A MINIMIZATION APPROACH FOR 3D RECOVERY IN REGION-BASED STEREO VISION

A Lopez and F Pla

Universitat Jaume I, Spain

INTRODUCTION

Stereoscopic vision is the set of techniques that try to recover three-dimensional information from two or more views of a scene. The usual technique consists of computing correspondences between points in both images in order to be able to calculate their depth, thus obtaining their three-dimensional location. Usually, the correspondences are represented by a disparity map from which we can obtain the depth map of the reference image.

An important problem to take into account while computing correspondences is the presence of depth discontinuities in the scene. Several techniques try to solve this problem during the matching process, such as correlation-based techniques using adaptive windows (7), dynamic programming approaches (5, 6, 9) and Bayesian approaches (1).

These techniques are based on the fact that depth discontinuities are usually located at intensity discontinuities, and the use of some smoothness constraint in the other areas. Here, we propose using regions as the matching primitive, given that they represent homogeneous areas limited by intensity discontinuities. We impose some constraint in the shape of depth inside the regions, while allowing depth discontinuities at the region boundaries. The used constraint is an important issue that will be discussed in the next section.

The use of regions has been already proposed in other works (3, 8, 10) usually as an initial stage of the correspondence problem. These techniques have to be able to handle segmentation errors, often by merging some adjacent regions in order to improve correspondence. In our approach, we propose considering the regions resulting from segmentation of a reference image and finding their correspondences without segmenting the other image.

The method proposed here directly computes depth of the regions without the intermediate calculation of disparities. This method consists of starting with an initial estimation of the depth of each region and computing depth increments/decrements towards the solution in a multiscale iterative scheme. Depth increments/decrements are computed by minimizing an energy function that represents the error in similarity between each region and its corresponding region at a given depth. However, pixels whose corresponding pixel is occluded in the other image include inaccuracy in the similarity error calculations. Detecting occluded areas and excluding them in the calculations increases the method accuracy.

In the next section, the basis of the method and a detailed explanation of the algorithm for region matching and occlusion detection are presented. In the following section, the experimental results of the method are shown. Finally, we discuss the conclusions extracted from the experiments carried out so far.

STEREO VISION BY MINIMIZATION

Let I_1 and I_2 be a stereo pair of images, and let I_1 be the reference image. Let us assume that the depth of all the pixels of image I_1 is known and is called Z. Then, a replica of I_1 could be obtained by using the calibration parameters, the depth map Z, and the intensity values of I_2. The replica, T, consists of assigning to each pixel the intensity of its corresponding pixel in I_2.

$$T(m) = I_2(m'), \quad \forall m \in I_1 \tag{1}$$

$$m' = f_{12}(m, Z(m)), \tag{2}$$

where m' is the pixel that corresponds to m at a given depth, $Z(m)$.

If Z is the ground truth depth map, T would be equal to I_1, except for the occluded areas of I_2 where no correspondence exists. Therefore, T is the most similar replica of I_1 that one can obtain from I_2 and Z. When Z is not far from the ground truth depth map, some calculations can be done in order to increment or decrement each pixel depth towards the solution. An iterative scheme can be designed, based on minimizing the differences between I_1 and T. However, this is an ill-posed problem that requires to apply some constraints to the solution in order to achieve convergence.

The basic idea consists of starting with an estimation of the depth map, Z_0, which is incremented or decremented iteratively by minimizing an energy function based on intensity differences. This idea is inspired

on restauration techniques and some related work is in Robert and Deriche (11) where a regularization term is needed in order to achieve convergence. The regularization term has to be able to smooth the depth function in homogeneous intensity areas, while allowing depth discontinuities where intensity discontinuities are found.

Here, we propose to impose some constraint in the depth of pixels within a region. For example, assuming a scene made of planar surfaces is applicable to robotics environments where the scenes are mainly composed of man-made objects. We propose calculating depth by minimizing an energy function based on correlation between a region in the reference image and its corresponding region in the other image. Depth is computed for each region, so that each pixel depth is derived by following the imposed constraint.

First, we will develop the method under the most severe assumption: constant depth in the regions. This constraint implies an implicit assumption about the surfaces which consists of a scene made of fronto-parallel planes. After that, we will discuss the influence of occlusions in the method, as well as an algorithm to detect and manage them.

The energy function

The minimization technique faces the stereo problem as the minimization of an energy function. We intend to calculate depth with respect to the regions obtained from the segmentation of the reference image so that the energy is expressed as a function of the scene depth.

$$E(Z) = \int_{R \in I_1} F(R, Z(R)) dR \tag{3}$$

Given a region R in the reference image, its corresponding region, R' in the second image can be calculated from the calibration parameters and the region depth,

$$R' = g_{12}(R, Z(R)), \tag{4}$$

where g_{12} means that each pixel $m_i \in R$ corresponds to each pixel $m'_i \in R'$ at a given depth $Z(m_i)$ by means of equation 2 and $Z(m_i)$ is a component of the region depth $Z(R)$ following the assumed constraint.

Once the current corresponding region has been obtained, we can compute the error in similarity between both regions by using some correlation measurement. For example, the Zero-Mean Normalized Cross-Correlation method (ZNCC) is a measurement function to be maximized, so that we can minimize the following similarity function,

$$\begin{aligned} F(R, Z) &= -ZNCC(R, R') = \\ &= -\frac{1}{N} \sum_{\forall m_i \in R} \mathcal{I}_1(m_i) \mathcal{I}_2(m'_i) \end{aligned} \tag{5}$$

where N is the size of R and $\mathcal{I}_1(m_i)$, $\mathcal{I}_2(m'_i)$ are the zero-mean normalized intensities of pixels m_i and m'_i with respect their region, respectively, which can be calculated from the mean intensity and standard deviation of each region,

$$\mathcal{I}_k(m_i) = \frac{I_k(m_i) - \overline{I_k(R)}}{\sigma(R)}, \quad m_i \in R \subset I_k. \tag{6}$$

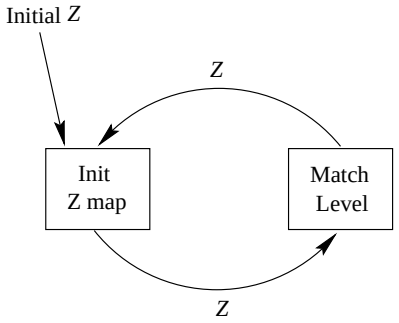


Figure 1: Multiscale scheme.

The proposed energy function depends on two independent variables $m = (u, v)$ and a functional, $Z(m)$. According to the Euler equation (4), the depth functional $z = Z(m)$ that minimizes the energy function is a solution of the following equation,

$$F_z - \frac{\partial}{\partial u}\{F_p\} - \frac{\partial}{\partial v}\{F_q\} = 0, \quad p = \frac{\partial u}{\partial z}, q = \frac{\partial v}{\partial z}. \tag{7}$$

where F_k is the partial derivative of F with respect to k. When constant depth in the regions is assumed $p = 0$ and $q = 0$, so that the equation becomes,

$$\sum_{i=1}^{N} \mathcal{I}_1(m_i) \left(\mathcal{I}_2(m'_i) \mathcal{H}(R') - \frac{\frac{\partial}{\partial z} I_2(m'_i)}{\sigma_2(R')} \right) = 0, \tag{8}$$

where $\mathcal{H}(R')$ is the *weighted mean gradient* of R',

$$\mathcal{H}(R') = \frac{1}{N} \sum_{k=1}^{N} \mathcal{I}_2(m'_k) \frac{\frac{\partial}{\partial z} I_2(m'_k)}{\sigma_2(R')}. \tag{9}$$

Function F_z has a zero-crossing at each local minimum of function F. When F_z is positive Z must be decremented and viceversa in order to achieve the zero-crossing. Then, we can use this equation for incrementing or decrementing iteratively the current depth towards the solution. The solution will be achieved when the current depth reaches $F_z = 0$. It is important to note that the initial depth should not be far from the solution in order to avoid other local minima different from the global solution. This is the reason why a multiscale scheme (figure 1) is needed, in order to avoid local minima different from the solution. In the iterative process we have to take into account the following issues:

- When F_z is positive Z must be decremented and viceversa.
- The increment/decrement amount in depth must produce disparity increments/decrements lower than 1 pixel in order to avoid exceeding the nearest zero-crossing and obtaining another one.
- If F_z is positive at iteration t and it is negative at iteration $t+1$, then the zero-crossing has been exceeded, and viceversa. Then, a progressively smaller increment/decrement should be used in order to reach the zero-crossing.

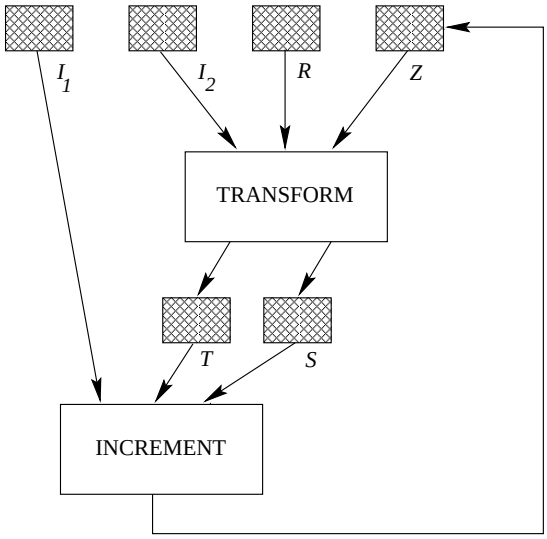


Figure 2: Process of matching at each level.

Occlusions

Let I_1, I_2 be the stereo pair of images and $\mathcal{R}$ be the segmentation of the reference image, I_1. At each level of the pyramids of these images a depth function is obtained in an iterative algorithm that minimizes the energy function. Each iteration can be divided in two steps, as shown in figure 2:

- *Transform* the second image into a replica T of the reference image at a given depth, Z.
- *Increment* (or decrement) Z by comparing images I_1 and T.

If the presence of occlusions were not considered, depths of different regions can be computed independently, so that the intersection of region correspondences would not be empty in the occlusion areas. Pixels of I_1 whose corresponding pixel is occluded in I_2, introduce inaccuracy in the calculation of the similarity error between regions. When considering occlusions, one of the possible correspondences should be selected and the other areas that intersect with it should be marked as occluded. In this case, all the region depths have to be computed at the same time, and depth has to be incremented or decremented slowly in each iteration in order to achieve convergence in a cooperative algorithm where matches and occlusions are found at the same time. The whole process stops when no region depth is either incremented or decremented, that is, when Z converges. The process is detailed in the matching algorithm 1.

Algorithm 1
MATCH level k of the I_1, I_2 pyramids.

Let $\mathcal{R}$ be the list of regions from I_1 segmentation.
for each region $R_i \in \mathcal{R}$ **do**
  $Z^0(R_i) \leftarrow$ extract depth from level $k-1$
end for
TRANSFORM I_2 into T and S, according to Z
for each region $R_i \in R$ **do**
  Compute $F_z^1(R_i)$ from R_i active pixels, I_1, T
  Determine Δd direction and magnitude
  $Z^1(R_i) \leftarrow \mathrm{IncZ}(Z^0, \Delta d, Z_{min}, Z_{max})$
end for
while Z^t not converged **do**
  TRANSFORM I_2 into T and S, accord. to Z^t
  $Z^{t+1} \leftarrow \mathrm{INCREMENT}^t\ Z^t$ according to I_1, T, S
end while

The iterative algorithm should take into account the following issues:

- When two regions have correspondences whose intersection is not empty, the intersected area corresponds to the nearest region, while the furthest region matches an occluded area.
- The comparison between regions should not take into account the pixels in R that match occluded points in R' (*occluded* pixels).
- The comparison between regions should not take into account the pixels in R that match points out of I_2 limits (*outbounded* pixels).
- The occluded and outbounded pixels may vary from one iteration to another.

In order to deal with all these cases, we define a *pixel status* such that at each iteration all the pixels in the reference image are marked as *active, occluded* or *outbounded*. Only the active pixels are considered in the calculation of region statistics (zero-mean normalized intensities, weighted mean gradient, etc.) and therefore, in the calculation of depth increments/decrements.

Let S be the status map for all the pixels in the reference image. The status can be computed while *transforming* image I_2 into T according to the current depth. The algorithm of the *transform* operation is detailed in algorithm 2.

Algorithm 2
TRANSFORM I_2 into T, S according to Z.

```
Sort R in increasing order of Z
for each region R_i ∈ R do
  for each pixel m ∈ R_i do
    m' ← f_12 (m, Z(m))
    S(m) ← status of m'
    if S(m) is ACTIVE then
      T(m) ← I_2(m')
    end if
  end for
end for
```

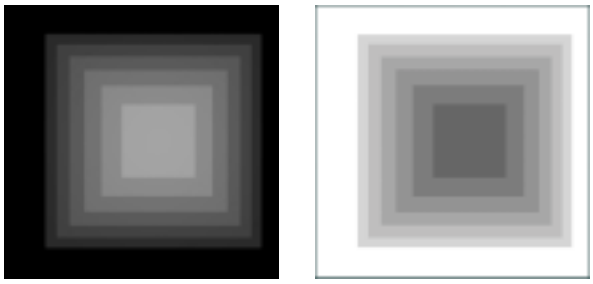
On the other hand, the *compare* operation is detailed in algorithm 3. This operation consists of recomputing F_z at each iteration as explained in the previous subsection and calculating the new depth increments by means of a function called IncZ which also applies depth limits, Z_{min} and Z_{max}.

Algorithm 3
INCREMENT Z according to I_1, T, S (iter. t)

```
for each region R_i ∈ R do
  Recompute F_z^t(R_i) from R_i active pixels
  if F_z^t(R_i) = 0 then
    R_i converges
  else if sign(F_z^t(R_i)) ≠ sign (F_z^{t-1}(R_i)) then
    Δd_i^t ← Δd_i^{t-1}/D
    Z_i^t ← IncZ(Z_i^{t-2}, Δd_i^t, Z_min, Z_max)
    if Z_i^t = Z_i^{t-2} then
      R_i converges
    end if
  else
    Δd_i^t ← Δd_i^{t-1}
    Z_i^t ← IncZ(Z_i^{t-1}, Δd_i^t, Z_min, Z_max)
    if Z_i^t = Z_i^{t-1} then
      R_i converges
    end if
  end if
end for
```

The method to recognize the pixels whose corresponding pixels at a given depth map may consist of maintaining a table of visited pixels. The corresponding pixel (in I_2) of each region pixel is marked as visited while performing the *transform* operation. As the regions are ordered in increasing Z, the first pixel in I_1 that marks each pixel in I_2 establishes the correspondence and achieves an *active* status. Afterwards, the status of the pixels of other regions that try to match the pixel, are marked as *occluded* pixels in S.

As the corresponding pixels in I_2 are obtained in non-integer coordinates, the accuracy of the table of visited pixels influences the accuracy of the method. In the next section we will discuss this influence by observing the experimental results.



(a) Left image (b) Ground truth depth map

Figure 3: An example of synthetic images.

EXPERIMENTAL RESULTS

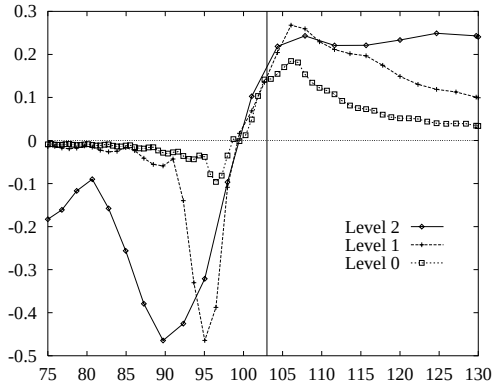
For the experiments performed so far, we assumed constant depth in each region, that implies an implicit assumption about the surfaces which consists of a scene made of fronto-parallel planes. We performed experiments with synthetic scenes (to evaluate the method exhaustively) and real scenes (to evaluate the method qualitatively).

The scene in figure 3 consists of a synthetic pyramid with depths ranging from 82 to 117 cm. The ground truth map is shown in figure 3(b), where clearer areas correspond to further points.When not considering occlusions in the method, the mean relative error of the obtained depth map is 3.78%, without considering the background of the image.

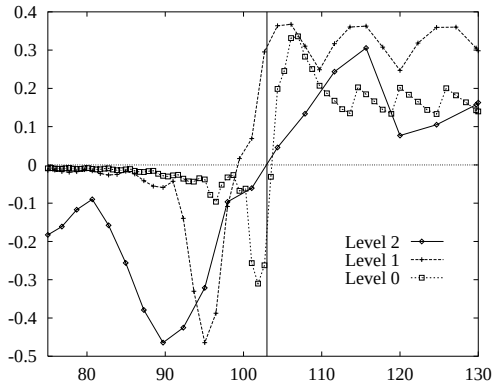
Let $R_1, R_2, ...R_6$ be the regions representing the planes of the pyramid example from the nearest to the furthest one. The evolution of F_z of one of the regions (i.e. R_4) at different levels of the multiscale structure of the images is shown in figure 4(a). Whatever the depth is initialized, level 2 obtains a zero-crossing quite close to the ground truth depth, represented by the vertical line.At level 0, the zero-crossing of R_1 is very accurate, while the zero-crossing of the other regions is displaced with respect to the ground truth depth. Pixels whose corresponding pixel is occluded introduce inaccuracy in the F_z calculations, and only the depth of the regions without occlusions is obtained accurately.

When considering occlusions in the matching process, the evolution of F_z of region R_1 is almost identical, while the evolution of F_z of the other regions is improved towards the ground truth solution. For calculation of the occlusions in this test, depths of the other regions are set to their ground truth depth. Figure 4(b) shows the results for region R_4. A nearest neighbour rule has been used when marking the visited pixels. This is the reason why a repetitive pattern appears in F_z. Using this rule, a mean relative error of 2.97% is obtained. The results can

be still improved by using a more accurate rule for occlusion detection. For example, when considering 3×3 subpixel sections, the mean relative error obtained is 2.30%. Figure 5 shows some results with real scenes. Clearer areas in the depth maps correspond to nearer points. Note that even in most of the sloped homogeneous regions in the "parking meter" example (images from the JISCT stereo test set (2)) the method provides reasonable results.



(a) without considering occlusions



(b) taking into account occluded areas

Figure 4: Evolution of $F_z, \forall Z$, of region R_4.

CONCLUSIONS

Experiments with synthetic scenes show that the method obtains reasonable results. Ground truth depth of a synthetic stereo pair allows calculating a mean relative error of 2.30% in depths ranging from 82 to 117 cm.

The results obtained so far encourage us to generalize the method in scenes with any type of planar surfaces. Further work is directed to perform this generalization in order to be able to do more experiments that include real scenes, to validate the proposed method.

Another issue may consist of improving the occlusion detection procedure, for example, by including percentage of occlusion of each pixel and considering this percentage in the statistical calculations.

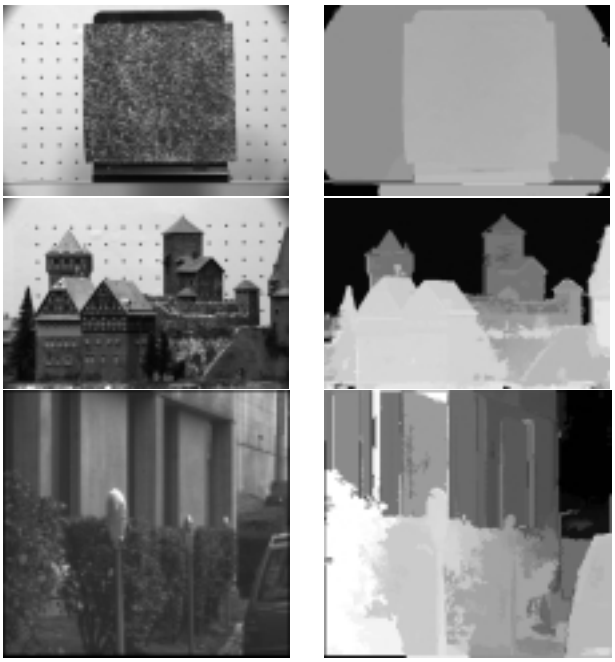


Figure 5: Real scenes: left images and depth maps.

REFERENCES

1. Belhumeur P. N., 1993, "A Bayesian Approach to the Stereo Correspondence Problem". PhD thesis, Electrical Engineering, Yale University.
2. R.C. Bolles and H.H. Baker and M.J. Hannah, 1993, "The JISCT stereo evaluation". Proc. IUW, 263–274.
3. Cohen L., Vinet L., Sander P., and Gagalowicz A., 1989, "Hierarchical region based stereo matching". Proc. CVPR, 416–421.
4. Elsgoltz L., 1977, "Ecuaciones diferenciales y cálculo variacional". Editorial MIR, Moscu.
5. Geiger D., Ladendorf B., and Yuile A., 1995, "Occlusions and binocular stereo". IJCV, 221–226.
6. Intille S. S. and Bobick A. F., 1994, "Disparity-space images and large occlusion stereo". Proc. ECCV, 674–677.
7. Kanade T. and Okutomi M., 1994, "A stereo matching algorithm with an adaptive window: Theory and experiment". IEEE Trans. on PAMI, 16(9), 920–932.
8. Marapane S. B. and Trivedi M. M., 1989, "Region-based stereo analysis for robotics applications". IEEE Trans. on Systems, Man, and Cybernetics, 19(6), 1447–1464.
9. Ohta Y. and Kanade T., 1985, "Stereo by intra- and inter-scanline search". IEEE Trans. on PAMI, 7, 139–154.
10. Randriamasy S. and Gagalowicz A., 1991, "Region based stereo matching oriented image processing". Proc. CVPR, 736–737. IEEE Comp. Soc. Press.
11. Robert L. and Deriche R., 1996, "Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities". Proc. 4th ECCV.

AKNOWLEDGEMENTS

This work was partially supported by grants GV97-TI-05-27 (Generalitat Valenciana) and TIC98-0677-C02-01 (CICYT, Ministerio de Educación y Ciencia)