www.ietdl.org

Published in IET Computer Vision
Received on 30th November 2010
Revised on 9th March 2012
doi: 10.1049/iet-cvi.2010.0212

ISSN 1751-9632

Multiresolution energy minimisation framework for stereo matching

Á. Arranz Á. Sánchez M. Alvar

ICAI School of Engineering, Comillas Pontifical University, Alberto Aguilera 23, 28015 Madrid, Spain
E-mail: alvaro.arranz@iit.upcomillas.es

Abstract: Global optimisation algorithms for stereo dense depth map estimation have demonstrated how to outperform other stereo algorithms such as local methods or dynamic programming. The energy minimisation framework, using Markov random fields model and solved using graph cuts or belief propagation, has especially obtained good results. The main drawback of these methods is that, although they achieve accurate reconstruction, they are not suited for real-time applications. Subsampling the input images does not reduce the complexity of the problem because it also reduces the resolution of the output in the disparity space. Nonetheless, some real-time applications such as navigation would tolerate the reduction of the depth map resolutions (width and height) while maintaining the resolution in the disparity space (number of labels). In this study a new multiresolution energy minimisation framework for real-time robotics applications is proposed where a global optimisation algorithm is applied. A reduction by a factor R of the final depth map's resolution is considered and a speed of up to 50 times has been achieved. Using high-resolution stereo pair input images guarantees that a high resolution on the disparity dimension is preserved. The proposed framework has shown how to obtain real-time performance while keeping accurate results in the Middlebury test data set.

1 Introduction

Stereo reconstruction has been intensively studied during the last decade. Stereo vision is considered to be more versatile and cheaper than other reconstruction methods. Recently, research has focused on dense depth map estimation because of its interesting application in view synthesis and image-based rendering [1, 2]. Although sparse feature reconstruction algorithms have been intensively studied in the past, especially for real-time applications, dense depth maps generally give much more information about the scene structure than sparse features. The challenge is to obtain accurate dense depth maps while running in real time.

Nowadays, retrieving the three-dimensional structure of a given scene has many applications. Some of them are oriented to computer graphics such as view synthesis [1], monument reconstruction and rendering, urban reconstruction [3, 4] and so on. In all these applications a high-resolution dense depth map is required in order to achieve an acceptable accuracy in the results. This high definition depth estimation is mainly used to produce some kind of high definition multimedia content. Generally, these applications do not need to be computed in real time, so off-line computation and more time-consuming algorithms may be used. Nevertheless, other applications are not oriented to high definition multimedia content, but to intelligent systems such as robot navigation and simultaneous localisation and mapping (SLAM). These applications generally need to create a general approximation of the structure of the environment, but no high definition content is needed. In these cases, the estimation of a low definition depth map is enough. Usually, these applications need to be computed in real time, hence fast stereo algorithms are mandatory.

For real-time-oriented applications, several solutions have been proposed in the literature. Traditionally, the most successful ones are those based on local methods, which are generally very computationally efficient. Although a lot of variations have been proposed, these methods have low accuracy as their main drawback. As a modification of the local methods, and in order to reduce their computational cost, several authors proposed to apply multiresolution methods [5–7]. Although these multiresolution algorithms use information from coarse resolutions to guide the search in the fine ones, the presented approach is totally different. This method, called multiresolution minimisation framework, analyses the disparities in a fine resolution image but the output depth map is of a coarser one. Moreover, this framework can be used with the well-known Markov random fields (MRFs) model, so more accurate results can be achieved compared with the local methods. A coarser resolution for the disparity image is used by other devices such as rgb-d devices or Kinect, and they still have plenty of applications.

In this paper a new energy minimisation framework applied to the MRFs is proposed where the size of the depth map obtained can differ from the size of the reference image. Using this framework has some interesting advantages. Firstly, the information contained in the high-resolution input is used to maintain a high resolution in the disparity space. Secondly, the size of the problem is reduced depending on the reduction factor applied to the final depth map. Consequently, the performance of the algorithm is enhanced.

In Section 2 an overview of the most popular dense stereo algorithms is presented. In Section 3 the classical energy minimisation framework is reviewed and a new multiresolution energy minimisation framework is proposed. In Section 4 the main results are presented and in Section 5 some conclusions are drawn.

2 Dense two-frame stereo algorithms overview

Dense stereo is one of the most researched topics in recent computer vision. Many overviews and surveys that analyse and compare most popular algorithms have been published. One of the most important surveys is the taxonomy and evaluation made in [2], where a thorough state of the art of dense two-frame algorithms and their comparison is presented. According to this taxonomy, the dense two-frame stereo algorithms can be classified into two groups: local methods and global optimisation methods.

2.1 Local and global stereo

Local methods estimate the depth for each pixel independently, only using the information surrounding them. Typically, they are simple algorithms that perform some kind of correlation based on photometric properties over a support window. They compute a cost for each available disparity and select for each pixel independently the disparity configuration that achieves a lower cost. The great advantage of these algorithms is that they usually can be computed in parallel and are much faster than the global optimisation ones [1, 2]. In fact, nowadays, local methods are the most popular for real-time applications. The most important algorithms that are included in this group are square window, shiftable window, boundary guided and adaptive weight [8]. All of them are reviewed and compared in [9]. The comparison is made in terms of accuracy and performance.

Global optimisation methods estimate the depth of each pixel of the reference image considering the depth and the cost value of every other pixel in the image. Global algorithms can be classified into three different groups: MRFs based, dynamic programming and cooperative algorithms. The MRF-based methods [10] are solved using an energy minimisation framework: an energy function that depends on depth estimation is defined and a minimum energy configuration is searched. The purpose of these algorithms is to obtain a disparity configuration that minimises the aforementioned energy function. This approach is very popular because it can be justified in terms of maximum a posteriori estimation of an ‘MRF’. During the last decade, many algorithms have been proposed to solve this NP-hard optimisation problem. Simulated annealing was the first algorithm used for solving them but it was demonstrated to be too computationally demanding and did not perform well in terms of accuracy. Recently, other algorithms have been presented such as iterated conditional modes (ICM) [11], expansion, Swap [12] and belief propagation [13]. A comparison of their accuracy and performance is studied in [14, 15]. Recently, parallel versions of the graph cuts and belief propagation have been successfully implemented in GPGPUs [16–18]. The main drawback of these algorithms is that generally powerful GPGPUs are not available for real-time applications such as robotic navigation. Modifications to the graph-cut algorithms for performance enhancement have been recently proposed in [19–23]. These techniques are complementary to the framework proposed in this paper, so they can be applied as the optimisation algorithm.

The dynamic programming [24–27] approach optimally solves each scanline independently. For this reason, no coherence is enforced between different scanlines. The size of the problem is reduced considerably and near real-time performance can be achieved. The accuracy of the solutions is similar to the ones obtained with local methods [2]. Cooperative algorithms [28, 29] use local non-linear operations that finally result in global optimisation behaviour. They achieve good accuracy but with high runtime.

On the one hand, global methods generally obtain better results than local methods especially in the discontinuity areas; on the other hand, global methods require more computational resources than local methods and are not suited for real-time applications.

In Middlebury’s website (http://vision.middlebury.edu/stereo/) a complete accuracy analysis for many state of the art algorithms is presented. The methods that achieve the best results are mainly based on global algorithms solved using belief propagation [30–32]. Apart from the basic algorithms mentioned earlier, every method presented in the evaluation includes other processing algorithms that improve the disparity accuracy. Usually, some kind of preprocessing algorithm is applied, such as image segmentation, which helps to obtain better disparity estimation.

2.2 Multiresolution approaches

Using multiresolution techniques in the two-frame stereo correspondence problem is not novel. The early work in [33] proposes a coarse-to-fine pyramidal approach applied to airborne imagery. Different resolution levels are created where coarser layer’s analysis is used to constraint the search in the finer ones. Owing to the reduction of search space and the parallelisation capabilities they report an enhancement of both accuracy and performance. In [34], the authors propose to use coarser resolution for objects that are near the camera and only apply fine resolution for objects that are far away. This approach has the main drawback that for robotic applications it is generally preferable to obtain high disparity resolution for near objects (potential obstacles) than for the rest of the objects in the environment. In [35], a similar approach to [33] is presented but information about edges is incorporated. Recently, a couple of interesting multiresolution algorithms have been successfully implemented in GPGPUs. In [6], some implementation-based modifications to the pyramidal approach are made in order to achieve real-time performance on commodity graphics hardware. In [5], a pyramidal algorithm is also implemented in a GPGPU which but uses adaptive windows local search and foreground detection.

Every multiresolution algorithm presented in the literature suffers the same problem: coarser pyramid levels also lead to coarser resolution in the disparity dimension (number of disparity labels, Fig. 1) and that can lead to bad correspondence that propagates to the finer levels. Hence, if high resolution in the disparity dimension is needed, the whole fine layer must be analysed. They also have the main drawback that only local methods are used, which leads to

doi: 10.1049/iet-cvi.2010.0212

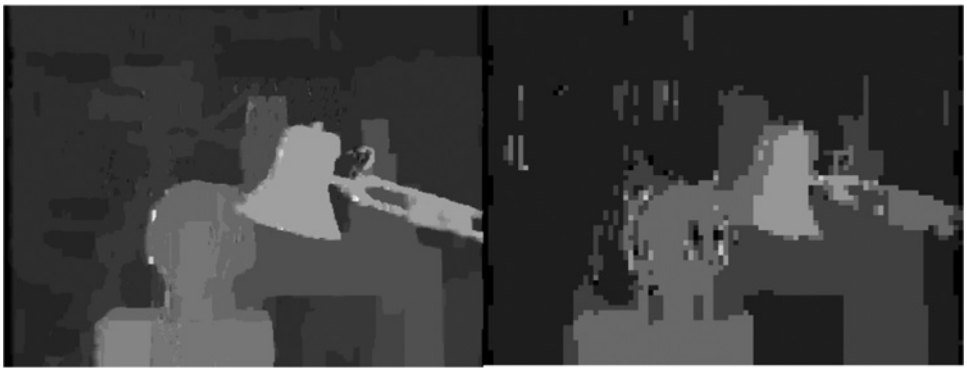

Fig. 1 *Quantisation error because of low-resolution image input*

less accurate results as stated in [2]. The only multiresolution algorithm applied to a global optimisation algorithm can be found in [36], but simulated annealing is used, which is an outdated approach compared with graph-cuts or belief propagation.

In this paper a completely different multiresolution approach is presented. Although other methods construct a pyramid from coarse to fine resolutions, our framework uses only two different resolutions: the one of the input stereo images and the one desired for the dense disparity depth map. The main contribution of this paper is developing a framework for reducing the final depth map resolution (width and height) while preserving the disparity resolution (number of disparity labels). The disparity pixels for the low-resolution depth map are modelled as an MRF, which considerably reduces the complexity of the problem. In contrast, the disparity range and the data term are preserved from the high-resolution images in order to achieve the same disparity resolution as the high-resolution input. The reduction of the depth map's resolution provides a great improvement in performance because of the great reduction in the complexity of the problem.

3 Energy minimisation frameworks

In this section the classical energy minimisation framework is reviewed and a new multiresolution energy minimisation framework that reduces the complexity of the problem and at the same time maintaining accuracy is proposed.

3.1 Classic energy minimisation framework

As described in [10], the stereo matching problem can be stated as the minimisation of an energy function that includes two different terms. The main objective of the global optimisation algorithms is to find the set of labels $\bar{f} = \{f_1, \ldots, f_n\}$, where n being the number of pixels in the reference image, that minimises the energy function $E(\bar{f})$. In the case of stereo correspondence, each label f_n can take any value of the discretised disparity space $\Lambda = 1, \ldots, L$, L being the value of the maximum disparity analysed

$$\min E(\bar{f}) \tag{1}$$

$$E(\bar{f}) = E_{\text{data}}(\bar{f}) + E_{\text{smooth}}(\bar{f}) \tag{2}$$

The $E_{\text{data}}(\bar{f})$ term measures the quality of the stereo correspondence between pixels of different images based on photometric similarity

$$E_{\text{data}}(\bar{f}) = \sum_{p\in H} D_p(f_p) \tag{3}$$

where H is the set of all pixels in the reference image. The $D_p(f_p)$ function describes how good the label f_p fits for a certain point p. In the case of stereo correspondence, it usually measures how different is the pixel corresponding to the point p in the reference image compared with the pixel matched with it in the other image of the stereo pair. The pixel matched to p is determined by the label f_p. In the typical situation, where the left image is the reference image, an expression for the $D_p(f_p)$ function could be

$$D_p(f_p) = \sum_{c\in C} \text{abs}(I_{\text{L}}(x(p), y(p), c) - I_{\text{R}}(x(p) - f_p, y(p), c)) \tag{4}$$

where I_{L} and I_{R} are the stereo pair images, C is the set of channels included in them and $x(p)$ and $y(p)$ are the x and y coordinates of the point p, respectively. It is important to note that this similarity measurement is very sensitive to bias difference between the two images. This difference is very common because of the different lenses and sensors used in the stereo pairs. However, this colour and shape issue can be largely overcome through a calibration process. If previous calibration between the stereo pair is not possible, other similarity measurements have been proposed. For example, Klaus *et al.* [30] use a combination of both gradient and absolute colour difference. Another important contribution was published in [37], where a colour similarity measurement that is insensible to image sampling is proposed.

The $E_{\text{smooth}}(\bar{f})$ term determines how smoothly the labels change throughout the disparity image

$$E_{\text{smooth}}(\bar{f}) = \sum_{\{p,q\}\in N} V_{\{p,q\}}(f_p, f_q) \tag{5}$$

N being the set of neighbouring pixels defined by

$$p = (x, y), \quad q = (i, j), \quad p, q \in H \tag{6}$$

$$p, q \in N \Leftrightarrow |x - i| + |y - j| = 1 \tag{7}$$

The $V_{\{p,q\}}(f_{\text{p}}, f_q)$ function measures how probable the label f_p is for the pixel p considering that its neighbouring pixel q has been assigned the label f_q. As for the $D_p(f_p)$ function, many functions for $V_{\{p,q\}}(f_p, f_q)$ have been proposed. One of the most popular ones is

$$V_{\{p,q\}}(f_p, f_q) = \min(K, |f_p - f_q|) \tag{8}$$

where K is a maximum constant limit for the cost function.

The complexity of the classical optimisation problem is determined by both the number of labels considered and the size of the reference image. It can be seen in the previous equations that the smoothing term and the data term have to be evaluated for each pixel in the reference image. Thus, high-resolution input images imply increasing the problem size and, as a result, higher runtimes are needed. Nevertheless, low-resolution images reduce the complexity of the problem but they also reduce the resolution in the disparity dimension (the disparity space Λ is reduced compared with high-resolution images).

3.2 Multiresolution energy minimisation framework

The aim of the new formulation presented in this section is to reduce the complexity of the problem in order to enhance the performance of the optimisation algorithm, while maintaining the accuracy of the results. The new formulation is very similar to the classical optimisation presented in the previous section, reducing the number of evaluations needed in both terms of the energy function. This reduction will be achieved by the introduction of a parameter R, which will also determine the final resolution of the disparity map. If R is increased, the performance of the algorithm will be improved but the resolution of the final disparity map will be decreased, involving a less accurate reconstruction.

As in the classical energy minimisation framework, an energy function made of a data term and a smoothing term is minimised

$$\min E'(\bar{\boldsymbol{f}}) \tag{9}$$

$$E'(\bar{\boldsymbol{f}}) = E'_{\text{data}}(\bar{\boldsymbol{f}}) + E'_{\text{smooth}}(\bar{\boldsymbol{f}}) \tag{10}$$

The final width w_d and height h_d of the disparity image, which is reduced by a ratio R compared with the original input images, is defined by

$$w_\text{d} = \text{floor}\left(\frac{w_\text{l}}{R}\right), \quad h_\text{d} = \text{floor}\left(\frac{h_\text{l}}{R}\right) \quad \forall R \in \mathbb{R},\ R \geq 1 \tag{11}$$

w_l and h_l being the width and height of the original left input image.

A new subset of pixels is defined for the new resolution coordinate system

$$p \in P \Leftrightarrow p = (x, y), \quad \forall x, y \in \mathbb{N}/x < w_\text{d}, y < h_\text{d} \tag{12}$$

The purpose of this new resolution for the disparity image is to solve the correspondence problem for a limited subset of pixels and, thus reduce the MRF complexity. The new data term for the multiresolution framework is

$$E'_{\text{data}}(\bar{\boldsymbol{f}}) = \sum_{p \in P} D'_p(f_p) \tag{13}$$

$$D_p(f_p) = \sum_{c \in C} \text{abs}(\psi(I_L, x(p)R, y(p)R, c) - \psi(I_R, x(p)R - f_{x,y}, y(p)R, c)) \tag{14}$$

where ψ is an interpolation function such as bilinear or bicubic functions. The most important thing to note in the new data term is that it uses the high-resolution input images I_L and I_R for generating a cost associated with each disparity level and for each point p of the low-resolution disparity image. This reduction from the H set to the P subset reduces in R^2 times the size of the corresponding MRF. Note that the number of labels is not reduced in this model. Thus, the depth resolution remains the same as in the classic optimisation framework.

Finally, the smoothing term is formulated as in the classical energy minimisation framework in the new coordinate system

$$E'_{\text{smooth}}(\bar{\boldsymbol{f}}) = \sum_{\{p,q\} \in M} V_{\{p,q\}}(f_p, f_q) \tag{15}$$

$$p = (x, y), \quad q = (i, j), \quad p, q \in P \tag{16}$$

$$p, q \in M \Leftrightarrow |x - i| + |y - j| = 1 \tag{17}$$

The reduction in the size of the problem depends on the value of the parameter R. Solving the data and smoothing terms using the presented framework is equivalent to solving the high-resolution problem with the same number of labels only for a subset of pixels of the whole reference image.

3.3 Subsampling issues

The multiresolution minimisation framework proposed in this section is based on the idea that subsampling the stereo input that is reducing it by a R factor does not substantially affect the informational content of the image. Obviously, subsampling generally involves a loss of information and, in certain situations, may absolutely change the image content. We have analysed the effects of subsampling in the disparity space and in the image space independently, always focusing on the real-time robotics application.

3.3.1 Determining the resolution in disparity space: Firstly, the resolution of the stereo inputs I_L and I_R will determine the maximum final resolution in disparity space. For a typical stereo camera with rectified images the formulas that determine this resolution are the following

$$Z = \frac{fB}{d} \tag{18}$$

where f is the focal length of the camera, B is the baseline, Z is the distance to the camera and d is the disparity. If a certain depth resolution ΔZ is required at a certain distance Z, the resolution needed in the image sensor would be

$$(d_1 - d_2) = \frac{fB}{Z} - \frac{fB}{Z + \Delta Z} = \frac{fB\Delta Z}{Z^2 + Z\Delta Z} \tag{19}$$

Therefore if the sensor is known to have a size pix_{size}, the decimation Dec, which can be applied to the original image, can be computed by the following equation

$$\mathrm{Dec} = \frac{(d_1 - d_2)}{\mathrm{pix}_{\mathrm{size}}} \tag{20}$$

Note that this is a complete decimation of the input images. For a typical stereo camera (i.e. PointGrey Bumblebee2) with parameters $f = 3.8$ mm, $B = 12$ cm, $\mathrm{pix}_{\mathrm{size}} = 4.65$ μm, if $\Delta Z = 3$ cm at $Z = 1$ m is acceptable, for example, for robotic navigation applications, Dec = 2.86 is obtained and a downsampling from 1032 to 361 pixels can be performed.

3.3.2 Determining R: The R parameter of the multiresolution framework establishes the final resolution of the disparity depth map. The relationship between the resolution of an image and the size of the object projected is given by the following equation

$$\mathrm{Obj}_{\mathrm{size}} = \frac{Z\,\mathrm{pix}_{\mathrm{size}} R\,\mathrm{Dec}}{f} \tag{21}$$

where $\mathrm{Obj}_{\mathrm{size}}$ is the minimum object size detectable at distance Z with camera parameters $\mathrm{pix}_{\mathrm{size}}, f$. R is the reduction factor of the model and Dec is the previous decimation of the input images. Therefore if the same camera of the previous example is used and a decimation of Dec = 2, a reduction of $R = 3$ and a distance of $Z = 5$ m are selected, a minimum object size of $\mathrm{Obj}_{\mathrm{size}} = 3.67$ cm can be detected.

Thus, the subsampling is limited by the application, the camera used and the surroundings of the system. Note that performing a subsampling in the main reference image using the multiresolution minimisation framework does not modify the resolution of the disparity space, which is the real valuable information.

3.4 Proposed implementation

In the previous section a new energy minimisation framework that incorporates a reduction factor for the final disparity map was described. In this framework, any energy function proposed in the literature can be used. Moreover, any energy minimisation algorithm that is appropriate for solving the classical framework is also appropriate for solving the new energy minimisation framework.

Many energy functions have been proposed in the literature. In order to keep the analysis as simple as possible, the simplest data and smoothing terms have been selected. For data and smoothing terms, (14) and (15) have been chosen, respectively.

The optimisation algorithm was selected based on the comparative study made in [14]. Several energy minimisation methods were analysed and compared, concluding that the alpha-expansion [12] and the TRW-S [38] algorithms obtain the lowest energies. Comparing the performance, expansion seems to be clearly the best between both methods. For these reasons, expansion has been used in all the experiments carried out in this paper.

4 Experimental results

In this section the new multiresolution energy minimisation framework is evaluated and compared with the classical

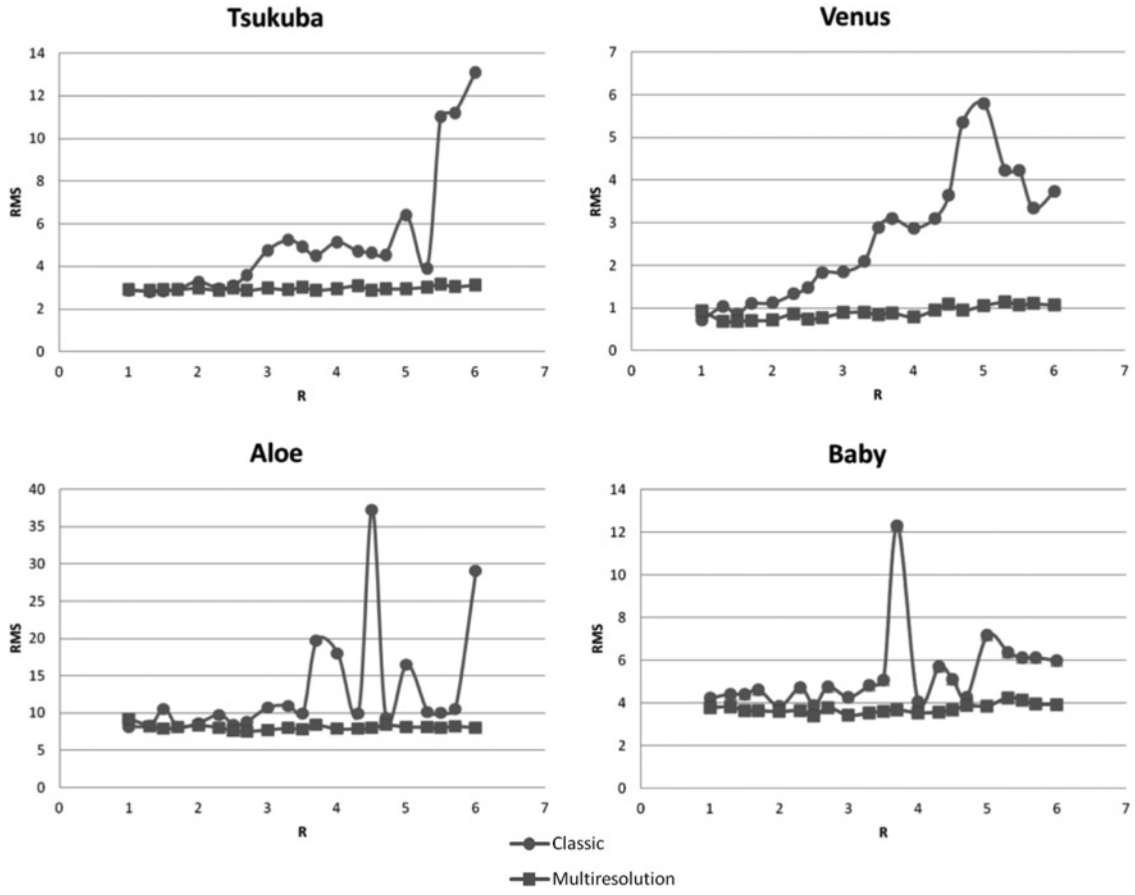

Fig. 2 *Evolution of the RMS error measure when the R factor is increased*

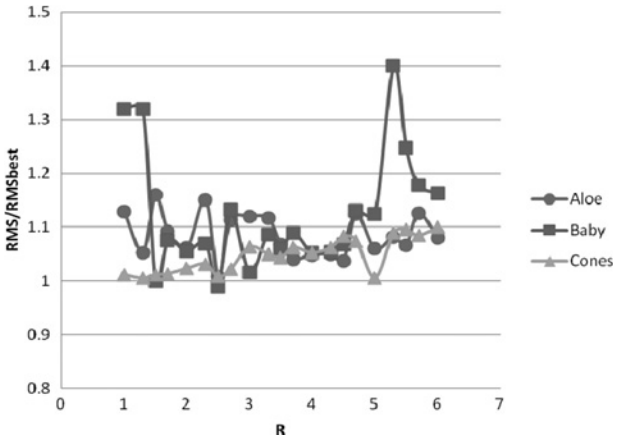

Fig. 3 *Evolution of the RMS error for the multiresolution framework depending on R compared to the best solution found for the classical framework (RMSbest1)*

energy minimisation framework. The data term and the smoothing term used for the energy function in both the new and the classical frameworks are described in Section 3. No preprocessing or refinement is used after the disparity estimation on either case.

The set of test images used for this analysis is the Middlebury data test. They are widely used for stereo research and can be found on Middlebury's website.

The frameworks have been evaluated in terms of accuracy and performance. The quality metrics used are the same as described in [2], the root-mean-squared error (RMS) and the bad pixel percentage (B)

$$\mathrm{RMS} = \left(\frac{1}{N} \sum_{(x,y)} |d_{\mathrm{C}}(x, y) - d_{\mathrm{T}}(x, y)|^2 \right)^{1/2} \qquad (22)$$

$$B = \frac{1}{N} \sum_{(x,y)} (|d_{\mathrm{C}}(x, y) - d_{\mathrm{T}}(x, y)| > \delta_{\mathrm{d}}) \qquad (23)$$

where d_{C} and d_{T} are the disparities calculated and the true disparities, respectively, and N is the total number of pixels. The statistics shown in Figs. 2 and 4 were calculated over occluded regions, textureless regions and regions of discontinuity. As recommended in [2], the δ_{d} term has been set to 1 in all the analysis made. For evaluating the RMS and B error measures, the true disparity image has been subsampled using bilinear interpolation in order to perform a pixel-by-pixel comparison.

The minimisation algorithms used were coded in C++ using the implementation of [14] as a reference. The algorithms were run on an Intel i7-2600 at 3.4 GHz CPU and 7 GB of RAM.

4.1 Accuracy analysis

Several benchmarks have been run in order to compare both frameworks. As stated in the previous subsection, two different statistics have been computed for each image to compare the accuracy of each solution. For each framework several solutions have been reached by using the implementation proposed and through different reduction factors R. The reduction factor for the classical framework

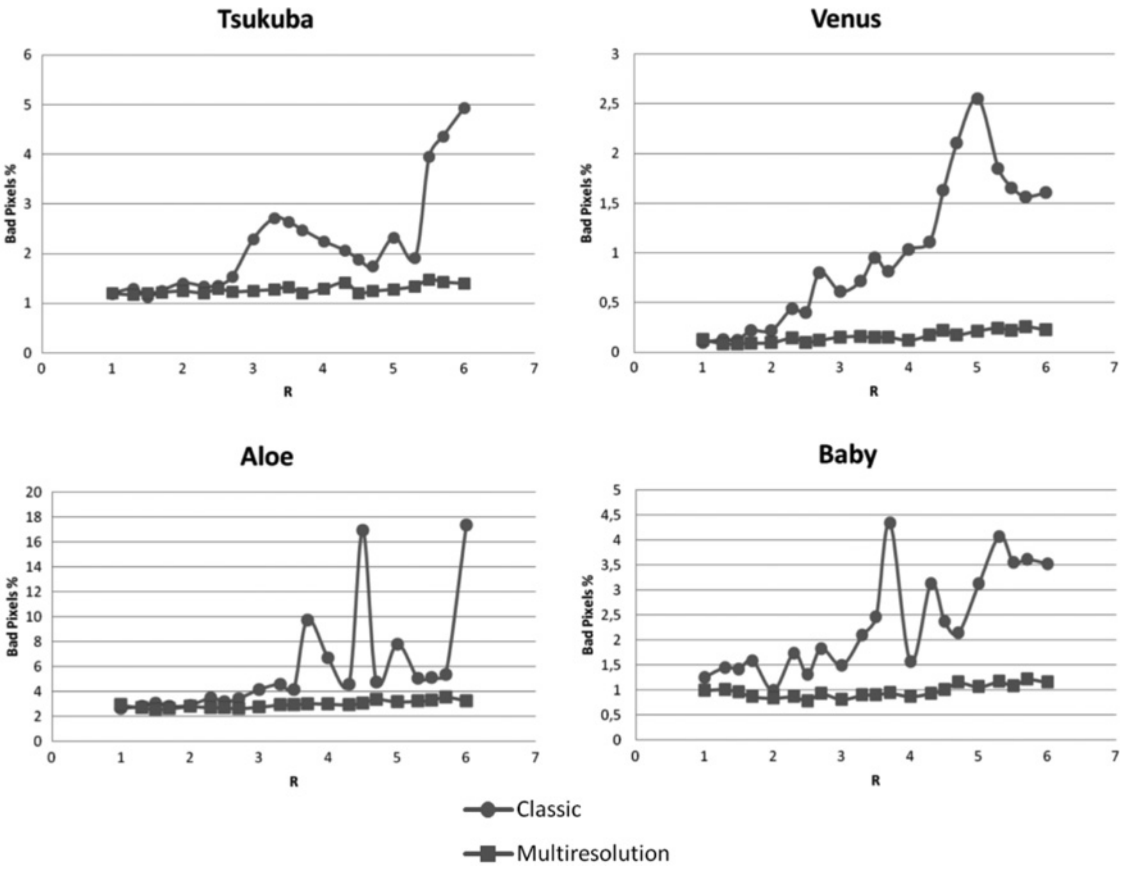

Fig. 4 *Evolution of the bad pixels error measure when R is increased*

will determine the resolution reduction performed over the input images to obtain a disparity map with lower resolution.

In Fig. 2 a benchmark analysis of the RMS error for several test images is shown. The main objective of this figure is to analyse the evolution of the RMS error when the R parameter is increased for both frameworks. For the multiresolution case, R is an internal parameter of the model and has the meaning explained before in Section 3.2. Meanwhile, for the classic framework, the R parameter is used to subsample both stereo input images in order to obtain the same final resolution in the disparity maps as in the multiresolution one. For subsampling, the bilinear interpolation function has been used. For the classic approach, this figure clearly shows that subsampling the input images generally involves an uncontrolled behaviour of the RMS error. Owing to the reduction of the resolution in disparity space, the error depends on the fortune of having the correct labels between the set of labels available in the model. That is, the cyclical increase and decrease of some of the errors is justified as a subsampling issue of disparity space. It is important to note that the best solution found for the classic framework is always the one with the higher resolution, that is, lower R. This result leads to generally prevention of using subsampling in the stereo input if good RMS error and also good resolution in the disparity domain are required, as is the case of robotics navigation. The Venus test case is probably the most important one given that it is made of non-fronto-parallel planes and thus every disparity value exists in the true disparity image. Fig. 2 shows that for a Venus data set, the error in the classic framework increases progressively with R.

In contrast, the multiresolution accuracy stays absolutely controlled in each of the tests made. In most cases, varying the R parameter only supposes an increase of less than 20%

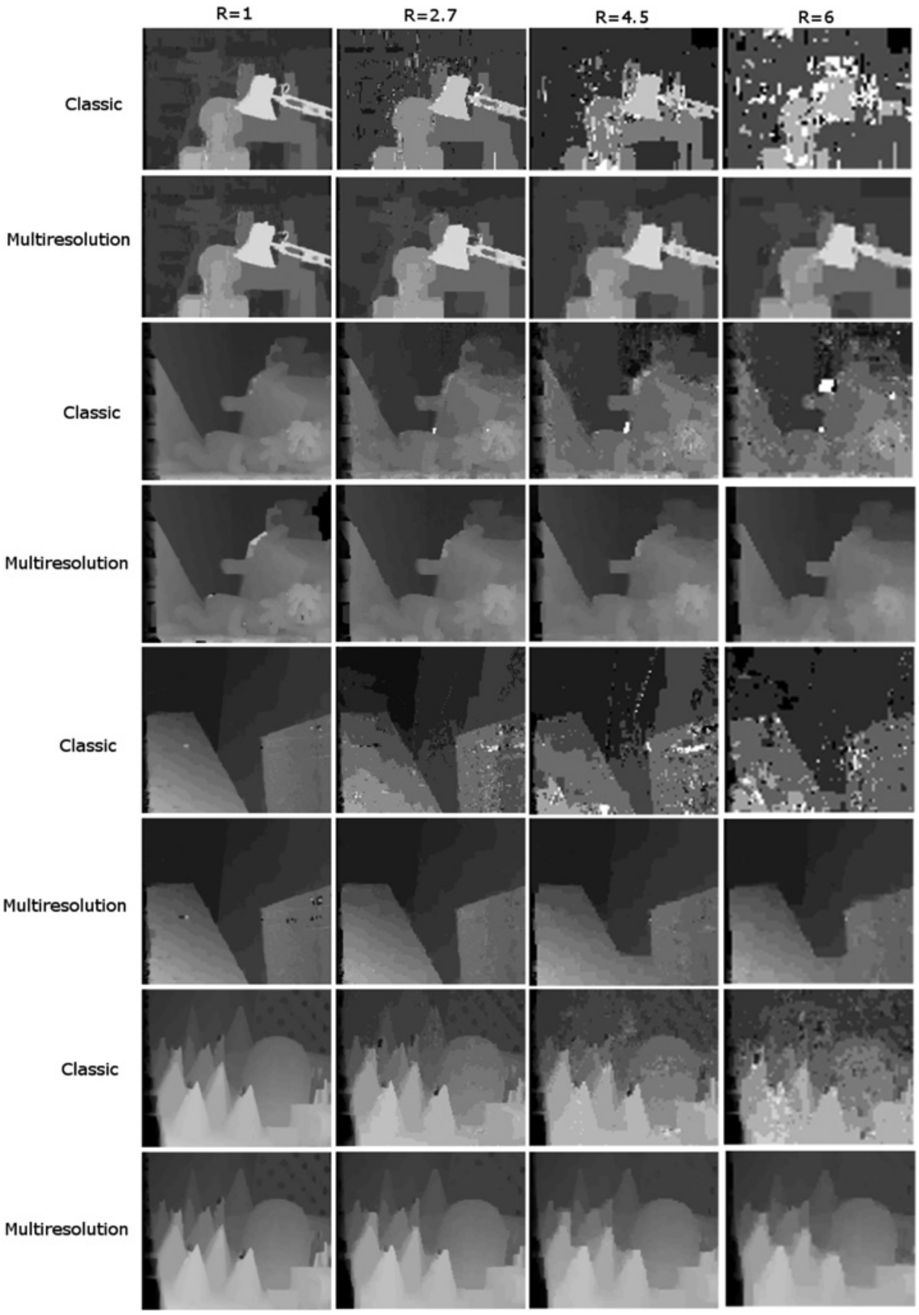

Fig. 5 *Comparison of depth maps obtained by using classic and multiresolution frameworks*

For each data set a solution using $R = 1$, 2.7, 4.5 and 6 is shown, respectively

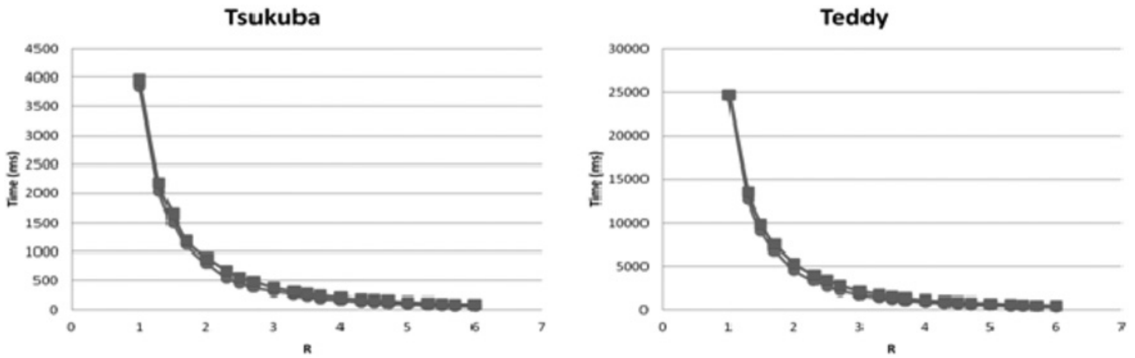

Fig. 6 *Evolution of the computing time depending on R*

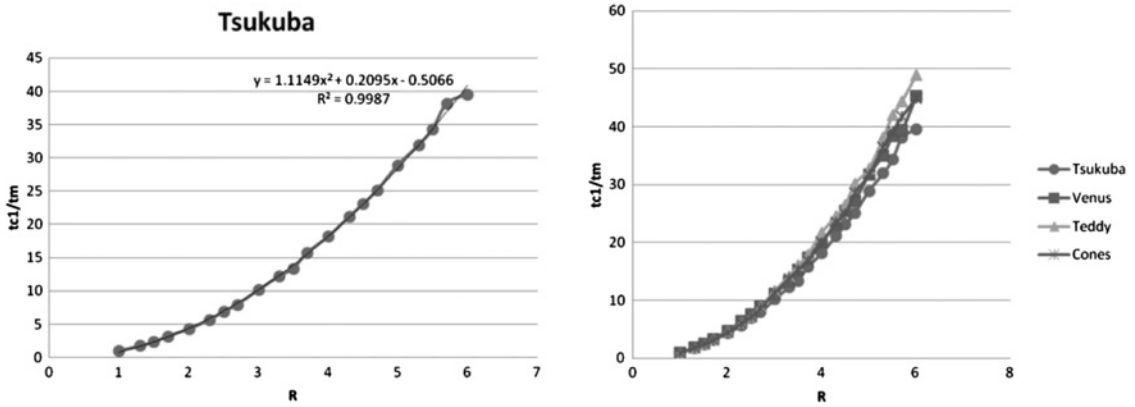

Fig. 7 *Performance enhancement of incrementing R*

T_{c1} is the computing time needed for the classical model to find a solution for $R = 1$. T_m is the computing time for multiresolution model to find a solution for each R on the graph

of the best RMS error found in the classic framework. However, in the worst case the error still stayed under 40% of the best RMS error (Fig. 3). As a result, using the multiresolution framework proposed in this paper guarantees maintaining control on the RMS error of the solution. The evolution of this relative RMS error is shown in Fig. 3. Given that alpha expansion is a stochastic algorithm that does not guarantee finding the global minima, the fluctuation found in this evolution is not surprising. However, the trend of the relative RMS error is to slowly increase with R.

A similar accuracy analysis can be made if the bad pixel percentage error measure is used. Same results as in the RMS case are shown in Fig. 4. The resulting depth maps for various test cases varying R are shown in Fig. 5.

4.2 Performance analysis

For the performance analysis presented in this section, ten iterations of the alpha-expansion algorithm for both classic and multiresolution frameworks have been run.

Firstly, the impact of R in the alpha-expansion performance is analysed. As previously mentioned in the accuracy analysis, R is used to perform a subsampling in the input stereo pair for the classic framework. In Fig. 6 the evolution of the elapsed time for various values of R is presented. Although only the Tsukuba and Teddy data set are actually shown, the same evolution has been observed in all the other data sets. The figure shows that even low R values, such as 1.3, can lead to a dramatic impact on the computational time. For both classic and multiresolution frameworks, a hyperbolic evolution of the performance with R can be deduced. Significant difference has not been found between the classic and the multiresolution framework in this section. As a consequence, for improving the performance of the alpha-expansion algorithm the R parameter should be chosen as high as possible.

Secondly, Fig. 7 shows a comparison between the elapsed time for the best solution using the classic framework and the spent time for the multiresolution case using different R values. This figure shows the time spent by the classic framework divided by the time spent by the multiresolution for both the Tsukuba and Venus data set. In this case the time spent by the classical with the original size as a reference is used. Moreover, the performance increases

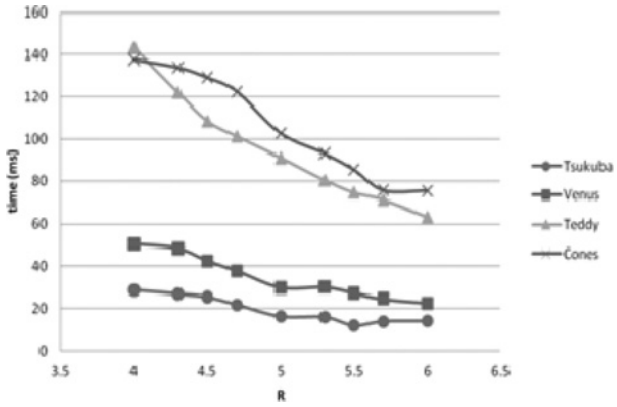

Fig. 8 *Real-time performance analysis for a single iteration of the multiresolution model, varying the R parameter*

monotonically with R, obtaining a speedup of up to 50. A parabolic function close to linear can be deduced from the data (polynomial regression parameters of second order are shown in Fig. 7).

For real-time applications, a maximum of milliseconds is available for stereo analysis. Fig. 8 shows the order of magnitude of computational time needed for high values of R for four different data sets. Both Teddy and Cones stereo input images have a higher bit resolution and much more disparity levels to analyse than the other two data sets, hence an impact on the performance is evident. However, ~15 fps are achieved in the Teddy and Cones case whereas 50 fps are achieved in the two easier ones, Tsukuba and Venus.

5 Conclusions

A new multiresolution energy minimisation framework for stereo matching has been proposed. This framework extends the traditional energy minimisation framework with the main objective of enhancing its performance. A new data and smoothing terms are defined including a new parameter R in the model. The main advantage of this new framework is the reduction of the computational complexity of the stereo matching problem while maintaining the resolution in the disparity domain. Moreover, traditional algorithms for solving MRFs, such as alpha-expansion and belief propagation, can be used. Note that the framework presented is compatible with new parallel algorithms such as CUDAcuts [16].

Traditionally, a reduction in the size of the problem is achieved by reducing the resolution of the stereo input. In order to avoid decimation in the disparity domain (disparity labels), a subsampling of the input stereo pair is not recommended. As is shown in the results section, performing subsampling in the input stereo image probably leads to uncontrolled behaviour of the final disparity error. In contrast, using the framework herein proposed, no decimation in the disparity space is performed, although a reduction in the size of the problem is still achieved. The experiments carried out, using the alpha-expansion algorithm, demonstrate that an RMS error bound of 20% has been achieved using the framework presented in this paper.

The main success of this framework is to obtain a second-order polynomial reduction in computing time. A speedup of 50 times has been obtained with high values of R, while maintaining accuracy. This result invites us to use the multiresolution framework in real-time applications where no high-resolution disparity maps are required, such as robotics navigation. This framework allows using high-quality global optimisation stereo algorithms in real-time applications without the requirement of parallel computing.

6 References

1 Rogmans, S., Lu, J., Bekaert, P., Lafruit, G.: 'Real-time stereo-based view synthesis algorithms: a unified framework and evaluation on commodity GPUs', *Image Commun.*, 2009, **24**, (1–2), pp. 49–64

2 Scharstein, D., Szeliski, R.: 'A taxonomy and evaluation of dense two-frame stereo correspondence algorithms', *Int. J. Comput. Vis.*, 2002, **47**, (1–3), pp. 7–42

3 Pollefeys, M., Nister, D., Frahm, J.M., *et al.*: 'Detailed real-time urban 3D reconstruction from video', *Int. J. Comput. Vis.*, 2008, **78**, (2–3), pp. 143–167

4 Cornelis, N., Leibe, B., Cornelis, K., Van Gool, L.: '3D urban scene modeling integrating recognition and reconstruction', *Int. J. Comput. Vis.*, 2008, **78**, (2–3), pp. 121–141

5 Zhao, Y., Taubin, G.: 'Real-time stereo on GPGPU using progressive multi-resolution adaptive windows', 2011, **29**, (6), pp. 420–432

6 Yang, R., Pollefeys, M.: 'Multi-resolution real-time stereo on commodity graphics hardware'. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2003, vol. 1, pp. I-211–I-217

7 Gong, M., Yang, Y.-H.: 'Multi-resolution stereo matching using genetic algorithm'. Proc. IEEE Workshop on Stereo and Multi-Baseline Vision, 2001 (SMBV 2001), 2001

8 Yoon, K.J., Kweon, I.S.: 'Adaptive support-weight approach for correspondence search', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (4), pp. 650–656

9 Wang, L., Gong, M., Gong, M., Yang, R.: 'How far can we go with local optimization in real-time stereo matching'. Int. Symp. on 3D Data Processing Visualization and Transmission, 2006, pp. 129–136

10 Geman, S., Geman, D.: 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1984, **PAMI-6**, (6), pp. 721–741

11 Besag, J.: 'On the statistical analysis of dirty pictures', *J. R. Stat. Soc.*, 1986, **B-48**, pp. 259–302

12 Boykov, Y., Veksler, O., Zabih, R.: 'Fast approximate energy minimization via graph cuts', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (11), pp. 1222–1239

13 Yedidia, J.S., Freeman, W.T., Weiss, Y.: 'Understanding belief propagation and its generalizations'. Exploring Artificial Intelligence in the New Millenium, 2003, pp. 239–269

14 Szeliski, R., Zabih, R., Scharstein, D., *et al.*: 'A comparative study of energy minimization methods for Markov random fields with smoothness-based priors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008, **30**, (6), pp. 1068–1080

15 Tappen, M.F., Freeman, W.T.: 'Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters'. ICCV'03: Proc. Ninth IEEE Int. Conf. on Computer Vision, 2003

16 Vineet, V., Narayanan, P.J.: 'CUDA cuts: fast graph cuts on the GPU'. Conf. on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08, 2008

17 Yang, Q., Wang, L., Yang, R. *et al.* (Eds.): 'Real-time global stereo matching using hierarchical belief propagation' (BMVC: British Machine Vision Association, 2006)

18 Liang, C.-K., Cheng, C.-C., Lai, Y.-C., Chen, L.-G., Chen, H.H.: 'Hardware-efficient belief propagation'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009

19 Kohli, P., Torr, P.H.S.: 'Dynamic graph cuts for efficient inference in markov random fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (12), pp. 2079–2088

20 Alahari, K., Kohli, P., Torr, P.H.S.: 'Dynamic hybrid algorithms for MAP inference in discrete MRFs', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (10), pp. 1846–1857

21 Juan, O., Boykov, Y.: 'Active graph cuts'. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2006

22 Wang, L., Jin, H., Yang, R.: 'Search space reduction for MRF stereo'. Proc. 10th European Conf. on Computer Vision: Part I, Berlin, 2008

23 Yu, T., Lin, R.-S., Super, B., Tang, B.: 'Efficient message representations for belief propagation'. IEEE 11th Int. Conf. on Computer Vision 2007, ICCV 2007, 2007

24 Bobick, A.F., Intille, S.S.: 'Large occlusion stereo', *Int. J. Comput. Vis.*, 1999, **33**, pp. 181–200

25 Salmen, J., Schlipsing, M., Edelbrunner, J., Hegemann, S., Luke, S.: 'Real-time stereo vision: making more out of dynamic programming' (CAIP, 2009)

26 Veksler, O.: 'Stereo correspondence by dynamic programming on a tree'. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition 2005, CVPR 2005, 2005, vol. 2, pp. 384–390

27 Wang, L., Liao, M., Gong, M., Yang, R., Nister, D.: 'High-quality real-time stereo using adaptive cost aggregation and dynamic programming'. Third Int. Symp. on 3D Data Processing, Visualization, and Transmission, 2006, pp. 798–805

28 Marr, D., Poggio, T.: 'Cooperative computation of stereo disparity', Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1976

29 Wang, Z.-F., Zheng, Z.-G.: 'A region based stereo matching algorithm using cooperative optimization'. IEEE Conf. on Computer Vision and Pattern Recognition 2008, CVPR 2008, 2008, pp. 1–8

30 Klaus, A., Sormann, M., Karner, K.: 'Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure'. 18th Int. Conf. on Pattern Recognition 2006, ICPR 2006, 2006, vol. 3, no. 3, pp. 15–18

31 Qx, Y., Wang, L., Yang, R.G., Stewenius, H., Nister, D.: 'Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **31**, (3), pp. 492–504

32 Zitnick, C., Kang, S.: 'Stereo for image-based rendering using image over-segmentation', *Int. J. Comput. Vis.*, 2007, **75**, (1), pp. 49–65

33 Zemerly, M.J., Holden, M., Muller, J.-P.: 'A multi-resolution approach to parallel stereo matching of airborne imaginery', *Syst. Data Process. Anal.*, 1992, **92**, pp. 330–357

34 Iocchi, L., Konolige, K.: 'A multiresolution stereo vision system for mobile robots', AIIA Workshop, 1998

35 Satorre, R., Compañ, P., Botía, A., Rizo, R.: 'Multiresolution scheme for stereo correspondence using correlation techniques'. Proc. 12th Portuguese Conf. on Pattern Recognition, 2002

36 Chang, C., Chatterjee, S.: 'Multiresolution stereo by simulated annealing'. Int. Joint Conf. on Neural Networks 1990, 1990 IJCNN, 1990

37 Birchfield, S., Tomasi, C.: 'A pixel dissimilarity measure that is insensitive to image sampling', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998, **20**, (4), pp. 401–406

38 Wainwright, M., Jaakkola, T., Willsky, A.: 'MAP estimation via agreement on (hyper)trees: message-passing and linear programming approaches', *IEEE Trans. Inf. Theory*, 2002, **51**, pp. 3697–3717

doi: 10.1049/iet-cvi.2010.0212