, y+\eta)^2 \sum\limits_{(\xi,\eta)} \mathbf{I}_r(x+\xi, y+\eta)^2}} \tag{13}$$

where, $\xi \in [-n, n]$ and $\eta \in [-n, n]$ represent the dimensions of the sliding window along horizontal and vertical directions, respectively. It can be observed that squared differences need to be computed only once for each disparity. Moreover, when the window moves by one pixel, the convolution sum can take advantage of the previous computation without requiring its computation from scratch.

Once the disparity d is computed for the position of the target from left and right images, the distance of the target Z_w from the camera along optical axis is computed using the following formula

$$Z_w = f\frac{b}{d} \tag{14}$$

where f is the focal length for the rectified pair of images and b denotes the distance between the optical centers of two PTZ cameras. Let (x_l, y_l) be the position of the target in the left camera image, then the position of the target in the plane normal to the optical axis of the camera is given by

$$X_w = \frac{x_l Z_w}{f} \qquad\qquad Y_w = \frac{y_l Z_w}{f}$$

The location (x_g, y_g) of a target in a ground plane map is calculated as follows

$$\begin{bmatrix} x_g \\ y_g \\ 1 \end{bmatrix} = \mathbf{H}_m^w \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix}$$

where $\mathbf{H}_m^w$ is the homography matrix between the homogeneous coordinates of ground plane position (X_w, Y_w) of some selected points and their respective position (x_g, y_g) on the given test map.

B. High Resolution Depth Map Estimation and Mosaic Construction

The second application of the proposed system is studied for scene understanding in the case of a large environment. Depth obtained from the stereo images can be a very crucial cue in scene understanding. In a scene having large variations in depths at various positions (like parking lot or a hill), it is necessary to use higher resolution images for obtaining depth map. In case of a flat region (like empty ground) where depths at different points have smooth variation, lower resolution images can be used to obtain the depth map. Such a multiresolution depth map based strategy is useful for establishing a trade off between accuracy and computational cost. Finally, the depth map of the whole environment can be obtained by making the mosaic of several overlapped and multiresolution depth maps.

1) Depth Map Estimation: A multi step process is proposed for selecting the optimal zoom values of the two cameras according to the earlier described strategy. When an event of interest has taken place in the FOV of any static camera (let say S_1), this camera delivers the information to the dual PTZ cameras for focusing on the region of interests. Let S_1 delivers the information to the PTZ cameras C_l and C_r. Firstly, the initial resolutions for C_l and C_r is set in such a way that it covers the whole scene, i.e. a low resolution more or less equivalent to the static camera. In the second step, the resolution for C_l is refined to acquire the selected region with maximum resolution. Then the disparity map is calculated between the high resolution image from left camera and a low resolution image of right camera. The variation of depths are checked from the disparity map to classify the associated region as flat or complex. In the former case, the process is stopped and the computed disparity map is used to compute depths. In the latter case, further processing is required. The following steps are proposed to obtain high resolution depth maps:

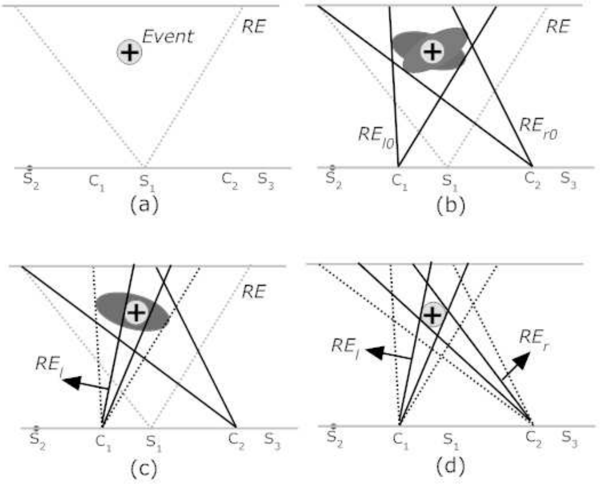

Fig. 6. Procedure for acquainting region of interests by PTZ cameras with maximal resolution.

1) Detect the event of interest in the static camera S_1.
2) Deliver the information to both PTZ cameras (C_l, C_r) system for focusing towards the regions of interest. Let the region visible in FOVs of these cameras initially be RE_l^0 and RE_r^0 and the corresponding images be $\mathbf{I}_l^0$ and $\mathbf{I}_r^0$.
3) Change the resolution of the left camera in such a way that the event of interest is acquired with best possible resolution to capture the image $\mathbf{I_l}$. Some priori information about various zoom settings of the left camera and its corresponding FOV are used to adopt the zoom setting for best possible resolution.
4) Compute the disparity map $\mathbf{D}^0$ (having disparities $d^0(x, y)$ for all (x, y)) from the images $\mathbf{I}_l$ and $\mathbf{I}_r^0$.
5) Check whether the disparities $d^0(x, y)$ have large variations for all (x, y) in the disparity map $\mathbf{D}^0$. If not, stop the algorithm and use $\mathbf{D}^0$ for computing its corresponding depth map $\tilde{\mathbf{D}}$. Otherwise, proceed to the next step.
6) Compute an image $\mathbf{I}_r^c$ as $\mathbf{I}_r^c = \mathbf{I}_l + \mathbf{D}^0$.
7) Change the resolution of the right camera based on the image $\mathbf{I}_r^c$ and acquire a new image $\mathbf{I}_r$ with such a resolution.
8) Find the higher resolution disparity map $\mathbf{D}$ from the images $\mathbf{I}_l$ and $\mathbf{I}_r$.
9) Compute the depth map $\tilde{\mathbf{D}}$ from the disparity map $\mathbf{D}$.

Fig. 6 provides a graphical representation of the process described in the above steps.

2) Construction of Depth Map Mosaic: In general, two approaches can be used to obtain a depth map mosaic of a large scene. The first approach [42] works by stitching the overlapped images for each camera separately to obtain the two stereo panoramic images and then performing the disparity estimation. The second way foresees to compute a depth map for each stereo image pair and then mosaic all the depth maps to construct the panoramic depth map. The main difficulty in the later one is the estimation of matching points between the depth maps of overlapped images, since it is very difficult to apply feature matching between depth maps. To cope this problem, we use the same transformation matrices which are used for stitching the images of left camera. However, the second approach has the following advantages when compared to the earlier one.

- Multiresolution depth cues can be easily maintained in final depth map mosaic.
- We obtain the final depth map mosaic (for a large region) by stitching several depth maps (of various small regions). In this context, the depth value for each pixel belonging to the overlapped regions in consecutive images is calculated by fusing two depth cues, so the robustness and accuracy of the final depth mosaic can be maintained.
- The final depth map can be updated anytime for a new image pair.

The use of disparity drift [34] compensates the uncertainty in the reading of pan, tilt and zoom parameters which is required for correct interpolation of rotation parameters associated with their corresponding rectification transformations. Assuming that there are n rectified pairs $(\mathbf{I}_l^i, \mathbf{I}_r^i)$ of stereo images captured at different pan, tilt and zoom settings. The following steps are adopted to construct the final depth map mosaic.

1) Perform stereo matching between all image pairs $(\mathbf{I}_l^i, \mathbf{I}_r^i)$, and obtain their corresponding disparity maps $\mathbf{D}^i$ for $1 = 1, 2, \ldots, n$.
2) Normalize the gray-level values between consecutive disparity maps. The process starts from the maps used to specify the reference panoramic image coordinate system. This process can be done by finding the linear regression parameters (α_i, β_i) between each consecutive pairs of disparity maps for all matching pixels (x_m, y_m).

$$\mathbf{D}^{(i+1)} = \alpha_i \ \mathbf{D}^i + \beta_i \tag{15}$$

where $i = 1, 2, \ldots, n-1$.
3) Calculate the disparity drift ρ^i for each disparity map $\mathbf{D}^i$.
4) Compute the modified disparity maps as

$$\mathbf{D}_r^i = \mathbf{D}^i + \rho^i \mathbf{I}_d$$

where $\mathbf{I}_d$ represents an identity matrix having the same size as the disparity map $\mathbf{D}$.
5) Compute the depth maps $\tilde{\mathbf{D}}^i$ from their corresponding disparity maps $\mathbf{D}_r^i$.
6) Construct the depth map mosaic $\mathbf{DMM}$ by stitching all depth maps $\tilde{\mathbf{D}}^i$ for $i = 1,\ 2,\ \ldots,\ n$, into the reference panoramic image coordinate system.

Sometimes for a complex scene, a fusion of several depth cues is required for a better representation of the depths for scene understanding. A weighted average method as in [34] can be used for fusing several depth cues together.

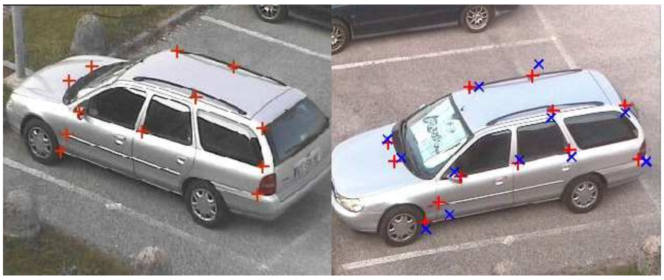

Fig. 7. SIFT ‘x’ and Chain of Homographies ‘+’ based correspondence between wide baseline stereo images.

VI. Results and Discussions

For the experimental validation of the proposed framework, a network of static cameras composed by AXIS 221 network cameras has been adopted. For the stereo unit, two different PTZ cameras (i.e., Axis 213 and Axis 233D) are used. Four different types of experiments have been performed to: 1) evaluate the correspondence between stereo images captured with the two cameras placed far away from each other (i.e., wide baseline stereo), 2) evaluate the proposed interpolation based rectification algorithm for various pairs of stereo images having unequal zoom, 3) evaluate the proposed algorithm for target localization on a given 2D map and 4) evaluate the computation of high resolution depth map mosaic for large scene understanding. Different criterions have been used for comparing the performance of the proposed framework in each case.

A. *Correspondence between wide baseline stereo images*

To show the importance of the chain of homographies based matching algorithm in case of wide baseline stereo images, correspondence between a pair of images has been considered. First, a homography H^d has been computed by using the matching points extracted with SIFT method between this pair of images. Then, the homography has been evaluated by using the proposed chain of homographies based approach. To do this, a pair of stereo images has been captured of a far scene where SIFT can be implemented accurately. Then, two different chains of homogrpahies have been computed using five different tilt positions in case of each cameras separately. Finally, the final homography H^n has been computed using (2). Corresponding points in the right image have been computed for 12 selected points in left image using the homographies (H^d and H^n). Fig. 7 shows the results for this experiment and it can be observed that the corresponding points obtained from proposed chain of homographies based approach are accurate enough, while the corresponding points obtained from direct method are erroneous.

B. *Rectification*

Mainly, the unequal zoom settings between the two PTZ cameras produce a distortion error in the rectified images. The distorted images produce the error in the final stereo based 3D localization. Fig. 8 shows the rectified image pairs obtained

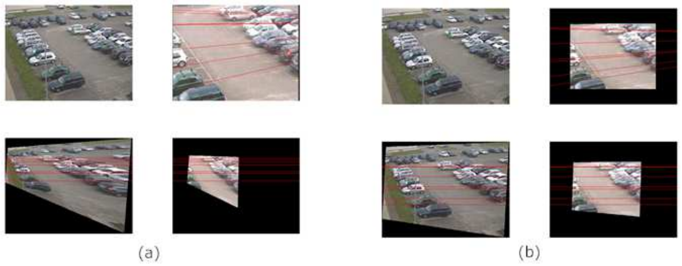

Fig. 8. Direct (a) and Proposed (b) rectification for a image pair having unequal zoom levels.

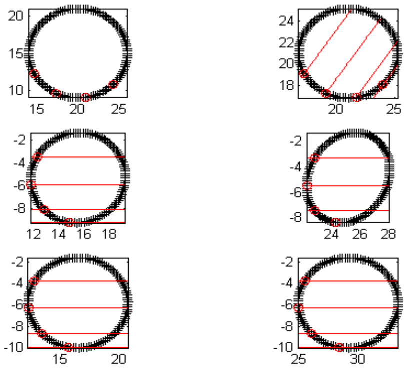

Fig. 9. Rectification of synthetic image pairs: Original pair (first row); direct rectification (middle row); proposed rectification (in bottom row).

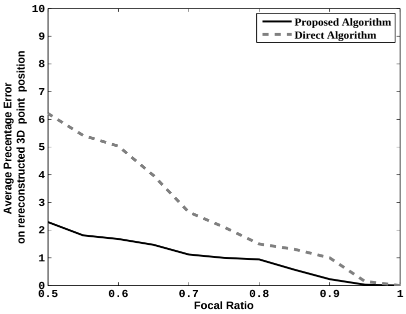

Fig. 10. Errors in the 3D position of synthetic points from direct and proposed rectified pairs of images.

with the direct rectification (without compensating the effect of unequal zoom) and proposed interpolation based method. It is clear from the given results that the rectified images obtained with direct method are distorted badly, while the rectified images are accurate enough for using in the computation of 3D position of the target. Moreover, the correctness of the proposed interpolation based rectification process has been evaluated on synthetic data in terms of 3D positions obtained with stereo vision based reconstruction. Fig.9 shows the results for the rectification of an image pair having points in a circle geometry. It is important to notice that the circle is distorted in the rectified images when direct rectification has been used, while the proposed algorithm is almost shape preserving. A performance comparison between the proposed rectification algorithm and a direct rectification algorithm is shown in Fig.10. The horizontal axis represents the focal ratio between the two images while the vertical axis represents the mean relative error in the 3D position of the reconstructed points using the rectified images. The error is calculated by taking the average over the sum of all three coordinate positions of all synthetic points. It is clearly visible that the reconstruction error is minor when images are rectified with the proposed algorithm. When the focal ratio decreases, the error drastically increases for the direct rectification while it is tolerable for the proposed rectification method.

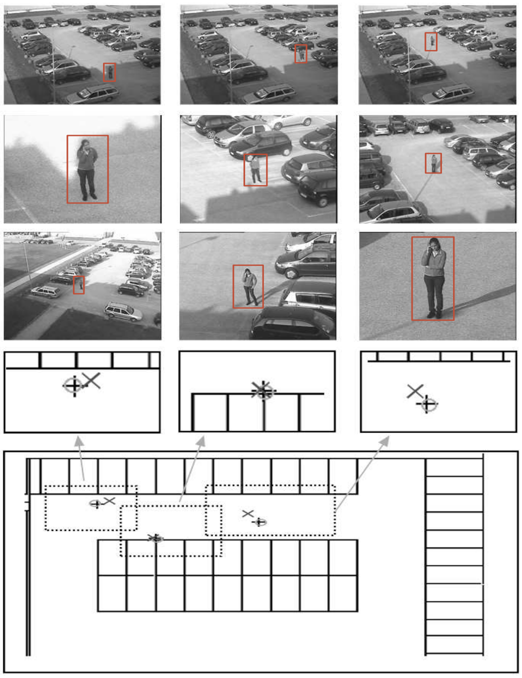

Fig. 11. Localization of a target on 2D map in case of unequal zoom between two PTZ camera with the proposed zoom compensation '+', without zoom compensation 'x' approaches. The 'O' represents the ground truth position of the target. First row: Master (static) camera image, second row: Left PTZ camera image, third row: right PTZ camera image and finally localization in bottom row.

C. *Target Localization*

Experimental studies with real video sequences have been carried out in order to test the performance of the proposed localization algorithm in a parking lot scenario. The experimental results have been obtained by considering different cases, i.e. with occluded and non-occluded targets, using different zoom levels for PTZ cameras, etc.

Two different criteria have been selected for this experiment. In the first set of experiments, a moving target has been detected in the different frames captured at different pan, tilt and zoom settings. These different pairs of frames have been rectified using the proposed LUT based rectification algorithm. Then, the moving target is localized using the proposed stereo vision based approach. Fig. 11 shows the achieved localization results in three different pairs of images with unequal zoom values. The first row contains images captured by a static camera (master), second and third rows contain the images captured by left and right PTZ cameras, respectively. The zoom levels are decreasing for the images captured by the

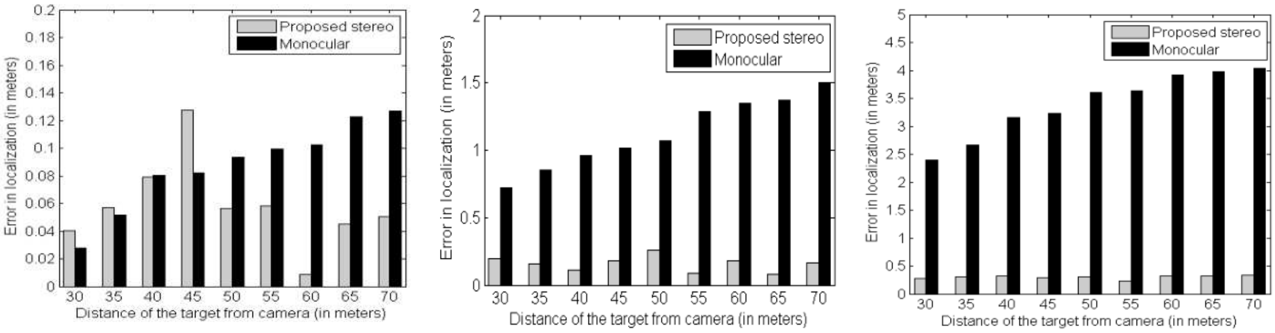

Fig. 13. Error in localization versus distance of target from camera at occlusion's height $0.0m$ (no occlusion), $0.5m$ and $1.0m$ (left to right).

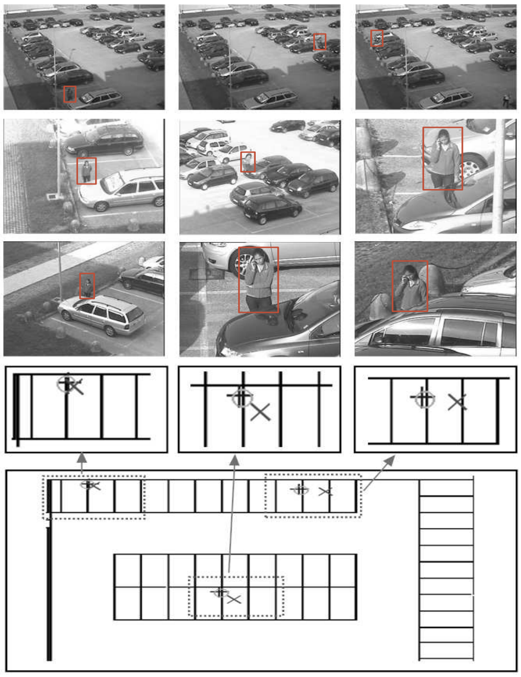

Fig. 12. Localization of a partially occluded target on 2D map with the proposed '+' and monocular 'x' approaches. The 'O' represents ground truth position of target. First row: Master (static) camera image, second row: Left PTZ camera image, third row: right PTZ camera image and finally localization in bottom row.

left PTZ camera from left to right and vice versa for the right PTZ camera in third row. The rectification is performed with and without compensating the effect of unequal zoom level for all three pairs of stereo frames and then localization is made from both kinds of rectified frames. It can be seen, from the presented results, that the localization accuracy is better when rectification has been performed after compensating the effect of unequal zoom.

The second set of experiments are presented in Fig. 12, to show the superiority of the proposed stereo framework over monocular camera based approach in the localization of partially occluded objects. Again, three different pairs of frames have been considered having different distances from cameras and occlusion's heights. The better performance of the proposed stereo vision based localization algorithm can be seen in Fig. 12. The proposed algorithm is capable to localize the target correctly even in presence of occlusions. In such situations, the proposed stereo vision based scheme produces better localization than a monocular camera based scheme. Fig. 13 shows three bar charts representing the localization error for different distances of the target from the camera. These bar charts (left to right) are plotted for three different values of the occlusion's height $0.0m$ (no occlusion), $0.5m$ and $1.0m$, respectively . It is observed that, if the distance of the object from the left camera or the height of occlusion is increasing, the error is also increasing in the case of monocular camera based scheme. Instead, the proposed stereo localization framework generates an error that is almost constant and not dependent to the occlusion height or object distance. In terms of quantitative analysis, a person, at $70m$ from the cameras and occluded from feet to $1.00m$ height, is localized by the proposed method with an error of $0.47m$. The monocular system, in the same conditions, has an error of $4.06m$. This analysis proves the better performance of the proposed localization framework over the existing monocular camera based techniques. A speed of 4 frame per second (fps) have been obtained for the localization of the target coupled with the interpolation based computation of rectification transformations.

D. Depth Map

High resolution depth maps are estimated for a far and large scene in a complex environment. Fig. 14 shows the depth map obtained for a building by adopting a coarse-to-fine strategy in two successive iterations. The top row represents the disparity map results obtained from a pair of images captured at low resolution of both cameras, i.e. in the first iteration of the process when both PTZ cameras are directed towards this region. Then the zoom level of left PTZ camera is selected with the proposed resolution strategy and a corresponding disparity map is obtained. From the obtained disparity map, it is found that the variation in depths are larger for the selected region of interests. In this context, the FOV of the right PTZ camera is refined to acquire high resolution image. The high resolution images for the selected region are given in the second row. Finally, the high resolution depth map (right most in the second row) is obtained with this pair of images. The higher zoom difference between these two pairs of images results in a better depth information for the selected region.

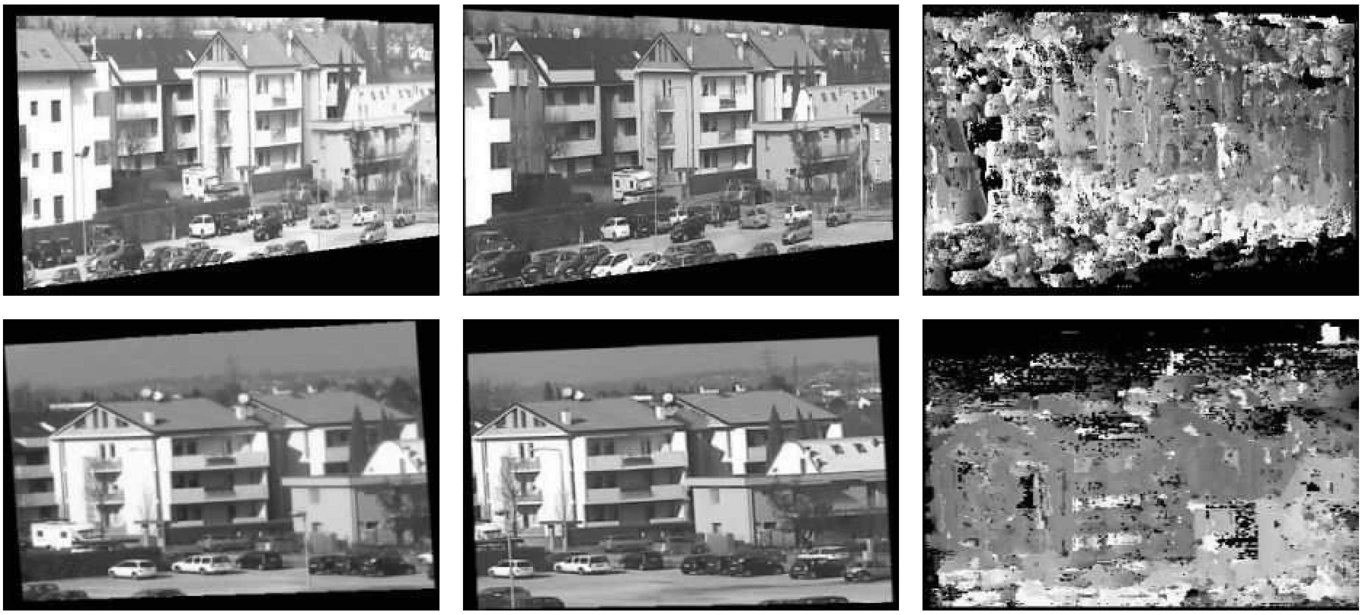

Fig. 14. Experiment for high resolution depth estimation in two successive iterations (top to bottom). In each row left camera image, right camera image and corresponding depth map image.

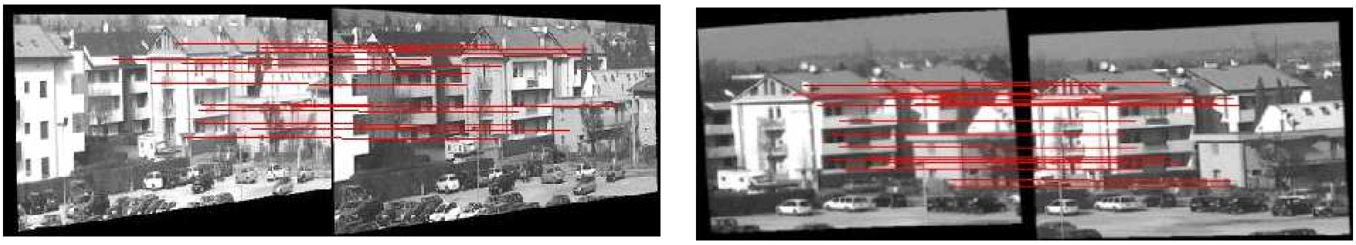

Fig. 15. Testing points with matching lines in two successive iterations (left to right).

To judge the accuracy of high resolution depth map over the low resolution depth map, three points at different depths have been selected within the region of interest. In the initial depth map (low resolution), the points appear to have the same depths, while these three points appear to be at different depths in the high resolution depth map. To give a quantitative evaluation of the results obtained with the iterative procedure, twenty points with ground truth depths information have been selected. The matching is performed for finding corresponding points in all three rectified image pairs (see Fig. 15) with good pixel precision. Five points have been randomly selected to compute the disparity drift for both pairs of images. Let the calculated and ground truth disparities for these five points be d_j and d'_j, respectively. The disparity drift has been calculated as

$$\rho = \frac{1}{5}\sum_{i=1}^{5} |(d'_j - d_j)| \qquad \text{for} \;\; j = 1,\ 2,\ \ldots,\ 5$$

For the other 15 points, we have estimated the depths $\tilde{d}_j = f\ (b/(d_j + \rho))$ and computed the mean and standard deviation of the absolute differences between ground truth and estimated depths in the two successive iterations. Moreover, we have compared the depth uncertainty $\delta = \tilde{d'}^2 (u/b)$ in support of our claim that the high resolution depth map is more accurate for the region having more depth variations. Here, $\tilde{d'}$ represents the average depth in a depth map image $\tilde{\mathbf{D}}$ and u is the horizontal resolution of the rectified images. Table I shows the comparison results based on the above mentioned criterion. It is important to notice that all the measures are improving with further iterations, i.e. the depth errors are very high in the first iteration while these reduce significantly in the final iteration. In the similar way, relative depth uncertainty and disparity drift iteratively improve. Finally, the depth mosaics from several low and high resolution depth maps have been generated. For this, 12 different pairs of images captured from both PTZ cameras at various zoom settings have been used. The zoom settings are automatically adapted by both cameras

TABLE I
COMPARISON OF DEPTH ESTIMATION IN TWO SUCCESSIVE ITERATIONS

Parameters	Itr-1	Itr-2
Mean Relative Error (%)	3.29	0.86
Standard Dev. Error (%)	3.50	1.04
Relative depth uncertainty δ (m)	0.797	0.307
Disparity drift ρ (m)	-.0088	-.0064

using the proposed scheme. All depth map images have been stitched in the coordinate frame of a priori selected image. Perspective transformation is used to align such depth maps for generating the mosaic. Fig. **??** shows the generated depth map mosaic, in which the gray value linearly reveals the magnitude of the depth value. The visual quality of the obtained depth map mosaic represents that the proposed method works well for a large and complex environment.

VII. Conclusions

A dual PTZ camera based stereo system has been presented for different video surveillance applications. First, a new real time rectification algorithm has been proposed. The real-time rectification transformations have been achieved by interpolating the rotation parameters for given orientations of the PTZ cameras. A process for compensating the unequal zoom effects between the images of a stereo pairs has been given to generate more accurate rectified images. The rectified frames have been used in two different applications: a) target localization and b) depth map mosaic. The evaluation of the proposed system performed on real sequence allows to derive the following considerations:

1) The proposed interpolation based rectification works very well for achieving real-time rectification.
2) For the localization of partially occluded targets, the proposed system outperforms monocular camera based systems.
3) The proposed framework is able to obtain high resolution depth maps for regions having larger variation in depths and the low resolution depth maps for flatter regions. Moreover, this process requires only limited a-priori information.

In the near future, the proposed framework will be used to develop a multispectral stereo active system (using visible and thermal PTZ cameras). This will allow to perform stereo tasks in environmental conditions (like foggy, rainy etc), where visible cameras do not perform well.

References

[1] B. Abidi, A. Koschan, S. Kang, M. Mitckes, and M. Abidi, *Automatic Target Acquisition and Tracking with Cooperative Static and PTZ Video Cameras*. Kluwer Academic, 2003, ch. Multisensors Surveillance Systems: The Fusion, pp. 43–59.

[2] S. Haritaoglu, D. Harwood, , and L. Davis, "Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 809–830, 2000.

[3] G. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance systems," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 25–37, Mar. 2005.

[4] N. Martinel, C. Micheloni, and G. Foresti, "Robust painting recognition and registration for mobile augmented reality," *Signal Processing Letters*, p. In Press, 2013.

[5] C. Micheloni, M. Lestuzzi, and G. Foresti, "Adaptive video communication for an intelligent distributed system: Tuning sensors parameters for surveillance purposes," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 1432–1769, Oct 2008.

[6] M. Brown, D. Burschka, and G. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25(8), pp. 993–1008, 2003.

[7] A. Jain, D. Kopell, K. Kakligian, and Y. Wang, "Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention," in *IEEE Int. Conf. of Computer Vision and Pattern Recognigation (CVPR*, 2006, pp. 537–544.

[8] D. Wan and J. Zhaou, "Stereo vision using two ptz cameras," *Computer Vision and Image Understanding*, vol. 112(2), pp. 184–194, 2008.

[9] J. Hart, B. Scassellati, and S. Zucker, "Epipolar geometry for humanoid robotic heads," in *International Cognitive Vision Workshop*, 2008, pp. 24–36.

[10] C. H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and ptz cameras for multiple object tracking," *IEEE Transaction on Circuit and Systems for Video Technology*, vol. 18(8),, pp. 1052–1063, 2008.

[11] S. Kumar, C. Micheloni, C. Piciarelli, and G. Foresti, "Stereo localization based on network's uncalibrated camera pairs," in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009., pp. 502–507.

[12] C. Micheloni, G. Foresti, and L. Snidaro, "A network of co-operative cameras for visual surveillance," in *IEE-proc. Vis. Image Signal Process*, vol. 152(2), 2005, pp. 205–212.

[13] C. Micheloni, B. Rinner, and G. Foresti, "Video analysis in pan-tilt-zoom camera networks," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 78–90, 2010.

[14] S. Kumar, C. Micheloni, and B. Raman, *Intelligent Multimedia Surveillance*. Springer, 2013, ch. Multiresolution Depth Map Estimation in PTZ Camera Network, pp. 149–169.

[15] S. Kumar, C. Micheloni, and G. Foresti, *Smart Cameras*. Springer US, 2009, ch. Stereo Vision in Cooperative Camera Networks, pp. 267–280.

[16] S. Kumar, C. Micheloni, and C. Piciarelli, "Stereo localization using dual ptz cameras," in *International Conference on Computer Analysis of Images and Patterns*, vol. 5702/2009, Munster, GE, Sep. 2-4 2009, pp. 1061–1069.

[17] C. Piciarelli, C. Micheloni, and G. Foresti, "Ptz camera network reconfiguration," in *ACM/IEEE International Conference on Distributed Smart Cameras*, Como,IT, Aug. 30 - Sep. 2 2009, pp. 1–7.

[18] S. Kumar, C. Micheloni, C. Piciarelli, and G. L. Foresti, "Stereo localization based on network uncalibrated camera pairs," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, Genova,IT, Sep. 2-4 2009.

[19] B. Morris and M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18(8), pp. 1114–1127, 2008.

[20] C. Micheloni, S. Canazza, and G. Foresti, "Audio-video biometric recognition for non-collaborative access granting," *Iournal of Visual Languages and Computing*, vol. 20, no. 6, pp. 353–367, Dec 2009.

[21] C. Piciarelli, C. Micheloni, and G. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, Nov. 2008.

[22] ——, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18(11), pp. 1544–1554, 2008.

[23] C. Soto, B. Song, , and A. Roy-Chowdhury, "Distributed multi-target tracking in a self-configuring camera network," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition,*, Miami, USA, 20-25 June 2009, p. 14861493.

[24] G. Foresti and C. Micheloni, "A robust feature tracker for active surveillance of outdoor scenes," *Electronic Letters on Computer Vision and Image Analysis*, vol. 1, no. 1, pp. 21–36, 2003.

[25] G. Foresti, C. Micheloni, and C. Piciarelli, "Detecting moving people in video streams," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2232–2243, 2005.

[26] C. Micheloni and G. Foresti, "Real time image processing for active monitoring of wide areas"," *Journal of Visual Communication and Image Representation*, vol. 17, no. 3, pp. 589–604, June 2006.

[27] C. Micheloni, G. Foresti, C. Piciarelli, and L. Cinque, "An autonomous vehicle for video surveillance of indoor environments," *IEEE Transactions on Vehicular Technologies*, vol. 56, no. 2, pp. 487–498, Mar 2007.

[28] S. Kumar, C. Micheloni, C. Piciarelli, and G. Foresti, "Stereo rectification of uncalibrated and heterogeneous images," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1445–1452, Aug. 2010.

[29] B. Dieber, C. Micheloni, and B. Rinner, "Resource-aware coverage and task assignment in visual sensor networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 10, pp. 1424–1437, Oct 2011.

[30] J. SanMiguel, C. Micheloni, K. Shoop, G. Foresti, and A. Cavallaro, "Self-reconfigurable smart camera networks," *IEEE Computer*, vol. In Press, 2014.

[31] G. Foresti and C. Micheloni, "Real-time video surveillance by an active camera," in *Ottavo Convegno Associazione Italiana Intelligenza Artificiale (AI*IA) - Workshop sulla Percezione e Visione nelle Macchine*, Sep 11-13 2002.

[32] C. Micheloni, G. Foresti, and F. Alberti, “A new feature clustering method for object detection with an activecamera,” in *IEEE International Conference on Image Processing*, Singapore, Oct. 24-27 2004, pp. 271–275.

[33] C. Piciarelli, C.Micheloni, and G. Foresti, “An autonomous surveillance veichle for people tracking,” in *13th International Conference on Image Analysis and Processing*, Cagliari, IT, Sep. 6-8 2005, pp. 1140–1147.

[34] D. Wan and J. Zhou, “Multiresolution and wide-scope depth estimation using a dual-ptz-camera system,” *IEEE Transaction on Image Processing*, vol. 18(3), pp. 677–682, 2009.

[35] D. Gallup, J. Frahm, P. Mordobhai, and M. Pollefeys, “Variable baseline/resolution stereo,” in *IEEE Int conf. on Computer Vision and Pattern Recogintion (CVPR)*, 2008, pp. 1–8.

[36] J. Meltzer and S. Soatto, “Edge descriptors for robust wide-baseline correspondence,” in *IEEE Int. Conf. on Computer Vision and Pattern Recogination*, 2008., pp. 1–8.

[37] F. Qureshi and D. Terzopoulos, “Planning ahead for ptz camera assignment and handoff,” in *Third ACM/IEEE International Conf. on Distributed Smart Cameras (ICDSC’09),*, Como, Italy, September 2009.

[38] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 2(60), pp. 91–160, 2004.

[39] A. Fusiello and L. Israra, “Quasi epipolar unclaibrated rectification,” in *IEEE Int. Conf. on Image processing (ICPR)*, 2008., pp. 1–4.

[40] S. Kumar, C. Micheloni, and C. Piciarelli, “Stereo localization using dual ptz cameras,” *Computer Analysis of Images and Patterns (LNCS-Springer*, pp. 1061–1069, 2009.

[41] M. Trajkovic, “Interactive calibration of a ptz camera for surveillance applications,” in *Asian Conference on Computer Vision (ACCV)*, 2002.

[42] Z. Zhu and A. R. Hanson, “Mosaic-based 3d scene representation and rendering,” *Signal processing: Image Communication*, vol. 21(9), pp. 739–754, 2006.