EXTENDED VISUAL SEQUENCE METRIC RECONSTRUCTION FROM UNCALIBRATED CAMERAS

Vicente José
Bustos Pablo, Bachiller Pilar
Departamento de Informática
Universidad de Extremadura
Avda. de la Universidad S/N
jvcrespo@unex.es
Geometría y Geometría Computacional

1.Abstract

This paper is focused on real-time algorithms for estimating dense 3D structure and camera motion from images sequence. Within this approach a practical method is proposed which can retrieve metric reconstruction from continuous image sequences obtained with uncalibrated cameras.

Our approach can be divided in two main stages. First, the projection matrices between non-differential camera displacement are estimated, it's known this process relies on correct estimations of matched points. In order to efficiently deal with correspondence searches, a continuous differential tracking of interest points is accomplished by a real-time parameter estimation procedure. Making use of those projection matrices, a self-calibration and camera ego-motion estimation are carried out. Secondly, a dense point reconstruction is performed combining optical flow constraint with the projective model of a moving camera that has been determined in previous stage. In both stages no correspondence searches are found in an explicit way, alternatively corresponding point searches are guided by a continuous tracking of interest points and optical flow constraint. Additionally, we implement a multi-resolution approach to avoid failures in employing differential models such as optical flow constraint, that is the case of large image displacements.

2.Introduction

One of the main goals in robot vision is recovering 3D spatial information from 2D image-projected points, this task grows in complexity when the unique information available is the data present in the image sequence, such as pixel grey values or color components.

It become clear that with uncalibrated camera and no additional constraints, only a projective 3D reconstruction stratum can be obtained from images correspondences, hence the projection matrices and the reconstructed points are related to real world by an indeterminate projective transformation. We can imagine that this structure is unsuitable in real applications, nevertheless it constitutes a proper starting structure, which needs a further enhancement.

To upgrade this basic projective 3D structure to the useful Affin and Metric ones, additional constraints over camera parameters must be imposed. Affin upgrade are related to locate and place the so call "plane at infinity" at his canonical position. On the other hand, metric upgrade is related to the estimation of camera internal parameters, such as focal length, central points and skew. Both structure enhancements comprise the entire self-calibration stage.

As the starting structure is 3D projective, and this estimation relies on correct calculation of matched points. An efficient algorithm to search matching points is necessary in real-time implementations, that decreases the heavy computational load. Early works have fixed those problems using a calibrated framework, in those methods a camera internal parameters and camera pose are computed ahead with the aid of a known metric pattern. Hence the major detriments of those methods is employing an off-line estimation, and be linked to fixed internal camera parameters. The Maybank and Faugeras[1] research opens a possibility of calibrating a camera by using a sequence of images, and without previous knowledge of scene points at which the camera is situated. In those kind of techniques, the calibration process can be carried out just with the information available from the images. Subsequent works have extended this process allowing varying internal camera parameters[2].

Present work differs from later approaches in the method used to obtain matching points. As we are interested in real-time algorithms to scene reconstruction, this search must be less expensive that former algorithms. Improvement in time requirement can be achieved by taking advantage from the differential image sequence properties. The key idea is to perform low cost calculations between differential images, light enough that can be carried out while camera is moving. In this way, to deal with correspondence search, a continuous differential tracking of select points is accomplished.

Finally, in order to retrieve dense reconstructed points, a dense correspondence search must be done somehow from the image sequence. This task can be accomplished by searches of homologous over epipolar lines and minimization criterions, such as normalized cross correlation. In this work, this dense reconstruction process is performed combining optical flow constraint with previously estimated projective model and image rectification.

3.Projective Model

The projection of a 3D scene point onto an 2D image point can be solved using a projective model, that is determined by a projective transformation matrix *P*.

The equation $m = PM$ relates the involved parameters, where $m=[x\ y\ 1]^T$ is an image point and $M = [X\ Y\ Z\ 1]^T$ is a scene point, both expressed in homogeneous coordinates, and *P* is the projection matrix of dimension *3x4*, that maps world point onto sensor points at a fixed position of a camera.

In a calibrated framework, the camera projection matrix can be factorized as follows:

$$K\,[\,R^T\,|\,-R^T\ T\,]$$

Where *T* represents a translation vector and *R* a rotation matrix, both performing a rigid euclidean transformation between the camera pose and world reference coordinate system. *K* is un upper triangular matrix, called calibration matrix, that encodes the intrinsic parameters of camera construction, it can be interpreted as an affin transformation between image plane and sensor array plane.

$$K = \begin{pmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

The parameters f_x and f_x represent the focal length in x-and y-sensor axe, [u_0 , v_0] is the principal point coordinates determined by the projection of the optical axe, and s is the skew or angle between *x* and *y* sensor axes. All of them depend on the camera construction and camera operation.

In a calibrated framework, an image sequence is acquired with different known calibrated-cameras (K_0, ..., K_j, ..., K_s), at unknown positions ($R_0\ T_0$, ... , $R_j\ T_j$, ..., $R_s\ T_s$). Therefore, an euclidean scene point M_0 is projected at different sensor positions S_o : (m_0, ..., m_j, ..., m_s), this set of points is called homologous points in the image sequence, and are determined by the following equations.

$$m_0 \;=\; P\,M_0 \;= K_0\ [\,R_0{}^T \,|\, - R_0{}^T\ T_0\,]\,M_0$$

$$m_j \;=\; P\,M_0 \;= K_j\ [\,R_j{}^T \,|\, - R_j{}^T\ T_j\,]\,M_0$$

$$m_s \;=\; P\,M_0 \;= K_s\ [\,R_s{}^T \,|\, - R_s{}^T\ T_s\,]\,M_0$$

A group of sets of homologous points (S_o, ..., S_p) can be determined in advance by means of normalized cross correlations over image sequence or another search mechanism. Consequently it constitute the input data for the search engine in the reconstruction problem. We can conclude that in this case, the reconstructions problem consists of searching the camera positions ($R_0\ T_0$, ... , $R_j\ T_j$, ..., $R_s\ T_s$) and the 3D euclidean points (M_0, ..., M_p) such that fit the sets of homologous points in the image sequence (S_o, ..., S_p).

The uncalibrated case is more general and complex, it assume that *K* matrices are unknown. Under this circumstance, it's convenient to start with a projective structure, that involves less restrictive projection matrices (P_0, ..., P_j, ..., P_s) and less restrictive reconstructed points *M*. In this case the reconstruction problem consists of searching a set of projection matrices (P_0, ..., P_j, ..., P_s) and 3D projective points (M_0, ..., M_p) such that fit the sets of homologous points in the image sequence (S_o, ..., S_p) .

It's known this problem has no unique solution due to the use of less restrictive model. The possible solutions to this problem are related one another by an indeterminate projective transformation *T*. The particular transformation T^* that upgrade projective reconstruction to euclidean one, constitute the search target of any calibration mechanism. It's convenient to explain that this target transformation is the unique that can factorize the result projection matrix in a proper euclidean form, as follows: $PT^*=P'=K[R|-RT]$

$$m_0 \;= P_0\,M \;=(P_0\,T^*)(T^{*-1}M) = P_0'\,M' = K_0\ [\,R_0{}^T\,|-R_0{}^T\ T_0\,]M'$$

$$m_j \;= P_j\,M \;=(P_j\,T^*)(T^{*-1}M) = P_j'\,M' = K_j\ [\,R_j{}^T\,|-R_j{}^T\ T_j\,]M'$$

$$m_s \;= P_s\,M \;=(P_s T^*)(T^{*-1}M) = P_s'\,M' = K_s\ [\,R_s{}^T\,|-R_s{}^T\ T_s\,]M'$$

4.3D Reconstruction

4.1.Interest Point Selection

The algorithm begins with a selection of points that are easily identifiable and with stable existence properties along the sequence, those points compose the support for the correspondence searching. With this aim, those points have to be stable under common transformations, such as camera translation, rotation and zoom, this property ensures a high degree of repeatability in the images. Another requirement is those points content high information around its image position that enables an easy identification.

In order to search such interest points, we employ a Harris[3] corner detector that has such desirable properties. It takes into account image gradient in x-direction I_x and y-direction I_y ,that allows determining image points that have significant changes in both directions (corner) around its position. The neighborhoods are weighted by a gaussian filter *G* to take into account less significant importance of distant regions, the interest Harris points are detected by the following expressions:

$$Eigenvalues\begin{pmatrix} G*I_x^2 & G*I_xI_y \\ G*I_xI_y & G*I_y^2 \end{pmatrix} = \begin{pmatrix} \lambda & \beta \end{pmatrix}$$

$$If\ min\ (\lambda,\ \beta) > Threshold \Rightarrow Interest\ point$$

4.2.Interest Point Tracking

The success of the initial projective reconstruction and hence the overall algorithm, relies on a correct estimation of matching points in the sequence. Therefore, the robustness of the algorithm depends on the capability of detecting and rejecting mismatches. We implement three hierarchical levels of confidence for matching points validation, the first is a motion estimation and tracking of interest points based on a Kalman filter. The second consists of a normalized cross correlation test over points that have been previously linked by the tracking procedure. Both levels have different sampling scope, the point tracking is accomplished continuously as frame is acquired. On the other hand, normalized cross correlation test is carried out when significant displacements of matched points are accumulated along frames.

We utilize the following Normalized Cross Correlation, where $\bar{I}$ is the grey level mean into a squared region of dimension *bxb* pixels.

$$NCC(P_x,P_y) = \frac{\sum_{i=-b}^{i=b}\sum_{j=-b}^{j=b}\left[I_L(i-p_{LX},j-p_{LY})-\bar{I}_L\right]\left[I_R(i-p_{RX},j-p_{RY})-\bar{I}_R\right]}{\sqrt{\sum_{i=-b}^{i=b}\sum_{j=-b}^{j=b}\left[I_L(i-p_{LX},j-p_{LY})-\bar{I}_L\right]^2\sum_{i=-b}^{i=b}\sum_{j=-b}^{j=b}\left[I_R(i-p_{RX},j-p_{RY})-\bar{I}_R\right]^2}}$$

The result of this process is a list of sets, where each set is composed of confident matched points that conform the projection of an unknown 3D point in the images. The third confidence test is accomplished using the RANSAC (RANdom SAmpling Consensus) paradigm explained in the next section.

4.3.Initial Projective Reconstruction a Six Point Solution

The goal of this section is to obtain an initial solution of camera projection matrices, and projective structure of 3D points that have a set of matched points. That allows employing a posterior reprojective mechanism linked with a RANSAC[4] method to validate both estimations.

Therefore, the input data of the algorithm is same sets of interest point trajectories (corner here) into the image sequence, trajectories that have passed the former confidence tests. Hence, camera projection matrices and structure estimation involve a parameter search which fits the data input correspondences. Once both are obtained, them conform the basic for calibration process and posterior structure upgrade.

Our approach use base sequences of three views, to this end it's necessary to split the overall sequence in sub-sequence of image triples, that correspond with the output of three consecutive normalized cross correlation level. To compute the camera projection matrix of each sub-sequence, we use a minimal method[5], those minimal approaches have the advantage of needing the least data that is necessary to estimate structure and camera pose, analogy to estimate the images transfer geometry such as trifocal tensor. Such is the case of computing epipolar geometry with the minimal set of seven points correspondences in two views.

We compute the projection matrices and 3D structure with the minimal set of six corner tracking over three views. In this calculation if a mismatch comes about into the selected six points, leads to an improper projections matrices estimation. Hence those minimal solutions are used as search engine in robust estimation methods, such as reprojection and RANSAC as summarize.

RANSAC:

1- Choose six sets of point correspondences and make up a 3D projective basis liked to five of them.

2- Compute projection matrices and 3D coordinates by means of this six points.

3- With former solution compute 3D coordinates of all remaining homologous sets.

3- Compute reprojection error of all 3D points

if 70 per cent are good enough exit

else go to step 1

4.4.Camera Calibration

Once camera projection matrices are correctly estimated in a projective framework, we can use them to camera calibration and hence to metric upgrade computation. It's convenient to remark that this initial projective structure is related to a metric one by two important transformations, which represent the overallcalibration process.

The first one involves the localization of the particular plane that holds the metric points at infinitum $\Pi^{p}=[\pi_x, \pi_y, \pi_z, 1]$ (where π_x, π_y, π_z unknown). Once this plane is localized the transformation T that recover its original metric position at $\Pi^{m} = [1, 1, 1, 0]$ also restore the affin properties of the original structure, that is $\Pi^{m} = T^{-1}\Pi^{p} \Rightarrow M_{(Affin)} = T\,M_{(Projective)}$. The second consists of an affin transformation $Tm = K^{-1}$ characterized by the camera internal parameters K and conform the final metric upgrade.

To estimate both transformations, we employ an approach based on the absolute dual quadric Ω^{*}, this concept was introduced by Triggs[6] and subsequently by Pollefeys[2] with varying camera intrinsic parameter. This quadric jointly encodes the position of the euclidean points at infinitum infinitum $\Pi_{\infty}^{p}=[\pi_x, \pi_y, \pi_z, 1]$ together with camera constructions parameters K_0 of first view in a compact form.

This fact can be explained by two qualities, this 3D quadric is always confined in a plane, and this plane is just the euclidean plane at infinitum. On the other hand its shape is characterized by the camera internal parameters

$K_0\,K_0^{-1}$. Those two properties and its associated unknown parameters are put on view in the absolute dual quadratic expression.

$$\Omega^* = \begin{pmatrix} K\ K^T & K\ K^T \Pi_\infty^{PT} \\ \Pi_\infty^{P} K\ K^T & \Pi_\infty^{P} K\ K^T \Pi_\infty^{PT} \end{pmatrix}$$

The link that allows camera self-calibration with this entity is another basic property, when is projected over any image by means of the corresponding known camera projection matrix $P_{j,}$ this quadric become a 2D conic $K_j\,K_j^{-1}$ that codifies the camera calibration matrix of this particular camera.

$$W_j = K_j K_j^T = P_j \Omega^* P_j^T$$

This property allows camera calibration, provided that same restrictions over camera internal parameters $K_j K_j^{-1}$ (left) are known, such as fixed focal length, zero skew, known principal point, each of those assumptions imposes a restriction in the quadric form, and hence in the unknown parameters, π_x, π_y, π_z and K_0. In this work zero skew, known principal point and $f_x = f_y$ have been assumed.

4.5.Image Rectification

Once the calibration matrices ($K_{0,}$ ..., K_j, ..., K_s) and rotation matrices of each camera ($R_{0,}$ ..., $R_{j,}$ ..., R_s) are known, it's to convenient transform the original image sequence Q in other sequence Q' where each new image is equivalent to another one taken at the same camera translation but with no rotation component.

This process is accomplished by warping the original image, thus each original point m is transformed to its new position by the warping transformation $m' = K\,R^{-1}K^{-1}m$. This transformation cancel the rotation component, as can be deduced by the following equations:

$$\begin{aligned} m_0' &= K_0\,R_0^{-1}K_0^{-1}m_0 = K_0\,R_0^{-1}K_0^{-1}K_0\ [\,R_0\ \mid -R_0\ T_0\,]M = K_0[I\ \mid -T_0]M \\ m_j' &= K_j\,R_j^{-1}K_j^{-1}m_j = K_j\,R_j^{-1}K_j^{-1}K_j\ [\,R_j\ \mid -R_j\ T_j\,]M = K_j[I\ \mid -T_j]M \qquad (1) \\ m_s' &= K_s\,R_s^{-1}K_s^{-1}m_s = K_s\,R_s^{-1}K_s^{-1}K_s\ [\,R_s\ \mid -R_s\ T_s\,]M = K_s[I\ \mid -T_s]M \end{aligned}$$

Generally this new position won't be integer, hence an image grey level interpolation will be necessary. As result of this process each image has its epipolar lines in horizontal position, it allow an efficient employment of region based optical flow mechanism in dense map reconstruction.

4.6.Dense Deep Map

In order to compute a dense deep map, we employ the optical flow restriction along with previously estimated image rectification. As previously commented those rectifications are equivalent to another sequence with no rotation and equivalent translation, that allows a more accurate calculation in region based optical flow.

Let $I=I(x,y,t)$ denote the time-varying image intensity function and let (u,v) denote the x-y-components of the instantaneous optical flow. The computation of the optical field using the classical image motion constraint equation $I_xu+I_yv+I_t=0$ is difficult to owing to the aperture problem an unstable. Instead we employ a region support adopting the approach of Lucas y Kanade[7] .

This approach obtains the necessary additional constraints from a finite region around the point, and combines its spatial and temporal derivatives by gaussian filter weight.

This region based optical flow constraint relate the position of an image point $m=(x, y, 1)$ to its homologous $m=(x+u, y+v, 1)$ in the next image. This equation can be easily manipulated to take into account points instead of displacement vectors, this yields the following linear system of two equations:

$$\begin{pmatrix} G*I_x^2 & G*I_xI_y & -xG*I_x^2-yG*I_xI_y-G*I_xI_t \\ G*I_xI_y & G*I_x^2 & -xG*I_xI_y-yG*I_y^2-G*I_yI_t \end{pmatrix} \begin{pmatrix} x+u \\ y+v \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad (2)$$

This linear system represents two lines in the projective plane s_x^j and s_y^j , which have the important property of containing just the homologous point m_j in the next image j, hence $s_x^j\, m_j=0$ and $s_y^j\, m_j=0$. In addition both lines cross one other in the homologous point m_j . The angle between them is related to the eigenvalues obtained with the Harris corner detector, and represent the amount of x-y-gradient information around the pixel position, this measure determines the quality of the implicit matching process.

$$\begin{aligned} s_x^j &= (\ G*I_x^2 \quad G*I_xI_y \quad -xG*I_x^2-yG*I_xI_y-G*I_xI_t) \\ s_y^j &= (G*I_xI_y \quad G*I_x^2 \quad -xG*I_xI_y-yG*I_y^2-G*I_yI_t) \end{aligned} \qquad (3)$$

In each consecutive image triples, the algorithm computes a dense deep map combining the two previous models: a) An Euclidean model obtained with a camera undergoing translational movements and their corresponding rectified images and b) the projective lines obtained by the optical flow equations. The first model depends on the following parameters: scene points coordinates (unknown), camera calibration (known) and translational component (known), on the other hand the second depends on image intensity gradients (known). Combining those models by means of equations (1)(2)(3) lead to the following four equations for each point in the image:

$$s_x^1\, m_0 = s_x^1\, K_1\, K_1^{-1} m_0 -(1\,/\,z)\, s_x^1\, K_1T_1=0$$

$$s_y^1\, m_0 = s_y^1\, K_1\, K_1^{-1} m_0 -(1\,/\,z)\, s_y^1\, K_1T_1=0$$

$$s_x^2\, m_0 = s_x^2\, K_1\, K_1^{-1} m_0 -(1\,/\,z)\, s_x^2\, K_1T_1=0$$

$$s_y^2\, m_0 = s_y^2\, K_1\, K_1^{-1} m_0 -(1\,/\,z)\, s_y^{12}\, K_1T_1=0$$

In order to calculate the inverse deep map structure *(1/z),* we minimize the contribution of each of the four equations in the unknown parameter *(1/z).* This approach lead to minimize the following expression in each image point:

$$Min\ [\ s_x^1\, m_0 + s_y^1\, m_0 + s_x^2\, m_0 + s_y^2\, m_0\ ]$$

5.Results

In the experiment a multi-scale approach has been employed to allow higher range of image displacements. We have built an hierarchical Laplacian pyramid of three levels, by this mechanism the estimations in higher levels have been reutilized as initial data to perform the estimation in the lower ones. With this approach we were able to manage displacements close to 40 pixels. The input sequence is showed in figure 3, it's composed of image triples in grey scale values. In the following figures the result inverse deep map is showed at different resolutions, figure 4 represents the lower resolution, figure 5 medium resolution and finally figure 6 illustrates the refined inverse deep map. In all of them the grey scale values represent the corresponding scaled values of the inverse depth.

We have used an AMD-K7 processor at 800MHz, the code has been constructed employing MMX and 3Dnow instructions building a set of computer vision libraries, that have allowed to accelerate the heavy operations. With this environment the result real Magaflops has been 600.

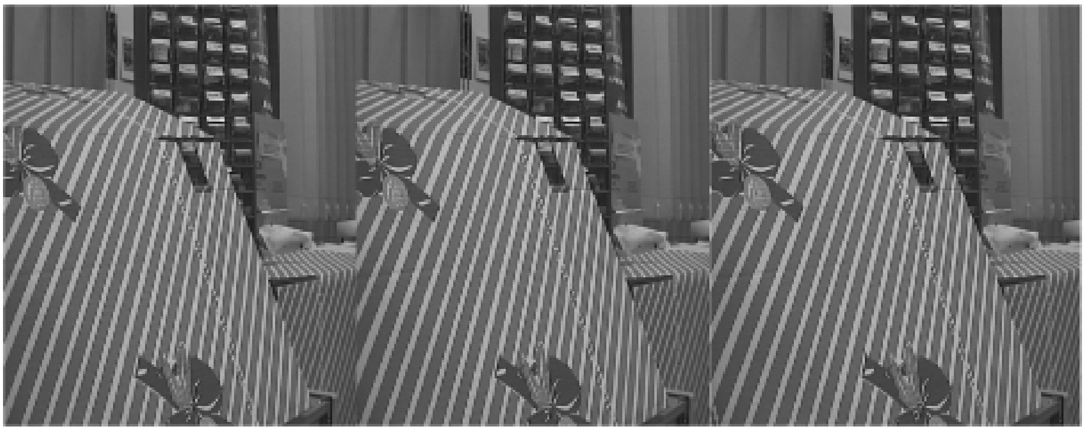

Figure 1 Real input sequence

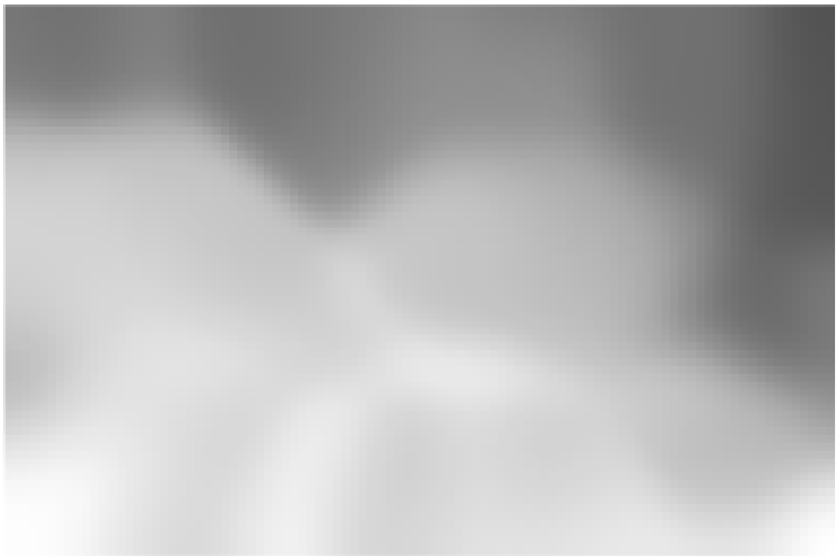

Figure 2 Deep map at lower resolution

Figure 3 Deep map at medium resolution

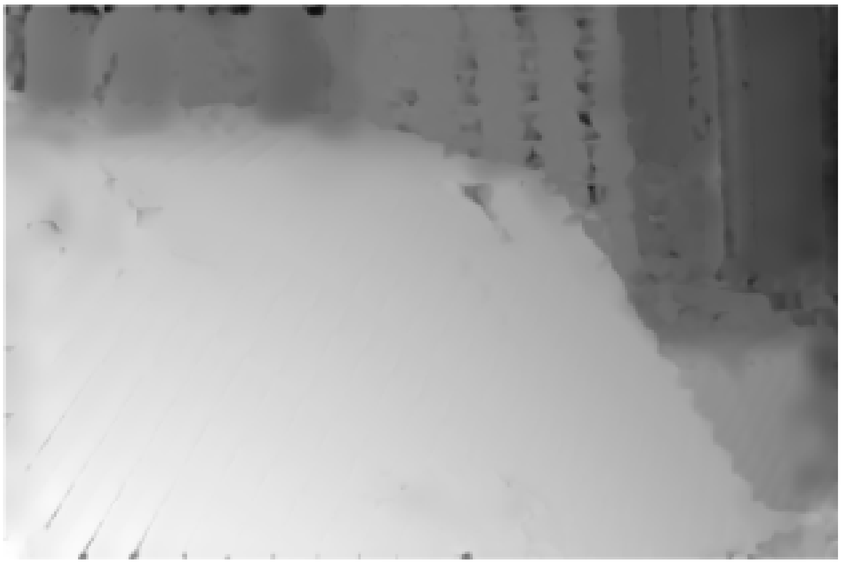

Figure 4 Deep map at higher resolution

6.References

[1] S. Maybank and O. D. Faugeras, "A theory of self-calibration of moving camera", International Journal of computer Vision, pp.123 -151, 1992.

[2] M. Pollefeys, R.Koch, and L.V. Gool, "Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters", IJCV, pp.7 - 27, 1999.

[3] C.Harris and M.Stephens, ``A combined corner and edge detector," In Proc. Alvey Conf., pp.189-192, 1987.

[4] P. Torr and A.Zisserman, "Robust parameterization and computation of the trifocal tensor", In Proc. British Machine Vision Conference, R.Fisher and E.Trucco, eds., pp. 655-664, BMVA, Sept 1996. Edinburgh.

[5] F.Schaffalitzky, A.Zisserman, R. Hartley, and P.Torr, "A six point solution for structure and motion ", In Proc. European Conference on Computer Vision, pp.632-648, Springer-Verlag, June 2000.

[6] B.Triggs, "The absolute quadric" , In Proc. Conference on Computer Vision and Pattern Recognition, pp.609-614, IEEE Computer Soc. Press, 1997.

[7] B.Lucas and T.Kanade, "An iterative image registration technique with an application to stereo vision", In Proc. of the 7th International Joint Conference on Artificial Intelligence, pp.674-679, 1981.