Fattening free block matching

G. Blanchet [*] A. Buades [†‡] B. Coll[†] J.M Morel [§]
B. Rougé [¶]

Abstract

Block matching along epipolar lines is the core of most stereovision algorithms in geographic information systems. The usual distances between blocks are the sum of squared distances in the block (SSD) or the correlation. Minimizing these distances causes the *fattening* effect, by which the center of the block inherits the disparity of the more contrasted pixels in the block. This fattening error occurs everywhere in the image, and not just on strong depth discontinuities. The fattening effect at strong depth edges is is a particular case of fattening, called *foreground fattening* effect. A theorem proved in the present paper shows that a simple and universal adaptive weighting of the SSD resolves the fattening problem at all smooth disparity points[1]. The optimal SSD weights are nothing but the inverses of the squares of the image gradients in the epipolar direction. With these adaptive weights, it is shown that the optimal disparity function is the result of the convolution of the real disparity with a prefixed kernel. Experiments on simulated and real pairs prove that the method does what the theorem predicts, eliminating surface bumps caused by fattening. However, the method does not resolve the foreground fattening.

1 Introduction

Stereovision consists in finding the depth of a scene from several views of it. This is one of the central problems in computer vision, and it has been an active object of research in the last forty years. Stereovision is based on the fact that differences of depth in a 3D scene create geometrical disparities between views of the same scene taken from different points of view.

Given two stereo rectified images u and v, the question reduces to finding a disparity function ϵ such that $u(x) = v(x + \epsilon(x))$. Like in motion estimation, the above equation presents the aperture problem, namely the ambiguity of the solution, even when some regularity is demanded for the disparity. For this reason, many stereovision algorithms do not look for a function ϵ matching the grey level intensity of each pixel. They prefer to compare the grey levels of an

[*] CNES, 18 avenue Edouard Belin, Toulouse 31401, France

[†] Dpt Matematiques Informatica, Universitat Illes Balears, Ctra Valldemossa km 7.5, Palma de Mallorca, Spain

[‡] MAP5, CNRS - Université Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France

[§] CMLA, ENS Cachan, 61 av. Président Wilson, Cachan 94235, France

[¶] CESBIO, 18 Av. Edouard Belin, Toulouse 31400, France

[1] A Spanish patent has been applied for by Universitat de Illes Balears [1]

entire block around each pixel. The simplest resulting algorithm is known as block matching by SSD (sum of squared distances).

The most important drawback of SSD is the well known “fattening effect”. According to Kanade et Okutomi [7],

> A central problem in stereo matching by computing correlation or sum of squared differences (SSD) lies in selecting an appropriate window size. The window size must be large enough to include enough intensity variation for reliable matching, but small enough to avoid the effects of projective distortion. If the window is too small and does not cover enough intensity variation, it gives a poor disparity estimate, because the signal (intensity variation) to noise ratio is low. If, on the other hand, the window is too large and covers a region in which the depth of scene points (i.e. disparity) varies, then the position of maximum correlation or minimum SSD may not represent correct matching due to different projective distortion in the left and right images. The fattening effect occurs when the selected window contains pixels at different depth. In that case we cannot find exactly the same window and the obtained disparity depends on the different disparities of the window and not only the central pixel itself.

The usual way to cope with the fattening effect is to use adaptive windows that avoid image discontinuities as was first proposed by Kanade et al [7]. Similar works pre-computing edge points and recursively growing a comparison window avoiding them were proposed by Lotti et al. [9] and recently by Wang et al. [23]. Patricio et al. [15] and Yoon et al. [25] select an adaptive window containing only pixels with a grey level similar to the reference one, like in neighborhood and bilateral filters [21, 24].

Other approaches do not try to avoid the discontinuities of the image. They select an adaptive window with a minimum distance criterion. The subjacent idea is that windows which do not contain discontinuities will be matched with a small window distance. Fusiello et al. [5] choose among all the windows containing the reference pixel the one which has a minimal distance with its corresponding one in the second image. Veksler [22] applied the same strategy but used in addition square windows of different sizes. A more elaborated version by Hirschmuller et al [6] adapts the shape of the window by dividing the comparison window into small sub-windows and taking those which attain the minimum distance. The Delon et al. [4] paper proposes a different strategy, the barycentric correction attributing the disparity of a window to the window barycenter pondered by the image gradients.

Point feature matching methods overcome the fattening problem at the cost of a drastic reduction of the match density. Matched features can also be curvilinear, which also circumvents the fattening problem to some extent. For instance, Schmid [20] describes a set of algorithms for automatically matching individual line segments and curves. Robert [16] presents an edge-based stereovision algorithm, where the primitives to be matched are cubic B-splines approximations of the 2-D edges. Musé et al. [14] and Cao et al. [3] discuss how to automatically match pieces of level lines and extract coherent groups of such matches. The Matas et al. [11] MSER method solves the problem by matching stable and homogeneous image regions, but their match set is again sparse. Even if features may seem more local, they depend anyway on a broad neighborhood. The same remark applies to the SIFT method (Lowe [10]) and their affine invariant extensions [13]. Even if the fine scale Laplacian extrema used (e.g.) in the SIFT method are very local, their descriptor around involves anyway a 8×8 window (see [2, 8, 12] for comparison on MSER and SIFT). Thus the fattening problem can occur anyway with these methods.

The fattening effect is not the sole obstacle to a correct disparity computation. Occlusions and moving objects make it a very difficult and sometimes ill-posed problem. Taking simultaneous snapshots with a low baseline avoids partially these drawbacks. However, when using a low baseline a larger precision in the disparity computation is needed to get the same depth precision. The use of a low B/H (where B is the baseline and H is the altitude) was proposed in satellite imaging by Delon and Rouge [4].

2 Mathematical analysis of SSD

Let us denote by $\mathbf{x} = (x, y)$ an image point in the continuous image domain, and by $u_1(\mathbf{x}) = u_1(x, y)$ and $u_2(\mathbf{x})$ the images of an ortho-rectified stereo pair. Assume that the epipolar direction is the x axis. The underlying depth map can be deduced from the disparity function $\varepsilon(\mathbf{x})$ giving the shift of an observed physical point $\mathbf{x}$ from the left image u_1 in the right image u_2. The physical disparity $\varepsilon(\mathbf{x})$ is not well-sampled. Therefore, it cannot be recovered at all points, but only essentially at points $\mathbf{x}$ around which the depth map is continuous. Following the formulation by Delon and Rouge [4] and Sabater [17], around such points, the deformation model from an image to the other is

$$\begin{aligned} u_1(\mathbf{x}) &= u(x + \varepsilon(\mathbf{x}), y) + n_1(\mathbf{x}) \\ u_2(\mathbf{x}) &= u(\mathbf{x}) + n_2(\mathbf{x}), \end{aligned} \tag{1}$$

where u the true scene image and $n_1(\mathbf{x})$ and $n_2(\mathbf{x})$ independent Gaussian white noises with standard deviation σ. (The captor noises are independent because the snapshots are different.) Block matching amounts to finding the disparity at $\mathbf{x}_0$ minimizing

$$e_{\mathbf{x}_0}(\mu) = \int_{[0,N]^2} \varphi(\mathbf{x} - \mathbf{x}_0)\big(u_1(\mathbf{x}) - u_2(\mathbf{x} + (\mu, 0))\big)^2 d\mathbf{x}. \tag{2}$$

where $\varphi(\mathbf{x}-\mathbf{x}_0)$ is a soft window function centered at $\mathbf{x}_0$. For a sake of compactness in notation, $\varphi_{\mathbf{x}_0}(\mathbf{x})$ stands for $\varphi(\mathbf{x}-\mathbf{x}_0)$, $\int_{\varphi_{\mathbf{x}_0}} u(\mathbf{x})d\mathbf{x}$ will be an abbreviation for $\int \varphi(\mathbf{x} - \mathbf{x}_0)u(\mathbf{x})d\mathbf{x}$; we will write $u(\mathbf{x} + \mu)$ for $u(\mathbf{x} + (\mu, 0))$ and ε for $\varepsilon(\mathbf{x})$. The minimization problem (2) rewrites

$$\min_{\mu} \int_{\varphi_{\mathbf{x}_0}} \big(u(\mathbf{x} + \varepsilon(\mathbf{x})) + n_1(\mathbf{x}) - u(\mathbf{x} + \mu) - n_2(\mathbf{x} + \mu)\big)^2 d\mathbf{x}.$$

Differentiating this energy with respect to μ implies that any local minimum $\mu = \mu(\mathbf{x}_0)$ satisfies

$$\int_{\varphi_{\mathbf{x}_0}} \Big(u(\mathbf{x}+\varepsilon(\mathbf{x}))+n_1(\mathbf{x})-u(\mathbf{x}+\mu)-n_2(\mathbf{x}+\mu)\Big)\times\Big(u_x(\mathbf{x}+\mu)+(n_2)_x(\mathbf{x}+\mu)\Big)\, d\mathbf{x} = 0. \tag{3}$$

One has by Taylor-Lagrange formula $u_x(\mathbf{x}+\mu) = (u_x(\mathbf{x}+\varepsilon)) + O_1(\mu-\varepsilon)$, with

$$O_1(\mu-\varepsilon) \leq |\mu-\varepsilon| \max |u_{xx}(\mathbf{x}+\varepsilon)| \tag{4}$$

and $u(\mathbf{x}+\varepsilon(\mathbf{x})) - u(\mathbf{x}+\mu) = u_x(\mathbf{x}+\varepsilon)(\varepsilon-\mu) + O_2((\varepsilon-\mu)^2)$, where

$$|O_2((\varepsilon-\mu)^2)| \leq \frac{1}{2} \max |(u_{xx}(\mathbf{x}+\varepsilon))| (\varepsilon-\mu)^2 \,.$$

Thus equation (3) yields

$$\begin{multline}\int_{\varphi_{\mathbf{x}_0}} \Big(u_x(\mathbf{x}+\varepsilon)(\varepsilon-\mu) + O_2((\varepsilon-\mu)^2) + n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu) \Big) \times \\ \Big(u_x(\mathbf{x}+\varepsilon) + O_1(\mu-\varepsilon) + (n_2)_x(\mathbf{x}+\mu) \Big)\, d\mathbf{x} = 0. \qquad (5)\end{multline}$$

and therefore

$$\mu \int_{\varphi_{\mathbf{x}_0}} (u_x(\mathbf{x}+\varepsilon))^2 d\mathbf{x} = \int_{\varphi_{\mathbf{x}_0}} (u_x(\mathbf{x}+\varepsilon))^2\, \varepsilon(\mathbf{x})\, d\mathbf{x} + \tilde{\mathcal{A}} + \tilde{\mathcal{B}} + \mathcal{O}_1 + \mathcal{O}_2, \tag{6}$$

where

$$\begin{aligned}
\tilde{\mathcal{A}} &= \int_{\varphi_{\mathbf{x}_0}} u_x(\mathbf{x}+\varepsilon)\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big) d\mathbf{x}; & (7)\\
\tilde{\mathcal{B}} &= \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu) d\mathbf{x}; & (8)\\
\mathcal{O}_1 &= \int_{\varphi_{\mathbf{x}_0}} u_x(\mathbf{x}+\varepsilon)(\varepsilon-\mu)(n_2)_x(\mathbf{x}+\mu) d\mathbf{x} & \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_1(\mu-\varepsilon)\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big) d\mathbf{x}; & (9)\\
\mathcal{O}_2 &= \int_{\varphi_{\mathbf{x}_0}} O_2(\varepsilon-\mu)^2 (u_x(\mathbf{x}+\varepsilon)) d\mathbf{x} & \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_2(\varepsilon-\mu)^2 [O_1(\mu-\varepsilon) + (n_2)_x(\mathbf{x}+\mu)] d\mathbf{x} & \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_1(\mu-\varepsilon)(u_x(\mathbf{x}+\varepsilon))(\varepsilon-\mu) d\mathbf{x}. & (10)
\end{aligned}$$

Denote by $\overline{\varepsilon}$ the average of ε on the support of $\varphi(\mathbf{x}-\mathbf{x}_0)$, denoted by $B_{\mathbf{x}_0}$. By the Taylor-Lagrange theorem we have

$$\tilde{\mathcal{A}} = \mathcal{A} + \mathcal{O}_{\mathcal{A}}$$

where

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} u_x(\mathbf{x}+\varepsilon)\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x} \tag{11}$$

and

$$\mathcal{O}_{\mathcal{A}} = (\bar{\varepsilon} - \mu) \int_{\varphi_{\mathbf{x}_0}} (u_x(\mathbf{x}+\varepsilon))(n_2)_x(\mathbf{x}+\tilde{\varepsilon}(\mathbf{x}))d\mathbf{x}, \tag{12}$$

where $\tilde{\varepsilon}(\mathbf{x})$ satisfies $\tilde{\varepsilon}(\mathbf{x}) \in [\min(\mu,\bar{\varepsilon}), \max(\mu,\bar{\varepsilon})]$. In the same way,

$$\tilde{\mathcal{B}} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x}$$

so that $\tilde{\mathcal{B}} = \mathcal{B} + \mathcal{O}_{\mathcal{B}}$, where

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\bar{\varepsilon})\big)(n_2)_x(\mathbf{x}+\bar{\varepsilon})d\mathbf{x} \tag{13}$$

and

$$\mathcal{O}_{\mathcal{B}} = (\mu - \bar{\varepsilon}) \int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_{xx}(\mathbf{x}+\tilde{\varepsilon}(\mathbf{x})) - (n_2(n_2)_x)_x(\mathbf{x}+\tilde{\varepsilon}(\mathbf{x}))d\mathbf{x}. \tag{14}$$

The terms $\mathcal{A}$ and $\mathcal{B}$ are stochastic and we must estimate their expectation and variance. The terms $\mathcal{O}_1$, $\mathcal{O}_2$, $\mathcal{O}_{\mathcal{A}}$, $\mathcal{O}_{\mathcal{B}}$ are higher order terms with respect to $\varepsilon - \mu$ and are negligible if $\varepsilon - \mu$ is small, and the noise samples bounded.

Lemma 1 *Consider the main error terms*

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} u_x(\mathbf{x}+\varepsilon(\mathbf{x}))\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\bar{\varepsilon})\big)d\mathbf{x}$$

and

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\bar{\varepsilon})\big)(n_2)_x(\mathbf{x}+\bar{\varepsilon})d\mathbf{x}$$

as defined above. One has $\mathbf{E}\mathcal{A} = \mathbf{E}\mathcal{B} = 0$ *and*

$$\begin{aligned}
\mathrm{Var}(\mathcal{A}) &= 2\sigma^2 \int \left[\varphi(\mathbf{x}-\mathbf{x}_0)u_x(\mathbf{x}+\varepsilon)\right]_N^2 d\mathbf{x} \\
&\le 2\sigma^2 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 u_x(\mathbf{x}+\varepsilon)^2; \\
\mathrm{Var}(\mathcal{B}) &\le \frac{2\pi^2\sigma^4}{3} \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}.
\end{aligned}$$

Proof: Notice that $n_1(\mathbf{x})$ and $n_2(\mathbf{x}+\bar{\varepsilon})$ are independent Gaussian noises with variance σ^2. Thus their difference is again a Gaussian noise with variance $2\sigma^2$. It therefore follows that

$$\mathrm{Var}(\mathcal{A}) = 2\sigma^2 \int \left[\varphi(\mathbf{x}-\mathbf{x}_0)u_x(\mathbf{x}+\varepsilon))\right]_N^2 d\mathbf{x} \le 2\sigma^2 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 (u_x(\mathbf{x}+\varepsilon))^2 d\mathbf{x}.$$

$$\begin{aligned}
\mathrm{Var}(\mathcal{B}) &\le 2\left[\mathrm{Var}(\int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_x(\mathbf{x}+\bar{\varepsilon}) + \mathrm{Var}(\int_{\varphi_{\mathbf{x}_0}} n_2(\mathbf{x}+\bar{\varepsilon})(n_2)_x(\mathbf{x}+\bar{\varepsilon}))\right] \\
&\le 2\left[\sigma^2 \times \frac{\pi^2\sigma^2}{3} \int \varphi^2(\mathbf{x}-\mathbf{x}_0) + \frac{\sigma^4}{2} \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2\right] \\
&= \frac{2\pi^2\sigma^4}{3} \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2.
\end{aligned}$$

Theorem 1 (Main disparity formula and exact noise error estimate) *Consider an optimal disparity $\mu(\mathbf{x}_0)$ obtained as any absolute minimizer of $e_{\mathbf{x}_0}(\mu)$ (defined by (2)). Then*

$$\mu(\mathbf{x}_0) = \frac{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x}+\varepsilon(\mathbf{x}))]^2 \varepsilon(\mathbf{x}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x}+\varepsilon(\mathbf{x}))]^2 d\mathbf{x}} + \mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0} + \mathcal{O}_{\mathbf{x}_0} \tag{15}$$

where

$$\mathcal{E}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} (u_x(\mathbf{x}+\varepsilon(\mathbf{x}))\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x}+\varepsilon(\mathbf{x}))]^2 d\mathbf{x}}$$

is the dominant noise term,

$$\mathcal{F}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x}+\varepsilon(\mathbf{x}))]^2 d\mathbf{x}}$$

and $\mathcal{O}_{\mathbf{x}_0}$ is made of smaller terms. In addition the variances of the main error terms due to noise satisfy

$$\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) = 2\sigma^2 \frac{\int [\varphi(\mathbf{x}-\mathbf{x}_0) u_x(\mathbf{x}+\varepsilon)]_N^2 \, d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u_x(\mathbf{x}+\varepsilon)^2 d\mathbf{x}\Big)^2}; \tag{16}$$

$$\mathrm{Var}(\mathcal{F}_{\mathbf{x}_0}) \le \frac{\frac{2\pi^2}{3}\sigma^4 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u_x(\mathbf{x}+\varepsilon)^2 d\mathbf{x}\Big)^2}. \tag{17}$$

Finally,

$$\mathcal{O}_{\mathbf{x}_0} = \frac{\mathcal{O}_1 + \mathcal{O}_2 + \mathcal{O}_{\mathcal{A}} + \mathcal{O}_{\mathcal{B}}}{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x}+\varepsilon(\mathbf{x}))]^2 d\mathbf{x}},$$

and

$$\mathbf{E}\mathcal{O}_{\mathbf{x}_0} = O(\max_{\mathbf{x}\in B_{x_0}} |\varepsilon(\mathbf{x}) - \mu|),$$

$$\mathrm{Var}(\mathcal{O}_{\mathbf{x}_0}) = O(\max_{\mathbf{x}\in B_{x_0}} |\varepsilon(\mathbf{x}) - \mu|^2).$$

Proof: This result is an immediate consequence of (6) completed with the variance estimates in Lemma 1. The estimates for the higher order terms $\mathcal{O}$ are a straightforward application of Cauchy-Schwartz inequality.

Remark Theorem 1 makes sense only when the optimal disparity $\mu(\mathbf{x}_0)$ is consistent, namely satisfies for $\mathbf{x}$ in the support $B_{\mathbf{x}_0}$ of $\varphi(\mathbf{x}-\mathbf{x}_0)$,

$$|\varepsilon(\mathbf{x}) - \mu(\mathbf{x}_0)| << 1. \tag{18}$$

Thus, one of the main steps of block matching must be to eliminate inconsistent matches.

Remark In all treated examples, it will be observed that $\mathrm{Var}(\mathcal{B}) \ll \mathrm{Var}(\mathcal{A})$, which by Lemma 1 directly follows from

$$\sigma^2 \left[\frac{2\pi^2}{3} \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 + \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2\right] \ll 2 \int [\varphi(\mathbf{x}-\mathbf{x}_0) u_x(\mathbf{x}+\overline{\varepsilon})]_N^2. \tag{19}$$

3 Mathematical definition of fattening, and its solution

The previous mathematical formulation tells us that the obtained minimizer for the SSD problem satisfies

$$\mu(\mathbf{x}_0) = \frac{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x} + \varepsilon(\mathbf{x}))]^2 \varepsilon(\mathbf{x}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u_x(\mathbf{x} + \varepsilon(\mathbf{x}))]^2 d\mathbf{x}} \tag{20}$$

up to the noise terms. In other terms, the obtained minimizer will be an center of mass of the disparities at each pixel in the correlation window, each being weighted by its squared image gradient.

This explains the fattening effect, which actually occurs at *every pixel*: Whenever a pixel or a cluster of pixels have a large gradient with respect to their neighboring ones, the estimated disparity for these neighboring pixels will be obtained by combining mainly the disparities of these few very contrasted pixels. It can even happen that a single pixel dominates the estimated disparity for all of its neighboring ones. This effect is strong in all textures and also near image edges, where a line of pixels dominates the SSD of all their neighboring ones. This case causes the so called *foreground fattening* phenomenon by which buildings looks fatter than they really are. Yet, the fattening effect happens everywhere, because a gradient barycenter is never exactly the center of the correlation window. Even if this is not very noticeable when looking at the disparity image, this effect becomes conspicuous when looking at the 3D reconstruction of the estimated depth, (Fig. 4).

The above calculations show that there is only one way to avoid the fattening: It is to remove the disparity imbalance in the comparison window. One can compensate the effect of the squared gradients in the above integral by directly modifying the values of the window function φ, making it adaptive. By taking $\varphi_{x_0}(\mathbf{x}) = \frac{\rho_{\mathbf{x}_0}(\mathbf{x})}{u_x(\mathbf{x}+\varepsilon(\mathbf{x}))^2}$ in equation (20) we obtain

$$\mu(\mathbf{x}_0) = \frac{\int_{\rho_{\mathbf{x}_0}} \varepsilon(\mathbf{x}) d\mathbf{x}}{\int_{\rho_{\mathbf{x}_0}} d\mathbf{x}}, \tag{21}$$

which is equivalent to

$$\mu(\mathbf{x}_0) = \int \rho(\mathbf{x} - \mathbf{x}_0) \varepsilon(\mathbf{x}) d\mathbf{x},$$

since the function ρ is normalized to have the integral equal to one. In that way the disparity becomes a weighted average of all disparities in the correlation neighborhood, which is no more weighted by the image gradient. Therefore, the computed disparity is the convolution of the ground truth disparity ε with a kernel, which can incidently be fixed at will. The most natural choice for the window ρ is an isotropic kernel, for example a Gaussian G_a. If we select such a kernel, the computed disparity writes $G_a * \varepsilon$, which can be interpolated and could even be deconvolved to some extent. The choice of the size of the window depends primarily on the noise variance. If there were no noise at all the window could be a Dirac. In presence of noise, the dominant disparity error term due to the noise given by Theorem 1 rewrites

$$\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) = 2\sigma^2 \int \frac{\rho(\mathbf{x}-\mathbf{x}_0)^2}{u_x(\mathbf{x}+\varepsilon(x))^2} d\mathbf{x}. \tag{22}$$

Thus, the size of the window must be large enough to ensure this value to be low enough to compensate for σ^2. Indeed, the integral of ρ being 1, the broader the support of ρ the smaller the integral will be, because of the presence of the ρ^2 term. This implies that the integral behaves like $1/n$, where n is the number of pixels in the window. A good point of the above result is that adaptive window can be larger without causing a fattening effect.

The discrete implementation of such an algorithm faces the problem of computing the true derivatives $u_x(\mathbf{x}+\varepsilon(\mathbf{x}))$ from the two available images u_1 and u_2. We can compute the derivative on the first image, obtaining

$$u_1'(\mathbf{x})^2 = (u'(\mathbf{x}+\varepsilon(\mathbf{x}))(1+\varepsilon'(\mathbf{x})) + n_1'(\mathbf{x}))^2.$$

Since this is a stochastic term, the right choice must be indicated by its mean

$$Eu_1'(\mathbf{x})^2 = u'(\mathbf{x}+\varepsilon(\mathbf{x}))^2(1+\varepsilon'(\mathbf{x}))^2 + 2\sigma^2.$$

This identity shows that, because of the noise term, we will be only able to compute the actual derivatives if and when $\varepsilon'(\mathbf{x})$ is small. We shall make this assumption, which means that the relief is smooth. In order to avoid too small gradients due mainly to noise, we shall use the following weighting function

$$\varphi_{\mathbf{x}_0}(\mathbf{x}) = \frac{\rho_{\mathbf{x}_0}(\mathbf{x})}{\max(u_x(\mathbf{x}+\varepsilon(\mathbf{x}))^2, 6\sigma^2)},$$

where σ is the noise standard deviation.

4 Comparative experiments

In order to illustrate and compare the performance of the classical SSD strategy and the proposed adaptive algorithm, several tests were performed on synthetic and real stereo pairs, and the proposed method was compared with the two most classic fattening correction strategies.

The first experiments were simulated pairs with a smooth disparity function. The disparity ε in Fig. 2 was applied to the reference texture images u of Fig. 1. Each image was warped by ε to obtain the image pair. Gaussian white noise was added to both images of the pair. Texture images were used to make sure that around each pixel there was enough information to permit its correct matching. The first ground truth disparity varies slowly and smoothly while the other two are more oscillatory.

Fig. 3 presents the disparity maps obtained by both strategies for the first image of the data base. In this case, a noise with standard deviation 1 has been added, yielding a signal to noise ratio of about one hundred. The results with SSD and with the proposed strategy are shown with prolate functions supported by 7×7 and 11×11 pixels. Observe that the disparity obtained with the proposed strategy is more similar to the ground truth than the classical SSD algorithm. This improvement is conspicuous when the 11×11 prolate is

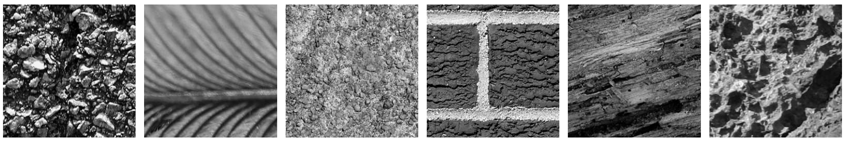


Figure 1: Reference image warped by a known disparity to obtain an image pair.

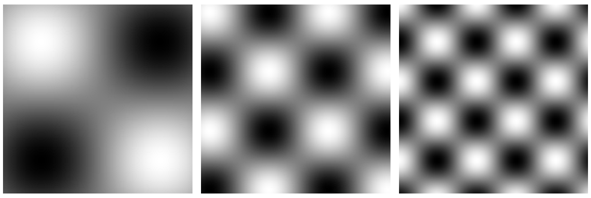


Figure 2: Ground truth disparities applied to images in Fig. 1.

used or when the disparity map is more oscillatory. This experimental fact is in agreement with the mathematical arguments and formulas developed in the previous section. The obtained disparity for the classical SSD strategy depends on the true disparity on the 7×7 or 11×11 neighborhood and is weighted by the square of the gradient. Thus, with a larger window the probability of having large gradients on the window is increased and the favored disparity by these large gradient points can be more different than the one of the reference pixel.

In Fig. 4 are displayed the three-dimensional representations of the central row in Fig. 3 with a 7×7 prolate function. One better evaluates with this representation the difference between the classical and the adaptive SSD. The surface obtained by the adaptive SSD is smooth and very similar to the ground truth. However, the surface by the classical SSD strategy presents many irregularities due to its dependence on the image gradients.

Table 1 shows the average Euclidean distance between the obtained disparity and the ground truth for the six images in Fig. 1. The error values are very similar when the prolate is small or when the disparity varies slowly, while they increase for the classical SSD algorithm when a larger prolate or an oscillating ground truth is applied. Table 2 shows the error committed by comparing the true normals to the surface of the ground truth with the normals to the surfaces of the obtained disparities. Are shown the ratio of points of the surface for which the normal has an error of more than 10 degrees with respect to the original normal. The accuracy gain is quite important by using the adaptive strategy. Notice that the distance of normals is the right measure to estimate how two renderings of the same object differ visually. Indeed, most 3D visualizations are done by a Lambertian model. The grey level of the rendered image is the scalar product of the surface normal with the solar direction. Thus the above error measure is the right one to estimate the visual gain.

It is observed in Table 2 that with a small correlation window the use of the adaptive strategy is more sensitive to noise. This is not easily explained by comparing the precision terms in Theorem 1, but it can be explained by simple probabilistic arguments. When computing the weighted Euclidean distance of two noisy patches, the influence of noise on the distance is proportional to

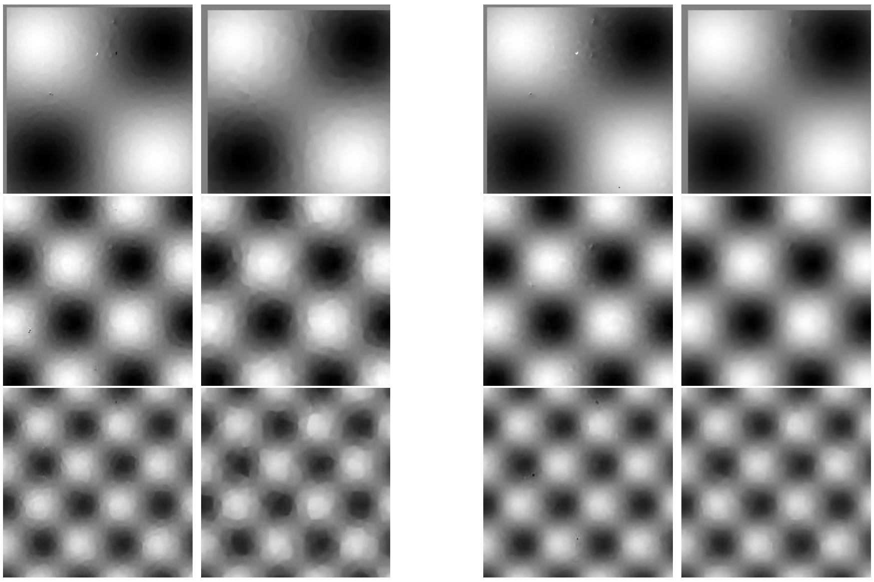


Figure 3: Obtained disparities for the first image in Fig. 1 and the three ground truth disparities in Fig. 2. The left column shows the disparities obtained with a classical SSD algorithm with an isotropic weighting window of size 7×7 and 11×11. In the right column same experiments, but with the proposed algorithm.

the energy of the window weight distribution. This influence is minimal when using a flat window or similarly an isotropic kernel. When using the proposed adaptive kernel, the weight of large gradient points is reduced and the weight of non gradient points increased. This makes the window weighting less uniform. This noise sensitivity is reduced by increasing the window size, as shown in the same table.

The next experiment was performed with a synthetic disparity map applied to a building image. The background has uniform disparity but the building has a sloped roof. Since the background has uniform disparity, we can only observe the fattening effect in and near the building. The ground truth disparity and the simulated image pair are shown in Fig. 5. Fig. 6 shows the estimated disparities with the classical SSD algorithm and with the proposed adaptive SSD, using again prolate windows of 7×7 and 11×11 pixels. The same figure shows the error image, namely the difference between the estimated disparities and the ground truth. With the proposed strategy the obtained image difference stands between the estimated disparities and the convolved ground truth by the same prolate. This is consistent with the formulation in the previous section, where we showed that the adaptive SSD estimates a convolved disparity, independent of the gradient of the image. For the SSD algorithm, we observe a prominent error near the boundaries of the building, while for the proposed strategy this error passed unnoticed.

The next experiment displays a more complicated case with occlusion and shadows containing nearly no information. Fig. 7 shows the image pair and its ground truth. In Fig. 8 are displayed the estimated disparities and the error image difference between the estimated disparities and the ground truth.

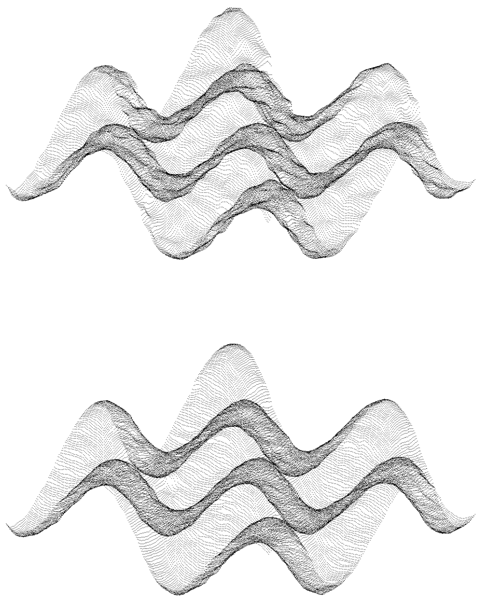


Figure 4: Three dimensional representation of the estimated disparity from the middle row of Fig. 3. Top: estimated disparity by SSD with a 7×7 correlation window. Bottom: proposed adaptive SSD with the same 7×7 correlation window. The fattening effect creates evident irregularities in the reconstructed surface.

7×7	$\sigma = 0.0$	$\sigma = 1.0$ (SNR=100)	$\sigma = 2.0$ (SNR=50)
SSD	0.118	0.121	0.138
Proposed	0.108	0.113	0.139

11×11	$\sigma = 0.0$	$\sigma = 1.0$ (SNR=100)	$\sigma = 2.0$ (SNR=50)
SSD	0.135	0.136	0.139
Proposed	0.107	0.109	0.116

Table 1: Average error on the disparity computation on the six images of Fig. 1 and the middle ground truth of Fig. 2. For the proposed method the distance is computed to the convolved ground truth as predicted by the formulas. The first table is obtained by using a correlation window of 7×7 pixels while the second table is obtained by using a correlation prolate of size 11×11. We observe that the SSD error increases when using a larger window. By using a larger window the ground truth disparity varies more and the possibility of having a large gradient increases, therefore making SSD more sensitive to adhesion. The obtained errors are quite similar for both algorithms, showing that the use of an adaptive SSD does not diminish the precision of SSD.

7×7	$\sigma = 0.0$	$\sigma = 1.0$ (SNR=100)	$\sigma = 2.0$ (SNR=50)
SSD	0.35	0.54	1.27
Proposed	0.04	0.20	1.26

11×11	$\sigma = 0.0$	$\sigma = 1.0$ (SNR=100)	$\sigma = 2.0$ (SNR=50)
SSD	0.48	0.50	0.64
Proposed	0.01	0.01	0.11

Table 2: Average on the six images of Fig. 1 and the middle ground truth of Fig. 2 of the percentage of points with an angular difference of the surface normal to the ground truth normal larger than 10 degrees. For the proposed method the distance is computed to the convolved ground truth as predicted by the formulas. The first table is obtained by using a correlation window of 7×7 pixels while the second table is obtained by using a correlation prolate of size 11×11. Observe that with a larger correlation window a surface more similar to the original one is obtained. This result is notable: the obtained percentage of points with a very different normal to the surface is much higher for the classical SSD than the proposed algorithm.

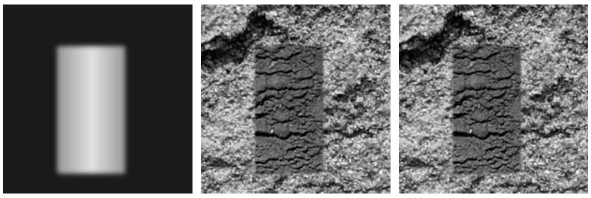


Figure 5: Synthetic image pair. Left: the disparity ground truth, the background has uniform disparity while the building simulates the slope of a roof. Center and right: image pair.

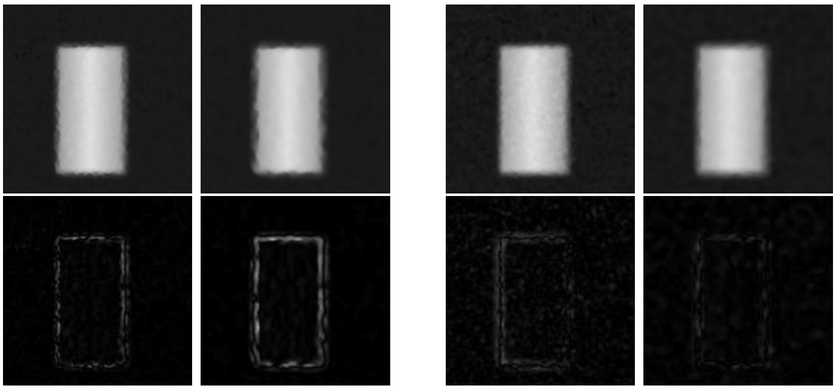


Figure 6: Obtained disparities for the synthetic image pair in Fig. 5. The top left columns display the disparities obtained with a classical SSD algorithm with an isotropic weighting window of size 7×7 and 11×11. The top right columns show the same experiments but with the proposed algorithm. Bottom: image difference between the estimated disparities and the ground truth. For the proposed strategy the displayed image difference stands between the estimated disparities and the convolved ground truth by the same prolate.

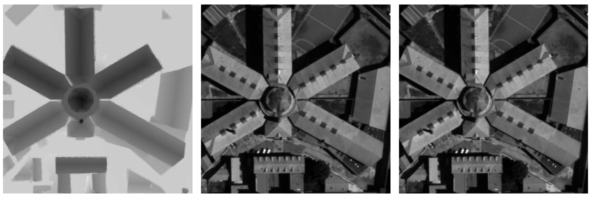
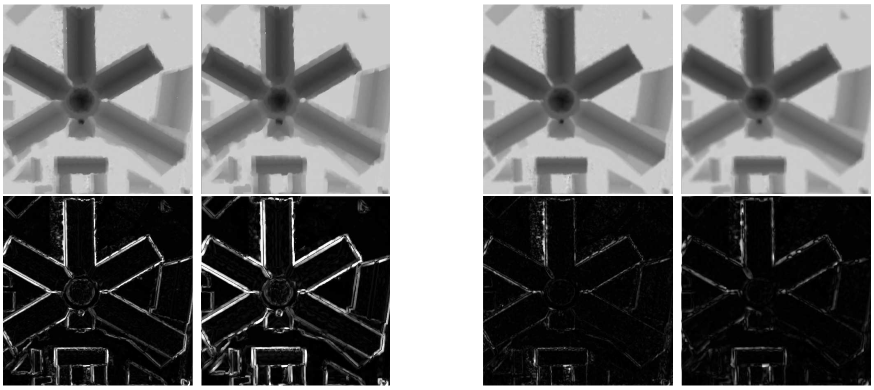


Figure 7: Synthetic image pair. Left: the disparity ground truth. The background has uniform disparity while the building has a sloped roof. Center and right: image pair.



a) Classical SSD

b) Adaptive SSD

Figure 8: Estimated disparities with the classical SSD (a) and adaptive SSD strategy (b) for the synthetic image pair in Fig. 5. Top: disparities obtained with a weighting window of size 7×7 and 11×11. Bottom: image difference between the estimated disparities and ground truth.

For the proposed strategy the image difference stands again between the estimated disparities and the convolved ground truth by the same prolate. Observe that the error is mainly concentrated near the edges of the building, where the foreground fattening effect is severe. Although in the synthetic case of Fig 6 we were able to nearly eliminate the error near the edges with the proposed strategy, this is not the case for this pair. The error committed by the SSD algorithm is reduced but not eliminated. This is due to the occlusions which make ε discontinuous, and to the fact that near most of the building boundaries the shadow has removed all possible information that could be used to correct the match. Surprisingly, the error is much smaller at non shadowed edges, even if occlusions and discontinuities of the disparity are still present.

4.1 Comparison with foreground fattening elimination strategies

As exposed in the introduction, many strategies have been proposed to remove the fattening effect and are beautifully reviewed and compared in [19]. Our goal now is to compare the proposed strategy with two of the more performing algorithms. Yoon et al. [25] selects an adaptive window containing only pixels

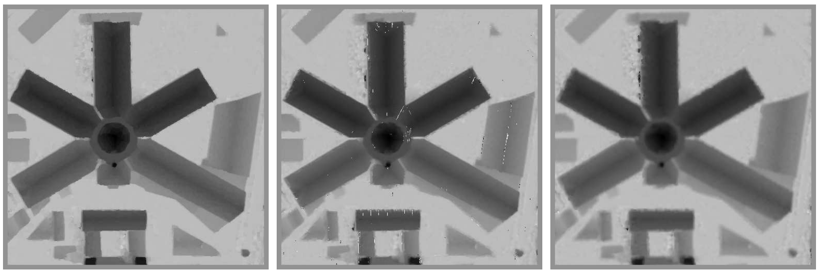


Figure 9: Estimated disparity on stereo pair in Fig. 7. From left to right: Fusiello et al [5] min-filter, Yoon et al [25] bilateral strategy, and the proposed adaptive window strategy. The three estimated disparities remove the dilatation of buildings due to the fattening effect. The estimated disparity by RAFA is more blurred than the other two, because the recovered disparity is a convolution of the original one with the correlation window.

with a grey level similar to the reference one, in the spirit of bilateral filters [21]. The main idea is to keep in the correlation window only points belonging to the same object, which are likely to have a similar grey level. The Fusiello et al. [5] classic *min-filter* chooses among all the windows containing the reference pixel the one which has a minimal distance with its corresponding one in the second image.

Fig. 9 compares the adaptive strategy with these two algorithms on the pair of Fig. 7. The three estimated disparities remove the dilatation of buildings due to the fattening effect. But the estimated disparity by RAFA is more blurred than the other two, since the recovered disparity is by our theorem a convolution of the original one with the correlation window.

In order to evaluate the subpixel precision of the three methods, we applied the algorithms to the first texture image in Fig. 1 and the second simulated ground truth disparities in Fig. 2. Fig. 11 displays the estimated disparity for the three algorithms. It is observed that the disparity estimated by the min-filter produces a shock effect which creates discontinuities of the estimated disparity. These shocks are not present when using the adaptive window of Yoon et al.. However, many irregularities are present in the estimated disparity, which are similar to the ones obtained by the classical SSD. The estimated disparity by RAFA algorithm is more similar to the ground truth. This can also be observed by looking at the 3D representation of the estimated disparities by the three algorithms.

The conclusion of these comparisons is that classical fattening removal techniques work correctly when for *foreground fattening* due to the presence of important disparity discontinuities. Nevertheless, as was pointed out, fattening occurs everywhere, even in the absence of strong depth discontinuities. In that case classical techniques do not attain the optimal precision obtained with the adaptive window technique. This might be due to the fact that most existing databases for stereo benchmarks furnish only a pixel precise ground truth. Therefore the benchmarks do not permit to detect or to evaluate the loss of precision due to the general surface fattening. To demonstrate this fact, Fig. 13 shows a detail of a stereo pair of the Middlebury dataset [18]. Are shown the estimated disparities by the classical SSD and by the adaptive RAFA. Both

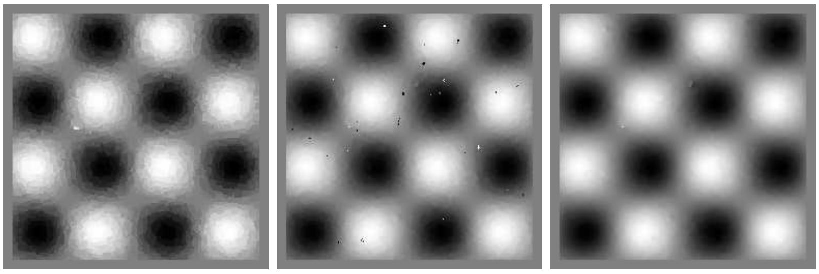


Figure 10: Estimated disparity on the first texture image in Fig. 1. Independent additive white noise with standard deviation 2 was added to both images before matching. From left to right: the min-filter Fusiello et al [5], the bilateral Yoon et al [25], and the proposed adaptive window strategy. The RMSE (in pixels) are respectively 0.18, 0.19, and 0.11.

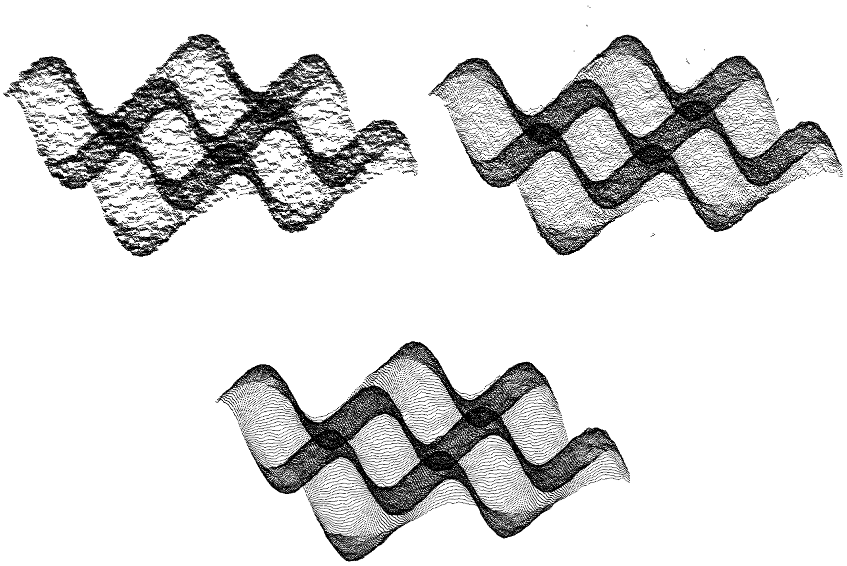


Figure 11: Three dimensional representation of the estimated disparity from Fig. 11. As predicted by the theorem, the fattening effect is optimally removed by the adaptive window.

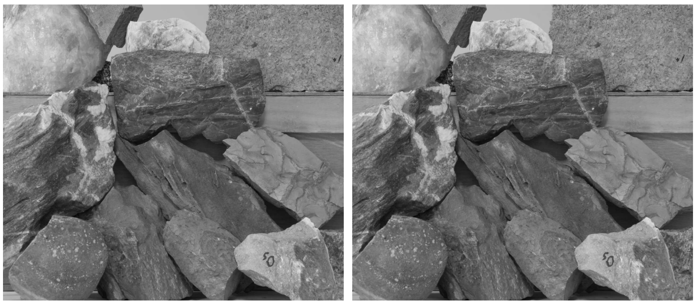


Figure 12: Stereo pair obtained from Middlebury dataset [18].

disparities present many irregularities probably present in the object (a rough stone). However, the smoothness of the furnished ground truth did not allow for a numerical comparison of the two estimated disparities.

5 Conclusion

This paper has shown that in block matching methods the fattening phenomenon occurs everywhere. A mathematical analysis has proved that fattening could be completely avoided in the regions with smoothly varying disparity, by introducing adaptive weights in the SSD block matching. Experimental evidence on simulated data has been provided to confirm that fattening is indeed avoided with the adaptive window. Yet the adaptive window does not correct the strong foreground fattening, particularly annoying near large building edges in aerial imaging. However, the adaptive window promises to be a valuable and simple correction the fixed windows used widely in block matching method. Future work will consider how to insert this correction in a complete stereo reconstruction chain.

References

[1] A. Buades, B. Coll, JM Morel, and B. Rouge. Procedimiento de establecimiento de correspondencia entre una primera imagen digital y una segunda imagen digital de una misma escena para la obtencion de disparidades. *Spanish Patent, Reference P25155ES00, UIB*, 2009.

[2] F. Cao. *A theory of shape identification.* Springer Verlag, 2008.

[3] F. Cao, J. Delon, A. Desolneux, P. Muse, and F. Sur. A unified framework for detecting groups and application to shape recognition. *Journal of Mathematical Imaging and Vision*, 27(2):91–119, 2007.

[4] J. Delon and B. Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007.

[5] A. Fusiello, V. Roberto, and E. Trucco. Symmetric stereo with multiple windowing. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(8):1053–1066, 2000.

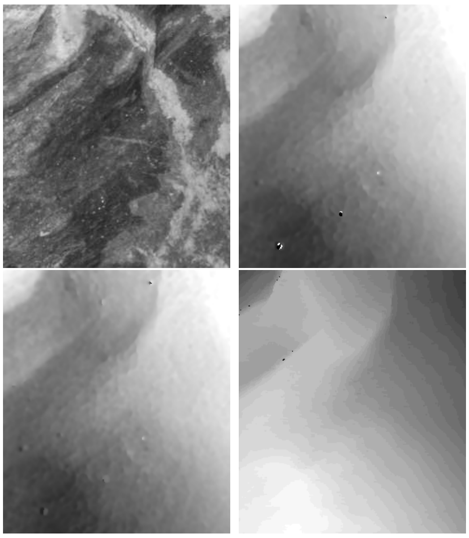


Figure 13: Obtained disparities by classical SSD and RAFA on a piece of Fig. 13. From top to bottom and left to right: detail in the reference image, classical SSD disparity, adaptive RAFA disparity and ground truth furnished in the same dataset [18]. The ground truth furnished is too smooth and quantized. Fattening effects cannot be compared on it.

[6] H. Hirschmuller, P.R. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246, 2002.

[7] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994.

[8] R. Kimmel, C. Zhang, A. M. Bronstein, and M. M. Bronstein. Are mser features really interesting? *IEEE Pami, in press*, 2010.

[9] J. Lotti and G. Giraudon. Correlation algorithm with adaptive window for aerial image in stereo vision. *In Image and Signal Processing for Remote Sensing*, 1:2315–10, 1994.

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

[11] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1):43–72, 2005.

[13] J.M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[14] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3):295–315, 2006.

[15] M.P. Patricio, F. Cabestaing, O. Colot, and P. Bonnet. A similarity-based adaptive neighborhood method for correlation-based stereo matching. In *International Conference on Image Processing*, volume 2, pages 1341–1344, 2004.

[16] L. Robert and O.D. Faugeras. Curve-based stereo: figural continuity and curvature. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1991.

[17] N. Sabater. *Reliability and accuracy in stereovision. Application to aerial and satellite high resolution images.* Ph.D. thesis, ENS Cachan, December 2009.

[18] D. Scharstein and R. Szeliski. Middlebury stereo vision page. *Online at http://www. middlebury. edu/stereo*, 2002.

[19] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(47(1/2/3)):7–42, 2002.

[20] C. Schmid and A. Zisserman. The geometry and matching of lines and curves over multiple views. *International Journal of Computer Vision*, 40(3):199–234, 2000.

[21] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, volume 846. Citeseer, 1998.

[22] O. Veksler. Fast variable window for stereo correspondence using integral images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:556–561, 2003.

[23] Liang Wang, Miao Liao, Minglun Gong, Ruigang Yang, and David Nister. High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 798–805, 2006.

[24] L. Yaroslavsky and M. Eden. Fundamentals of digital optics, 2003.

[25] S. Yoon, K.-J.and Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006.