Binocular stereo pipeline

Andres Almansa, Neus Sabater, Pascal Monasse, Jean-Michel Morel

November 22, 2010

Abstract

From a pair of images of a scene, the stereo pipeline recovers the 3D geometry by a succession of algorithms. This geometry is concretely a disparity map, associating to pixels their apparent motion (called disparity) between both images. This amount is inversely proportional to the depth of a 3D point. Finding correspondences between images is the fundamental problem that the stereo pipeline addresses. The most popular method is based on block matching: each pixel is coded by a neighborhood and the best correlation of such a block is searched in the other image. We describe two complementary filters that eliminate ambiguous matches: one is based on *a contrario* methodology and the other one rejects blocks with self-similarity. Finally, we discuss an optimal refinement that computes sub-pixel disparity, giving better accuracy to the recovered 3D geometry.

1 Preliminaries

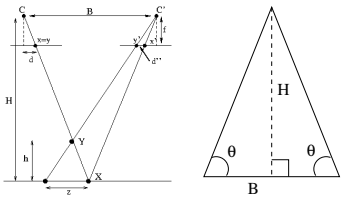
We assume throughout this document that images are in the rectified situation, meaning that the apparent motion of pixels is horizontal everywhere. As has been seen, there is always the possibility to simulate this situation with no extra information necessary: this is the task of the epipolar rectification process.

1.1 The B/H dilemma

The apparent motion of a point in the image due to camera displacement is illustrated in Fig. 1(a): estimating the altitude h of the 3D point y relative to the ground amounts to measuring the shift d'' w.r.t. the projection of the point X (on the ground in the same light ray as Y in first camera) in the right image. Besides, all points on the ground have the same apparent motion $x' - x = f\frac{B}{H}$, if f is the focal length expressed in pixels. We see that by recovering the apparent motion of the 3D point Y, we can estimate its altitude by triangulation. We have:

$$\Delta x := x' - x = f\frac{B}{H} \tag{1}$$

$$\Delta y := y' - y = f\frac{B}{H-h} \tag{2}$$



(a) (b)

Figure 1: (a) Apparent motion due to camera displacement. (b) Triangulation formula: $H = \frac{B}{2} \tan\theta$.

from which we get

$$h = fB(\frac{1}{\Delta x} - \frac{1}{\Delta y}) = H - f\frac{B}{\Delta y}. \tag{3}$$

We see that the depth $H - h$ of point Y is inversely proportional to Δy, called the disparity of Y. The *disparity map* is the image whose intensity at a pixel is its disparity (provided it can be computed, which implies in particular that the 3D point is visible in both images).

Assuming h is small with respect to H, we have at first order:

$$h \sim \frac{H}{\Delta x} d''. \tag{4}$$

The displacement is all the more important that the point is closer to the camera. This is the familiar experience when looking through the window of a train in motion: hills in the distance seem to move slowly while mid-distance trees move faster.

Fig. 1(b) illustrates a simple case (isocele) of triangulation. The optical centers of the two views are the endpoints of the base segment of length B. The formula yielding H as a function of θ is

$$H(\theta) = \frac{B}{2} \tan\theta.$$

Deriving this with respect to θ we get

$$H'(\theta) = \frac{B}{2}(1 + \tan^2\theta) = \frac{B}{2}\left(1 + 4\left(\frac{H}{B}\right)^2\right).$$

We see that as B/H increases linearly, this derivative increases quadratically. Therefore, the precision on H is better when B/H is large. However, in that case, the images look really different with occlusions, stronger perspective distortions, etc. Therefore the correspondence problem is all the more difficult that B/H is large. That is why we prefer a small B/H (for example 0.1) and recover a good precision through sub-pixel matching.

1.2 Two classes of methods

Two main classes exist to estimate the disparity map:

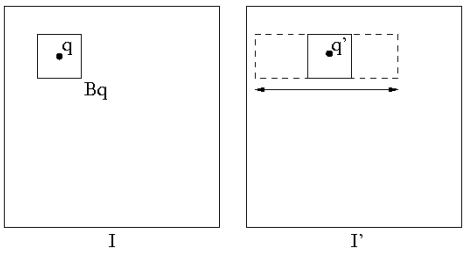


Figure 2: The search for the best correlation. To the pixel $\mathbf{q}$ in image I is associated a square window $B_{\mathbf{q}}$ and this window is translated horizontally in image I' in search of the best correlation (typically lowest L^2 distance).

1. global methods try to estimate globally the map by minimizing an energy with a term penalizing image intensity mismatch plus some prior model of regularity of the map. This energy is typically minimized through dynamic programming or graph cuts algorithms.
2. local methods estimate independently the apparent motion of a point by considering a neighborhood around it (typically a square window) and finding the best motion for this neighborhood in the second image.

Global methods require an *a priori* regularity assumption about the disparity map. Finding such a realistic prior is very difficult, and we will not consider this class of methods.

Local methods encode each point by a neighborhood (fixed or variable) and search for the best horizontal translation of this neighborhood in the other image, with respect for example to an L^2 norm. The simplest is to use some square (for example 9×9) centered around the pixel, as illustrated in Fig. 2. The best translation found is the estimated disparity for the pixel.

1.3 The adhesion effect

The major drawback of local methods is the adhesion effect, due to the fixed window used as neighborhood. This happens when such a window overlaps two objects with different disparities. Such a depth discontinuity is often coupled with an intensity discontinuity. The correlation will tend to align the discontinuities in the images, therefore assigning to window center one of the disparities present in the windows. What we get is actually a dilation of the highest disparities equal to the radius of the window.

In what follows, we do not try to correct the adhesion artifact. Actually, pixels near edges could be simply masked from the disparity map since they are at risk of suffering from adhesion. We choose not doing that, to show really what our filters provide.

2 A contrario block matching

2.1 Background model for patch comparison

Our goal can be formulated in one single question, that clearly depends on the observed set of patches in one particular image and not on the probability space of *all* patches. The question is:

> What is the probability that given two images and two patches in these images, their similarity arises just by chance?

The "just by chance" implies the existence of a stochastic *background model*, often called *a contrario* model.

There is an interesting simplification in *a contrario* models with respect to classic Bayesian ones. In the Bayes model, a model of the set of patches (the background model) would be required, but also a model of the patch itself. The H_1 alternative would be that patches 1 and 2 arise from the same patch model, and the H_0 or null alternative would be that a patch similarity like the one observed by patches 1 and 2 is likely to happen in the background model. A bayesian algorithm would then choose for each patch the hypothesis (either H_0 or H_1) with the higher probability. In the simpler *a contrario* framework, the decision is based solely on the probability under H_0, which is intuitively the probability of a match occurring just by chance. If this probability is small enough the background model (H_0 hypothesis) is rejected, and this is enough to claim that a meaningful match has been found. Thus, in the *a contrario* setting, the background model is enough to gain a strict control of the number of wrong matches, as Thm. 1 will show, and experimental evidence will confirm.

When trying to define a well suited model for image blocks, many possibilities open up. Simple arguments show, however, that over-simplified models do not work. Let H be the gray-level histogram of the second image I'. The simplest *a contrario model* of all might simply assume that the observed values $I'(x)$ are instances of i.i.d. random variables $\mathcal{I}'(x)$ with cumulative distribution H. This would lead to declare that pixels q in image I and q' in image I' are a meaningful match if their gray level difference is unlikely small,

$$\mathbb{P}[|I(q) - \mathcal{I}'(q')| \leq |I(q) - I'(q')| := \theta] \leq \frac{1}{N_{tests}}.$$

As we shall see later, the number of tests N_{tests} is quite large in this case ($N_{tests} \approx 10^7$ for typical image sizes), since it must consider all possible pairs of pixels (q, q') that may match. But such a small probability can be achieved (assume that H is uniform over $[0, 256]$) only if the threshold $\theta = |I(q) - I'(q')| < 128 \cdot 10^-7$. On the other hand, $|I(q) - I'(q')|$ cannot be expected to be very small because both images are corrupted by noise, among other distortions. Even in a very optimistic setting, where there would be only a small noise distortion between both images (of about 1 gray level standard deviation), such a small difference would only happen for about a tiny proportion ($3.2 * 10^{-5}$) of the correct matches.

This means that a pixel-wise comparison would require an extremely strict detection threshold to ensure the absence of false matches, but this

leads to an extremely sparse detection (about thirty meaningful matches per mega-pixel image). This suggests that the use of local information around the pixel is unavoidable. The next simplest way could be to compare blocks of a certain size $\sqrt{s} \times \sqrt{s}$ with the usual ℓ^2 norm, and with the same background model as before. Thus, we could declare blocks B_q and $B_{q'}$ as meaningfully similar if

$$\mathbb{P}\left[\frac{1}{|B_0|}\sum_{x\in B_0}|I(q+x)-\mathcal{I}'(q'+x)|^2 \leq \right.$$
$$\left. \frac{1}{|B_0|}\sum_{x\in B_0}|I(q+x)-I'(q'+x)|^2 := \theta\right] \leq \frac{1}{N_{tests}} \quad (5)$$

Now the test would be passed for a more reasonable threshold ($\theta = 6, 28, 47$ for blocks of size 3×3, 5×5, 7×7 respectively), which would ensure a much denser response. However, this *a contrario* model is by far too naive, and produces many false matches. Indeed, blocks stemming from natural images are much more regular than the white noise generated by the background model. Considering all pixels in a block as independent leads to overestimating the similarity probability of two observed similar blocks. It therefore leads to an over-detection.

In order to fix this problem, we need a background model that actually reflects the statistics of natural image blocks. But directly learning such a probability distribution from a single image in dimension 81 (for 9×9 blocks) is hopeless.

Fortunately, shape high-dimensional distributions can be approximated by the tensor product of their adequately chosen marginal distributions. Such marginal laws, being one-dimensional, are more easily learned from a single image. Ideally, ICA should be used to learn which marginal laws are the most independent, but the simpler PCA will show accurate enough for our purposes. Indeed, it ensures that the principal components are decorrelated, a first approximation to independence. Fig. 4 gives a visual assessment of how well a local PCA model simulates image patches in a class.

3 The *a contrario* Model for Block-Matching

We shall denote by $\mathbf{q}$=(q_1, q_2) a pixel in the reference image I and by $B_{\mathbf{q}}$ a block centered at $\mathbf{q}$. To fix ideas, the block will be a square throughout this paper, but this is by no means a restriction. A different shape (rectangle, disk) would possible, and even a variable shape. Given a point $\mathbf{q}$ and its block $B_{\mathbf{q}}$ in the reference image, block-matching algorithms look for a point $\mathbf{q}'$ in the second image I' whose block $B_{\mathbf{q}'}$ is similar to $B_{\mathbf{q}}$.

3.1 Principal Component Analysis

For building a simple *a contrario* model the principal component analysis can play a crucial role. Indeed, it allows a strong dimension reduction

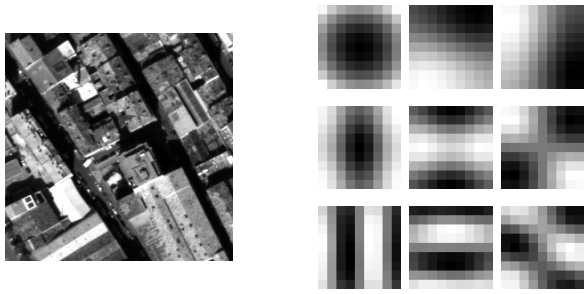


Figure 3: Left: Reference image of a stereo pair of images. Right: the nine first principal components of the 9×9 blocks.

and decorrelates these dimensions, giving a first approximation to independence. This permits to build up a probabilistic density function for the blocks as a tensor product of its marginal densities.

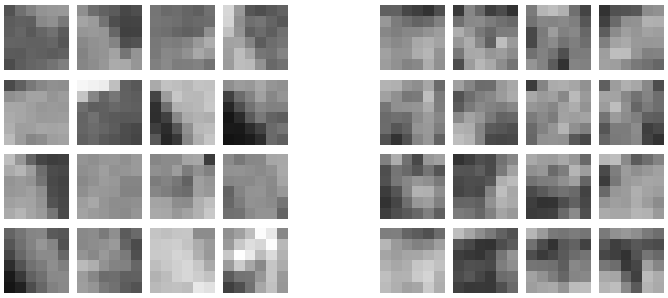
Let $B_{\mathbf{q}}$ be the block of a pixel $\mathbf{q}$ in the reference image and $(x_1^{\mathbf{q}}, \ldots, x_s^{\mathbf{q}})$ the intensity gray levels in $B_{\mathbf{q}}$, where s is the number of pixels in $B_{\mathbf{q}}$. Let n be the number of pixels in the image. Consider the matrix $X = (x_i^j)$ $1 \leq i \leq s,\ 1 \leq j \leq n$ consisting of the set of all data vectors, one column per pixel in the image. Then, the covariance matrix is $C = \mathrm{Cov}(X) = \mathbb{E}(X - \bar{x}\mathbf{1})(X - \bar{x}\mathbf{1})^T$, where $\bar{x}$ is the column vector of size $s \times 1$ storing the mean values of matrix X and $\mathbf{1} = (1, \cdots, 1)$ a row vector of size $1 \times n$. Notice that $\bar{x}$ corresponds to the block whose k-th pixel is the average of all k-th pixels of all blocks in the image. Thus, $\bar{x}$ is very close to a constant block, with the constant equal to the image average. The eigenvectors of the covariance matrix are called principal components and are orthogonal. They give the new coordinate system we shall use for blocks. Fig. 3 shows the first principal blocks.

Usually, the eigenvectors are sorted in order of decreasing eigenvalue. In that way the first principal components are the ones that contribute most to the variance of the data set. By keeping the first $N < s$ components with larger eigenvalues, the dimension is reduced but the significant information retained. While this global ordering could be used to select the main components, a local ordering for each block will instead be used for the statistical matching rule. In other words, for each block, a new order for the principal components will be established given by the corresponding ordered PCA coordinates (the decreasing order is for the absolute values). In that way, comparisons of these components will be made from the most meaningful to the least meaningful one for this particular block.

Now each block is represented by N ordered coefficients

$$(c_{\sigma_{\mathbf{q}}(1)}(\mathbf{q}), \ldots, c_{\sigma_{\mathbf{q}}(N)}(\mathbf{q})),$$

where $c_i(\mathbf{q})$ is the resulting coefficient after projecting $B_{\mathbf{q}}$ onto the principal component $i \in \{1, \ldots, s\}$ and $\sigma_{\mathbf{q}}$ the permutation representing the final order when ordering the absolute values of components for this particular $\mathbf{q}$ in decreasing order. By a slight abuse of notation we will write



(a) (b)

Figure 4: (a) Patches of the reference image, chosen at random. (b) Simulated random blocks following the law of the reference image. This experiment illustrates the adequacy of the *a contrario* model.

$c_i(\mathbf{q})$ instead of $c_{\sigma_\mathbf{q}(i)}(\mathbf{q})$ knowing that it represents the local order of the best principal components. But notice that $\sigma_\mathbf{q}(1) = 1$ for most $\mathbf{q}$ because of the dominance of the first principal component. Moreover notice that this first component has a quite different coefficient histogram than the other ones (see Fig. 6), because it approximately computes a mean value of the block. Indeed, the barycenter of all blocks is roughly a constant block whose average grey value is the image average grey level. The set of blocks is elongated in the direction of the average grey level and, therefore, the first component computes roughly an average grey level of the block. This explains why the first component histogram is similar to the image histogram.

3.2 *A Contrario* similarity measure between blocks

Definition 1. *We call a contrario block model associated with a reference image a random block* $\mathbf{B}$ *described by its (random) components* $\mathbf{B} = (\mathbf{c}_1, \ldots, \mathbf{c}_s)$ *on the PCA basis of the blocks of the reference image, satisfying*

- *the components* $\mathbf{c}_i$*,* $i = 1, \ldots, s$ *are independent random variables;*
- *for each* i*, the law of* $\mathbf{c}_i$ *is the empirical histogram of the* i*-th PCA component* $c_i(\cdot)$ *of the blocks of the reference image.*

The reference image will be the secondary image I'. Fig. 4 shows patches generated according to the above *a contrario* block model and compares them to blocks picked at random in the reference image. The *a contrario* model will be used for computing a block resemblance probability as the product of the marginal resemblance probabilities of the $\mathbf{c}_i$ in the *a contrario* model, which is justified by the independence of $\mathbf{c}_i$ and $\mathbf{c}_j$ for $i \neq j$. There is a strong adequacy of the *a contrario* model to the empirical model, since the PCA transform ensures that $\mathbf{c}_i$ and $\mathbf{c}_j$ are uncorrelated for $i \neq j$, a first approximation of the independence requirement.

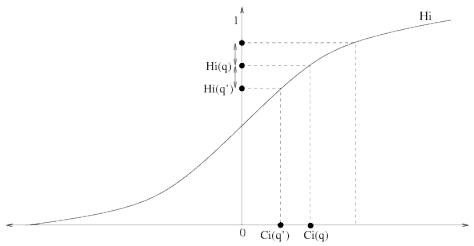


Figure 5: Normalized cumulative histogram of i-th PCA coordinates of the secondary image. $c_i(B_{\mathbf{q}})$ is the i-th PCA coordinate value in the first image. The resemblance probability for the i-th component is twice the distance $|H_i(B\mathbf{q}) - H_i(B')|$ when $H_i(B_{\mathbf{q}})$ is not too close to the values 0 or 1.

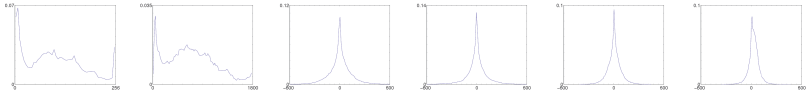


Figure 6: Histogram of the reference image, followed by the first five histograms of the block PCA coordinates. The first principal component roughly computes a mean of the block, which explains why its histogram is so similar to the image histogram.

We start by defining the resemblance probability between two blocks for a single component. Denote by $H_i(.) := H_i(c_i(.))$ the normalized cumulative histogram of the i-th PCA block component $c_i(.)$ for the secondary image I'.

Definition 2 (resemblance probability)**.** *Let $B_{\mathbf{q}}$ be a block in I and B' a block in I'. Define the probability that a random block $\mathbf{B}$ of I' be as similar to $B_{\mathbf{q}}$ for its i-th component as B' is, by*

$$\widehat{p^i}_{B_{\mathbf{q}}\,B'} = \begin{cases} H_i(B') & \text{if } H_i(B') - H_i(B_{\mathbf{q}}) > H_i(B_{\mathbf{q}}); \\ 1 - H_i(B') & \text{if } H_i(B_{\mathbf{q}}) - H_i(B') > 1 - H_i(B_{\mathbf{q}}) \\ 2|H_i(B_{\mathbf{q}}) - H_i(B')| & \text{otherwise.} \end{cases}$$

Fig. 5 illustrates how the resemblance probability is computed, and Fig 6 shows empirical marginal densities.

3.3 Robust similarity distance

The first principal components of $B_{\mathbf{q}}$, being in decreasing order, contain the relevant information on the block. Thus, if two blocks are not similar for one of the first components, they should not be matched, even if their other components are similar: we do not want a bad match for one component to be compensated by a better match for the other components.

Therefore, we will consider the worst match of the first k components:

$$\widehat{p}_{B_{\mathbf{q}}\,B'}(k) = \max_{i=1\dots k} \widehat{p^i}_{B_{\mathbf{q}}\,B'}. \tag{6}$$

We consider a quantization function π of probabilities, which maps a probability to its closest upper bound in $\Pi := \{\pi_j = 1/2^{j-1}\}_{j=1,\dots,Q}$.

The probability of having $\widehat{p}_{B_{\mathbf{q}}\,B'}(k) \leq \pi_j$ (for a given j) is that of having $\widehat{p^i}_{B_{\mathbf{q}}\,B'} \leq \pi_j$ for all i. Assuming these are independent events, we get

$$P_{B_{\mathbf{q}}\,B'}(k) = \pi\left(\widehat{p}_{B_{\mathbf{q}}\,B'}(k)\right)^k. \tag{7}$$

Since we do not know how many components k should be compared, we try a range of sensible values:

$$P_{B_{\mathbf{q}}\,B'} = \min_{k=k_{\min}\dots k_{\max}} P_{B_{\mathbf{q}}\,B'}(k). \tag{8}$$

Definition 3 (quantized probability)**.** *Let $B_{\mathbf{q}}$ be a block in I. Assume the components of the blocks are sorted in decreasing order:*

$$c_1(B_{\mathbf{q}}) \geq c_2(B_{\mathbf{q}}) \geq \cdots \geq c_{k_{\max}}(B_{\mathbf{q}}).$$

The quantized probability sequence that the random block $\mathbf{B}$ be as similar to $B_{\mathbf{q}}$ as B' is, is defined by

$$P_{B_{\mathbf{q}}\,B'} = \min_{k=k_{\min}\dots k_{\max}} \pi\left(\max_{i=1\dots k} \widehat{p^i}_{B_{\mathbf{q}}\,B'}\right)^k. \tag{9}$$

3.4 Number of tests

The number tests for comparing all the blocks is the product of three terms. The first one is the image size $\#I$. The second one is the size of the search region. We mentioned before that the search is done on the epipolar line. In practice, a segment of this line is enough. If $\mathbf{q} = (q_1, q_2)$ is the point of reference it is enough to look for $\mathbf{q}' = (q_1', q_2)$ such that $q_1' \in [q_1 - R, q_1 + R]$ where R is a fixed integer larger than the maximal possible disparity. The third factor is the number of possible probability distributions that can be envisaged. Of course all of these tests are not performed, but only the one indicated by the observed block $B_{\mathbf{q}'}$. Yet, the choice of this unique test is steered by an *a posteriori* observation, while the calculation of the expectation of the number of false alarms (NFA) must be calculated *a priori*. Thus we must compute the NFA as though all comparisons for all quantized decreasing probabilities were effectuated. This term is $Q \cdot (k_{\max} - k_{\min} + 1)$ (choice of a quantization level π_j and choice of a number k of components).

Definition 4. *With the above notation we call number of tests for matching two images I and I' the integer*

$$N_{test} = \#I \cdot \#S' \cdot Q \cdot (k_{\max} - k_{\min} + 1) \tag{10}$$

We are now in position to define a number of false alarms, which will control the overall number of false detections on the whole image.

Definition 5 (Number of false alarms)**.** *Let $B_{\mathbf{q}} \in I$ and $B_{\mathbf{q}'} \in I'$ be two observed blocks. Assume the principal components $i \in \{1, 2, \ldots, s\}$ are reordered so that $c_1(\mathbf{q}) \geq c_2(\mathbf{q}) \geq \cdots \geq c_s(\mathbf{q})$. We define the Number of False Alarms of the event*

> *"the random block $\mathbf{B}$ has components as similar to components of $B_{\mathbf{q}}$ as components of $B_{\mathbf{q}'}$ are"*

by

$$NFA_{B_{\mathbf{q}}, B_{\mathbf{q}'}} = N_{test} \cdot P_{B_{\mathbf{q}}\, B_{\mathbf{q}'}}. \tag{11}$$

Definition 6 (ϵ-meaningful match)**.** *A pair of pixels $\mathbf{q}$ and $\mathbf{q}'$ in a stereo pair (I, I') is an ϵ-meaningful match if*

$$NFA_{B_{\mathbf{q}}\, B_{\mathbf{q}'}} \leq \epsilon.$$

3.5 The main theorem

The NFA of a match actually gives a security level: the smaller the NFA, the more meaningful the match intuitively is. But Thm. 1 will give the real meaning of the NFA. To state it, we will use a clever trick used by Shannon in his information theory, namely to consider the probability of a random event as random variable. Here the NFA will become a random variable, by replacing $B_{\mathbf{q}'}$ by $\mathbf{B}$ in its definition.

In the *a contrario* model, each comparison of $B_{\mathbf{q}}$ with some $B_{\mathbf{q}'}$ is interpreted as a comparison of $B_{\mathbf{q}}$ to a trial of the random block model $\mathbf{B}$. Since $2R + 1$ comparisons are involved for each $\mathbf{q} \in I$, we are led to distinguish for each $\mathbf{q}$ $(2R + 1)$ trials which are as many i.i.d. random blocks $\mathbf{B}^{\mathbf{q},j}$, $j \in \{1, 2, \ldots 2R+1\}$, all with the same law as $\mathbf{B}$. They model *a contrario* the $(2R+1)$ trials by which $B_{\mathbf{q}}$ is matched to $(2R+1)$ blocks in I'. We are interested in the expectation of the number of such trials being successful (i.e. ε-meaningful), "just by chance".

Consider the event $E_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}}$ that a random block $\mathbf{B}^{\mathbf{q},j}$ in the *a contrario* model with reference image I' meaningfully matches $B_{\mathbf{q}}$. If this happens, it is obviously a *false alarm*. We shall denote by $\chi_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}}$ the random characteristic function associated with this event, with the convention that $\chi_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} = 1$ if $E_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}}$ is true, $\chi_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} = 0$ otherwise.

Theorem 1. *Let $\Gamma = \Sigma_{\mathbf{q} \in I, j \in \{1,\ldots,2R+1\}} \chi_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}}$ be the random variable representing the number of occurrences of an ϵ-meaningful match between a deterministic patch in the first image and a random patch in the second image. Then the expectation of Γ is smaller than ϵ.*

Proof. We have

$$\chi_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} = \begin{cases} 1, & \text{if } NFA_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} \leq \epsilon; \\ 0, & \text{if } NFA_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} > \epsilon. \end{cases}$$

Then, by the linearity of the expectation

$$\mathbb{E}[\Gamma] = \sum_{\mathbf{q},j} \mathbb{E}[\chi_{\mathbf{q},i}] = \sum_{\mathbf{q},j} \mathbb{P}\left[NFA_{B_{\mathbf{q}}, \mathbf{B}^{\mathbf{q},j}} \leqslant \epsilon\right].$$

The probability inside the above sum can be computed by Definitions 5 and 3:

$$\mathbb{P}\left[NFA_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j} \leqslant \epsilon\right] = \mathbb{P}\left[\min_k \pi(\max_{i\leq k} p^i_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j})^k \leq \frac{\epsilon}{N_{test}}\right]$$

There are many probability N-uples $p = (p^i_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j})_{i=1,\dots,N}$ permitting to obtain the inequality inside the above probability. Nevertheless, the probabilities having been quantized, we can reduce it to a (non-disjoint) union of events, namely all $p \in \Upsilon$ such that $\prod_i p_i \leqslant \epsilon/N_{test}$. By the Bonferroni correction, the considered probability can be upper-bounded by the sum of their probabilities sum. In addition the intersection below involves only independent events according to our background model. Thus

$$\begin{aligned}
\mathbb{P}\left[\prod_i^N p^i_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j} \leqslant \frac{\epsilon}{N_{test}}\right] &= \mathbb{P}\left[\bigcup_{\substack{p\in\Upsilon \\ \prod_i p_i \leqslant \epsilon/N_{test}}} \bigcap_i \left(p^i_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j} \leqslant p_i\right)\right] \\
&\leqslant \sum_{\substack{p\in\Upsilon \\ \prod_i p_i \leqslant \epsilon/N_{test}}} \prod_i p_i \\
&\leqslant \frac{\epsilon}{\#I\,\#S'},
\end{aligned}$$

where we have also used $N_{tests} = \#I\,\#S'\,\#\Upsilon$. So we have shown that

$$\mathbb{E}[\Gamma] = \sum_{\mathbf{q},i} \mathbb{E}\left[\chi_{B_{\mathbf{q}},\mathbf{B}^{\mathbf{q}},j}\right] \leqslant \sum_{\mathbf{q},i} \frac{\epsilon}{\#I\,\#S'} = \epsilon.$$

□

The ϵ parameter is the only parameter of the method, the other ones being fixed one and for all. The question of how many false alarms should be acceptable in a stereo pair depends on the size of the images. In all experiments with moderate size images, of the order of 10^6 pixels, the decision was to fix $\varepsilon = 1$, which makes the method into a parameterless method for all moderately sized images.

4 Self-similarity filter

Human environments contain many periodic local structures (for example the windows on a façade). Since, in general, the number of repetitions is insignificant with respect to the number of blocks that have been used to estimate the empirical *a contrario* probability distributions, the *a contrario* model does not learn this repetition, and can be fooled by such repetitions, thus signaling a significant match for each repetition of the same structure. Of course, one of those significant matches is the correct one, but chances are that the correct one is not the most significant one. In such a situation two choices are left: *(i)* try to match the whole set

of self-similar blocks of I as a single multi-block (typically, global methods such as graph-cuts do that implicitly); or *(ii)* remove any (probably wrong) response in the case where the stroboscopic effect is detected. The first alternative would lead to errors anyway, if the similar blocks have not the same height, or if some of them are out of field in one of the images. Fortunately, stereo pair block-matching yields a straightforward adaptive threshold. A distance function d between blocks being defined, let $\mathbf{q}$ and $\mathbf{q}'$ be points in the reference and secondary images respectively that are candidates to match with each other. The match of $\mathbf{q}$ and $\mathbf{q}'$ will be accepted if the following self-similarity (SS) condition is satisfied:

$$d(B_{\mathbf{q}}, B_{\mathbf{q}'}) < \alpha \min\{d(B_{\mathbf{q}}, B_{\mathbf{r}})|\ \mathbf{r} \in I \cap S(\mathbf{q})\} \tag{12}$$

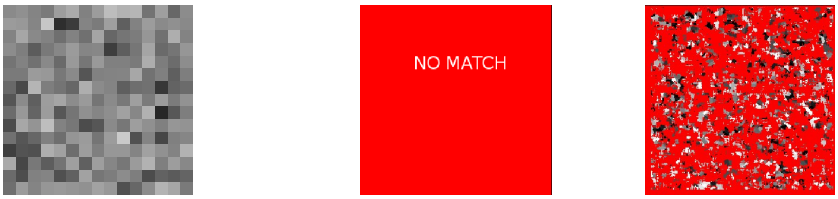
where $S(\mathbf{q}) = [q_1 - R\,,\, q_1 + R] \setminus \{q_1,\, q_1 + 1,\, q_1 - 1\}$ and R is the search range. The parameter α can be chosen as 0.6, as in the SIFT method [1]. As noted earlier, the search for correspondences can be restricted to the epipolar line. This is why the automatic threshold is restricted to $S(\mathbf{q})$.

Computing the similarity of matches in one of the images is not a new idea in stereovision. In [2] the authors define the *distinctiveness* of an image point x as the perceptual distance to the most similar other point in the search window. In particular, they study the case of the auto-SSD function (Sum of Squared Differences computed in the same image). The flatness of the function contains the expected match accuracy and the height of the smallest minimum of the auto-SSD function beside the one in the origin gives the risk of mismatch. They are able to match correctly ambiguous points by matching intrinsic curves [6]. However, the proposed algorithm only accepts matches when their quality is above a certain threshold. The obtained disparity maps are rather sparse and the accepted matches are completely concentrated on the edges of the image. According to [3], the ambiguous correspondences should be rejected. In this work a new *stability property* is defined as a condition a set of matches must satisfy to be considered unambiguous at a given confidence level. The stability constraint and the tuning of two parameters permits to take care of flat or periodic autocorrelation functions.

4.1 *A Contrario* vs Self-Similarity

Is the self-similarity (SS) threshold really necessary? One may wonder whether the *a contrario* decision rule to accept or reject correspondences between patches would be sufficient by itself. Conversely, is the self-similarity threshold enough to reject false matches in a correlation algorithm? This section addresses both questions and analyzes some simple examples enlightening the necessity and complementarity of both tests. For each example we are going to compare the result of the *a contrario* test and the result of a classic correlation algorithm combined with the self-similarity threshold alone.

First consider two independent Gaussian noise images (Fig 7). It is obvious that we would like to reject any possible match between these two images. As expected, (this is a sanity check!) the *a contrario* test rejects all the possible patch matches. On the other hand, the correlation



(a) (b) (c)

Figure 7: (a) Reference noise image. (b) No match at all has been accepted by the *a contrario* test! (c) Many false correspondences have been accepted by the self-similarity threshold.

	Bad matches	Total matches
SS	3.35%	85.86%
ACBM	0.37%	64.85%
ACBM+SS	0.36%	64.87%

Table 1: Quantitative comparison of several algorithms on Middlebury's Map image: the block matching algorithm with the self-similarity threshold (SS), the *a contrario* algorithm (ACBM) and the algorithm combining both (ACBM+SS). The percentage of matches for each algorithm is computed in the whole image and among these the number of wrong matches is also given. A match is considered wrong if its disparity difference with the ground truth disparity is larger than one pixel.

algorithm combined with the self-similarity is not sufficient: many false matches are accepted.

The second comparative test is about occlusions. If a point of the scene can be observed in only one of the images of the stereo pair, then an estimation of its disparity is simply impossible. The best decision is to reject its matches. A good example to illustrate the performance of both rejection tests ACBM and SS is the map image (Middlebury stereovision database, Fig 8) which has a large baseline and therefore an important number of occluded pixels. ACBM gives again the best result (see Table 1). The table indicates that the self-similarity test only removes a few additional points. Yet, even if the proportion of eliminated points is tiny, such mismatches can be very annoying and the gain is not negligible at all.

The *a contrario* methodology cannot detect the ambiguity inherent in periodic patterns. Indeed, periodicity certainly does not occur "just by chance". The match between a window and another identical window on a building façade is obviously non casual and is therefore legally accepted by an *a contrario* model. In this situation, the self-similarity test is necessary. A synthetic case has been considered in Fig. 9, where the accepted correspondences are completely wrong in the *a contrario* test for the repeated lines. On the contrary, the self-similarity threshold is able

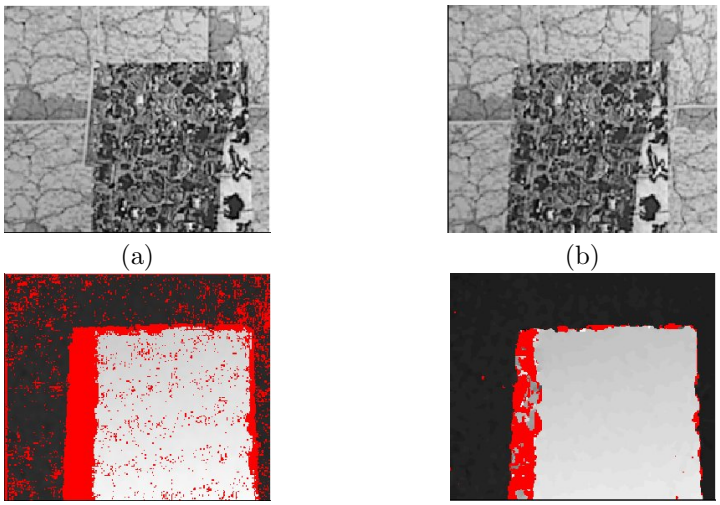


Figure 8: (a) Reference image (b) Secondary image. The rectangular object occludes part of the background (c) The *a contrario* test does not accept any match for pixels in the occluded areas. (d) With the self-similarity threshold the disparity map is denser, but wrong disparities remain in the occluded region.

to reject matches in this region of the image.

In short, ACBM and SS are both necessary and complementary. SS only removes a tiny additional number of errors, but even few outliers can be very annoying in stereo. From now on a possible match $(\mathbf{q}, \mathbf{q}')$ will therefore be accepted only if it is a meaningful match (ACBM test in Def. 6) and satisfies the SS condition given by (12).

5 Sub-pixel refinement

In this section we show how to refine the disparity map to reach sub-pixel accuracy. This has a direct consequence on the precision of the recovered 3D, avoiding a staircase effect of the depth function. The algorithm presented here is optimal in the sense that it relies on exact interpolation under usual assumptions concerning the sampling of the digital image.

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. Around such points, a deformation

(a) (b) (c)

Figure 9: (a) Reference image with a texture and a stripes periodic motif. The secondary image is a 2 pixels translation of the reference image. The obtained disparity map should be a constant image with value 2. (b) The *a contrario* test gives the right disparity 2 everywhere, except in the stripes region. (c) The repeated stripes are locally similar, so the self-similarity threshold rejects all the patches in this region.

model holds:

$$u_1(\mathbf{x}) = u_2(x + \varepsilon(\mathbf{x}), y). \tag{13}$$

The deformation model (13) is *a priori* valid when the angle of the 3D surface at $\mathbf{x}$ with respect to the camera changes moderately, which is systematically true for small (0.02 to 0.15) baseline stereo systems. The restriction brought by (13) is moderate. Indeed, the trend in stereo vision is to have multiple views of the 3D object to be reconstructed and therefore many pairs with small base line.

5.1 Preliminaries on sub-pixel interpolation

This section proves a discrete correlation formula which is faithful to the continuous image interpolates. Thanks to it, an accurate sub-pixel matching becomes possible. Without loss of generality, all considered images are defined on a square $[0, a]^2$ and are supposed to be square integrable. Thus, the Fourier series decomposition applies

$$u(x, y) = \sum_{k,l \in \mathbb{Z}} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}}, \tag{14}$$

where the $\tilde{u}_{k,l}$ are the Fourier series coefficients (or shortly the Fourier coefficients) of u. By the classic Fourier series isometry, for any two square integrable functions $u(\mathbf{x})$ and $v(\mathbf{x})$ on $[0, a]^2$,

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x})d\mathbf{x} = a^2 \sum_{k,l \in \mathbb{Z}} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}}. \tag{15}$$

The digital images are usually given by their N^2 samples $u(\mathbf{m})$ for $\mathbf{m}$ in the grid

$$\mathbb{Z}^1_a = [0, a]^2 \cap \left((\frac{a}{2N}, \frac{a}{2N}) + \frac{a}{N}\mathbb{Z}^2 \right).$$

Similarly, the over-sampling grid with four times more samples is denoted by

$$\mathbb{Z}^{\frac{1}{2}}_a = [0, a]^2 \cap \left((\frac{a}{4N}, \frac{a}{4N}) + \frac{a}{2N}\mathbb{Z}^2 \right).$$

N is always an even integer. In all that follows we shall assume that the images obtained by a stereo vision system are band-limited. This assumption is classical and realistic, the aliasing in good quality CCD cameras being moderate. As classical in image processing, under the (forced) a-periodicity assumption a band-limited image becomes a trigonometric polynomial. This periodicity assumption is not natural, but it only entails a minor drawback, namely a small distortion near the boundary of the image domain $[0,a]^2$. The payoff for the *band-limited + periodic assumption* is that the image can be interpolated, and its Fourier coefficients computed from discrete samples. Indeed, given N^2 samples $u_{\mathbf{m}}$ for $\mathbf{m}$ in $\mathbb{Z}^1_a$, there is a unique trigonometric polynomial in the form

$$u(x,y)= \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}} \tag{16}$$

such that $u(\mathbf{m}) = u_{\mathbf{m}}$. We shall call such polynomials *N-degree trigonometric polynomials*. The coefficients $\tilde{u}_{k,l}$ are the *Fourier coefficients* of u in the Fourier basis $e^{\frac{2i\pi(kx+ly)}{a}}$, $k,l \in \mathbb{Z}$. The map $u_{\mathbf{m}} \to u_{k,l}$ is nothing but the *2D Discrete Fourier Transform (DFT)*, and the map $(u_{\mathbf{m}}) \to N(\tilde{u}_{k,l})$ is an isometry from $\mathbb{C}^{N^2}$ to itself. The function $u(x,y)$ is therefore usually called the *DFT interpolate of the samples* $u_{\mathbf{m}}$. In consequence, there is an isometry between the set of N-degree trigonometric polynomials endowed with the $L^2([0,a]^2)$ norm, and $\mathbb{C}^{N^2}$ endowed with the usual Euclidean norm:

$$\int_{[0,a]^2} |u(x,y)|^2 = a^2 \sum_{k,l=-N/2}^{N/2-1} |\tilde{u}_{k,l}|^2 = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}^1_a} |u(\mathbf{y}+\mathbf{m})|^2 , \tag{17}$$

where the N^2 samples grid can have an arbitrary origin $\mathbf{y}$. If $u(\mathbf{x})$ and $v(\mathbf{x})$ are two N-degree trigonometric polynomials, we therefore also have

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) = a^2 \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}^1_a} u(\mathbf{y}+\mathbf{m})\overline{v(\mathbf{y}+\mathbf{m})}, \tag{18}$$

where $\overline{v}$ is the complex conjugate of v. Taking four times more samples, it follows from (18) that

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) = a^2 \sum_{k,l=-N}^{N-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}^{\frac{1}{2}}_a} u(\mathbf{m})\overline{v(\mathbf{m})}. \tag{19}$$

which is also valid if $u(\mathbf{x})$ and $v(\mathbf{x})$ are up to $2N$-degree trigonometric polynomials in $\mathbf{x}$.

This last fact has a first important consequence in block-matching. Consider two images $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ on $[0,a]^2$ and a window function $\varphi(\mathbf{x})$. Block-matching is the search for a value of μ minimizing the continuous quadratic distance

$$e_{\mathbf{x}_0}(\mu) := \int_{[0,a]^2} \varphi(\mathbf{x}-\mathbf{x}_0)\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2 d\mathbf{x}. \tag{20}$$

5.2 Arbitrary accuracy with $\times 2$ zoom of images

Proposition 1. (Equality of the discrete and the continuous quadratic distance) *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials on $[0,a]^2$ and let $\varphi(\mathbf{x})$ be a window function which we assume to be a $2N$-degree trigonometric polynomial. Then*

$$e_{\mathbf{x}_0}(\mu) = e^d_{\mathbf{x}_0}(\mu), \quad \textit{where} \tag{21}$$

$$e^d_{\mathbf{x}_0}(\mu) := \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}^{\frac{1}{2}}_a} \varphi(\mathbf{m}-\mathbf{x}_0)\big(u_1(\mathbf{m}) - u_2(\mathbf{m}+(\mu,0))\big)^2. \tag{22}$$

The proof follows from (19). Indeed, $\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2$ and $\varphi(\mathbf{x}-\mathbf{x}_0)$ are both $2N$-degree trigonometric polynomials in $\mathbf{x}$, so according to (19) the discrete scalar product defining $e^d_{\mathbf{x}_0}(\mu)$ equals the continues scalar product defining $e_{\mathbf{x}_0}(\mu)$. Thus *the continuous block distance is a finite sum of discrete samples!*

The block distance function $\mu \to e_{\mathbf{x}_0}(\mu)$, whose minimization is our main objective here, is also easily sampled. By (22) it is a $2N$-degree trigonometric polynomial with respect to μ. This proves:

Proposition 2 (Sub-pixel correlation requires $\times 2$ zoom)**.** *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials. Then the quadratic distance $e^d_{\mathbf{x}_0}(\mu)$ is well-sampled provided it has at least $2N$ successive samples. Thus the computation of $e^d_{\mathbf{x}_0}(\mu)$ at half samples $\mu \in \frac{a\mathbb{Z}}{2}$ (via zero-padding) allows the exact reconstruction of $e^d_{\mathbf{x}_0}(\mu)$ for any real μ by DFT interpolation.*

Remark that the last proposition does not require any assumption on the window function $\varphi(\mathbf{x})$. Prop. 2, which opens the way to rigorous block-matching with sub-pixel accuracy, has been noticed in [5]. It is also used in the MARC method used by the French space agency (CNES). The above simple proof of Prop. 2 is new.

From the above formulation, it seems possible to reach arbitrary accuracy on the disparity. Actually, there is a limitation: the samples $u(\mathbf{m})$ are known only with fixed precision, and more importantly they are polluted by some noise. So in practice there is a positive lower bound on the minimum error that can be expected, this bound being linked to the level of noise in the images.

6 Projects

These projects are under the supervision of Toni Buades, Pascal Monasse and Jean-Michel Morel

6.1 Project 1: IPOL implementation of various disparity filters

Various filters up to 2002 were listed in the famous paper [4]. This was accompanied with the Middlebury benchmark `http://vision.middlebury.edu/stereo/`, still used as a reference for comparing various stereo algorithms. The goal

of this project is to implement in IPOL some of the filters listed in that paper. These include:

- Bidirectional filter: validate the best match from a window W of I to a window W' of I' only if W is the best match when looking for W' in I.
- The min-filter, which associates to a pixel P the disparity evaluated from all windows containing P (not just the centered window) and yielding the lowest L^2 distance. This can be computed efficiently by the following algorithm:
  1. Attribute the best match at each pixel of the window centered around it, as usual, and keep track of the resulting error.
  2. At each pixel P, put the disparity associated with minimal error among all pixels within the window W centered on P.

6.2 Project 2: RAFA algorithm for adhesion correction

A recent algorithm named RAFA (Réduction de l'Adhérence par Fenêtre Adaptative) attempts to correct the adhesion effect by weighing each pixel in the L^2 distance by the inverse of the square image derivative along epipolar direction (x):

$$e(\mu) = \iint_W \frac{(u_2(x+\mu, y) - u_1(x,y))^2}{\left|\frac{\partial u_1}{\partial x}(x,y)\right|^2}. \tag{23}$$

The adhesion effect occurs when pixels with different disparities are present within the correlation window. The simple weight above is shown to yield the best correction under certain conditions in zones where the disparity is continuous. In particular, it does not solve the adhesion problem near disparity discontinuities.

References

[1] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[2] R. Manduchi and C. Tomasi. Distinctiveness maps for image matching. *Proceedings of the International Conference on Image Analysis and Processing*, pages 26–31, 1999.

[3] R. Sara. Finding the largest unambiguous component of stereo matching. In *Proceedings of the European Conference on Computer Vision-Part III*, pages 900–914. Springer-Verlag, 2002.

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

[5] R. Szeliski and D. Scharstein. Sampling the disparity space image. *PAMI*, 26(3):419–425, 2004.

[6] C. Tomasi and R. Manduchi. Stereo matching as a nearest-neighbor problem. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(3):333–340, 1998.