STEREO VISION USING HETEROGENEOUS SENSORS FOR COMPLEX SCENE MONITORING

Sanjeev Kumar and Claudio Piciarelli
Department of Mathematics and Computer Science, University of Udine
Via Della Sceinze 206, Udine- 33100, Italy

Keywords: Disparity, Epipolar Geometry, Focal Ratio, PTZ Camera, SIFT Matching, Stereo Vision, Zero Padding.

Abstract: The intelligent monitoring of complex scenes usually requires the adoption of different sensors depending on the type of application (i.e. radar, sonar, chemical, etc.). From the past few years, monitoring is mainly represented by visual-surveillance. In this field, the research has proposed great innovation improving the surveillance from the standard CCTV to modern systems now able to infer behaviors in limited contexts. Though, when environments allow the creation of complex scenes (i.e. crowds, clutter, etc.) robust solutions are still far to be available. In particular, one of the major problems is represented by the occlusions that often limit the performance of the algorithms. As matter of fact, the majority of the proposed visual surveillance solutions processes the data flow generated by a single camera. These methods fail to correctly localize an occluded object in the real environment. Stereo vision can be introduced to solve such a limit but the number of needed sensors would double. Thus, to obtain the benefits of the stereo vision discharging some of its drawbacks, a novel framework in stereo vision is proposed by adopting the sensors available in common visual-surveillance networks. In particular, we will focus on the analysis of a stereo vision system which is build from a pairs of heterogeneous sensors, i.e., static and PTZ cameras with a task to locate objects accurately.

1 INTRODUCTION

The problems of understanding complex scenes and detecting different moving objects (Foresti *et al.*, 2005) are hard to solve without an accurate localization of the objects moving in the scene. Such a process requires that any point in the image is associated to a point in the real environment. This is possible only by imposing some constraints (i.e. ground hypothesis) that are not maintained in case of occluded objects. To solve such a limit, stereo vision can be performed better than a single camera processing.

Stereo vision is used to recover 3-D shape information of a real world scene from two or more images taken from different viewpoints (Faugeras, 1993). In the existing literature, there are two research categories related to stereo vision. The first category relies on the use of a monocular camera with known scene information, and the second is the traditional stereo vision using dual cameras systems. The best example of latter one is human eyes system and a lot of researches have been done on this category (Forsyth and Ponce, 2003). The present work also belongs to the second category.

In general, researchers use static and ideal cameras (or homogeneous sensors) in the stereo vision for their low cost and relative simpleness in modeling. The problem of matching and rectification becomes simple using homogeneous cameras (Brown *et al.*, 2003). As PTZ cameras are able to obtain more degrees of freedom and cover large field of view, a combination of static and PTZ cameras is able to develop more significant results when compared to results obtained from a traditional stereo vision (Wan and Zhou, 2008). Apart from this fact, the precession of depth can be increased by improving the image resolution, since PTZ camera posses multi-resolution properties. However, there are many difficulties in the implementation of a vision system which contains a combination of heterogeneous sensors for stereo vision task, such as the variation in the internal and external parameters of PTZ camera in utility, variation in intensities in two images for the corresponding pixels and the most importantly the difference in resolution of two images due to the different zoom setting of both cameras. Therefore, the use of heterogeneous sensors is more challenging than the traditional homogenous approach, even though it can lead to better results.

Kumar S. and Piciarelli C. (2009).
STEREO VISION USING HETEROGENEOUS SENSORS FOR COMPLEX SCENE MONITORING.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 551-556
DOI: 10.5220/0001821105510556

In this paper, a novel method for the stereo vision is presented using pairs of heterogeneous images. To achieve such a result, the focal ratio between the focal lengths of the two images is computed for resizing the narrower image. The resized image has homogeneous focal information with respect to the wider image and to make it homogenous in terms of image resolution, zero padding is performed around the resized image. Once the images are made homogeneous by these two steps, then rectification process is run. Scale invariant features (Lowe, 2004) and (Micheloni and Foresti, 2003) are detected from both images to obtain pairs of matching points . Rectifying transformations are obtained by solving a nonlinear constrained minimization problem (Fusiello and Irsara, 2006), (Isgro and Trucco, 1999). The gray-level values are normalized in stereo images based on the intensities information of matching pairs. Disparity values have been computed to build range images from the given pairs of stereo images (Scharstein and Szeliski, 2002). In the disparity estimation, SSD criterion (Tao *et al.*, 2001) is used to find the best candidate for matching.

The rest of the paper is organized as follows: Section 2 is devoted to the detailed description of transforming process from heterogeneous to homogeneous pair of images. In section 3, SIFT matching is explained. Section 4 contains the stereo matching process. In section 5, experimental results using our methodology are given and finally in section 6, the concluding remarks are given.

2 TRANSFORMING PAIR INTO HOMOGENEOUS IMAGES

The images captured by a pair of heterogenous cameras have different imaging parameters. These make the acquired images heterogeneous due to camera positions, orientations, zoom and illumination. If we directly perform the further operations like SIFT, rectification and stereo matching on these images, the results would be affected by major performance degradation. To overcome this difficulty, the pair of images is made homogeneous before performing further operations. The process to make the heterogeneous pair of images as homogeneous is shown in Figure 1.

Let f_s and f_d be the focal lengths of the static and the PTZ cameras respectively when images are captured. The focal ratio is $R = \frac{f_s}{f_d}$ is computed and the image captured by the PTZ camera is shrunk by a factor of R. The shrunk image is then made homogeneous with respect to the static image by performing zero padding. Pairs of corresponding points (m_i, m_i') are then extracted by exploiting a SIFT matching algorithm. Such points are therefore used to compute the rectification transformations H and H' by minimizing

$$\sum_i (m_i'^T H'^T F_\infty H m_i)$$

where F_∞ is the fundamental matrix for rectified pair. To perform this minimization we choose the Levenberg-Marquardt algorithm because of its effectiveness and popularity. However, rectification process is performed to simplify a stereo matching procedure, and if the first row of H and H' is not chosen carefully in minimization, it may lead to a larger error and so failure in matching. Therefore, it is necessary to introduce some constraints in minimization process. Here, we have used the constraint that the distance between corresponding epipolar lines along vertical axis should be zero or very close to zero.

3 SIFT MATCHING

The process to obtain the matching points from the pair of stereo images is divided into two steps. First, we detect the scale invariant features in each image separately. In the next step, matching process of these features is performed between stereo pair of images.

The process of identifying locations in image scale space that are invariant with respect to image translation, scaling and rotation is based on the localization of a key. This task can be performed in following steps:

1. Perform the convolution operation on input image I with the Gaussian function with variance $\sigma = \sqrt{2}$. Let this operation gives an image I_1.
2. Repeat the step 1 on image I_1 to get a new image I_2.
3. Subtract image I_2 from image I_1 to obtain the difference of Gaussian function as $\sqrt{2}$.
4. Resample the image I_2 using bilinear interpolation with a pixel spacing of 1.5 in each direction. A 1.5 spacing means that each new sample will be a constant linear combination of 4-adjacent pixels. From this we generate a new pyramid level.
5. Determine the maxima and minima of this scale-space function by comparing each pixel in the pyramid to its neighbors.
6. Select key locations at maxima and minima of a difference of Gaussian function applied in scale space.

The scale invariant features can be detected from the locations of these keys. These features are detected

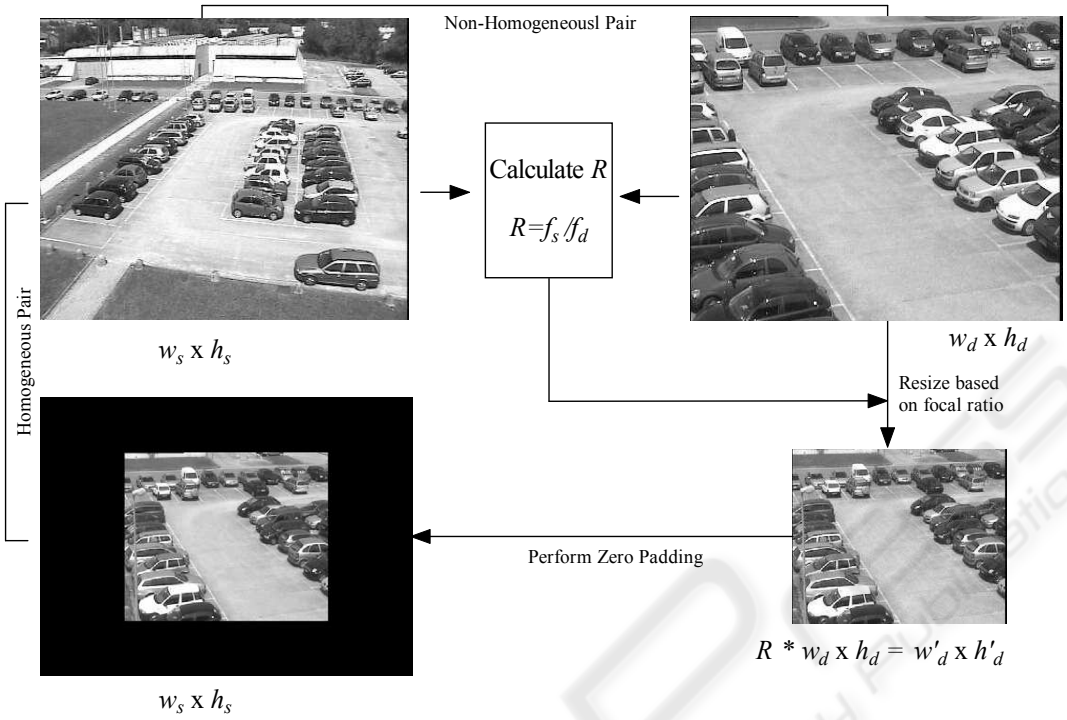

Figure 1: Overall process to obtain the pair of homogeneous images.

on the exact key locations as well as locations around it so that reliable matching between different views of an object or scene can be performed. These features are invariant to not only image orientation but also image scale, and provide robust matching across a substantial range of affine distortion, change in 3-D viewpoint, addition of noise, and change in illumination. For stereo image matching, SIFT features are extracted from left image and stored in a database. The right image features are matched by individually comparing each feature to this database and finding candidate matching features based on Euclidean distance of their feature vectors. We have performed features matching between stereo pair using the process given in (Lowe, 2004). The process of rectification is performed based on these pairs of matching points.

The RANSAC is used to remove the outliers from the pairs of matching points obtained from SIFT. More generally speaking, the basic assumption is that the data consist of inliers,i.e., data points which can be explained by some set of model parameters, and outliers which are data points that do not fit the model. In addition, the data points are subject to noise. An advantage of RANSAC is its ability to robustly estimate the model parameters. It finds reasonable estimates of the parameters even if a high percentage of outliers are present in the data set.

4 STEREO MATCHING

Once the pair of stereo images is rectified, the next step is to compute disparity between the matching pair. There are two approaches to obtain stereo matching, i.e., feature based and pixel (region) based methods. Due to the difference in intensities of stereo images captured from heterogeneous sensors, these methods can not be applied directly to obtain the stereo matching. To avoid this problem, here we perform this process in two steps, i.e., a combination of feature based and pixel based methods.

The first step is related to normalize the intensities in two images for the matching pairs. In order to perform this task we detect the matching pixels using SIFT matching from the pair of images. Then image can be normalized by a simple algorithm, which computes the parameters α, β of the gray level global transformation

$$S_r(x,y) = \alpha\, S_l(x,y) + \beta$$

by fitting a straight line between the intensities of all matching pixels which are obtained using SIFT. Once the values of α and β are computed then the left image can be normalized in the range of right image.

For each pixel in the left image (reference image I_l), similarity scores are computed by comparing a fixed, small window of size 5×5 centered on

the pixel to a window in the right image (I_r), shifting along the corresponding horizontal scan line. Windows are compared through the normalized SSD measure, which quantifies the difference between the intensity patterns:

$$C = \frac{\sum_{(\xi,\eta)}[I_l(x+\xi,y+\eta) - I_r(x+d+\xi,y+\eta)]}{\sqrt{\sum_{(\xi,\eta)} I_l(x+\xi,y+\eta)^2 \sum_{(\xi,\eta)} I_r(x+\xi,y+\eta)^2}}$$

where $\xi \in [-n,n]$ and $\eta \in [-m,m]$. The disparity estimate for pixel (x,y) is the one that minimizes the SSD error:

$$d_0(x,y) = \arg\ \min\ C(x,y,d)$$

However we can observe that squared differences need to be computed only once for each disparity, and the sum over the window need not be recomputed from scratch when the window moves by one pixel.

5 RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed framework, several experiments have been conducted on the images captured by a fixed and a PTZ cameras. For each experiment, the left image of the stereo pair does not change as it is captured by the static camera, while the right image is different as it is captured by the PTZ camera with four different zoom levels. The focal length information has been acquired online for these images and so the focal ratio. The SIFT matching is performed to obtain the pairs of corresponding points from the left and right images. The rectification error has been computed in the rectified pair of images using the criterion of the mean of the error index $r_i = |(Hm_i)_y - (H'm'_i)_y|$, i.e., the vertical distance between corresponding epipolar line.

The given results aim to show the improved performance of the proposed solution (using homogeneous images) over a method in which stereo process is applied directly on heterogeneous images. The first set of experiments consists in computing the error in rectification when the pair of images has been obtained using identical focal lengths. Since the image size of the two images is the same that implies the two images are homogenous. In this context, Figure 2 presents the results for a pair of images having focal ratio 1. The mean pixel error for rectified pair of images is 0.0696 when 10 pairs of matching points are used.

Since the main goal of the proposed algorithm is to perform stereo process on heterogeneous images, we run a set of experiments by progressively reducing the focal ratio. In Figures 3, 4 and 5 , the results

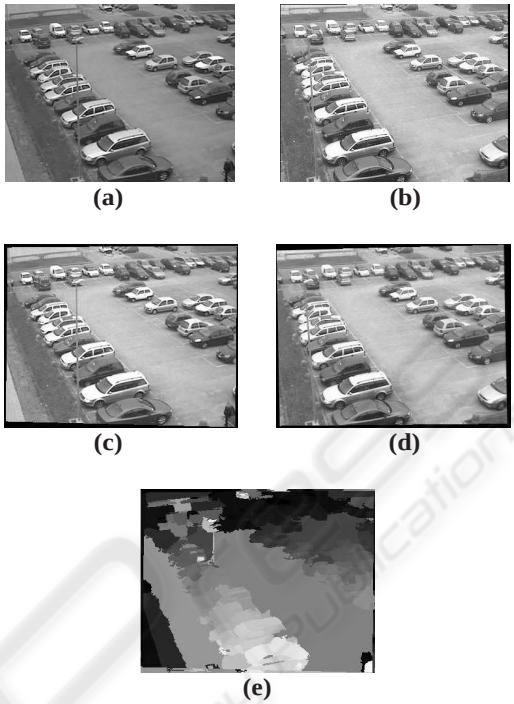

Figure 2: Results for a pair of stereo images having focal ratio 1.0. Left stereo image (a); right stereo image (b); rectified stereo images (c & d); range image (e).

are shown for a pair of stereo images having focal ratio 0.97, 0.94 and 0.90, respectively. As it can be seen from these figures and from the results in table 1, the proposed solution much better than standard stereo matching on heterogeneous images.

The error in the rectified pairs is shown in table 1 for both the cases, i.e., when the rectification process has been performed on heterogeneous and homogeneous pairs of stereo images. The error has been estimated for different pairs which have different values of focal ratio. It is clear from the table that the error is high when the rectification process has been performed directly using heterogeneous pairs of images while the error is very small when the pairs of images are made homogeneous before performing the rectification process. Apart from this comparison, one more thing is noticed about the difference between the quality of range images obtained from heterogeneous and homogeneous pairs of images. In the range images obtained from pairs of homogeneous images, the density is regularly decreasing as the distance of the object is increasing along the optical axis, i.e., objects near to camera have brighter intensity compared to the far ones. This phenomenon is not so regular in the range images which are obtained from the pairs of heterogeneous images.

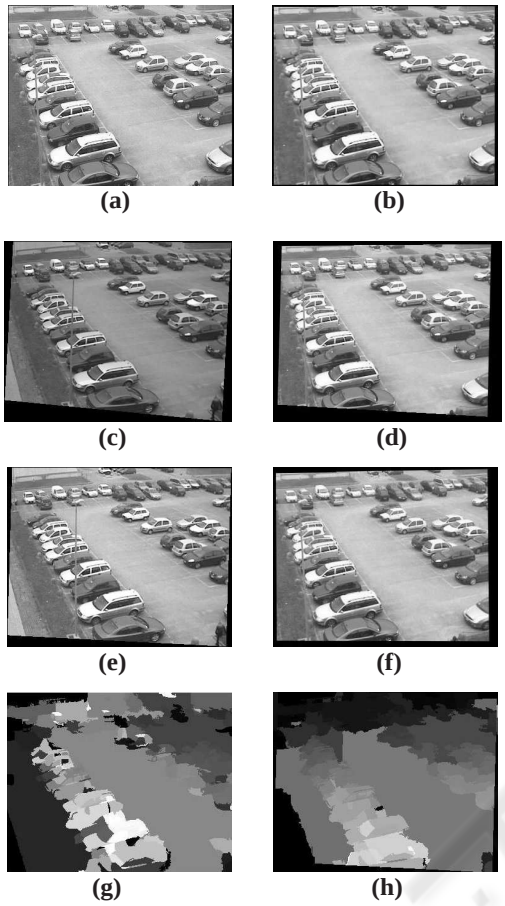

Figure 3: Results for a pair of stereo images having focal ratio 0.97. Heterogeneous right image (a); homogeneous right image (b); heterogeneous pair of rectified images (c & d); homogeneous pair of rectified images (e & f); range image using heterogeneous rect. pair of images (g); range image using homogeneous rect. pair of images (h). Left image of input stereo pair is same as in Figure 1(a).

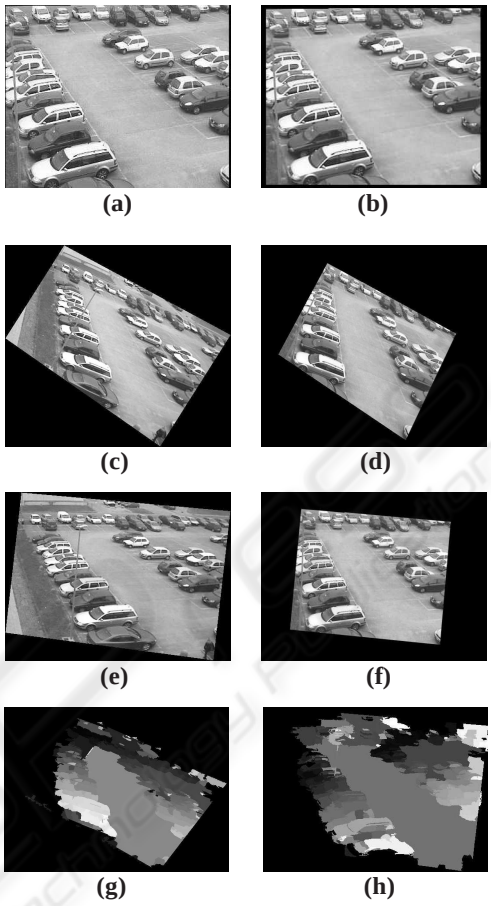

Figure 4: Results for a pair of stereo images having focal ratio 0.94. Heterogeneous right image (a); homogeneous right image (b); heterogeneous pair of rectified images (c & d); homogeneous pair of rectified images (e & f); range image using heterogeneous rect. pair of images (g); range image using homogeneous rect. pair of images (h). Left image of input stereo pair is same as in Figure 1(a).

Table 1: Mean Pixel-Error between corresponding epipolar lines between the rectified pairs of images.

Focal Ratio	Mean Error	
	Homo. Case	Hetero. Case
1.00	0.0696	0.0696
0.97	0.0714	0.0698
0.94	0.2891	0.0743
0.90	0.3367	0.0984

6 CONCLUSIONS

We have presented a framework for stereo vision using heterogeneous sensors to monitor a complex scene. The pair of images has been made homogeneous based on a focal ratio information and then by performing zero padding on the shrunk image. The pairs of corresponding points have been obtained using SIFT matching in stereo pair of images. The rectification transformations have been obtained by solving a nonlinear optimization problem. Experimental results show that the combination of static and PTZ cameras gives good results only if the captured images are made homogenous. This approach thus leads to better results if compared to a traditional stereo vision system in terms of depth accuracy when monitoring a complex scene.

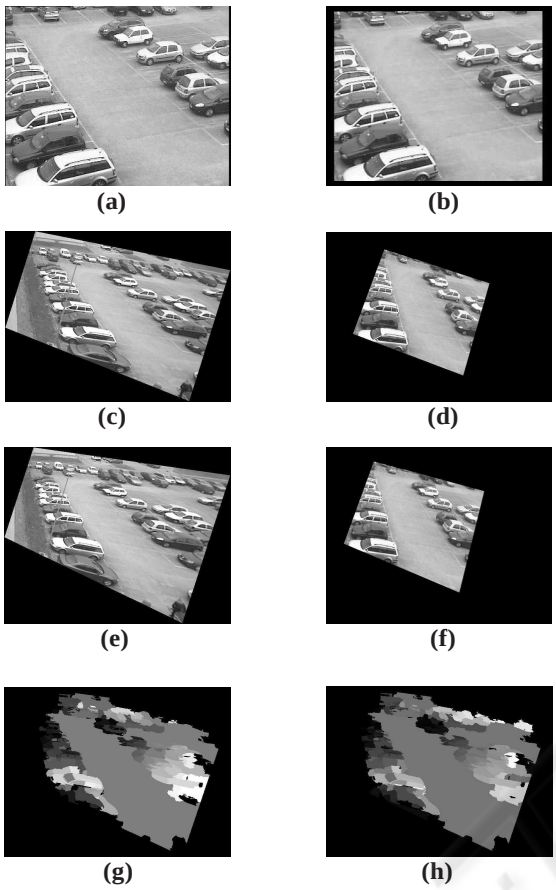

Figure 5: Results for a pair of stereo images having focal ratio 0.90. Heterogeneous right image (a); homogeneous right image (b); heterogeneous pair of rectified images (c & d); homogeneous pair of rectified images (e & f); range image using heterogeneous rect. pair of images (g); range image using homogeneous rect. pair of images (h). Left image of input stereo pair is same as in Figure 1(a).

ACKNOWLEDGEMENTS

This work was partially supported by the Italian Ministry of University and Scientific Research within the framework of the project entitled **Ambient Intelligence: event analysis, sensor reconfiguration and multimodal interfaces** (2006-2008). Sanjeev Kumar is also thankful to Department of Mathematics and Computer Science, University of Udine for the financial support during this work under the grant MIUR n. 179 dd. 29/01/2007 provided by Italian Ministry of University and Scientific Research.

REFERENCES

Foresti,G.L., Micheloni, C. and Piciarelli, C. (2005). Detecting Moving People in Video Streams, *Pattern Recognition Letters*, 26(15), 2232–2243.

Faugeras, O. (1993). Three-Dimensional Computer Vision, MIT Press, Cambridge, MA, USA.

Forsyth, D.A. and Ponce, J. (2003). Computer Vision: A Modern Approach, Prentice Hall.

Brown, M.J., Burschka, D. and Hager, G.D. (2003). Advances in computational stereo, *IEEE trans. of Pattern Analysis and Machine Intelligence*, 25(8), 993–1008.

Fusiello, A., Irsara, L. (2006). Quasi-euclidean uncalibrated epipolar rectification, Research Report RR 43/2006, Department of Computer Scince, Univ. of Verona.

Isgro, F., Trucco, E. (1999). On robust rectification for uncalibrated images, *in proc. of IEEE International Conference on Image Analysis and Processing*.

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

Micheloni, C. and Foresti, G.L. (2003). Fast Good Features Selection for Wide Area Monitoring, *in proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance*, Miami (FL), USA.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *Int. Journal of Computer Vision*, 47(1-3), 7–42.

Tao, H., Sawhney, H.S. and Kumar, R. (2001), A global matching framework for stereo computation, *in proc. of IEEE Int. Conferece on Computer Vision*, 532–539.

Wan, D. and Zhou, J. (2008). Stereo vision using PTZ cameras, *Computer vision and Image Understanding*, article in press.