Occlusion Detection in Dense Stereo Estimation with Convex Optimization

Antonin Chambolle, Pascal Monasse, Pauline Tan

OCCLUSION DETECTION IN DENSE STEREO ESTIMATION WITH CONVEX OPTIMIZATION

Antonin Chambolle[1] *Pascal Monasse*[2] *Pauline Tan*[3*]

[1] CMAP, École polytechnique, CNRS, Université Paris-Saclay, 91128, Palaiseau, France
[2] Université Paris Est, LIGM, CNRS, Marne-la-Vallée, France
[3] Onera, Dota, Palaiseau, France

ABSTRACT

In this paper, we propose a dense two-frame stereo algorithm which handles occlusion in a variational framework. Our method is based on a new regularization model which includes both a constraint on the occlusion width and a visibility constraint in nonoccluded areas. The minimization of the resulting energy functional is done by convex relaxation. A post-processing then detects and fills the occluded regions. We also propose a novel dissimilarity measure that combines color and gradient comparison with a variable respective weight, to benefit from the robustness of the comparison based on local variations while avoiding the fattening effect it may generate.

Index Terms— Stereo-matching, occlusion detection, variational method, convex relaxation

1. INTRODUCTION

Stereo matching has been a popular research topic in computer vision for many years [1], stimulated by the Middlebury evaluation and its datasets. Two-frame stereo algorithms aim at estimating the scene depth map given two images taken from two different points of view and knowing the cameras parameters and their relative position. The distance to the camera (depth) of any 3D point which is visible from both points of view can be deduced from the estimation of the displacement of its projections from a view to the other. The latter is called *disparity*.

Given a stereo pair, the disparity estimation is equivalent to matching each pixel in the reference view to its *homologue pixel*, namely the projection in the other view of the same 3D point. Without any reliable scene model, stereo matching methods mainly rely on the visual similarity of homologue pixels. Following Scharstein and Szeliski's survey [1], stereo algorithms can be classified into two families. Local methods look for the most similar pixel given a dissimilarity measure. However, they visually compare not only the pixels, but also a neighborhood, in order to make the matches more robust. They are usually easy to implement, but lack global consistency. Global methods aim at making up this loss by using some *a priori* knowledge about the disparity map. In particular, they embed in an energy functional the expected properties (e.g. the global smoothness) of the scene and aim at minimizing them. Such methods allow to exploit more information than local methods, but a serious counterpart is the huge algorithmic complexity. Moreover, some functionals are even not minimizable.

Unfortunately, the stereo matching problem is ill-posed. Indeed, some pixels are only visible from one viewpoint. Thus, they have no homologue pixel and the disparity map is not defined everywhere. This phenomenon is called *occlusion*. The affected pixels are said to be *occluded* and the hidding object (and its pixels) are called *occluding*. Despite its unavoidable occurence and the effect on the matching process, it is often ignored, or treated as noise. However, many methods detect the occluded areas in addition to the disparity estimation. A simple way to proceed is to apply a left-right-cross-check filter to detect matches which differ when the images are switched in the matching process [2, 3]. This implies to compute two disparity maps and decouples the disparity estimation and the occlusion detection. Other methods rely on the uniqueness constraint (which enforces a one-to-one matching) and/or the ordering constraint (objects keep their relative position from a view to the other) to prevent some pixels to be matched. Such methods usually provide a special label to mark these pixels [4, 5] and allow a joint estimation of the disparity map and the occlusion map.

2. RELATED WORK

In 2010, Pock *et al.* proposed a variational method which aims at solving the stereo matching problem [6]. The energy functional is minimized thanks to a convex relaxation technique. However, despite the generic form considered in the general framework, the stereo functionals handled in [6] only have two terms, namely the data term and the smoothness term. Functionals of this particular form are classical in

*The third author performed the work while at CMAP, École polytechnique.

stereo methods, but prevent from a proper occlusion handling. Furthermore, the variational framework enforces a dense disparity estimation, which means that the occluded pixels cannot be treated differently from non-occluded pixels and need to be matched.

An analysis of the occlusion phenomenon [7] shows that the occlusion obeys some specific rules. In the particular case where the ordering constraint holds, the occlusion horizontal size is equal in the left image to the disparity jump between the occlusion borders. The ordering-constraint case applies as soon as there is no object of extent less than the distance between the two camera centers. Bobick and Intille's dynamic programming method exploits this analysis [8] to estimate the disparity map line by line. Therefore, no smoothness along the columns is enforced, which leads to streak-like artefacts. Other attempts [9] added a two-pass strategies to keep the 1D settings while adding some 2D smoothness but such approaches lead to an anisotropic smoothness treatment.

In this work, we propose a variational method which enforces an isotropic TV smoothness for the disparity estimation, with an occlusion handling in a dense framework. We estimate a dense disparity map which is constrained to be *linearly interpolated* in occluded areas. This constraint is managed by a *visibility term*, which also allows occlusion detection in a second stage. An occlusion filling is then used to extend the disparity estimation in occluded areas. Our method relies on the convex approach proposed by [6]. Unlike almost all other methods the occlusion is handled together with the stereo matching and not in a separate process.

3. ENERGY FUNCTIONAL

Let $(\mathrm{I_L}, \mathrm{I_R})$ be a pair of stereoscopic images, of rectangular domain Ω. We make the classical assumption that they are generated by a camera with a fronto-parallel motion, which basically means that homologue pixels are located on the same horizontal line. We also assume that the 1D disparity range $I_{\text{disp}} = [u_{\min}, u_{\max}]$ is known. Hence, if a pixel $p \in \mathrm{I_L}$ in the left view in nonoccluded, its homologue pixel q is given by $q = p - u(p)$, with $u(p) \in I_{\text{disp}}$, where we abusively denote $p - u = (p_X - u, p_Y)$ if $p = (p_X, p_Y) \in \mathbb{R}^2$. The ordering constraint implies that $q_X := p_X - u(p)$ is nondecreasing. Hence, the horizontal slope of u cannot exceed 1.

We consider a three-term energy functional

$$E(u) := E_{\text{data}}(u) + E_{\text{smooth}}(u) + E_{\text{vis}}(u) \tag{1}$$

defined on the set of differentiable functions $u : \Omega \to \mathbb{R}$.

3.1. Data term

The data term aims at penalizing the visual dissimilarity between two matched pixels. It is thus defined thanks to a dissimilarity measure $D(p, q)$, which is designed to be large when the compared pixels p and q are not visually similar, and small otherwise. Classical choices for D include the Absolute Difference (AD), which compares the intensity (or the color) values $\mathrm{I_L}(p)$ and $\mathrm{I_R}(q)$ thanks to the Euclidean norm. The AD measure is easy to implement [10], but it is sensitive to noise. A comparison based on local variations [11] could be used to make it more robust. However, such a measure can be interpreted as a block-matching, which is known to favor fattening effect at object edges [12]. This is why we propose a variable weighted mixed comparison, by defining

$$D(p, q) = \alpha(p)\, D_{\text{AD}}(p, q) + \big(1 - \alpha(p)\big)\, D_{\text{grad}}(p, q) \tag{2}$$

with $D_{\text{AD}}(p, q) := \|\mathrm{I_L}(p) - \mathrm{I_R}(q)\|$ (where $\|\cdot\|$ stands for the Euclidean norm) and $D_{\text{grad}}(p, q) := \|\nabla \mathrm{I_L}(p) - \nabla \mathrm{I_R}(q)\|$. The variable weight $\alpha \in [0, 1]$ is chosen so that it is large when p is near scene discontinuities. Since the latter are included in the set of image discontinuities, we can for instance set it as $\alpha(p) = (1 + G_\gamma \star \|\nabla \mathrm{I_L^{ROF}}\|^2(p)/a)^{-1}$ where G_γ is a zero-centered normalized Gaussian of standard variation γ, the symbol $\star$ denoting the convolution product with $\mathrm{I_L^{ROF}}$ a smoothed version of $\mathrm{I_L}$ (smoothed by the Rudin-Osher-Fatemi model of parameter β [13]) and $a > 0$ a parameter. This filter aims at removing the texture, which leads to large gradients but does not correspond to scene discontinuities. When α is constant, (2) is close to the dissimilarity measure proposed in [2]. If we set $g(x, t) := D(x, x - t)$ for any $(x, t) \in \Omega \times \mathbb{R}$, then the data term is given by

$$E_{\text{data}}(u) := \mu \int_\Omega g\big(x, u(x)\big)\, \mathrm{d}x \tag{3}$$

where $\mu > 0$ is a weighting parameter.

3.2. Regularization term

This term encodes a smoothness *a priori* on the disparity map u. We define it by the TV semi-norm since it has been shown [14] to be a natural choice in image processing. This regularization does not penalize sharp discontinuities, which are expected in the disparity map. Hence, we set

$$E_{\text{smooth}}(u) := \mathrm{TV}(u) := \int_\Omega |Du|. \tag{4}$$

where $Du = \nabla u$ when u is smooth.

3.3. Visibility term

The occlusion analysis recalled in Section 2 ensures that if the disparity is linearly interpolated in the occluded areas, then its horizontal slope is equal to 1 where occlusion occurs. Moreover, the horizontal slope cannot exceed 1 elsewhere. Hence, we introduce the visibility term below:

$$E_{\text{vis}}(u) := \int_\Omega r_{\text{vis}}(Du) \tag{5}$$

where $r_{\text{vis}} : p^x = (p^x_X, p^x_Y) \in \mathbb{R}^2 \to \{0; +\infty\}$ is the characteristic function of $(-\infty, 1] \times \mathbb{R}$. In other terms, $r_{\text{vis}}(p^x)$ equals 0 when $p^x_X \leq 1$ and is infinite otherwise.

4. MINIMIZATION BY CONVEX RELAXATION

Although the functional E is nonconvex, it has been proved in [6] that its minimization can be exactly done by solving an auxiliary convex problem. Namely, there exists a convex functional F and a convex set $\mathcal{C}$

$$F(\mathbb{1}_u) = E(u) \quad \text{with} \quad \mathbb{1}_u(x,t) = \begin{cases} 1 & \text{if } u(x) \geq t \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathcal{C} := \left\{ v \in \mathrm{BV}(\Omega \times [0,1]) \,\middle|\, \begin{matrix} \forall\, t \leq u_{\min},\ v(x,\cdot) = 1 \\ \forall\, t \geq u_{\max},\ v(x,\cdot) = 0 \end{matrix} \right\} \tag{7}$$

such that [6, Theorem 3.1] if v^* is a minimizer of F over the convex set $\mathcal{C}$ then, for any $s \in [0,1)$, the s-thresholded map

$$\mathbb{1}_{\{v^* > s\}} : (x,t) \mapsto \begin{cases} 1 & \text{if } v^*(x,t) > s \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

is a also minimizer of F and the function u^s defined by

$$u^s(x) := \sup \left\{ t \in \mathbb{R} \mid \mathbb{1}_{\{v^* > s\}}(x,t) = 1 \right\} \tag{9}$$

is a global minimizer of the nonconvex functional E. The relaxed convex problem can be shown to be equivalent to the following the primal-dual problem, with $C(p) = p_X^x + p^t$ for any $p = (p_X^x, p_Y^x, p^t)$

$$\min_{v \in \mathcal{C}} \sup_{\substack{\phi \in \mathcal{K}_{\mathrm{TV}}(\mu) \\ \Lambda \in \mathrm{BV}(\Omega;\mathbb{R}^+)}} \int_{\Omega \times \mathbb{R}} \phi\, Dv + \Lambda\, C(Dv) \tag{10}$$

with $\mathcal{K}_{\mathrm{TV}}(\mu)$ given in [6, Section 4.2] and Λ the Lagrange multiplier associated to the convex constraint $\partial_X v + \partial_T v \leq 0$ which arises when computing the visibility part of F. The discrete counterpart of Problem (10) is given by

$$\min_{v^h \in \mathcal{C}^h} \sup_{\substack{\phi^h \in \mathcal{K}_{\mathrm{TV}}^h \\ \Lambda^h \in (\mathbb{R}^+)^{N_X N_Y N_T}}} \langle \phi^h, \nabla^h v^h \rangle + \langle \Lambda^h, \delta_X^h v^h + \delta_T^h v^h \rangle \tag{11}$$

where the gradient operator $\nabla^h : v^h \to (\delta_X^h v^h, \delta_Y^h v^h, \delta_T^h v^h)$ is a bounded operator, defined by the forward differences, with a multiplicative factor $1/h$ where h is the data accuracy ($h = 1$ for pixel accuracy, $h = 0.5, 0.25, ...$ for subpixel accuracy, obtained by upsampling). It is solved by a primal-dual algorithm [15] which alternates a projected gradient ascent on the dual variables (ϕ^h, Λ^h) and projected gradient descent on the primal variable v^h, with an overrelaxation step on v^h.

5. OCCLUSION HANDLING

5.1. Occlusion detection

According to the occlusion analysis, the occluded pixels are located where the horizontal slope of u_N^h is 1. Hence, we define the occlusion mask

$$M_{i,j}^{\mathrm{occ}} = \begin{cases} 1 & \text{if } (\delta_X^h u_N^h)_{i-1,j} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Note the x-shift introduced in this definition. It aims at compensating the systematic 1-pixel fattening that occurs on the left side of any disparity discontinuity, since the fattened solution avoids occluded pixels to be matched to occluding pixels, which in general do not belong to the same object.

To avoid holes in the detected occluded areas, which are due to bad placement of the disparity variations (caused by more favorable cost matching for instance), we enhanced the detection thanks to a filtering. The central idea is that, if two close pixels belonging to the same object are occluded, this is also reasonably the case of any pixels located in-between. This leads to the enhanced mask (with R the search radius):

$$\tilde{M}_{i,j}^{\mathrm{occ}} = \begin{cases} 1 & \text{if } \exists\, j - R \leq j'' < j < j' \leq j + R \\ & \text{and } M_{i,j'}^{\mathrm{occ}} = 1 \text{ and } M_{i,j''}^{\mathrm{occ}} = 1 \\ & \text{and } \|\mathrm{I}_{\mathrm{L}}^{\mathrm{ROF}}(i,j) - \mathrm{I}_{\mathrm{L}}^{\mathrm{ROF}}(i,j')\| \leq T \\ & \text{and } \|\mathrm{I}_{\mathrm{L}}^{\mathrm{ROF}}(i,j) - \mathrm{I}_{\mathrm{L}}^{\mathrm{ROF}}(i,j'')\| \leq T \\ M_{i,j}^{\mathrm{occ}} & \text{otherwise.} \end{cases} \tag{13}$$

5.2. Occlusion filling

Under the current hypotheses (left-right camera motion), if we assume that objects are only partially occluded, it is possible to use the known disparities to guess the unknown ones [16]. Hence, like [2], the occluded pixels are supposed to have same disparity as the object located on their left-side:

$$(\tilde{u}_N^h)_{i,j} = \begin{cases} (u_N^h)_{i,j} & \text{if } \tilde{M}_{i,j}^{\mathrm{occ}} = 0 \\ (u_N^h)_{i_0,j} & \text{if } \tilde{M}_{i,j}^{\mathrm{occ}} = 1 \text{ and} \\ & i_0 = \max\{i' \leq i \mid \tilde{M}_{i',j}^{\mathrm{occ}} = 0\}. \end{cases} \tag{14}$$

6. EXPERIMENTAL VALIDATION

6.1. Validation of the data term

We first tested our data term on a synthetic example. The stereo pair consists in a fixed textured background in front of which a textured rectangle moves horizontally. Figure 1 showx the disparity estimation when using a color comparison, a gradient comparison, the mixed dissimilarity measure used in [2] and our variable weighted mixed comparison. The parameters used for α are $a = 100$, $\gamma = 8$ and $\beta = 1/50$. The parameter μ is chosen equal to 50 (for image values between 0 and 255). We observed many errors with the color single comparison, whereas the fattening effect occurs on the right edge when the dissimilarity measure is based on a gradient comparison of constant weight. It has been removed with our data term.

6.2. Disparity estimation and occlusion detection

We tested our methods on images from the Middlebury benchmark[1]. We used Version 2 of this benchmark as it

[1] http://vision.middlebury.edu/stereo/eval/

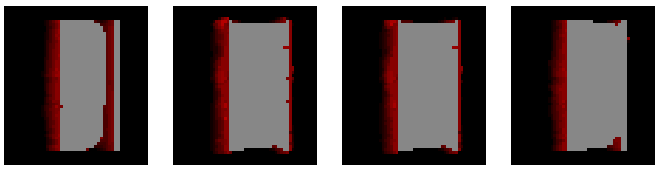

Fig. 1. Fattening effect removal. From left to right: color dissimilarity, gradient dissimilarity, mixed measure of [2], our dissimilarity measure. In red, the estimation error.

	None	using M^{occ}	using $\tilde{M}^{\text{occ}}$
Tsukuba	6.09%	5.64%	5.66%
Venus	2.69%	2.28%	2.18%
Teddy	20.44%	19.64%	19.38%
Cones	15.92%	15.37%	15.62%

Table 1. Disparity estimation error.

provides occlusion maps, necessary for quantitive evaluation. We evaluate the quality of the occlusion detection by the precision and recall rates (see Tab. 2), which are respectively defined as the number of correct detections over the number of detections and the number of correct detections over the number of occluded pixels (according to the groundtruth). The disparity estimation is evaluated thanks to the groundtruth (see Tab. 1).

To avoid side effects of the method, which may occur on the left and right borders of the scene, we horizontally extend the discrete cost volume defined by $g^h_{i,j,k} := g((i,j),k)$ by replicating the first and last raws $N_t := u_{\max} - u_{\min}$ times.

We compare our results with two others occlusion detectors. We first ran the graph cuts method of Kolmogorov and Zabih[2] [4], which we refer to as KZ2. We also applied the left-right-cross-check filter to the disparity maps generated by [6], namely the TV-regularization of the cost volume g^h. Except for one pair, KZ2 did better for both precision and recall rates. However as their method labels as occluded pixels for which matching is too costly, this sometimes leads to false detections. On the opposite, our low rates are often due to wrong placement of the detection (see for instance Venus in Figure 2). The disparity estimation error rates are displayed in Table 1 and at a threshold of 0.5px.

6.3. Occlusion filling

We filled the detected occlusion areas thanks to formula (14). Columns 2 and 3 in Table 1 display the disparity estimation error rate after the filling done from the occlusion maps given by the first mask M^{occ} and by the enhanced mask $\tilde{M}^{\text{occ}}$. We chose $R = 9$. In all cases, the disparity estimation is improved after the occlusion filling process. Sometimes the enhanced mask leads to better results than the initial detection.

[2]We used the code found in [17].

	Our method	KZ2 [4]	TV+LR check
Tsukuba	47.46%	60.30%	36.48%
Venus	26.06%	28.80%	14.52%
Teddy	29.38%	62.73%	20.87%
Cones	43.69%	48.85%	31.21%
	Our method	KZ2 [4]	TV+LR check
Tsukuba	53.94%	60.61%	46.58%
Venus	46.28%	71.34%	42.74%
Teddy	61.50%	63.14%	59.25%
Cones	45.11%	41.95%	61.79%

Table 2. Occlusion detection: precision (top)/recall (bottom).

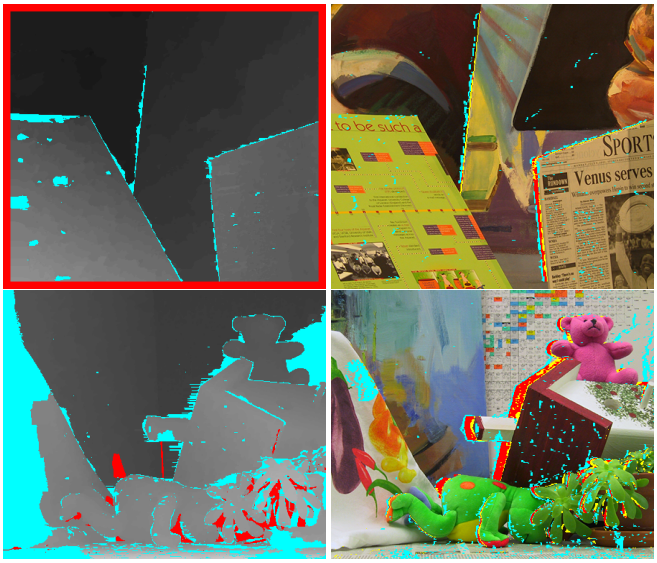

Fig. 2. Disparity map (left) and occlusion map (right). Left: in red, the unknown disparity value and in cyan, the estimation error. Right: in red, the true positive, in yellow, the false negative and in cyan, the false positive.

7. CONCLUSION

In this paper, we have proposed a variational method which handles the occlusion phenomenon in a dense framework. The occlusion detection is parameter-free and relies on disparity slope condition when the ordering constraint is satisfied. A noteworthy feature of our method is that it can be adapted to any dissimilarity measure and various smoothness terms, which makes it very versatile.

8. REFERENCES

[1] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[2] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3017–3024.

[3] Qingxiong Yang, Liang Wang, Ruigang Yang, Henrik Stewénius, and David Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 492–504, 2009.

[4] Vladimir Kolmogorov and Ramin Zabih, "Computing visual correspondence with occlusions using graph cuts," in *IEEE International Conference on Computer Vision*. IEEE, 2001, vol. 2, pp. 508–515.

[5] Jong Dae Oh and C-C Jay Kuo, "Robust stereo matching with improved graph and surface models and occlusion handling," *Journal of Visual Communication and Image Representation*, vol. 21, no. 5, pp. 404–415, 2010.

[6] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle, "Global solutions of variational models with convex regularization," *SIAM Journal on Imaging Sciences*, vol. 3, no. 4, pp. 1122–1145, 2010.

[7] Davi Geiger, Bruce Ladendorf, and Alan Yuille, "Occlusions and binocular stereo," *International Journal of Computer Vision*, vol. 14, no. 3, pp. 211–226, 1995.

[8] Aaron F Bobick and Stephen S Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.

[9] Heiko Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[10] Marsha J Hannah, "Computer matching of areas in stereo images," Tech. Rep., DTIC Document, 1974.

[11] Daniel Scharstein, "Matching images by comparing their gradient fields," in *nternational Conference on Pattern Recognition*. IEEE, 1994, vol. 1, pp. 572–575.

[12] Julie Delon and Bernard Rougé, "Le phénomène d'adhérence en stéréoscopie dépend du critère de corrélation," in *18^e Colloque sur le traitement du signal et des images*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2001.

[13] Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[14] Leonid Iakov Rudin, "Images, numerical analysis of singularities and shock filters," 1987.

[15] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[16] Shafik Huq, Andreas Koschan, and Mongi Abidi, "Occlusion filling in stereo: Theory and experiments," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 688–704, 2013.

[17] Vladimir Kolmogorov, Pascal Monasse, and Pauline Tan, "Kolmogorov and Zabih's Graph Cuts Stereo Matching Algorithm," *Image Processing On Line*, vol. 4, pp. 220–251, 2014.