Meaningful Matches in Stereovision

Neus Sabater*, Andrés Almansa** and Jean-Michel Morel*
*ENS Cachan, CNRS-CMLA. France.
**Telecom ParisTech, CNRS-LTCI. France.

***Abstract*—This paper introduces a statistical method to decide whether two blocks in a pair of images match reliably. The method ensures that the selected block matches are unlikely to have occurred "just by chance." The new approach is based on the definition of a simple but faithful statistical *background model* for image blocks learned from the image itself. A theorem guarantees that under this model not more than a fixed number of wrong matches occurs (on average) for the whole image. This fixed number (the number of false alarms) is the only method parameter. Furthermore, the number of false alarms associated with each match measures its reliability. This *a contrario* block-matching method, however, cannot rule out false matches due to the presence of periodic objects in the images. But it is successfully complemented by a parameterless *self-similarity threshold*. Experimental evidence shows that the proposed method also detects occlusions and incoherent motions due to vehicles and pedestrians in non simultaneous stereo.**

***Index Terms*—Stereo vision, Block-matching, Number of False Alarms (NFA), *a contrario* detection.**

I. Introduction

Stereo algorithms aim at reconstructing a 3D model from two or more images of the same scene acquired at different angles. This work only considers previously stereo-rectified image pairs. In that case the 3D reconstruction requires that the matched points in both images belong to the same horizontal epipolar line. The matching process of stereo image pairs has been studied in depth for more than four decades. [1] and [35] contain a fairly complete comparison of the main methods. According to these surveys there are roughly two main classes of algorithms in binocular stereovision: local matching methods and global methods.

Global methods aim at a coherent solution obtained by minimizing an energy functional containing matching fidelity terms and regularity constraints. The most efficient ones seem to be Belief Propagation [15], [42], Graph Cuts [16], Dynamic Programming [9], [28] and solvers of the multi-label problem [13], [30]. They often resolve ambiguous matches by maintaining a coherence along the epipolar line (DP) or along and across epipolar lines (BP & GC). They rely on a regularization term to eliminate outliers and reduce the noise. They give a match to all points which are not detected as occluded. Global methods are, however, at risk to make or propagate errors if the regularization term is not adapted to the scene. A classic example is when a large portion of the scene is nearly constant, for example a scene including a cloudless sky, since there is no information in such a region to compute reliable matches (see Fig. 11 for an example). On such ambiguous regions, global methods perform an interpolation by using the informative pixels. This interpolation can be lucky, as it is the case in most images of the Middlebury benchmark[1]. But it can also fail, as is apparent in the above example and in many outdoor scenes. Furthermore, the energy in global methods, has at least two terms and one parameter weighting them (and sometimes three terms and two parameters [16]). These parameters are difficult to tune, and even to model. Thus, it remains a valid question how to rule out by a parameterless method the dubious regions where the matches cannot be scientifically demonstrated.

On the other hand local methods are simpler, but equally sensitive to local ambiguities. Local methods start by comparing features of the right and left images. These features can be blocks in block-matching methods, or even local descriptors [21] like SIFT descriptors [18], [31], curves [36], corners [3], [11], etc. The drawback of local methods is that they do not provide a dense map as global methods do (meaning that the percentage of matched points is lower than 100%).

Recent years have therefore seen a blooming of global methods, which reach the best performance in recent benchmarks such as the Middlebury dataset [35]. But our purpose is to show that local methods can also be competitive. This paper considers the common denominator of most local methods, block-matching. It shows that this tool is amenable to a local statistical decision rule telling us whether a match is reliable. In fact, not all the pixels in an image pair can be reliably matched in real scenes (40 to 80% of pixels). The lack of corresponding points in the second image or the ambiguity in certain points stirs up gross errors in dense stereovision. In particular block-matching methods suffer from two mismatching causes that must be tackled one by one:

1) The main mismatch cause in local methods is the absence of a theoretically well founded threshold to decide whether two blocks really match or not. Our main goal here will be to define such a threshold by an *a contrario* block-matching (ACBM) rejection rule, ensuring that two blocks do not match "just by chance."
2) A second minor mismatch cause is the presence on the epipolar line of repetitive shapes or textures, a problem sometimes called "stroboscopic phenomenon," or "self-similarity." The proposed ACBM only rules out stochastic similarities, not deterministic ones. While the ACBM rule mismatches repetitive patterns, these types of mismatches are easily eliminated by a simple self-similarity rule (SS). We shall, however, verify that a self-similarity rule by itself is far from reaching the ACBM performance. Both rules are necessary and complementary.

The elimination of these two sorts of mismatches is a key

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

issue in block-matching methods. The problem of sifting out matching errors in stereovision has of course been addressed many times. We shall discuss a choice of the significant contributions for each cause of mismatch.

Occlusions are still an open problem in stereovision and one of the main causes of mismatch. For this reason numerous stereo approaches focus on detecting them. Global energy methods [16] address occlusions by adding a penalty term for occluded pixels in their energy function. In [14] the major contribution is the reasoning about visibility in multi-view stereo. [42] computes two disparity maps symmetrically and verifies the left-right coherence to detect occluded pixels. [28] asserts that if two points in the epipolar line match with two points with a different order then there is an occlusion. Again this can lead to errors if there are narrow objects in the scene. See also [7], which compares a choice of methods to detect occlusions.

Matching pixels in *poorly textured regions*, where noise dominates signal, is clearly the main cause of error. Based on local SNR estimates, [5] has proposed to reject matches by thresholding the second derivative of the correlation function: the flatter the correlation function, the less reliable the match. In [34], the mismatches due to weakly textured objects or to *periodic structures* are considered. The author defines a confidently stable matching in order to establish the largest possible unambiguous matching at a given confidence level. Two parameters control the compromise between the percentage of bad matches and the match density of the map. Yet, the match density falls dramatically when the percentage of mismatches decreases. We will see that the method presented here is able to get denser disparity maps with less mismatches. Similarly, [20] tries to eliminate errors on repeated patterns. Yet their matches seem to concentrate mainly on image edges and therefore have a low density. A more primitive version of the rejection method developed here was applied successfully to the detection of *moving and disappearing objects* in [33]. This is a foremost problem in the quasi-simultaneous stereo usual in aerial or satellite imaging where vehicles and pedestrians perturb strongly the stereo matching process. The extended method presented here deals with a much broader class of mismatches, including those due to poor signal to noise ratio.

A. Anterior Statistical A Contrario Decision Methods

Because of the above mentioned reasons one cannot presuppose the existence of uniquely determined correspondences for all pixels in the image. Thus, a decision must be taken on whether a block in the left image actually meaningfully matches or not its best match in the right image. This problem will be addressed by the *a contrario* approach initiated by [6]. This method is generally viewed as an adaptation to image analysis of classic hypothesis testing. But it also has a psychophysical justification in the so-called Helmholtz principle, according to which all perceptions could be characterized as having a low probability of occurring in noise. Early versions of this principle in computer vision are [17], [10], [38].

A probabilistic *a contrario* argument is also invoked in the SIFT method [18], which includes an empirical rejection threshold. A match between two descriptors S_1 and S_1' is rejected if the second closest match S_2' to S_1 is actually almost as close to S_1 as S_2' is. The typical distance ratio rejection threshold is 0.6, which means that S_2 is accepted if $dist(S_1', S_1) \leq 0.6 \times dist(S_2', S_1)$ and rejected otherwise. Interestingly, Lowe justifies this threshold by a probabilistic argument: if the second best match is almost as good as the first, this only means that both matches are likely to occur casually. Thus, they must be rejected. Recently, [31] proposed a rigorous theory for this intuitive method. SIFT matches are accepted or rejected by an *a contrario* methodology involving the Earth mover distance. The *a contrario* methodology has also already been used in stereo matching. [23] proposed a probabilistic criterion to detect a rigid motion between two point sets taken from a stereo pair, and to estimate the fundamental matrix. This method, ORSA, shows improved robustness compared to a classic RANSAC method. In the context of foreground detection in video, [22] proposed an *a contrario* method for discriminating foreground from background pixels that was later refined by [29]. Even though this problem has some points in common with stereo matching, it is in a way less strict, since it only needs to learn to discriminate two classes of pixels. Hence they do not need to resort to image blocks, but rely only on a 5 dimensional feature vector composed of the color and motion vector of each pixel.

Among influential related works, Robin *et al.* [32] describe a method for change detection in a time series of Earth observation images. The change region is defined as the complement of the maximal region where the time series does not change significantly. Thus, what is controlled by the *a contrario* method is the number of false alarms (NFA) of the no-change region. This method can therefore be regarded as an *a contrario* region matching method. It is fundamentally different from the method we shall present. Indeed, Robin's method assumes (in addition to the statistical background model) a statistical image model that the time series follows in the regions where no change occurs, which is not feasible in stereo matching.

The method in [27] is also worth mentioning. It is an *a contrario* method for detecting similar regions between two images. This method is a classic statistical test rather than an *a contrario* detection method in the sense of [6]. Indeed, the role of the background model (H_0 hypothesis) and the structure to be tested (H_1 hypothesis) are reversed: This method only controls the false negative rate and not the false positive rate (as in typical *a contrario* methods). Furthermore the significance level of the statistical test is set to $\alpha \approx 0.1$ in accordance with classical statistical testing, whereas as demonstrated in [6] the significance level can be made much more secure, of the order of 10^{-6}.

The *a contrario* model for region matching in stereo vision used in [12] is simple and efficient. The gradient orientations at all region pixels are assumed independent and uniformly distributed in the background model. A more elaborate version learns the probability distribution of gradient orientation differences under the hypothesis that the disparity (or motion) is zero, and uses this distribution as a background model. Still, pixels are all considered as independent under the background

model. Once this background model is learned, a given disparity (or motion model) is considered as meaningful if the number of aligned gradient orientations is sufficiently large within the tested region. This region matching method works well, but requires an initial over-segmentation of the gray-level image which is later refined by an *a contrario* region merging procedure. Because of the rough background model, false positive region matches can be observed.

The key to a good background or *a contrario* model in block-matching would be to learn a realistic probability distribution of the high-dimensional space of image patches. The seminal works [25] and [4] in the context of shape matching (where shapes are represented as pieces of level lines of a fixed size) showed that high-dimensional shape distributions can be efficiently approximated by the tensor product of (well chosen) marginal distributions. The marginal laws are one-dimensional, and therefore easily learned. In [26] these marginals are learned along the orientations of the principle components. The present work can be viewed as an extension of this curve matching method to block-matching.

[2] proposed an alternative way of choosing detection thresholds such that the number of false detections under a given background model is ensured to stay below a given threshold. The procedure does not require analytical computations or decomposing the probability as a tensor product of marginal distributions. Instead, detection thresholds are learned by Monte-Carlo simulations in a way that ensures the target NFA rate. This method, that was developed in the context of image segmentation, involves the definition of a set of thresholds to determine whether two neighboring regions are similar. However, as in [27], the detected event whose false positive rate is controlled is *"the two regions are different,"* and not the one we are interested in in the case of region matching, namely *"the two regions are similar."*

In conclusion, the *a contrario* methodology is expanding to many matching decision rules, but does not seem to have been previously applied to the block-matching problem. We shall now proceed to describe the *a contrario* or background model for block-matching. The proposed model is the simplest that worked, but the reader may wonder if a still simpler model could actually work. In the next section we analyze a list of simpler proposals, and we explain why they must be discarded.

B. Choosing an Adequate A Contrario Model for Patch Comparison.

The goal of this section is is to reject simpler alternatives to the probabilistic block model that will be used. In recent years, patch models and patch spaces are becoming increasingly popular. We refer to [19] and references therein for algorithms generating sparse bases of patch spaces. Here, 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 similar patches in these images, this similarity arises just by chance?*"
The "just by chance" implies the existence of a stochastic *background model*, often called the *a contrario* model.

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 $I'(\mathbf{x})$ are instances of i.i.d. random variables $\mathscr{I}'(\mathbf{x})$ with cumulative distribution H. This would lead us to affirm that pixels $\mathbf{q}$ in image I and $\mathbf{q}'$ in image I' are a meaningful match if their gray level difference is unlikely small,

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

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 $(\mathbf{q},\mathbf{q}')$ that may match. But such a small probability can be achieved (assume that H is uniform over $[0,255]$) only if the threshold $\theta = |I(\mathbf{q})-I'(\mathbf{q}')| < 128 \cdot 10^{-7}$. On the other hand, $|I(\mathbf{q})-I'(\mathbf{q}')|$ 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 approach could be to compare blocks of a certain size $\sqrt{s} \times \sqrt{s}$ with the ℓ^2 norm, and with the same background model as before. Thus, we could declare blocks $B_{\mathbf{q}}$ and $B_{\mathbf{q}'}$ as meaningfully similar if

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

where B_0 is the block of size $\sqrt{s} \times \sqrt{s}$ centered at the position (0,0). 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 better reflecting 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, as pointed out in [25], high-dimensional distributions of shapes can be approximated by the tensor product of their adequately chosen marginal distributions. Such

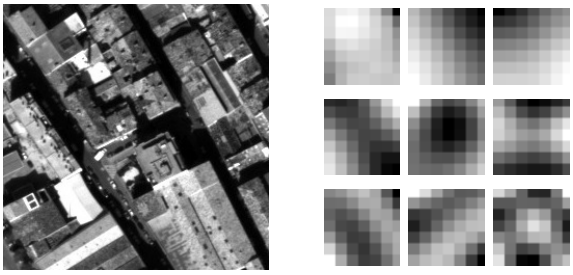


Fig. 1. Left: Reference image of a stereo pair of images. Right: the nine first principal components of the 7×7 blocks.

marginal laws, being one-dimensional, are easily learned from a single image. Ideally, ICA (Independent Component Analysis) should be used to learn which marginal laws are the most independent, but the simpler PCA analysis will show accurate enough for our purposes. Indeed, it ensures that the principal components are decorrelated, a first approximation to independence. Fig. 2 gives a visual assessment of how well a local PCA model simulates image patches in a class. Nevertheless, the independence assumption will be used as a tool for building the a-contrario model. This independence is not an empirical finding on the set of patches.

C. Plan

Section II introduces the stochastic block model learned from a reference image. Section II-B presents the *a contrario* method applied to disparity estimation in stereo pairs and treats the main problem of deciding whether two pixels match. Theorem 1 is the main result of this section, ensuring a controlled number of false detections. Section III tackles the stroboscopic problem by a parameterless method, and demonstrates the necessity and complementarity of the *a contrario* and self-similarity rejections. Experimental results and comparison with other methods are in Section IV. Section V is conclusive. An appendix summarizes the algorithm and gives its complete pseudo-code.

II. 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 be 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 $B_{\mathbf{q}}$.

A. Principal Component Analysis

For building a simple *a contrario* model the principal component analysis can play a crucial role, as shown in [26]. Indeed, it allows for effective dimension reduction and decorrelates these dimensions, giving a first approximation to independence. This facilitates the construction of 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 $(x_1^{\mathbf{q}},\dots,x_s^{\mathbf{q}})$ 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 $X=(x_i^j)$ $1\le i\le s$, $1\le j\le n$ consisting of the set of all data vectors, one column per pixel in the image. Then, the covariance matrix of the block is $C=\mathbb{E}(X-\bar{\mathbf{x}}\mathbf{1})(X-\bar{\mathbf{x}}\mathbf{1})^T$, 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 $\bar{\mathbf{x}}$ corresponds to the block whose k-th pixel is the average of all k-th pixels of all blocks in the image. Thus, $\bar{\mathbf{x}}$ 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. 1 shows the first principal blocks.

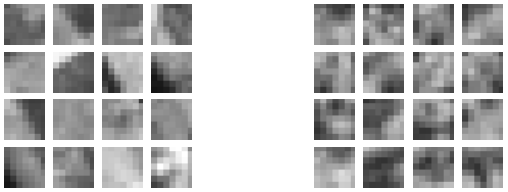
Usually, the eigenvectors are sorted in order of decreasing eigenvalue. In that way the first principal components are the ones that contribute most to the variance of the data set. By keeping the first $N<s$ components with larger eigenvalues, the dimension is reduced but the significant information 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.

Each block is represented by N ordered coefficients $(c_{\sigma_{\mathbf{q}}(1)}(\mathbf{q}),\dots,c_{\sigma_{\mathbf{q}}(N)}(\mathbf{q}))$, where $c_i(\mathbf{q})$ is the resulting coefficient after projecting $B_{\mathbf{q}}$ onto the principal component $i\in\{1,\dots,s\}$ 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 $c_i(\mathbf{q})$ 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. 4), 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.

B. A Contrario Similarity Measure between Blocks

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

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



(a) (b)

Fig. 2. (a) Patches of the reference image, chosen at random. (b) Simulated random blocks following the law of the reference image. This experiment illustrates the (relative) adequacy of the *a contrario* model. Nevertheless, the PCA components are empirically uncorrelated, but of course not independent.

The reference image will be the secondary image I'. Fig. 2 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 $\mathbb{C}_i$ in the *a contrario* model, which is justified by the independence of $\mathbb{C}_i$ and $\mathbb{C}_j$ for $i \neq j$. There is a strong adequacy of the *a contrario* model to the empirical model, since the PCA transform ensures that $\mathbb{C}_i$ and $\mathbb{C}_j$ are uncorrelated for $i \neq j$, a first approximation of the independence requirement.

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

Definition 2 (Resemblance probability): Let $B_{\mathbf{q}}$ be a block in I and $B_{\mathbf{q}'}$ a block in I'. Define the probability that a random block $\mathbb{B}$ of I' resembles $B_{\mathbf{q}}$ as closely as $B_{\mathbf{q}'}$ does in the i-th component by

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

Fig. 3 illustrates how the resemblance probability $\widehat{p^i}_{\mathbf{q}\mathbf{q}'}$ is computed and Fig. 4 shows empirical marginal densities.

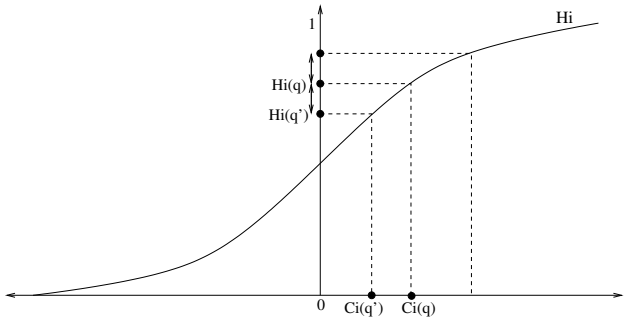


Fig. 3. Normalized cumulative histogram of i-th PCA coordinates of the secondary image. $c_i(\mathbf{q})$ is the i-th PCA coordinate value in the first image. The resemblance probability $\widehat{p^i}_{\mathbf{q}\mathbf{q}'}$ for the i-th component is twice the distance $|H_i(\mathbf{q}) - H_i(\mathbf{q}')|$ when $H_i(\mathbf{q})$ is not too close to the values 0 or 1.

C. 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 next components are similar. Due to this fact, the components of $B_{\mathbf{q}}$ and another block $B_{\mathbf{q}'}$ must be compared with a non-decreasing exigency level. In addition, in the *a contrario* model, the number of tested correspondences should be as small as possible to reduce the number of false alarms. A quantization of the tested resemblance probabilities is therefore required to limit the number of tests.

These two remarks lead to define the quantized resemblance probability as the smallest non-decreasing sequence of quantized probabilities bounding from above the sequence $\widehat{p^i}_{\mathbf{q}\mathbf{q}'}$.

Definition 3 (Quantized probability): Let $B_{\mathbf{q}}$ be a block in I. Let $\Pi := \{\pi_j = 1/2^{j-1}\}_{j=1,\dots,Q}$ be a set of quantized probability thresholds and let

$$\Upsilon := \left\{ \mathbf{p} = (p_1, \dots, p_N) \mid p_i \in \Pi, \quad p_i \leqslant p_j \text{ if } i < j \right\}$$

be the family of non-decreasing N-tuples in Π^N, endowed with the order $\mathbf{a} \geqslant \mathbf{b}$ if and only if $a_i \geqslant b_i$ for all i. The quantized probability sequence associated with the event that random block $\mathbb{B}$ resembles $B_{\mathbf{q}}$ as closely as $B_{\mathbf{q}'}$ does in the ith component is defined by

$$(p^i_{\mathbf{q}\mathbf{q}'})_{i=1,\dots N} = \inf_{t \in \Upsilon} \{ t \mid t \geqslant (\widehat{p^i}_{\mathbf{q}\mathbf{q}'})_{i=1,\dots N} \}. \qquad (2)$$

Notice that the infimum $(p^1_{\mathbf{q}\mathbf{q}'}, \dots, p^N_{\mathbf{q}\mathbf{q}'})$ is uniquely defined and belongs to Υ. Put another way the quantized probability vector $(p^1_{\mathbf{q}\mathbf{q}'}, \dots, p^N_{\mathbf{q}\mathbf{q}'})$ is the smallest upper bound of the resemblance probabilities $(\widehat{p^1}_{\mathbf{q}\mathbf{q}'}, \dots, \widehat{p^N}_{\mathbf{q}\mathbf{q}'})$ that can be found in Υ. Fig. 5 illustrates the quantized probabilities in two cases.

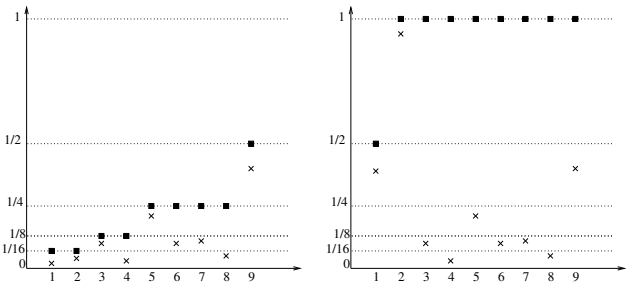


Fig. 5. Two examples of probabilities with $Q = 5$ and $N = 9$. The probability thresholds are in ordinate and the features in abscissa. The resemblance probabilities are represented with small crosses and quantized probabilities with small squares. The example on the left has a final probability of $1/(16^2 \cdot 8^2 \cdot 4^4 \cdot 2)$. The right example has the same resemblance probabilities excepting for features 1 and 2, but the final probability is $1/2$. Only the configuration on the left corresponds to a meaningful match.

Proposition 1 (Quantized resemblance probability): Let $B_{\mathbf{q}} \in I$ and $B_{\mathbf{q}'}$ be two blocks. Assume the principal components $i \in \{1, 2, \dots, s\}$ are reordered so that $|c_1(\mathbf{q})| \geqslant |c_2(\mathbf{q})| \geqslant \cdots \geqslant |c_s(\mathbf{q})|$. The probability of the event "the random block $\mathbb{B}$ has its N first components as

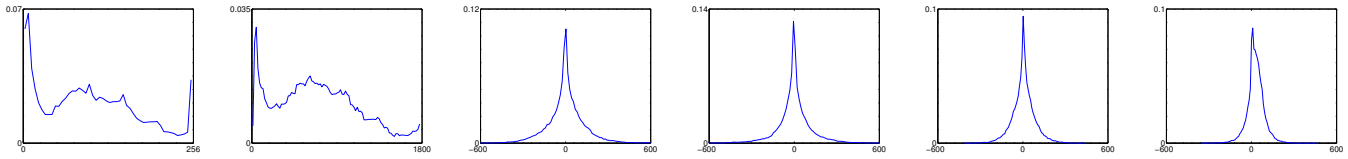


Fig. 4. Histogram of the reference image, followed by the first five histograms of the block PCA coordinates. The first principal component roughly computes a mean of the block, which explains why its histogram is so similar to the image histogram.

similar to those of $B_{\mathbf{q}}$ as to those of $B_{\mathbf{q}'}$" is

$$Pr_{\mathbf{q}\mathbf{q}'} = \prod_{i=1}^{N} p^i_{\mathbf{q}\mathbf{q}'} \,. \tag{3}$$

This is a direct consequence of Def. 1, the principal components of $\mathbb{B}$ being independent. The resemblance probability is the product of the marginal resemblance probabilities. As classic in statistical decision, we could stop and use the above resemblance probability. But, despite having a low resemblance probability for each $Pr_{\mathbf{q}\mathbf{q}'}$, the large number of resemblance tests allows for a very large number of false matches. Our next goal therefore is to define a number of false alarms, and not a probability, as the right criterion. To this aim, we need to estimate the number of tests.

D. Number of Tests

The number of tests for comparing all the blocks of image I with all the blocks in image I' is the product of three factors. The first one is the image size $\#I$. The second is the size of the search region denoted by $S' \subset I'$. 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 $\mathbf{q}' = (q'_1, q_2) \in I'$ 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 and most important factor is the number of different non-decreasing probability distributions $FC_{N,Q} = \#\Upsilon$ that can be envisaged. Of course not all of these tests are 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. A test can never be defined *a posteriori*, it cannot be steered by the observation. Thus the number of tests is not the number of tests effectively performed. There are $\#\Upsilon$ ways each couple of blocks could *a priori* be compared. In other terms $\#\Upsilon$ different distances are *a priori* tested. Theorem 1 will ultimately justify the following definition.

Definition 4 (Number of tests): With the above notation we call the number of tests for matching two images I and I' the integer $N_{test} = \#I \cdot \#S' \cdot \#\Upsilon = n\,(2R+1)\,FC_{N,Q}$.

Lemma 1: With the above notation,

$$FC_{N,Q} = \sum_{t=0}^{Q-1} (t+1) \cdot \binom{N+Q-t-3}{Q-t-1} \,, \tag{4}$$

where

$$FC_{N,Q} := \#\{\, f : [1,N] \to [1,Q] \mid f(x) \leqslant f(y), \forall x \leq y \,\}.$$

In order to prove this result we write

$$\overline{FC}_{N,Q} := \#\{\, f : [1,N] \to [1,Q] \mid f(1) = 1,\ f(N) = Q; \\ f(x) \leqslant f(y), \forall x \leqslant y \,\} \,.$$

Since $FC_{N,Q} = \sum_{t=0}^{Q-1} (t+1)\overline{FC}_{N,Q-t}$ and $\overline{FC}_{N,Q} = \binom{N+Q-3}{Q-1}$ the result follows.

We are now in a 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 $|c_1(\mathbf{q})| \geqslant |c_2(\mathbf{q})| \geqslant \cdots \geqslant |c_s(\mathbf{q})|$. We define the Number of False Alarms associated with event "the random block $\mathbb{B}$ has its N first components as similar to those of $B_{\mathbf{q}}$ as those of $B_{\mathbf{q}'}$ are" by

$$NFA_{\mathbf{q},\mathbf{q}'} = N_{test} \cdot Pr_{\mathbf{q}\mathbf{q}'} = N_{test} \cdot \prod_{i=1}^{N} p^i_{\mathbf{q}\mathbf{q}'},$$

where N_{test} comes form Def. 4 and $Pr_{\mathbf{q}\mathbf{q}'}$ is the probability that the random block $\mathbb{B}$ have its first N PCA components as similar to those of $B_{\mathbf{q}}$ as those of $B_{\mathbf{q}'}$ are (Prop. 1).

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

$$NFA_{\mathbf{q}\mathbf{q}'} \leqslant \varepsilon \,. \tag{5}$$

E. The Main Theorem

As it is computed above the NFA dimensionality is that of a number (of false alarms) *per image*. An alternative would be to measure the NFA as a number of false alarms per pixel, in which case the number of tests would not contain the cardinality of the image factor $\#I$. With the proposed NFA, it is up to the users to decide which number of false alarms per image they consider tolerable. 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 [37], page 22-23, namely to treat the probability of an event as random variable and to play with its expectation. Here the NFA will become a random variable, replacing $B_{\mathbf{q}'}$ with $\mathbb{B}$ 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 $\mathbb{B}$. In total, $B_{\mathbf{q}}$ is compared with $2R+1 otherblocks$ for each $\mathbf{q} \in I$. So, we are led to distinguish for each $\mathbf{q}$ $(2R+1)$ trials which are as many i.i.d. random

blocks $\mathbb{B}^{\mathbf{q},j}$, $j \in \{1,2,\dots 2R+1\}$, all with the same law as $\mathbb{B}$. 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 $E_{\mathbf{q},j}$ that a random block $\mathbb{B}^{\mathbf{q},j}$ 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 $\chi_{\mathbf{q},j}$ the random characteristic function associated with this event, with the convention that $\chi_{\mathbf{q},j}=1$ if $E_{\mathbf{q},j}$ is true, $\chi_{\mathbf{q},j}=0$ otherwise. Similarly $NFA_{\mathbf{q},j}$ and $p^i_{\mathbf{q},j}$ are the NFA and quantized probabilities associated with the event $E_{\mathbf{q},j}$.

Theorem 1: Let $\Gamma = \Sigma_{\mathbf{q}\in I, j\in\{1,\dots,2R+1\}}\chi_{\mathbf{q},j}$ 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 less than or equal to ε.

Proof:

We have

$$\chi_{\mathbf{q},j} = \begin{cases} 1, & \text{if } NFA_{\mathbf{q},j} \leqslant \varepsilon; \\ 0, & \text{if } NFA_{\mathbf{q},j} > \varepsilon. \end{cases}$$

Then, by the linearity of the expectation

$$\mathbb{E}[\Gamma] = \sum_{\mathbf{q},j} \mathbb{E}[\chi_{\mathbf{q},j}] = \sum_{\mathbf{q},j} \mathbb{P}\left[NFA_{\mathbf{q},j} \leqslant \varepsilon\right].$$

The probability inside the above sum can be computed by Definitions 5 and 2:

$$\mathbb{P}\left[NFA_{\mathbf{q},j} \leqslant \varepsilon\right] = \mathbb{P}\left[\prod_i^N p^i_{\mathbf{q},j} \leqslant \frac{\varepsilon}{N_{test}}\right]$$

There are many probability N-tuples $p = (p^i_{\mathbf{q},j})_{i=1,\dots,N}$ 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 $\prod_i p_i \leqslant \varepsilon/N_{test}$. 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

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

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

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

■

The ε parameter is the only legitimate parameter of the method, the other ones namely the block size $\sqrt{s}$, the number of principal components N and the number of quantized probability thresholds Q can be fixed once and for all for a given SNR (Signal to Noise Ratio). All experiments are made with a common SNR, but a lower SNR would allow smaller blocks and consequently a different set of parameters. 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 $\varepsilon = 1$. Thanks to Theorem 1 this means that it is expected to find one false alarm in average for images with 10^6 pixels. Then, fixing ε makes the method into a parameterless method for all moderately sized images.

III. The Self-Similarity Threshold

Urban 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 also the most significant. 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 do not have 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:

$$d(B_{\mathbf{q}}, B_{\mathbf{q}'}) < min\{d(B_{\mathbf{q}}, B_{\mathbf{r}}) |\, r \in I \cap S(\mathbf{q})\} \tag{6}$$

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. As noted earlier, the search for correspondences can be restricted to the epipolar line. This is why the automatic threshold is restricted to $S(\mathbf{q})$. The distance used in the self-similarity threshold is the sum of squared differences (SSD) of all the pixels in the block and the block size is the same than the block size use for ACBM.

Computing the similarity of matches in one of the images is not a new idea in stereovision. In [20] the authors define the *distinctiveness* of an image point $\mathbf{q}$ as the perceptual distance to the most similar point other than itself 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 ambiguous points correctly by matching intrinsic curves [39]. 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 [34], the ambiguous correspondences should be rejected. In this work a new *stability property* is defined. This property is one 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. The comparison of this last algorithm with our results will be done in section IV.

A. 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. 6). 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 algorithm combined with the self-similarity is not sufficient: many false matches are accepted.

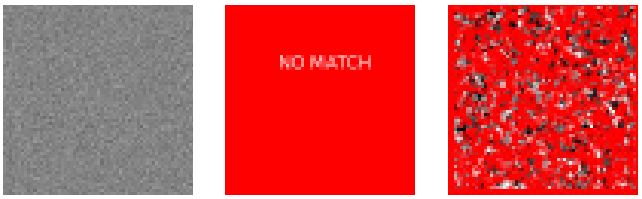


Fig. 6. (a) Reference noise image. (b) No match at all has been accepted by the *a contrario* test! (c) Many false correspondences have been accepted by the self-similarity threshold.

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. 7) which has a large baseline and therefore an important number of occluded pixels. ACBM gives again the best result (see Table I). 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. 8, 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 to reject matches in this region of the image.

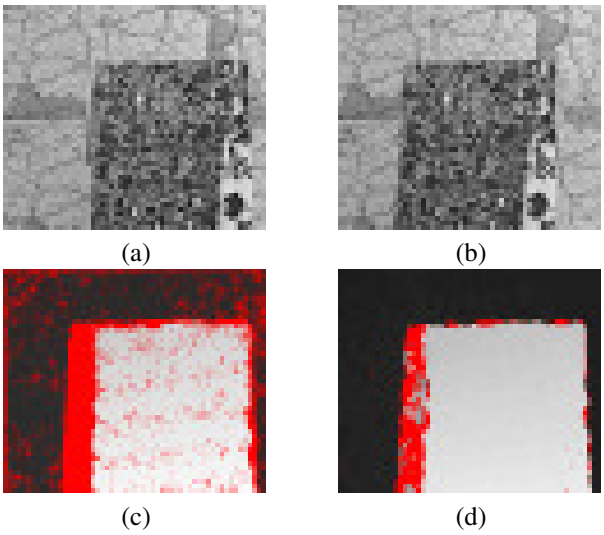


Fig. 7. (a) Reference image (b) Secondary image. The rectangular object occludes part of the background (c) The *a contrario* test does not accept any match for pixels in the occluded areas. (d) With the self-similarity threshold the disparity map is denser, but wrong disparities remain in the occluded region.

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

TABLE I
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.

In short, ACBM and SS are both necessary and complementary. SS only removes a tiny additional number of errors, but even a 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 (6).

IV. COMPARATIVE RESULTS

The algorithm parameters are identical for all experiments throughout this paper. The comparison window size is 9×9, the number of considered principal components is 9, the number of quantum probabilities is 5. The previous section showed how the proposed method (ACBM + SS)

Fig. 8. (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.

deals with noise, occlusions and repeated structures. The detection method is also adapted to quasi-simultaneous stereo from aerial or satellite images, where moving objects (cars, pedestrians) are a serious disturbance. Essentially, this is the same problem as the occlusion problem, but the occlusion is caused by camera motion in presence of a depth difference instead of object motion. Figure 9 shows a stereo pair of images of the city of Marseille (France). In both cases, several cars have changed position between the two images. They are duly detected. The shadow regions, which contain more noise than signal, have also been rejected. We have also compared our results with the Kolmogorov's graph cut implementation [16] which rejects *a posteriori* incoherent matches and are labeled as occlusions. In these examples, graph cuts is able to reject some mismatches due to the moving objects in the scene but a lot of conspicuous errors remain in the final disparity map. Likewise, OpenCV's stereo matching algorithm [8] fails completely on this kind of pairs, even though it obtains correct results in more simple examples like the one in figure 7.

The proposed algorithm will now be compared with the non-dense algorithms of [34], [40], [41] and [24], whose aims are comparable. All of these papers have published experimental results on the first Middlebury dataset [35] (Tsukuba, Sawtooth, Venus and Map pair of images), on the non-occluded mask. These four algorithms compute sparse disparity maps and propose techniques rejecting unreliable pixels. We also show some additional comparison with the block matching method implemented in the OpenCV library version 2.2.0 [8], because it is possibly the most widely used one since it comes close to real-time performance.

The authors of [24] compute an initial classic correlation disparity map and select correct matches based on the support these pixels receive from their neighboring candidate matches in 3D after tensor voting. 3D points are grouped into smooth surfaces using color and geometric information and the points which are inconsistent with the surface color distribution are removed. The rejection of wrong pixels is not complete, because the algorithm fails when some objects appear only in one image, or when occluded surfaces change orientation. A variation of the critical rejection parameters can lead to quite different results.

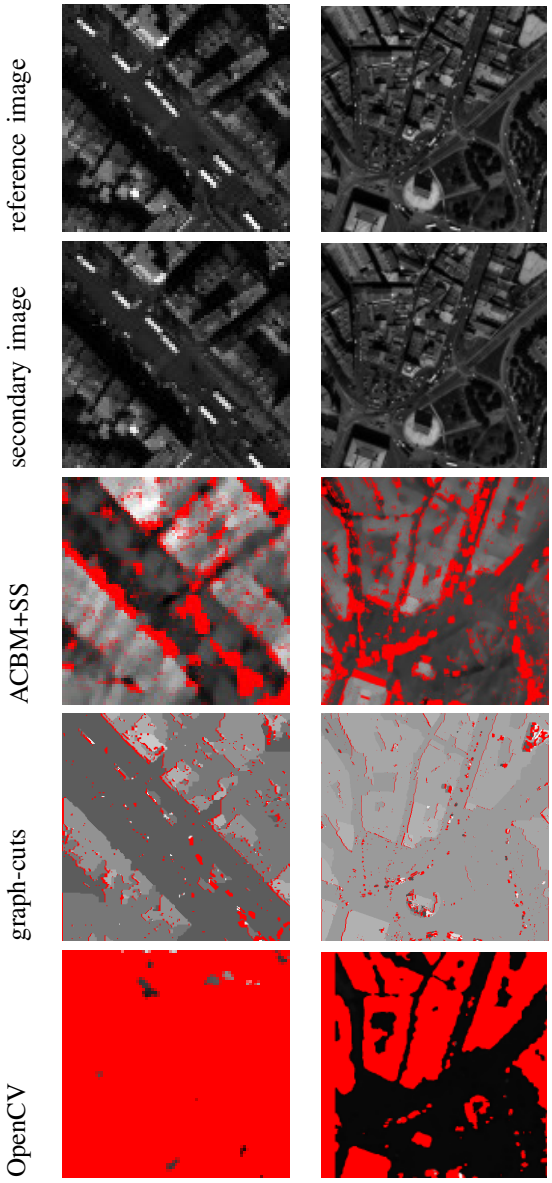


Fig. 9. From top to bottom: reference image, secondary image, ACBM+SS disparity map, graph cuts disparity map, and OpenCV disparity map. In our disparity map, red points are points which haven't been matched. Notice that patches containing a moved car or bus haven't been matched. Poorly textured regions (shadows) where noise dominates have also been rejected. Red points in the graph cuts disparity map are rejected *a posteriori* and considered as occlusions. The graph cuts disparity map is denser and smoother but several mismatches appear in the low textured areas and regions with moving objects.

[40] detects and matches so called "dense features" which consist of a connected set of pixels in the left image and a corresponding set of pixels in the right image such that the intensity edges on the boundary of these sets are stronger than their matching error on the boundary (which is the absolute intensity difference between corresponding boundary pixels). They call this the "boundary condition." The idea is that even the boundary of a non textured region can give a correspondence. Then, each dense feature is associated with a disparity. The main limitation is the way dense features are extracted. They are extracted using a local algorithm which processes each scan line independently from the others. As a

result, top and bottom boundaries are lost. On the contrary, [41] uses graph cuts to extract "dense features" (which of course does not necessarily imply a dense disparity map) thus enforcing the boundary conditions. The results in [40] are rather dense and the error rate is one of the most competitive ones. Yet these good results are also due to the particularly well adapted structure of the benchmark. Indeed, the Sawtooth, Venus and Map scenes consist of piecewise planar surfaces, with almost fronto-parallel surface patches. The ground truth of Tsukuba is a piecewise constant disparity map with six different disparities.

Table II summarizes the percentage of matched pixels (density) and the percentage of mismatches (where the estimated disparity differs by more than one pixel from the ground truth). This table reports first the result of ACBM+SS, whose error rate is very small and yields larger match densities than Sara's results [34]. To compare with other algorithms yielding denser disparity maps, the results of ACBM+SS have been densified by the most straightforward proximal interpolation (a 3×3 spatial median filter). Doing this, the match density rises significantly while keeping small error rates. Still, large regions containing poor textures, typically shadows in aerial imaging, are impossible to fill in because they contain no information at all. Besides the compared algorithms in Table II [14] also published non-dense results for the Tsukuba image (error rate of 2.1% with a density of 45%) but since non-dense results on other images are not published it does not appear in our table.

Fig. 10 compares the ACBM+SS results with opencv, graph cuts and the Sara published results on the classic CMU Shrub pair[2]. Sara's disparity map has several mismatches and the ACBM+SS results are obviously denser. On the other hand, Kolmogorov's graph cut implementation is denser but the mismatches have risen considerably. OpenCVs disparity map is more dense than Kolmogorov's, and less dense than Sara's, but it has also the highest number of wrong matches. So, the proposed algorithm ACBM+SS has a better trade-off between density and mismatches. In the Kolmogorov graph cuts implementation the occlusions are detected, providing a non-dense disparity map. It is clear that detecting occlusions in real images is not enough to avoid mismatches. Another example is shown in Fig. 11, where the almost dense disparity map obtained with graph cuts is compared with the ACBM+SS disparity map. The top left of the image gets by Graph Cuts a completely wrong disparity: the sky and the tree branches are clearly not at the same depth in the scene. This type of error is unavoidable with global methods. The depth of the smooth sky is inherently ambiguous. By the minimization process it inherits the depth of the twigs through which it is seen.

An interesting question arises out of the comparative results about the duality error/density. We have seen that our algorithm gives very low error percentages with densities between 40% and 90%. The parameter ε can be increased but then the error rate will rise. Our goal is to match with high reliably the points between two images and reject any possible false match. So the choice of one expected false alarm ($\varepsilon = 1$) is a conservative choice but ensures a very small error percentage.

Discussion on the other parameters: We have mentioned that the number of considered principal components N and the number of quantum probabilities Q can be increased without noticeable alteration of the results. In practice, the two values are chosen (for computational reasons) to the minimal values not affecting the quality of the result. They are fixed once and for all to $N = 9$ and $Q = 5$ respectively. Another parameter is the search region size $(2R+1)$ but it is easy to find since we only need R to be larger than the largest disparity in the image, which is a classic assumption in stereovision algorithms (in practice R can be estimated from the sparse matching of interest points that was previously obtained for the epipolar rectification step). Finally, the last parameter is the size of the block. We know that very small blocks are affected by image noise but at the same time, the bigger the block, the bigger the fattening error (also named adhesion error). This error becomes apparent at the object borders of the scene causing a dilation of their real size, which is proportional to the block-size. The fattening phenomenon is not the object of the this paper but different solutions have already been suggested to avoid it [5]. Fixing the size of the block to 9×9 seems to be a good compromise between noise and fattening for a realistic SNR conditions, ranging from 200 to 20 (the SNR is measured as the ratio between the average grey level and the noise standard deviation.)

Computational time: For the sake of computational speed, the PCA basis is previously learnt on a set of representative images and stored once and for all.[3] Then, this basis is used to compute all image coefficients. Notice that only the image coefficients of the second image need to be sorted in order to compute the resemblance probability between all possible matches. With our implementation, which is still not highly optimized for speed, an experiment with a pair of images of size 512×512 with disparity rang = [-5,5], takes 4.5 seconds running on a 2.4 GHz Intel Core 2 Duo processor.
A similar experiment with the OpenCV stereo algorithm takes between 5 and 500 miliseconds. This is much closer to real-time requirements, but results are also much more data-dependent, producing good results in easy examples like the Middlebury pair, but much less dense and less reliable results than our method in more difficult scenes like shrub, marseille or even the stereo pairs provided with OpenCV.

V. Conclusion

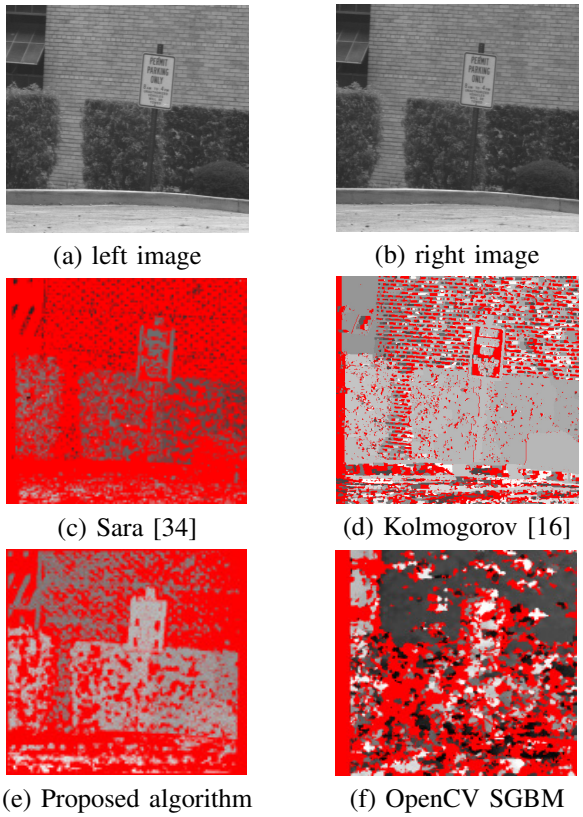
Wrong match thresholds were, in our opinion, the principal drawbacks for block-matching algorithms in stereovision. The *a contrario* block-matching threshold, that was the principal object of the present paper, combined with the self-similarity threshold is able to detect mismatches systematically, by an algorithm which is essentially parameter-free. Indeed, the only user parameter is the expected number of false matches, which can be fixed once and for all in most applications. The method indiscriminately detects occlusions, moving objects and poor or periodic textured regions.

[2]http://vasc.ri.cmu.edu/idb/html/jisct

[3]In our experience the (computationally intensive) choice of this basis does not significantly affect the results, but the (computationally fast) learning of marginal distributions for a particular image on this basis does.

	Tsukuba		Sawtooth		Venus		Map	
	Error(%)	Density(%)	Error(%)	Density(%)	Error(%)	Density(%)	Error(%)	Density(%)
ACBM + SS	**0.31**	45.6	**0.09**	65.7	0.02	54.1	**0.0**	84.8
ACBM + SS + Median filter	0.33	54.3	0.14	77.9	**0.0**	66.6	**0.0**	93.0
Sara [34]	1.4	45	1.6	52	0.8	40	0.3	74
Veksler 02 [40]	0.38	66	1.62	76	1.83	68	0.22	87
Veksler 03 [40]	0.36	75	0.54	87	0.16	73	0.01	87
Mordohai and Medioni [24]	1.18	74.5	0.27	78.4	0.20	74.1	0.08	94.2

TABLE II
QUANTITATIVE RESULTS ON THE FIRST MIDDLEBURY BENCHMARK DATA SET. THE ERROR STATISTICS ARE COMPUTED ON THE MASK OF NON OCCLUDED PIXELS. ANY ERROR LARGER THAN 1 PIXEL IS CONSIDERED A MISMATCH. ACBM+SS OBTAINS LESS MISMATCHES IN ALL FOUR IMAGES.

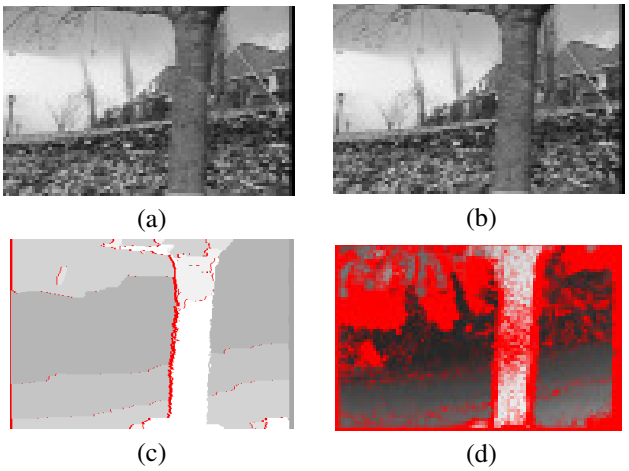


(a) left image (b) right image

(c) Sara [34] (d) Kolmogorov [16]

(e) Proposed algorithm (f) OpenCV SGBM

Fig. 10. CMU Shrub scene. (a) and (b) Reference and secondary images. (c) Method of Sara [34]. Red points are rejected. Density: 24% (d) Kolmogorov's Graph-Cuts [16]. Red points are points detected as occlusions. Density: 77% (e) ACBM+SS. Red points are rejected points. Density: 42%. Sara's disparity map has a lower density and has several evident mismatches. Kolmogorov's disparity map is denser but has many obvious errors. (f) The block matching algorithm included in OpenCV is also not very dense AND contains many errors. It is only provided as a reference of what can be easily obtained with a freely available quasi-real-time block matching algorithm. (e) Proposed method ACBM+SS.



(a) (b)

(c) (d)

Fig. 11. Flower-garden scene. (a) and (b) Reference and secondary images. (c) Graph Cuts (method of [16]). Red points are occluded points. (d) ACBM+SS. Red points are rejected points. Density: 59%. Most rejected points are obviously mismatched by the graph cut algorithm, which equates the depths of trees, sky and house.

Mismatches in block-matching have led to the overall dominance of global energy methods. However, global methods have no validation procedure, and the proposed *a contrario* method must be viewed as a validation procedure, no matter what the stereo matching process was. Block-matching, together with the reliability thresholds established in this paper, gives a fairly dense set of reliable matches (from 50% to 80% usually). It may be objected that the obtained disparity map is not dense.

This objection is not crucial for two reasons. First, having only validated matches opens the path to benchmarks based on accuracy, and to raise challenges about which precision can be ultimately attained (on *validated* matches only). Second, knowing which matches are reliable allows one to complete a given disparity map by fusing several stereo pairs. Since disposing of multiple observations of the same scene by several cameras and/or at several different times is by now a common setting, it becomes more and more important to be able to fuse 3D information obtained from many stereo pairs. Having almost only reliable matches in each pair promises an easy fusion. A straightforward solution in our case would be the following: Given $m > 2$ images, the disparity map between each possible pair of images is computed with ACBM+SS. Then the final disparity map is the accumulated disparity map considering all meaningful matches computed with all the image pairs whenever all the computed disparities for the same pixel are coherent.

VI. ACKNOWLEDGEMENTS

The authors thank Pascal Getreuer for helpful comments on this work. Work partially supported by the following projects FREEDOM (ANR07-JCJC-0048-01), Callisto (ANR-09-CORD-003), ECOS Sud U06E01 and STIC Amsud (11STIC-01 - MMVPSCV).

REFERENCES

[1] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.

[2] N. Burrus, T. M. Bernard, and J.-M. Jolion. Image segmentation by a contrario simulation. *Pattern Recognition*, 42(7):1520–1532, July 2009.

[3] F. Cao. Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualisation in Science*, 7:3–13, 2004.

[4] F. Cao, J.-L. Lisani, J.-M. Morel, and P. Musé F. Sur. *A Theory of Shape Identification*. Springer, 2008.

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

[6] A. Desolneux, L. Moisan, and J.M. Morel. *From Gestalt Theory to Image Analysis. A probabilistic Approach*. Springer, 2007.

[7] G. Egnal and R. P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Transactions of Pattern Analysis and Machine Intellignece*, 24(8):1127–1133, 2002.

[8] Victor Eruhimov. stereo_match.cpp sample program. provided with OpenCV software library version 2.2.0 http://opencv.willowgarage.com/wiki/, January 2010.

[9] S. Forstmann, Y. Kanou, J. Ohya, S. Thuering, and A. Schmitt. Real-time stereo by using dynamic programming. *In Conference on Computer Vision and Pattern Recognition Workshop*, 3:29–36, 2004.

[10] W.E.L. Grimson and D.P. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(12), 1991.

[11] C. G. Harris and M. Stephens. A combined corner and edge detector. *4th Alvey Vision Conference*, pages 147–151, 1988.

[12] L. Igual, J. Preciozzi, L. Garrido, A. Almansa, V. Caselles, and B. Rougé. Automatic low baseline stereo in urban areas. *Inverse Problems and Imaging*, 1(2):319–348, 2007.

[13] H. Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions of Pattern Analysis and Machine Intellignece*, 25(10):1333–1336, 2003.

[14] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. *Computer Vision Pattern Recocgition*, 1:103–110, 2001.

[15] A. Klaus and K. Sormann, M.and Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the International Conference on Pattern Recognition*, pages 15–18, 2006.

[16] V. Kolmogorov and R. Zabih. *Graph Cut Algorithms for Binocular Stereo with Occlusions*. Mathematical Models in Computer Vision: The Handbook, Springer-Verlag, 2005.

[17] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.

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

[19] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

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

[21] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:257–263, 2003.

[22] A. Mittal and N. Paragios. Motion-based background subtraction using adaptive kernel density estimation. *Computer Vision and Pattern Recognition*, 2:302–309, 2004.

[23] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218, 2004.

[24] P. Mordohai and G. Medioni. Stereo using monocular cues within the tensor voting framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:968–982, 2006.

[25] 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.

[26] P. Musé, F. Sur, and J.-M. Morel. Sur les seuils de reconnaissance des formes. *Traitement du Signal*, 20(3):279–294, 2003.

[27] G. Née, S.e Jehan-Besson, L. Brun, and M. Revenu. Significance tests and statistical inequalities for region matching. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 350–360, 2008.

[28] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):139–154, 1985.

[29] K. A. Patwardhan, G. Sapiro, and V. Morellas. Robust foreground detection in video using pixel layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):746–751, 2008.

[30] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *Proceedings of the European Conference on Computer Vision*, pages 792–805, Berlin, Heidelberg, 2008. Springer-Verlag.

[31] J. Rabin, J. Delon, and Y.Gousseau. A contrario matching of sift-like descriptors. In *International Conference on Pattern Recognition*, 2008.

[32] A. Robin, L. Moisan, and S. Le Hégarat-Mascle. An a-contrario approach for sub-pixel change detection in satellite imagery. Technical Report MAP5 Nro. 2009-15, Université Paris Descartes, 2009.

[33] N. Sabater, J.-M. Morel, A. Almansa, and G. Blanchet. Discarding moving objects in quasi-simultaneous stereovision. In *IEEE International Conference on Image Processing, ICIP*, 2010.

[34] 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.

[35] 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.

[36] 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.

[37] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[38] C.V. Stewart. Minpran: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938, 1995.

[39] 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.

[40] O. Veksler. Dense features for semi-dense stereo correspondence. *International Journal of Computer Vision*, 47(1-3):247–260, 2002.

[41] O. Veksler. Extracting dense features for visual correspondence with graph cuts. In *Computer Vision and Pattern Recognition*, volume 1, pages 689–694, 2003.

[42] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2354, 2006.

Neus Sabater received the BSc degree (2005) from Universitat de Barcelona, Spain, and the MSc degree (2006) and PhD degree (2009) in image processing from the Ecole Normale SupÃrieure de Cachan, France. She is currently a postdoctoral researcher at the California Institute of Technology, USA.

Andrés Almansa is currently a CNRS Research Scientist at Telecom ParisTech. He received his HDR, PhD and MSc/Engineer degrees in Applied Mathematics and Computer Science from Université Paris-Descartes, ENS Cachan (France) and Universidad de la República (Uruguay) respectively, where became Associate Professor in 2004. His current research interests include image restoration and analysis, subpixel stereovision and applications to earth observation, high quality digital photography and film restoration. Most of his recent research has focused on a long term collaboration with the French Space agency (CNES), which strives towards the development of mathematical tools for designing more accurate earth observation and photogrammetric systems.

Jean-Michel Morel was born France in 1953. He received the PhD degree in applied mathematics from University Pierre et Marie Curie, Paris, France in 1980. He started his career in 1979 as assistant professor in Marseille Luminy, then moved in 1984 to University Paris-Dauphine where he was promoted professor in 1992. He is Professor of Applied Mathematics at the Ecole Normale Supérieure de Cachan since 1997. His research is focused on the mathematical analysis of image analysis and processing. He has coauthored with S. Solimini a book on *Variational Methods in Image Segmentation* (Birkhäuser 1994). He has also co-authored with Agnès Desolneux and Lionel Moisan *From Gestalt theory to image analysis: a probabilistic approach* (Springer, 2008), and is also coauthor of *A Theory of shape identification* (Springer LNM, 2008).