AN ENERGY-BASED FRAMEWORK USING GLOBAL SPATIAL CONSTRAINTS FOR THE STEREO CORRESPONDENCE PROBLEM

P-M. Jodoin [‡] *M. Mignotte*[‡]

[‡] Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal,
P.O. Box 6128, Stn. Centre-Ville, Montréal, Québec, H3C 3J7.

E-MAIL : JODOINP@IRO.UMONTREAL.CA

ABSTRACT

This paper investigates the use of a region-based approach for the stereo matching problem. We have stated this problem in a commonly adopted global energy-based framework. Our energy-based model mixes a local and robust regularization term with global spatial constraints. These constraints are related to a (precomputed) partition into homogeneous regions with identical disparity. In practice, our approach assigns a single disparity to regions instead of individual pixels. These regions, used to globally constrain the ill-posed nature of our minimization problem, are estimated by combining an unsupervised Markovian segmentation and a roughly estimated disparity map. This disparity map is computed with a basic Winner-Take-All (WTA) procedure. The proposed global energy function seems to be well suited to find good disparity discontinuities at object boundaries, especially when the number of disparities is large. An Iterated Conditional Modes (ICM) algorithm is used to optimize this global energy function. We provide experimental results on real stereo image pairs. A quality measure, based on ground truth data, is used to evaluate the performance of our algorithm. Results indicate that our approach is fast and performs well compared to other existing methods.

1. INTRODUCTION

Stereo matching is one of the most active research areas and challenging topics in computer vision. A stereo vision setup commonly used is one involving two images separated by a distance D on a linear path, perpendicular to their optical axis [1, 2]. The goal of a stereo correspondence algorithm is to estimate a *disparity map* d with respect to a *reference* input image or a *cyclopean* view located between the two input images [2]. In this context, because the epipolar lines are horizontal, the disparity is seen as a difference in location between features seen in the two images [3]. In this work, d_p (the disparity at pixel p) makes a correspondence between pixel p in a *reference* image (I_{ref}) and a pixel q in a *matching* image (I_{mat}), when $q = p + d_p$.

In most previous work, the disparity map computation is stated in a heuristic energy-minimization framework. Among the simple energy-based methods, we can cite the *local* ones. Fast, greedy and easy to implement, they are known to generate more or less accurate results. Based on a matching cost, these methods generally use the *Winner-Take-All* (WTA) [1] procedure which selects the disparity with minimum matching cost (possibly aggregated [1]) for each individual pixel.

Another class of energy-based models are the *global* ones [1, 2, 4], which include *dynamic programming* [2, 5], *graph-cut* [6, 7], *simulated annealing* [8, 9], and *highest confidence first* [10]. These algorithms are generally slower then WTA but generate smoother and more accurate results [1].

In these latter methods, the energy function $E(d)$ often involves two components. The first one ($E_{\text{data}}(d)$) is called the *likelihood* energy term and measures the disagreement between the disparity map d and the two input stereo images I_{ref} and I_{mat}. As for the second one ($E_{\text{smooth}}(d)$), called the *prior* energy term, it encodes constraints on the desired solution (essentially the smoothness assumption of the disparity map). $E_{\text{data}}(d)$ is pixel-based or window-based and is generally seen as a matching cost function such as a sum of absolute or squared difference. $E_{\text{smooth}}(p)$ is often restricted to measure differences between pixel disparity. A simple function often used is one provided by the Potts model which simply counts the number of adjacent pixels with different disparities.

In this paper, we propose a novel global energy-based model efficiently mixing a local and robust regularization term with global spatial constraints. In practice, our model assumes that the disparity map is piecewise smooth and assigns single disparity to regions instead of pixels. Estimating a good set of regions $\hat{\mathcal{R}}$ is a delicate task since ideally, all regions $r \in \hat{\mathcal{R}}$ must be spatially homogeneous with identical disparity. This partition into regions $\hat{\mathcal{R}}$ is estimated by combining an unsupervised Markovian segmentation with a roughly estimated disparity map. The latter disparity map is obtained by using a basic WTA procedure. The proposed global energy function is well suited to finding good disparity discontinuities at object boundaries, especially when the number of disparities is large. An Iterated Conditional Modes (ICM) [11] algorithm is used to optimize this discontinuity-preserving global energy function.

The remainder of this paper is organized as follows. In section 2, our spatially constrained energy-based model is described. Section 3 describes the region partition model while Section 4 is devoted to the deterministic optimization algorithm used in our application. Finally, Section 5 reports some experimental results on real images and Section 6 concludes.

2. SPATIALLY CONSTRAINED ENERGY-BASED MODEL

The energy function used in this work involves two terms,

$$E(d) = E_{\text{data}}(d_{\hat{\mathcal{R}}}) + \lambda E_{\text{smooth}}(d), \tag{1}$$

i.e., a linear combination of a *likelihood* and *prior* term. In our application, the likelihood term is constrained by a set of spa-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

tially homogeneous regions with identical disparity $\hat{\mathcal{R}} \triangleq \{r_n, n = 1, \ldots, N_{r_{\max}}\}$. A constant disparity label is associated with each site belonging to a detected region $r_n \in \hat{\mathcal{R}}$. The way $\hat{\mathcal{R}}$ is computed will be further explained in Section 3. For the likelihood energy term $E_{\text{data}}(d_{\hat{\mathcal{R}}})$, we choose to use a (region-constrained) absolute difference, namely,

$$E_{\text{data}}(d_{\hat{\mathcal{R}}}) = \sum_{n=1}^{N_{r_{\max}}} \sum_{p \in r_n} |I_{\text{ref}}(p) - I_{\text{mat}}(p + d_p)|, \quad (2)$$

with the constraint $(\forall r_n \in \hat{\mathcal{R}}, \forall s \in r_n, d_s = C_r^{\text{st}})$ which means that every pixel within a region r_n is assigned a constant disparity C_r^{st}. $I_{\text{ref}}(p)$ and $I_{\text{mat}}(p + d_p)$ designate respectively the gray (or color) level value at pixel p and $p+d_p$ (on the same line) in the two stereo images. Experiments have shown that squared difference gives somewhat similar results [1].

Concerning $E_{\text{smooth}}(d)$, the choice of a good prior model is crucial, mainly to avoid poor results over object boundaries[1]. The Potts model is among the simplest edge-preserving models since it penalizes pairs of disparity equally [7, 12] :

$$E_{\text{smooth}}(d) = \sum_{p} \sum_{v \in \mathcal{N}_p} \delta_K(d_p, d_v), \quad (3)$$

where $\mathcal{N}_p$ is the set of neighboring pixels (in the disparity map) around p and $\delta_K(d_p, d_v)$ is the Kronecker delta function that returns 1 when $d_p \neq d_v$ and 0 otherwise. The Potts model performs well mainly when the number of disparity levels is small. However, when this value is large, the Potts model tends to oversmooth the disparity labels. To minimize this problem, we used a *robust* Potts model by replacing δ_K in Eq. (3) by a robust function $\rho(d_p, d_v) \in [0, 1]$. For simplicity, we use the Leclerc function, namely,

$$\rho(d_p, d_v) = 1 - \exp\Big(-\frac{(d_p - d_v)^2}{\sigma^2}\Big), \quad (4)$$

where σ is a constant value. Notice that Eq. (4) is isotropic and does not explicitly represent the discontinuities in disparity at the boundaries of objects. So, to capture this phenomenon, we made $E_{\text{smooth}}(d)$ also depend on the intensity difference of the input reference image I_{ref},

$$\phi(I_{\text{ref}}(p), I_{\text{ref}}(v)) = \exp\Big(-\frac{|I_{\text{ref}}(p) - I_{\text{ref}}(v)|}{\gamma^2}\Big), \quad (5)$$

in order to make depth discontinuities in d correspond to spatial discontinuities in I_{ref} (γ is a constant value). Similar (but different) discontinuity-preserving constraints were proposed by Kolmogorov and Zabih [7] and Belhumeur [2].

Combining together Eqs. (1) to (5), the global energy function to be minimized can be written as

$$E(d_{\hat{\mathcal{R}}}) = \sum_{n=1}^{N_{r_{\max}}} \sum_{p \in r_n} |I_{\text{ref}}(p) - I_{\text{mat}}(p + d_p)| + \lambda \sum_p \sum_{v \in N_p} \rho(d_p, d_v)\, \phi(I_{\text{ref}}(p), I_{\text{ref}}(v)). \quad (6)$$

With this global energy function, the estimated disparity map can be expressed by

$$\hat{d} = \arg\min_{d} E(d_{\hat{\mathcal{R}}}).$$

[1] Let us note that this characteristic is also taken into account and enforced by the global spatial constraints of our model.

3. REGION PARTITION MODEL

Two steps are required to obtain a reliable partition $\hat{\mathcal{R}}$ into spatially homogeneous regions with identical disparity. First, the reference image I_{ref} has to be spatially segmented into a label field I_{seg}. To this end, we use an unsupervised Markovian model designed for a gray level segmentation into m spatially homogeneous classes. Second, I_{seg} must be combined with a roughly estimated disparity map $d^{[k]}$ in order to make sure that each region $r \in \hat{\mathcal{R}}$ is likely to contain a single disparity.

3.1. Unsupervised Spatial Markovian Segmentation

Let $(I_{\text{ref}}, I_{\text{seg}})$ be a pair of random fields where $I_{\text{seg}} = \{I_{\text{seg}}(s), s \in S\}$ and $I_{\text{ref}} = \{I_{\text{ref}}(s), s \in S\}$ represent respectively the label field (related to the segmented image) and the observation field. They are both defined on $S = \{s = (i, j)\}$, a 2D lattice of N sites. Each $I_{\text{ref}}(s)$ takes a value in $\{0, ..., 255\}$ (256 gray levels) and each $I_{\text{seg}}(s)$ takes a value in $\{1, \ldots, m\}$, where m corresponds to the number of classes of the segmentation map.

In this framework, the segmentation problem in m classes can be viewed as a statistical labeling problem according to a global Bayesian formulation in which the posterior distribution $P(I_{\text{seg}}/I_{\text{ref}}) \propto \exp -U(I_{\text{seg}}, I_{\text{ref}})$ has to be maximized [11]. By assuming independence between each random variable $I_{\text{ref}}(s)$ given $I_{\text{seg}}(s)$, and an isotropic Potts model with a second-order neighborhood, the corresponding posterior energy to be minimized is [11]

$$U(I_{\text{seg}}, I_{\text{ref}}) = \sum_{s \in S} \Psi_s(I_{\text{seg}}(s), I_{\text{ref}}(s)) + \beta \sum_{<s,t>} \big[1 - \delta(I_{\text{seg}}(s), I_{\text{seg}}(t))\big], \quad (7)$$

where $\Psi_s(I_{\text{seg}}(s), I_{\text{ref}}(s)) = -\ln P(I_{\text{ref}}(s)/I_{\text{seg}}(s))$ and δ is the Kronecker function. This segmentation step leads to the minimization of $U(I_{\text{seg}}, I_{\text{ref}})$ which is also an energy function of the form of Eq. (1), i.e., $E_{\text{data}}(.) + \beta\, E_{\text{smooth}}(.)$.

We model the conditional distribution $P(I_{\text{ref}}(s)/I_{\text{seg}}(s))$ of each class by a Normal law. The parameter vector $\Phi = [(\mu_1, \sigma_1), \ldots, (\mu_m, \sigma_m)]$ of this distribution mixture is estimated with the Iterative Conditional Estimation (ICE) algorithm [13] which gives the best estimation $\hat{\Phi}$ in the least-squares sense. Once $\hat{\Phi}$ is estimated, Eq. (7) can be optimized using a classical ICM relaxation technique [11].

3.2. Partition Into Regions

To obtain a reliable partition $\hat{\mathcal{R}}$, a roughly estimated disparity map $d^{[0]}$ is needed. This is obtained with a basic Winner-Take-All (WTA) procedure. A set of regions $\hat{\mathcal{R}}$ is then defined by combining $d^{[0]}$ and I_{seg} with the following operation : every pixel $s \in \hat{\mathcal{R}}$ is assigned to the class label $d^{[0]}(s) + m \times I_{\text{seg}}(s)$, where m is the number of classes in I_{seg}. In this way, every region $r_n \in \hat{\mathcal{R}}$ is both spatially homogeneous (in the MAP sense according to the model defined in Section 3.1) and belonging to a unique disparity class (according to the WTA model). This set of regions is exploited by our region-based energy model defined in Section 2 which will be optimized by our minimization strategy defined in Section 4.

4. MINIMIZATION STRATEGY

To minimize the global energy function of Eq. (6), we used a region-constrained version of the classical ICM relaxation algorithm [11] (cf. Algorithm 1). This deterministic algorithm is not guaranteed to find the global *minima*; nevertheless, it drastically reduces computational time compared to stochastic relaxation techniques such as simulated annealing [8]. In our application, two factors ensure a fast and good convergence. The first one derives from the fact that the initial depth map, obtained with a WTA procedure, is relatively close to the global *minima*. A local optimization technique will efficiently improve this rough estimation. The second reason comes from the fact that our problem is correctly and strongly constrained by the precomputed partition map into regions $\hat{\mathcal{R}}$. It guarantees the existence and the uniqueness of a consistent solution which continuously depends on the data.

Region-based ICM Algorithm

d	The disparity map to be estimated
d_p	Disparity at pixel p taking values $\in [1, \ldots, N_d]$
$E(.)$	A real-valued function to be minimized
I_{ref}	Input reference image
I_{seg}	Label field of I_{ref} returned by the Markovian segmentation procedure
k	The iteration step
$\hat{\mathcal{R}}$	Partition into regions

1. Initialization

$d^{[0]}$ is obtained with a simple WTA procedure

$I_{\text{seg}} \leftarrow$ Segmentation of I_{ref} into m classes

$k \leftarrow 0$

2. ICM optimization

while $d^{[k+1]} \neq d^{[k]}$ **do**

  $\hat{\mathcal{R}} \leftarrow d^{[k]} + m \times I_{\text{seg}}$

  for *each region* $r \in \hat{\mathcal{R}}$ **do**

    for *every possible disparity* $p \in [1, \ldots, N_d]$ **do**

      Compute $E(d^{[k]}_{\hat{\mathcal{R}},[d_s=p\ \forall s \in r]})$, by considering a constant disparity label p for each site $s \in r$,

    Assign to every site $s \in r$ the disparity p that most minimizes $E(.)$, i.e.

$$\forall s \in r, \quad d_s = \hat{p} \quad \text{with}$$

$$\hat{p} = \arg\min_p E(d^{[k]}_{\hat{\mathcal{R}},[d_s=p\ \forall s \in r]})$$

  $k \leftarrow k + 1$

  $d^{[k]} \leftarrow d^{[k-1]}$

Algorithm 1: Region-based ICM algorithm

5. EXPERIMENTAL RESULTS

To reduce the effect of the noise in $d^{[0]}$, the matching cost $(C_{d_i}(x, y))$ for every disparity d_i is *aggregated* over a two-dimensional support region [1]. This operation is performed by making a 2D convolution,

$$C_{d_i}(x, y) = w(x, y) * C^0_{d_i}(x, y),$$

where $w(x, y)$ is a 3×3 separable box filter and $C^0_{d_i}(x, y)$ is the absolute difference over all pixels when disparity equals d_i.

To evaluate our stereo algorithm, we implemented a quality measure, based on a ground truth disparity map d^{Truth} which computes the percentage of bad matching pixels, namely,

$$B = \frac{1}{N} \sum_p (|d_p - d^{\text{Truth}}_p| > \delta),$$

where N is the total number of pixels in disparity map d, and δ is a disparity tolerance that we set to 1. We also provide the percentage of bad matching pixels for all pixels located over non-occluded regions. We call this measure $B_{\bar{O}}$ [1]. Furthermore, since processing time is an important factor when comparing stereo algorithms, we added a time measure reported in seconds. Table 1 summarizes results for five different algorithms including our *region-based ICM* (RB-ICM) approach.

The *WTA* method that we have implemented simply selects for each pixel the disparity with lowest matching cost. The *PB-ICM* method proposed here is a basic pixel-based ICM algorithm [11] applied to minimize the global energy function of Eq. (6). As for *SA*, it is a simulated annealing algorithm involving a Gibbs sampler [8] procedure also used to minimize Eq. (6). The number of iterations for SA is 500. For the graph cut (*GC*) method, we used the Maxflow / min-cut algorithm of Roy and Cox [6] which is one of the first graph cut algorithms applied to the problem of stereo matching. For these five algorithms, we chose parameters that give the best statistics. In this perspective, aggregation of the matching cost is made with a 9×9 sliding window [1] for WTA, ICM and SA but not for GC and RB-ICM. Results from *Tsukuba* and *Cones* are presented in Figure 5.

For every example, outside λ, our algorithm had constant parameters. We set $\gamma^2 = 64$, $\sigma = nbDisparities \times 0.1$, $\beta = 1$ and $m = 9$. For *Tsukuba* and *Venus* examples, we set $\lambda = 3$, and for *Sawtooth* and *Cones*, we set $\lambda = 12$. All stereo algorithms were executed on a 2 GHz Pentium IV with 512 MB of memory. Let us note that all examples used in this work were downloaded from Middlebury's stereo vision web site [14] (thanks to D. Scharstein and collaborators).

Table 1 shows that graph cut (GC) and *region-based-ICM* (RB-ICM) are clearly superior which is in part consistent with other published results [7, 1]. When comparing *RB-ICM* and GC, it should be noted that the latter minimizes a different global energy function since the prior term $E_{\text{smooth}}(d)$ is replaced by a single smoothing constant [6]. However, in most cases, *RB-ICM* happens to be quicker and more accurate. In all cases, *RB-ICM* converged between 5 and 12 iterations.

6. CONCLUSION

We have presented an original framework for inferring scene geometry from a pair of stereo images. The spatially-constrained global energy function introduced here is made of a combination

	Tsukuba			Venus			Sawtooth			Cones		
Dimensions	384x288			434x383			434x380			225x187		
Disparity Levels	16			20			20			28		
	B	$B_{\bar{O}}$	Time	B	$B_{\bar{O}}$	Time	B	$B_{\bar{O}}$	Time	B	$B_{\bar{O}}$	Time
WTA	6.5%	5.3%	9.3s	7.7%	5.4%	18.1s	4.3%	1.3%	18.2s	18.6%	16.4%	6.2s
ICM	6.0%	4.9%	19.5s	9.3%	6.8%	44.6s	4.2%	1.2%	46.6s	18.6%	17.1%	16.9s
SA	4.0%	2.3%	503s	4.4%	2.4%	1004s	4.8%	2.7%	982s	18.0%	15.1%	350s
GC	3.6%	2.0%	44.5s	3.2%	1.3%	98.1s	5.8%	2.3%	78.5s	16.4%	11.7%	25.8s
RB-ICM	3.2%	1.56%	26.7s	3.1%	1.4%	48.2s	3.7%	0.91%	65.3s	13.1%	8.9%	15.1s

TAB. 1 – *Comparative performance of stereo algorithms using three different measures. B represents the percentage of bad matching pixels over the entire disparity map, $B_{\bar{O}}$ represents the percentage of bad matching pixels over the non-occluded regions and Time is the time in seconds needed by each algorithm to converge.*

of a *likelihood* term and a *prior* term involving a *robust* Potts model with a discontinuity-preserving constraint.

The disparity map d is inferred by minimizing $E(d)$. Such an operation could have been done by a pixel-based stochastic or deterministic relaxation method. However, these approaches are not sufficiently constrained. Consequently, they turn out to be slow and/or prone to fall into local minimas. To overcome these limitations, we have proposed a deterministic *region-constrained* procedure. In practice, it assumes that the disparity map is piecewise smooth and assigns a single disparity to regions $r \in \hat{\mathcal{R}}$ instead of pixels. The computation of a partition $\hat{\mathcal{R}}$ into spatially homogeneous regions with unique disparity is made by a two-step procedure. The first step allows us to obtain a Markovian segmentation (I_{seg}) of an input reference image I_{ref}. This result is then combined with a roughly estimated disparity map $d^{[0]}$. In order to ensure reliable results, the *region-based-ICM*(RB-ICM) minimization procedure must be well initialized. This implies that $d^{[0]}$ must not be far from global minima, or RB-ICM will converge over inaccurate results. We have shown that a simple WTA is sufficient. Results indicate that our approach is fast and performs well compared to other existing methods.

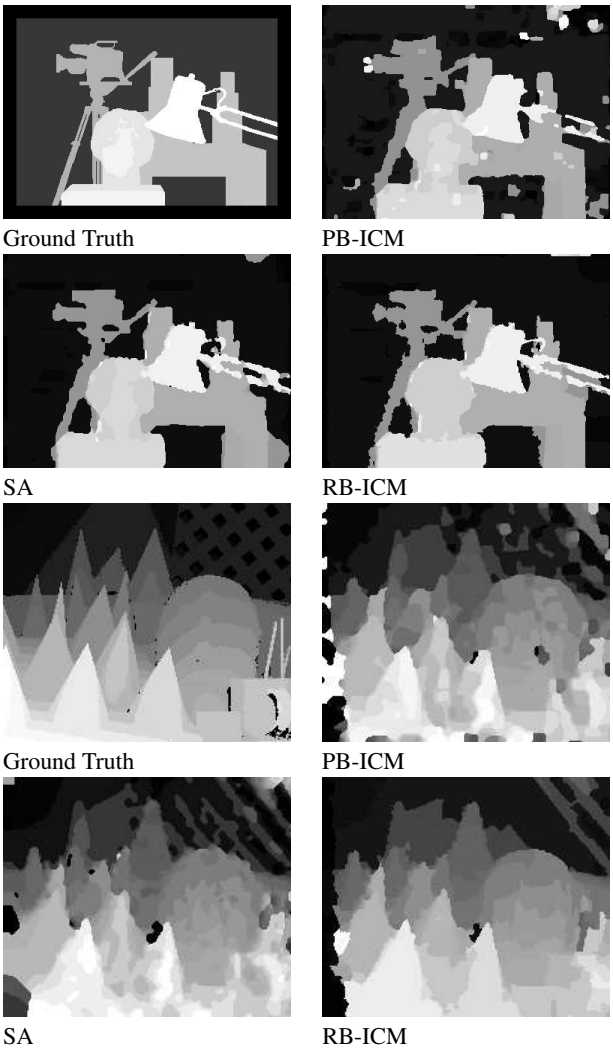

FIG. 1 – *Results obtained by three algorithms over the "Tsukuba" and "Cones" example.*

7. REFERENCES

[1] Szeliski R. Scharstein, D. and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *In Proc. of the IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, HI*, 2001.

[2] P. Belhumeur. A bayesian-approach to binocular stereopsis. *Intl. J. Comp. Vision*, 19(3) :237–260, 1996.

[3] D. Marr. *Vision*. W. H. Freeman and Company, San Francisco, CA, USA, 1982.

[4] R. Szeliski and R. Zabih. An experimental comparison of stereo algorithms. 1999.

[5] Hingorani S. Rao S. Cox, I. and B. Maggs. A maximum likelihood stereo algorithm. *Comput. Vis. Image Underst.*, 63(3) :542–567, 1996.

[6] S. Roy and J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *In Proc. of ICCV*, pages 492–502, 1998.

[7] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions via graph cuts. In *In Proc. of ICCV*, pages 508–515, 1999.

[8] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6) :721–741, 1984.

[9] S. Beernard. Stochastic stereo matching over scale. *Int. J. Comp. Vis.*, 3(1) :17–32, 1989.

[10] P. Chou and C. Brown. The theory and practice of bayesian image labeling. In *In Proc. of ICCV*, pages 185–210, 1990.

[11] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc.*, 48(3) :259–302, 1986.

[12] Geman S. Graffigne C. Geman, D. and P. Dong. Boundary detection by constrained optimization. *IEEE Trans. Pattern Anal. Machine Intell.*, 12(7) :609–628, 1990.

[13] Pieczynski W. Statistical image segmentation. *Machine Graphics and Vision*, 1(1) :261–268, 1992.

[14] www.middlebury.edu/stereo.