Noname manuscript No.
(will be inserted by the editor)

Rapid and accurate face depth estimation in passive stereo systems

Amel AISSAOUI · Jean MARTINET · Chaabane DJERABA

Received: date / Accepted: date

Abstract In this paper, we introduce a novel approach for face depth estimation in a passive stereo vision system. Our approach is based on rapid generation of facial disparity map, requiring neither expensive devices nor generic face models. It consists of incorporating face properties into the disparity estimation process to enhance the 3D face reconstruction. We propose a model-based method which is independent from the specific stereo algorithm used. Our method is a two-step process. First, an algorithm based on the Active Shape Model (ASM) is proposed to acquire a disparity model specific to the concerned face. Second, using this model as guidance, the dense disparity is calculated and the depth map is estimated. Besides, an original algorithm of post processing is proposed in order to detect holes and spikes in the generated depth maps caused by false matching and uncertainties. It is based on smoothness proprieties of the face and a local and global analysis of the image. Experimental results are presented to demonstrate the reconstruction accuracy and the rapidity of the proposed method.

Amel AISSAOUI
Laboratoire d'Informatique Fondamentale de Lille
50 Halley avenue, 59650 Villeneuve d'Ascq, France.
Tel.: +33-3-62531581
E-mail: amel.aissaoui@lifl.fr

Jean MARTINET
Laboratoire d'Informatique Fondamentale de Lille
50 Halley avenue, 59650 Villeneuve d'Ascq, France.
Tel.: +33-3-62531615
E-mail: jean.martinet@lifl.fr

Chaabane DJERABA
Laboratoire d'Informatique Fondamentale de Lille
50 Halley avenue, 59650 Villeneuve d'Ascq, France.
Tel.: +33-3-62531552
E-mail: Chabane.djeraba@lifl.fr

Keywords Depth estimation · Stereo matching · Active Shape Model · Face analysis

1 Introduction

Depth estimation for faces is an important problem that has been conjointly studied with face animation [28], facial analysis and face recognition [23,18].

In the past few decades, many approaches have been proposed for face depth estimation, including 3D from stereo matching, 3D morphable model based methods [3], Shape from Shading (SfS)[4], Shape from Motion techniques (SfM) [15,26] and statistical techniques [8,25]. A 3D morphable model is generally built from a registered set of 3D laser-scanned heads. Principal Component Analysis (PCA) is then generally applied on the shape and texture features in order to create the feature subspace which constitutes a generic 3D model of the face. Given one or multiple images, a deformation step of the 3D model is applied according to the given images in order to obtain the 3D model of the face. The Shading information is also explored in some works in order to recover the 3D shape of the face by dealing with the reflectance models. The shape from motion method is explored by many researchers in order to give a solution to face depth estimation problem. Authors in [15] propose a similarity transform based method to derive the 3D structure of a human face from a group of face images under different poses. Unfortunately, the high cost of time processing due to the genetic algorithm process used to estimate the depth and how to design a feasible gene operation scheme remain difficult problems. To reduce the computation of the method in [15], the nonlinear least-squares (NLS) model-based methods are proposed in [26]. A basic NLS model and a symmetric NLS model and are proposed to estimate the depth values of facial features point using one frontal-view face image and one non-frontal-view face. For cases when multiple non-frontal-view face images are available, a model-integration approach is proposed to improve the depth estimation accuracy. Some works based on multiple images use statistical techniques for depth recovery. By considering the observations as mixing signals, a novel algorithm for maximizing the posterior shape was developed in [8] to estimate the shape from perspective of blind source separation (BSS). In the same way, authors in [25] propose a model based on Independent Principle Component (ICA) in order to estimate face depth from one image. As in [26], an integration scheme is proposed in cases where more images are available.

However, how to efficiently acquire facial depth information from stereo images is still a challenging problem, especially in binocular passive systems, where only one image pair is used and neither structural lighting is available nor morphable model is needed.

In the literature, a wide spectrum of works dealing with the problem of 3D reconstruction using a binocular passive stereo system is proposed [22]. The major problem of these approaches lays in the definition of a stereo matching scheme for a given image pair. Indeed, this problem is more crucial with poorly-textured face images. This leads to ambiguities and additional complexities for matching algorithms. Therefore, a few approaches for dense 3D face reconstruction in binocular passive stereo system have been proposed, compared to those of active and multi-view stereo vision.

In this paper, we propose an improved method for determining the disparity information of a human face from stereo matching in a binocular vision system in real-time application by considering the topological properties of the face and its shape smoothness. An algorithm of post processing of the depth map is proposed which consists of detecting and removing holes and spikes.

The remainder of this paper is organized as follows. In section 2, we describe the general depth estimation process based on binocular images, the existing approaches in this field and we highlight some of the recent works on stereo depth estimation for human faces. An important step after depth estimation consists of the post processing consisting of depth map denoising in order to remove holes and spikes. We introduce the problem of depth map denoising and we present how works in face reconstruction are addressing this problem. In Section 3, we introduce the proposed method for disparity estimation consisting of incorporating prior knowledge on the face shape in the estimation process and we show how this could enhance the state of the art methods. Then, we present the proposed algorithm for post processing which use local and global analysis and assume the smoothness of the face shape in order to deal with noise detection (holes and spike). We finish this section by some experimental results to demonstrate the accuracy of the proposed denoising method. The experimental results are presented in Section 4, where we evaluate the accuracy and the rapidity of our method qualitatively and quantitatively. Finally, Section 5 concludes the paper and gives some perspectives of this work.

2 Related works

Depth estimation process in binocular system consists of estimating a so-called *disparity map* of a scene captured from two different points of view. In order to calculate this map, a step called *stereo matching* is applied. It consists on finding corresponding pixels in the both images representing the projection of the same real word 3D point. Since, initially, we do not know where we might find a corresponding point; the search space for matching a point is relatively large. A rectification process which consists in projecting the stereo pair onto a common image plane is applied to constrain the size of the search to 1D dimension.

Once the disparity map is estimated, the depth of a point $p(x, y, z)$ with a disparity value d is calculated as:

$$z = \frac{fb}{d}. \tag{1}$$

where f and b are the camera focal and baseline respectively.

In order to estimate the disparity map, many algorithms have been proposed [22]. They can be classified into two categories: global and local methods. Global methods resolve an optimization problem on the disparity map by including complex energy minimization methods. Some popular paradigms are, graph cut [14], belief propagation [24] and dynamic programming [27]. The global methods give a good accuracy since they process the pixels in a dependent way. However, they encounter difficulties in determining the correspondence pairs at non-edge pixels. This is because images may be ambiguous locally. For example, there may be many patches with similar appearance. Another feedback of global methods is that they require a long processing time.

Local methods (also called *block-matching methods*) are based on intensity correlation and they can be used in real time applications because of the low complexity of the algorithms. Correlation-based stereo matching algorithms typically produce dense depth maps by calculating the matching costs for each pixel at each disparity level in a certain range. Afterwards, the matching costs for all disparity levels can be aggregated within a certain neighborhood window. Finally, the algorithm searches for the lowest cost match for each pixel. Different similarity measures are used in correlation-based methods. The most common ones are: Sum of Absolute Differences (SAD), Sum of Squared Differences (SSD), Normalized Cross Correlation (NCC) and Sum of Hamming Distances (SHD). However, these methods suffer from 2 essential problems:

- Implicit hypothesis: all points within a correlation window move with same motion, which is incorrect at discontinuities and leads to blurred object boundaries.
- Aperture problem: the context can be too small in certain regions (homogenous regions), lack of information.

2.1 Face depth estimation

For face depth estimation, the stereo matching process is more complex (including using local or global method) because of the homogeneity of the face regions especially when the system is binocular and only one stereo pair is used. The homogenous aspect of the face present a limitation in all stereo methods (local or global) since all the patch have close intensities and therefore it is difficult to find the exact corresponding pixel when all the pixels give the same or close similarity values. This homogeneity leads to obtain more holes, spikes and many uncertain disparities in the depth map.

In the literature, most of the existing methods for face depth estimation from stereo systems are based on a fitting step of the estimated depth to a generic 3D model [16,20,32]. Le et al. [16] have built a coarse shape estimation based on 3D key points, and then used a linear morphable model to efficiently match the detailed shape and texture. Authors in [20] fit a sparse reconstruction of manually selected points to a generic model using a Thin-Plate Spline

(TPS) method. In [32], a reference 3D face is used as an intermedium for correspondence calculation. The virtual face images with known correspondences are first synthesized from the reference face. Then the known correspondences are extended to the incoming stereo face images, using face alignment and warping. The complete 3D face can thus reliably be reconstructed from stereo images. The main problems of these methods are the large processing time related to the fitting step (due to the high algorithmic complexity) and the manual initialization requirement e.g. in [32], for the face alignment step. Another disadvantage of these methods is the fact that the resulting faces are more similar to the generic model than to their specific model. In [17], Lengagne et al. proposed to apply an iterative algorithm to refine the face model resulting from the fitting process of the sparse disparity to the 3D generic model using differentials constraints. This method has an additional computation cost since it includes an iterative deformation step plus the calculation of the principle curvatures on each vertex. Also, it is very sensitive to noise because it uses the second derivative for calculating curves.

Some attempts have been proposed to use the Shape from Shading (SfS) method to enhance the stereo matching process. Cryer et al. [6] propose to merge the dense depth maps obtained separately from shape from stereo and shape from shading in the frequency domain. The merging process is based on the assumption that shape from stereo is good at recovery of high frequency information and shape from shading is good at recovery of low frequency information. However, they formulate the shading model using orthographic projection which is far from reality. Chow et al. [4] propose to enhance the work of Cryer et al. by using a rectification process to convert any lighting direction from oblique to orthographic. However, in all these method both processes (stereo and shading) are very sensitive to the lighting conditions. Besides, SfS methods are based on many assumptions as the direction of the light source and the surface reflection.

2.2 Face depth map denoising

The depth data resulting from the face reconstruction process is generally affected by two types of noise which are holes and spikes. Holes are pixels with undefined depth values. The disparity values for these pixels are set to zero in the process of disparity estimation. They occur in cases of occlusion or bad illumination condition. Spikes are pixels with wrong estimated depth value. They are generally caused by a wrong matching and they occur mostly in homogenous areas where pixels have approximately identical intensity value.

In the literature, different methods are proposed to face depth map denoising. We propose to classify them in two classes: global and local. Global methods consist in applying a noise reduction filters on the hole depth image to remove spikes and fill holes. The median filter is commonly applied for this purpose. In the work of [13], [1], and [11], a pass over the range image with a median filter smoothes the data and removes spikes in the z-coordinate.

Authors in [29] used three Gaussian filters with different variances to remove spikes, fill small holes, and smooth the data. Applying these kinds of filters can give good results for small noises. However, if the noisy area is large, these filters cannot eliminate the noise, but can only change the pixels values according to their neighborhood. The accuracy of these filters is much related to the kernel size which is also related to the noise size. Thus, the kernel size cannot be fixed and generalized for all the data. Another drawback of these methods is the fact that they can cause the loss of the exact data because they affect even pixels with a precisely estimated depth value.

Local methods consist in a local processing of the face depth map, which can essentially be divided into two sub-problems: identifying the holes and finding appropriate parameterizations that allow the reconstruction of the missing parts using the available data. Identifying holes consist of finding regions in the disparity map with zero-values (i.e. unidentified depth values). In [7], this is done by processing the data row by row, where boundary pixels are initially determined by a sweep through the depth image, to find the first and last non-zero pixels in each row. This process is repeated until no additional pixel is created. After identifying the boundary of holes, the filling step generally consists in applying an interpolation algorithm or a local median filter. This approach is more precise than the global one since it process only noises and preserve the non-noisy data. However, it can only handle holes since they have a known value (zero or undefined value), and therefore cannot be applied for removing spikes since they usually have a random values.

2.3 Main contribution

In our work, we propose to enhance the stereo matching process without using neither 3D generic model nor shading techniques. Our method consists of using a reconstructed disparity model of the face independently from the stereo matching method and incorporating smoothness and topological face properties in the estimation process to improve its result in real time application. We choose to use the block-matching method to estimate the disparity map of the face for its rapidity and low complexity. Assuming the rigidity of the face and its smoothness, the first problems of block-matching method will not influence the estimation results. However, the aperture problem is a major problem for face disparity estimation since the face has many homogenous areas. We propose to overcome this problem by incorporating the smoothness and the topological properties of the face while maintaining a low computational complexity. Some preliminary statistical works [12] reveals that the differences between pixels depth are smaller than 7 for a neighborhood of radius equal to 2 for 93% of the pixels of the depth map. Assuming this property of smoothness of the face depth, we propose also a free-parameterization algorithm for addressing noise (holes and spikes) detection in face depth maps, which give better results than the median filter that is commonly used in the state of the art as in [1] and [11].

3 Proposed method for face depth estimation

In order to estimate the disparity map, photo-consistency measures used in block-matching methods are not always sufficient to recover the precise geometry, particularly in low-textured scene regions (aperture problem). Since the face is a specific object having its proper structure and properties, it can therefore be helpful to incorporate face properties that bias the reconstruction to have desired characteristics. For this purpose, we search to construct a disparity model of the face from the stereo images using some prior knowledge as the smoothness of the face and the topological properties of its shape. In order to build the disparity model, we propose to use an Active Shape Model (ASM) fitting process. Then, the dense disparity map for the whole face is estimated using the common block-matching method considering the topological information obtained by the reconstructed disparity model and the smoothness properties of the face. Finally, the post processing proposed algorithm is presented.

3.1 Disparity model construction

The first step of our method consists of constructing a disparity model of the given face. This model gives a holistic representation of the disparity distribution of the face points which will be used as guidance in the disparity map calculation step. In order to establish the disparity model, we start by applying an ASM [19] fitting algorithm on both images to localize a set of corresponding points of high confidence (Fig.1.a). The ASM is a statistic shape model obtained from a learning process on an annotated face database. Fitting the ASM on a new face image consists of estimating the shape parameters of the model by minimizing a cost function defining how well a particular instance of the model describes the evidence in the image. For a great source of information on ASM-related research, the reader is referred to [19] and [5].

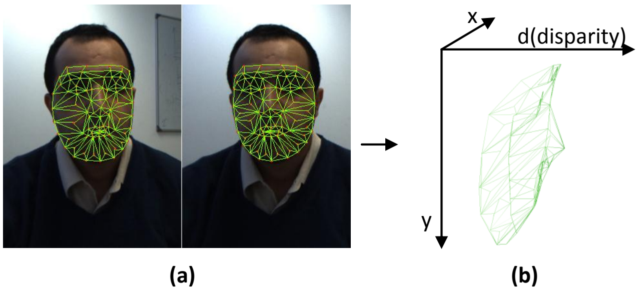


Fig. 1 Disparity model construction.

We use the ASM fitting, because in addition to the color information used in the block-matching method, it includes the shape information obtained by the off-line learning process, which guarantees a good face features localization in the stereo pair, and therefore a high disparity confidence for those points.

After fitting the ASM on both the left and right images separately, we obtain the 2D coordinates of n face feature points in the right image $R = \{(x_i, y_i), i \in [1, n]\}$ and in the left image $L = \{(x_i^{'}, y_i^{'}), i \in [1, n]\}$, which are then used to obtain the final set of 3 coordinates: $P = \{p_i(x, y, d), i \in [1, n]\}$, that represents the disparity model of the face under consideration (Fig.1.b). The disparity d of the points is calculated using the Euclidian distance as follows:

$$d_i = \sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2}. \tag{2}$$

Since we use a calibrated system and rectified stereo pairs, the y coordinates of each corresponding points should be identical. So the disparity is calculated using the distance between only the x coordinates as:

$$d_i = \sqrt{(x_i - x_i^{'})^2}. \tag{3}$$

If the y coordinates of the corresponding points resulting from the fitting step are not the same due to a fitting error -generally caused by noises in the stereo images-, we normalize it to their mean value.

The disparity model of the face constructed in this step is used in the next step as guidance for calculating the dense disparity map.

3.2 Disparity map calculation

In this step, we calculate the dense disparity map in a two-step process. In the

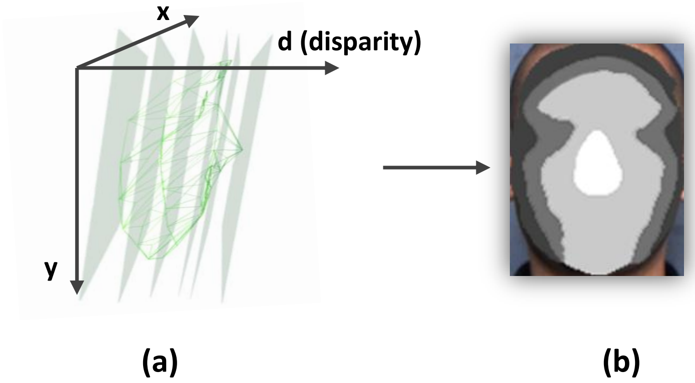


Fig. 2 Disparity model decomposition. (a) The decomposition process. (b) The 2D projection process

first step, a process consisting in decomposing the face disparity model into different ranges is performed with a set of *level planes* with associated disparity values (Fig.2.a). Using the assumptions of the depth face smoothness and concavity, the level planes are defined to be perpendicular to the normal vector centered on the point with the smallest disparity given by the face disparity model. This point is usually the nose tip when the head pose orientation is not too large.

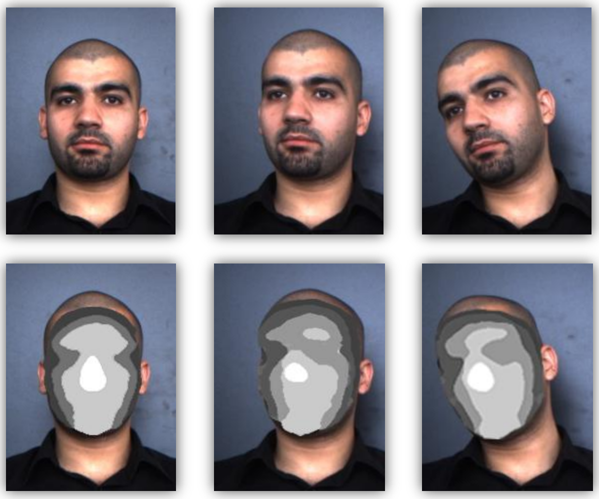


Fig. 3 Disparity model decomposition with different head pose orientations.

After the decomposition step, we can define different areas in the face image with different disparity ranges. In Fig. (2.b), we generate shapes regions in the right (or left, chosen arbitrarily) stereo image that corresponds to the intersection of the level planes and the points of the disparity model. A disparity range is assigned to each shape according the disparity values of the points belonging to the level plane. In figure 3, we demonstrate how the decomposition step of the disparity model considers the pose variation of the face.

The decomposition step guaranties the smoothness of the final estimated disparity map and also reduces the search area, instead of the entire epipolar line to just a small segment (inside the shape). It also reduces the number of spikes since it limits the disparity range of each face part to the minimum and the maximum of the neighbor ranges. Another advantage of the decomposition step is considering the face orientation (See Fig. 3).

In order to obtain the disparity value of a given pixel p belonging to a given shape S_i, we define the disparity interval as $[DispMin_p, DispMax_p]$ where $DispMin_p$ is the disparity value associated to the shape S_i and $DispMax_p$ is that of S_{i+1}.

In the second step, we calculate the disparity of the face points, using their disparity ranges to initialize the block-matching algorithm.

Given a face point p, with the right projection p_r and the left projection p_l, a correlation window w and a disparity interval $[DispMin_p, DispMax_p]$, we aim to obtain the disparity $d \in [DispMin_p, DispMax_p]$, which maximizes the correlation equation $E(d)$:

$$E(d) = Similarity(p_l(x, y), p_r(x + d, y)) \tag{4}$$

For the similarity function, we have used the Sum of Absolute Differences (SAD) [10] measure defined as:

$$\begin{aligned} SAD_{I_l(x,y),I_r(x',y'))} = \sum_{u=0}^{m}\sum_{v=0}^{n} &|I_l(x+u, y+v) \\ &- I_r(x^{'} + u + d, y^{'} + v)|. \end{aligned} \tag{5}$$

where:

I_l,I_r are the left and the right images.
$m \times n$ is the correlation window size.

Finally, using the estimated disparity, the depth map is obtained by applying Eq.1.

3.3 Depth map denoising

After depth estimation, a process of post processing is needed in order to remove holes and spikes caused by uncertainties and false matching. Since the global methods affect all the data and cause a data losing, we choose to follow a local methodology. Local methods consist of two steps: Noise detection and Noise removing.

3.3.1 Noise detection

Local method for depth face denoising are strongly depending to the noise detection step which is easily performed for holes and small spikes but it became difficult when it comes to detect wrong data represented by large spikes caused by false matching in homogenous surfaces of the face. As in the face depth estimation process, we aim to incorporate knowledge on face in order to propose a new method for addressing this problem. We propose an algorithm for automatically identifying holes and spikes in face depth images by incorporating the smoothness propriety of the face. A method consisting of local and global analysis is used in order to classify the depth map into noisy and non-noisy parts. In our algorithm, we do not distinguish between spikes and holes, and therefore we consider both to be *noise*. Our method differs from those of the state of the art in its ability to identify holes but also small and large spikes in order that the filling step impacts only the detected noise and doesn't affect the rest of the data. Another advantage of our algorithm is that it does not require any parameterizations and it is fully automatic. It is conceptually very simple and its implementation is straightforward. After identifying and removing noises, any state-of-the-art method can be used for filling the missing data. We choose to use a simple interpolation between the noise borders.

The process of the algorithm consists in searching the noisy parts by scanning the depth map row by row. Since the face surface is smooth horizontally and vertically, we can process the data row by row or column by column. In this work, we choose arbitrarily using the rows. It is a two-steps process. First, it consists in segmenting each depth row into different slices using the gradient. Then, slices are classified into noisy and non-noisy using local and global analysis of the depth row.

The segmentation step consists of cutting the depth rows in different slices. Considering the depth line as a smooth function, we use its first derivative to detect the main cut-points which segment it into a set of slices.

We consider a given depth as a function f defined as follows:

$$f : \mathcal{N}^2 \rightarrow \mathcal{R}$$
$$f(x, y) \rightarrow z$$

where $x \in [1, N-1]$ and $y \in [1, M-1]$ are the coordinates of the pixels in a depth image with $N \times M$ size and z is the depth value at those coordinates.

In this work, we process the depth row by row separately as the work in [7], because of that, we limit the function only to x coordinates and we consider y as a constant. Therefore the derivative $f^{'}$ can numerically be approximated as:

$$f^{'}(x, y) = \frac{f(x+h, y) - f(x, y)}{(x+h) - x} \quad (6)$$

Since our processing use the direct neighborhood of the pixel, we put $h = 1$ and therefore the equation is given as:

$$f^{'}(x, y) = f(x+1, y) - f(x, y) \quad (7)$$

In order to find the cut-points, we calculate the first derivative using Eq.7 with and we solve the equation below:

$$|f^{'}(x, y)| \geq t \quad (8)$$

where t is a threshold consisting of the mean of the derivative function $f^{'}(x)$. It is calculated as:

$$t = \frac{\sum_{i=0}^{n} f^{'}(x, y)}{n} \quad (9)$$

where n is the number of pixels in the depth row.

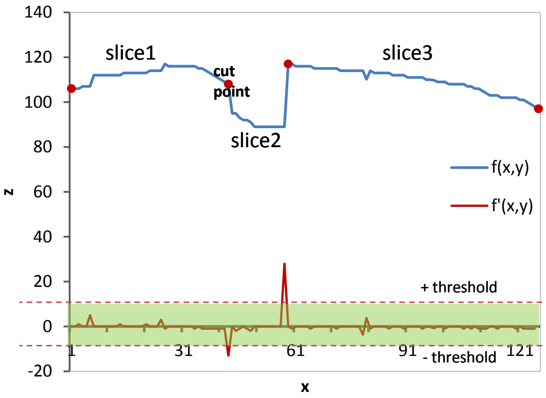


Fig. 4 Detection of cut points and depth line decomposition.

In the figure 4, we illustrate how to find the cut-points and define the main slices in a given depth-row.

The second step consists in noise-identification. In this step, we will classify the slices detected previously for each row into "noisy" and "non-noisy" slice. For this purpose, we use the standard deviation σ to measure the dispersion of the slices from their mean (See Eq (10)) and therefore identify the noisy slices. In Fig 5, the slice in red is identified as "noisy". The standard deviation is calculated as follows:

$$\sigma = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (z_i - \overline{z})^2} \tag{10}$$

where $z_1, z_2, \ldots, z_n$ are the depth row values and $\overline{z}$ is their mean value.

Given a depth raw with a mean μ, a standard deviation σ, and a set of segmented slices $s_1, s_2, ..., s_n$ (obtained in the row decomposition step) with associated means $m_1, m_2, ..., m_n$, we identify a given slice s_i with a mean m_i as a noise if $m_i \notin [\mu - \sigma, \mu + \sigma]$

In figure 5, we illustrated the step of noise identification given a noisy depth row.

3.4 Noise removing

After noise detection, all noisy parts are set to zero and then a hole-filling step using an interpolation algorithm is applied. The cubic interpolation is used because it can accurately fill in the holes which are to some extent large. In Fig.6, we illustrate an example of our depth map denoising process.

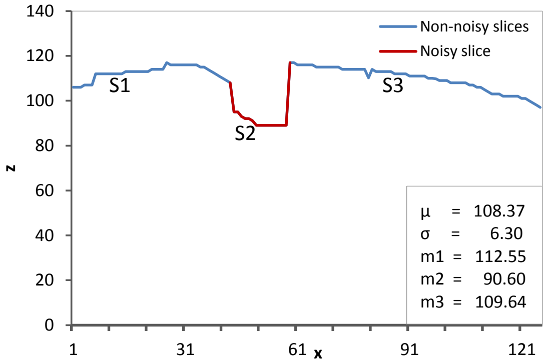


Fig. 5 Detection of noisy slices. Each m_i corresponds to the mean of the slice s_i.

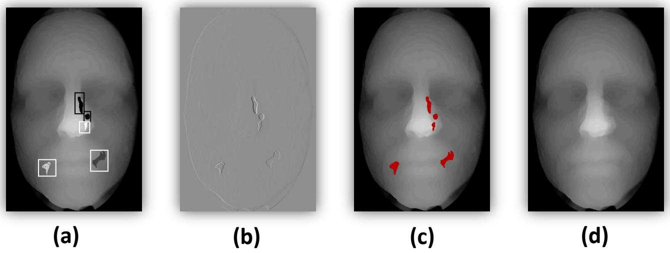


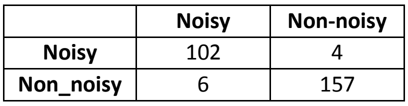
Fig. 6 Depth map denoising: (a) depth map with holes, (b) gradient, (c) noisy slices detection, (d) corrected depth map.

4 Experiments, results and discussion

In this section, we evaluate qualitatively and quantitatively the results of the proposed framework. First, the noise detection algorithm proposed for post processing of face depth map is evaluated. Then different experiments are performed on three public databases in order to evaluate the results of the proposed face depth estimation method and to compare it to the state-of-the-art methods, using different measures. The two first databases are Texas 3D face *Texas 3D database* [9] and Bosphorus database [21]. One prominent advantage with using these two databases is that the 3D coordinates of each face image of are available. Then, the estimated depth values of a reconstructed 3D face structure can be compared to its ground true values. Consequently, the performances of 3D reconstruction algorithms can be evaluated and compared more accurately. The third database consists of stereo images that we have created using a stereo camera (The point gray bumblebee stereo camera). This database allows only a qualitative validation since it doesn't contain the ground truth of the face models.

4.1 Noise detection

The noise detection algorithm is evaluated in this section in order to show how this step contribute in filling the missing data with keeping the original value of the boundary of the noise. In order to evaluate the "noisy-to-non-noisy" classification used in noise detection algorithm, the confusion matrix (See Fig. 7) is constructed from a set of 100 depth rows selected randomly from different face depth maps on which synthetic noises with different sizes are generated. The holes are easily detected since their values are null. Therefore, we generate more spikes then holes in the depth rows used in this experiment. The total number of the slices obtained by the segmentation step (See Seq. 3.3.1) is 269 were 108 are noisy and 161 are non-noisy-slice.



	Noisy	Non-noisy
Noisy	102	4
Non_noisy	6	157

Fig. 7 Confusion matrix for noisy-to-non-noisy classification

The confusion matrix shows the ability of the algorithm to pick out the noisy parts of a given depth row. The accuracy is 96.28%. We note that the false classification given by the algorithm correspond to cases when the noisy part is larger than the non-noisy part in the depth row. Indeed, in this case the statistic values calculated on the depth row are influenced more much by noise than by the correct data and therefore the classification is inverted.

In order to show how our algorithm contribute in depth filling step, we compare, in figure 8, a depth map denoised by applying a cubic interpolation for the detected noise defined by our proposed algorithm Fig. 8.(c) to two method of the state of the art : global method based on median filter (Fig. 8.(a)) like works of Kakadiaris et al. [13] and Berretti et al. [1], and local method based on identifying and removing steps proposed in [7] Fig. 8.(b). We can see that local methods (Faltemier at al. [7], and our method) correct the noisy parts with preserving the depth information and details unlike the global method where the entire range image is modified and the exact depth information of pixels locating around noise are lost. Holes in the depth map are perfectly removed using method of Faltemier et al. [7] However, spikes are not detected and still present in the depth map. Our proposed algorithm was able to detect all the spikes and holes in the same way since it does not search only the zero-value region but calculate statistical measures on the depth map locally and globally to identify the different noises (spikes and holes). In addition, unlike the filter-based methods, the proposed algorithm is parameterization-free so it is independent from noisy part size.

In order to measure the accuracy of the different depth denoising methods, we calculate the RMS (root-mean-square) [22] between the ground truth and the denoised depth maps obtained by the different methods. We calculate also

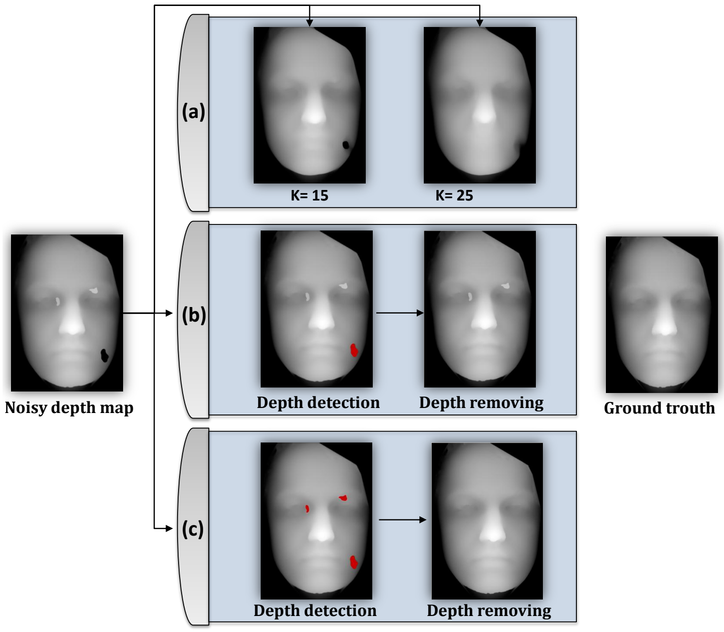


Fig. 8 Depth map denoising comparison : (a) Global method based on median filter (used in [13] and [1]), (b) Local method [7] , (c) Proposed method

the RMS error between the ground truth and the noisy depth map to use it as reference in the comparison. We can see in Fig. 9, that the RMS obtained for the global method is bigger than the RMS of the noisy map. Although the results of the global method based on the median filter seem visually satisfying and the noise removed, the RMS error obtained is bigger than that obtained by the noisy map since the global process affect all the map and therefore, the exact data is lost.

The smallest error is obtained by our method which proves its accuracy.

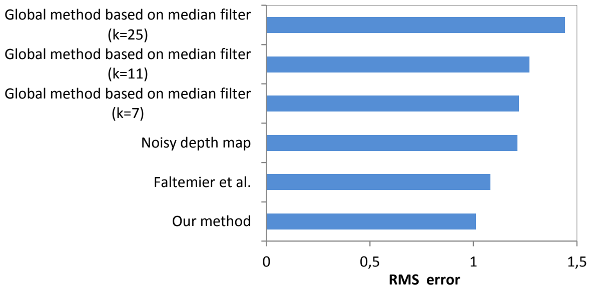


Fig. 9 RMS mesure between denoised depth map and the ground truth

4.2 Face depth estimation

In this work, we choose to evaluate our results on three databases. First a comparative study is performed using the Texas database between the results obtained by our method and the conventional Block-matching method in order to show how much the disparity model construction and the integration of prior information about faces can improve the depth estimation process. A qualitative evaluation is then performed on the database that we reconstructed under different poses and comparison with the state of the art is given.

4.3 Evaluation on the Texas database

In order to compare our results quantitatively, we have synthesized a binocular stereo database of 105 faces from the Texas 3D Face Database [9] in order to compare the estimated results to a ground truth. Disparity maps are estimated from the stereo pairs of faces from different persons using our method, the standard block-matching method on which our method is based [2] in order to show how the integration of prior information obtained by the disparity model construction step enhanced the depth estimation process using neither 3D morphable model nor additional processing time. The graph-cut based method [14] is also used in the comparison as an example of global methods. The depth maps are then generated by applying the equation Eq. 1 on the disparity maps. Finally, a post processing step consisting of filling holes and removing spikes in the depth maps estimated by the three methods is applied using our proposed method for depth map denoising. Faces in the Texas database are 501 × 751 pixels in size, with a resolution of 0.32 mm along the x, y, and z dimensions. For FRGC, the size is 501 × 751. Figure 10 shows the reconstructed depth maps compared to the ground truth depth maps of sample faces from Texas database.

In order to compare the results illustrated in Fig. 10, we calculate The RMS (root-mean-square) [22] error (Eq. 11) and the PBM (Percentage of Bad Matching pixels) (Eq. 12), between the ground truth maps and the estimated maps of faces obtained from block-matching (BM), graph-cut based method (GC), and our method.

$$RMS = (\frac{1}{np} \times \sum_{(x,y)} |d_E(x,y) - d_T(x,y)|^2)^{\frac{1}{2}} \tag{11}$$

$$PBM = \frac{1}{np} \times \sum_{(x,y)} D(x,y) \; where \; D(x,y) = \begin{cases} 0 \; if(|d_E(x,y) - d_T(x,y)| \leq \delta_d) \\ 1 \qquad\qquad otherwise \end{cases} \tag{12}$$

where :

- np is the number of pixels in the depth map.
- $d_E(x,y), d_T(x,y)$: are the estimated disparity and the ground truth disparity of the pixel (x,y), respectively.

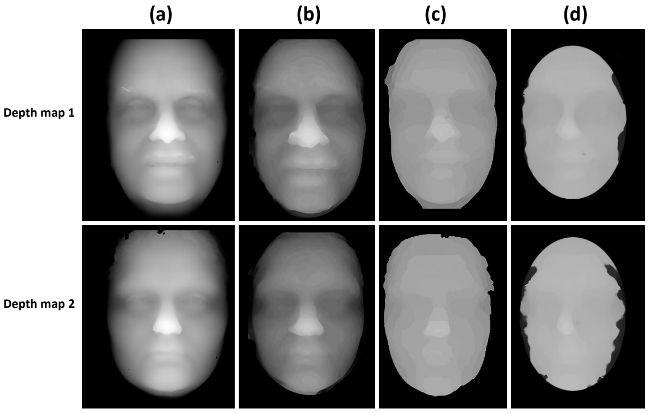


Fig. 10 Depth maps:(a) original (b) our method (c) graph-cut (d) block-matching.

- δ_d : is a disparity error tolerance. For the experiments in this paper we use $\delta_d = 1.0$ since it is the most used value in the majority of previously published studies [22].

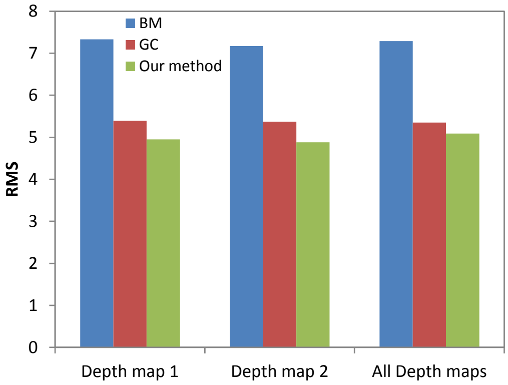


Fig. 11 RMS measure

Figure 11 shows that the RMS error is reduced from 7.33 in Block-matching results to 4.95 in our method that incorporates the disparity model in the block-matching process. The PBM graph (Fig. 12) shows how the percentage of the bad matching pixels is very small comparing to that obtained using the block-matching method. Although the block-matching method is rapid, our method is faster and requires less time than block-matching method and more less time than graph-cut method, since in our method the disparity interval is fixed automatically as a small segment from the epipolar line using the disparity model.

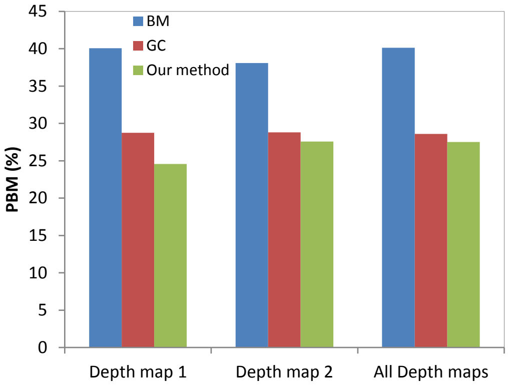


Fig. 12 PBM measure

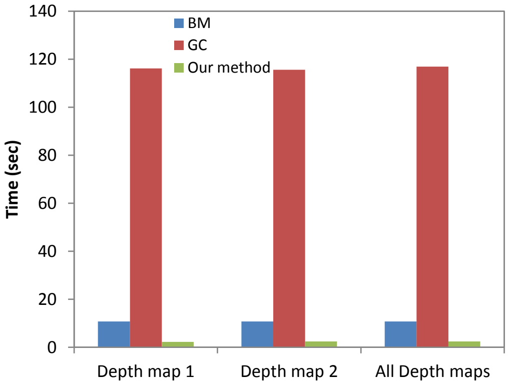


Fig. 13 Processing time

The results of the graph-cut based method are less noisy since the estimation of each pixel disparity is performed relatively to the neighbor pixels. However, we can see that the estimated depth range is very small compared to our results which give larger depth range and more details on topological areas (nose, eyes, etc.), and which are very similar to the original 3D depth map. The RMS and PBM obtained by our method and graph-cut method are close, however, as shown in figure 13, our method requires about 50 times less processing time than the graph-cut based method, for an image of 501 × 751 pixels.

We can see clearly that integrating the shape properties into the estimation process enhances the results of the estimation in terms of accuracy and in terms of reducing noises in the depth maps.

In order to evaluate the mesh characteristics of the resulting 3D faces, we constructed a dissimilarity matrix of the estimated and the original models (Fig. 14). First, we have generated 3D mesh from the ground truth depth maps and the estimated depth maps, obtained from the three methods, using graphics tools. Then, each mesh is matched with all the ground truth meshes in the database using the ICP (Iterative Closest Point) algorithm [31]. The ICP algorithm computes the residual error between the estimated mesh (resulting from our method, block-matching method and graph-cut based method) and the 3D ground truth mesh from the database. Finally, the distance between an estimated mesh and the original mesh is given by the Mean of the Point-Wise Distance (MPWD) which is calculated by the Equation 13 as:

$$MPWD(m_t, m_e) = \sum_{i=0,j=0}^{n} (D((P_i)^{m_t}, (P_j)^{m_e}))/n \tag{13}$$

Where :

- m_t, m_e : are the ground truth mesh and the estimated mesh, respectively.
- $D((P_i)^{m_t}, (P_j)^{m_e})$: is the distance calculated by ICP between each pair of corresponding points.
- n : is the number of points used in the ICP process.
- N : is the number of points in the mesh points ($n \leq N$).

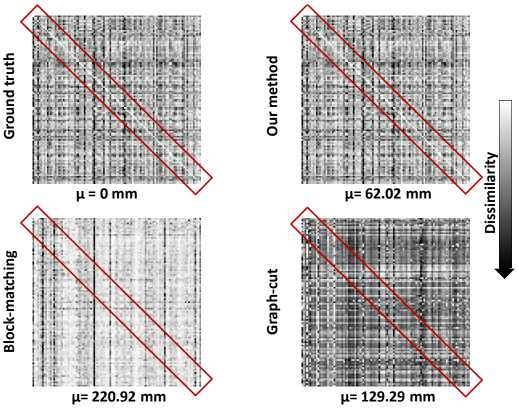


Fig. 14 Similarity matrix ($\overline{E}$: mean error of the diagonal).

Figure 14 shows the 4 dissimilarity matrices calculated for the ground truth and the estimated 3D models. Each cell (i, j) of the matrix represents the $MPWD(m_t^i, m_e^j)$. We can see that the matrix diagonal for our method has lower values (light gray line) than for the block-matching method and the graph-cut based method, meaning that the models reconstructed with our method are more accurate and closer to the original 3D models. Besides, these low values ($\overline{E} = 62,02mm$) are significantly different from the other higher values in the matrix in our method. The diagonal line is darker in the other two methods. This difference shows the specificity of the reconstructed model of our method, which is guaranteed by using a disparity model for each person (obtained by applying the ASM) when calculating the depth map.

4.4 Evaluation on our database

We have built a database of 60 stereo pairs of faces with different pose and expression variations using a Bumblebee stereoscopic system composed of two CDD pre-calibrated cameras. The images' size is 640 × 480 pixels.

In Figure 15, we compare the disparity maps estimated with the block-matching method, graph-cut based method, and our method of an example from our database. The face size is 120×180 pixels, the window aggregation size is 11×11. Since the block-matching and graph-cut method are very related to the disparity interval, we applied some experimental tests with different values and we fixed it to $[0, 127]$ because it gives the best results. However, for our method, this interval is defined automatically depending on the disparity model of the face (see Section 4.1). The disparity maps showed here are smoothed and an elliptical mask is applied automatically to remove the background.

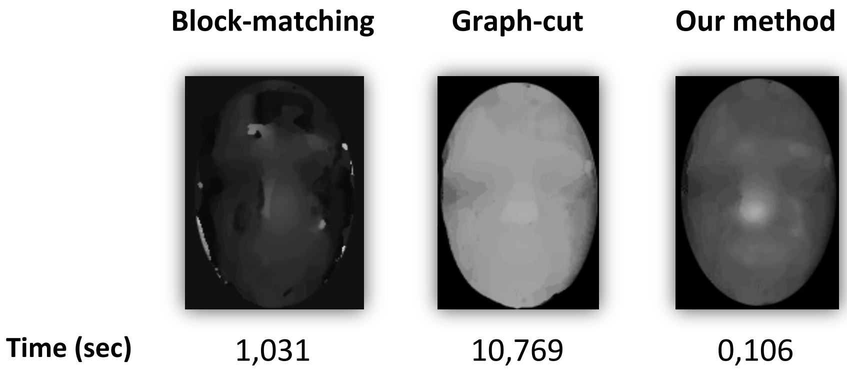


Fig. 15 Disparity maps.

The results show that considering the face shape and its properties, the reconstructed disparity model of the face can enhance the disparity map in terms of smoothness and also in terms of reducing the noise (holes and spikes) occurring due to insufficient texture in homogenous face areas.

We can see that our method gives better results than the block-matching method in terms of smoothness, disparity range estimation and noises. Although the block-matching method is fast, our proposed method need less time of processing since we associate for each point a limited disparity interval defined according to the topological part of the face to which the point is belonging.

When comparing our results to those of a global method based on the graph-cut optimization [14], which is supposed to give very good estimation results, we can see that our method, which includes face shape information and considers face proprieties, gives better results in terms of both disparity range exploration and continuity while requiring much less time (about 10 times faster) than the graph-cut based method processing time. The graph-cut-based method represents the image as a graph and tries to find cuts with an iterative process in this graph, which correspond to different disparities. When the image consists of a scene containing different objects, the method gives good estimation. However, since the face is one continuous and smooth surface, results of the graph-cut show flat regions with abrupt cuts, because they originate from a segmentation process, which leads to an information loss in depth and of smoothness loss of the disparity map.

In order to study the sensitivity of our method to pose variation, we applied our algorithm on a set of stereo faces with different poses. Figure 16 shows the ASM fitting step and the disparity map reconstructed with our method for these faces with pose variations in yaw, pitch and roll.

We can see that the disparity estimation process gives an accurate and less noisy result for the frontal view. In case of small pose variations (less than 30°), our method still gives a good estimation since the ASM fitting process is well done. However, when the pose variation is large, the ASM cannot find all necessary points in order to be fitted to the face and consequently disparity model can not be built. In that case, the disparity map cannot be calculated.

4.5 Evaluation on the Bosphorus database

Majority of the previous stereoscopic face reconstruction methods are evaluated qualitatively only. Typically, the results are shown in different poses in order to prove how the reconstruction is realistic. However, judging by human eyes is not the best way to evaluate the reconstruction results. Some researchers have evaluated there results indirectly by recognition accuracy. This evaluation gives an idea about the reconstruction step but does not give precise information about the reconstruction error and how much the results are closer to the ground truth and thus it is difficult to compare the accuracy of the reconstruction algorithm. Therefore we are not able to compare the accuracy of our reconstruction method to these works. However, the result of some recent works of depth face estimation [25,26] are evaluated quantitatively on the public Bosphorus database. Thus, we follow the same configuration used in these works in order to compare our results.

Five different models with different pose variation (annotated as PR_D, PR_SD, PR_SU, PR_U and YR_R10) are used for each subject in the database. The figure 17 shows an example of the used images. Using the depth images provided in the database, the stereo pairs corresponding to these images are synthesized in order to apply our method for reconstruction. In order to measure the accuracy of the methods, the correlation coefficient is calculated between the reconstructed and the ground truth model using 22 feature points.

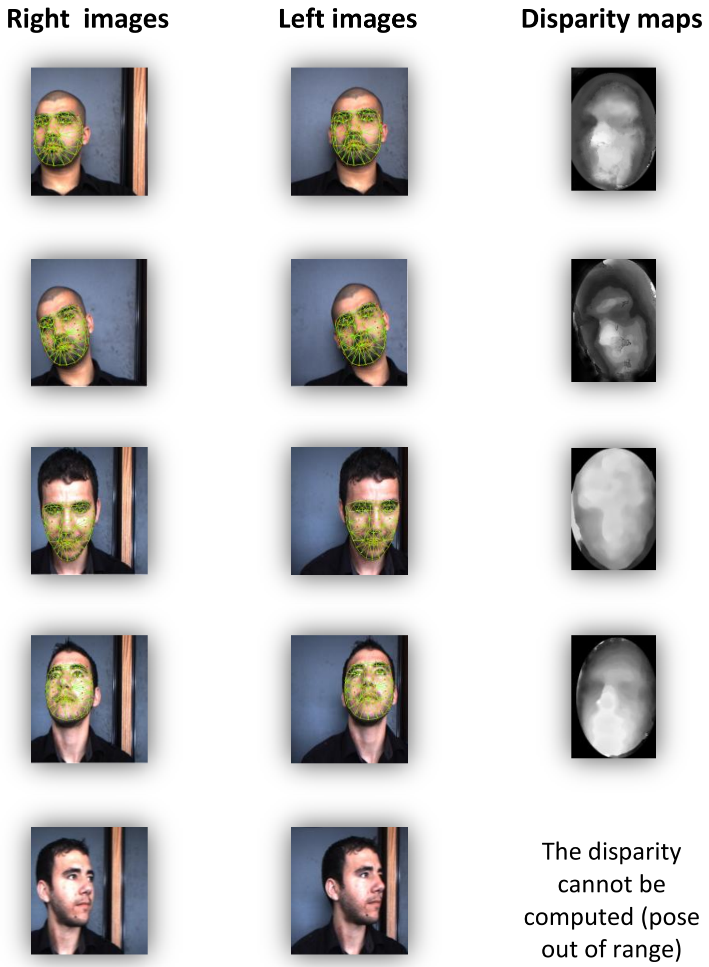


Fig. 16 Disparity maps for frontal and non-frontal view.

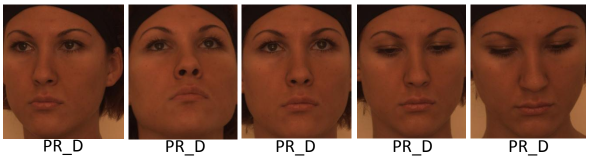


Fig. 17 Examples of images used in the experiments.

In figures 19 and 20, we compare our method to some state-of-the-art methods taking the first 30 and 20 subjects respectively. The number of images used in the depth estimation process for all the methods is given in table 18

	Number of images	Type of images
SM	NM	NM
cICA	2	Frontal + Non Frontal
cICA_MI	More than 4	Frontal + Non Frontal
NLS1_R_MI	More than 4	Frontal + Non Frontal
BSS_SMF	NM	NM
Our method	2	Stereoscopic pair

Fig. 18 Inputs of different methods used in comparison.

In Figure 19, we compare our result to the ICA based methods proposed in [25] and named as $cICA$ and $cICA_MI$ and the similarity transform based method SM proposed in [15]. The results of SM method are reported from [25]. The figure shows the correlation coefficient of the first 30 subjects of the Bosphorus database. We can see that the results of SM method are very low comparing to the $cICA$ and $cICA_MI$ results. The results obtained by $cICA_MI$ are more correlated to the ground truth than those obtained by $cICA$ however our results are the highest which prove the accuracy of the depth estimation of our method. In addition, the obtained coefficients show that our method is more robust to the identity of the person unlike the other methods where the results are strongly different for different subjects (especially for SM based method). This can be explained by the fact that we use a specific disparity model for each face in the estimation process.

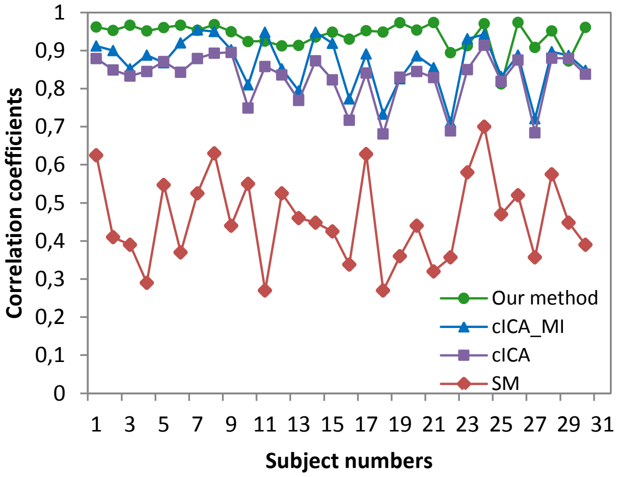


Fig. 19 Correlation coefficients of the first 30 subjects from Bosphorus database.

In figure 20, the NLS based methods proposed in [26] are used for comparison. The $BSS-SfM$ method [8] is also used for the comparison using the results reported in [26]. The NLS based method give the lowest correlation coefficient and is the most sensitive to person identity. The correlation coefficient are slightly improved by using the symmetry property of the face in NLS_MS and highly improved when more than two images are used in the model integration step used in NLS_MI method. Our method gives very high correlation coefficients for all the subjects and they are comparable to those obtained by $BSS-SfM$ and NLS_MI which use a training process with a set of images.

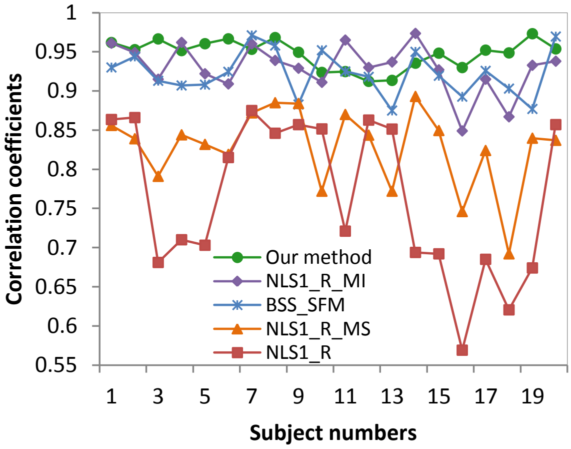


Fig. 20 Correlation coefficients of the first 20 subjects from Bosphorus database.

Taking subject 1 as an example, we compare the estimated values of the 22 features point to their ground truth values as shown in Fig. 21. All the values are normalized between 0 and 1.

In order to evaluate the methods in case of pose variation, we report, in table 22, the coefficients of correlation between the reconstructed and true depth values of five images with different pose variation (as images shown in Fig. 17) taking subject 1 as example. As we can see, the results of the proposed method are approximately the same for the different poses. The mean obtained by our method is better than the other results and with a small standard deviation.

For the convenience of results displaying, only a part of the database (20 and 30 subjects) is used above. However, to evaluate the methods on a large number of examples, we show in table 23 the mean and the standard deviation of the correlation coefficient obtained using 105 subjects.

The quantitative and qualitative evaluation of the results of our method on different databases and using a variety of accuracy measures shows that the proposed strategy is robust to homogenous surfaces of the face and to the small pose variations, accurate and fast.

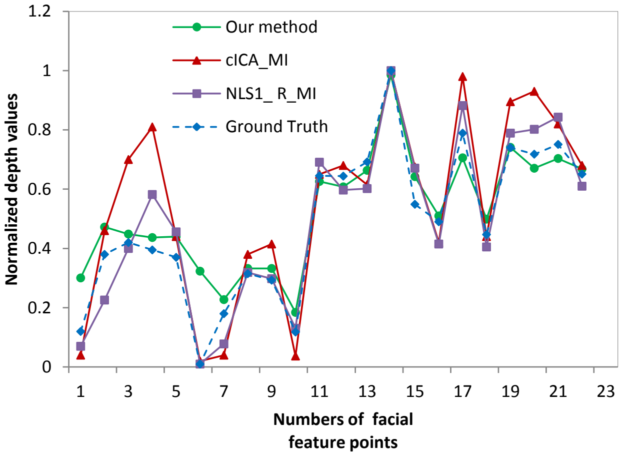
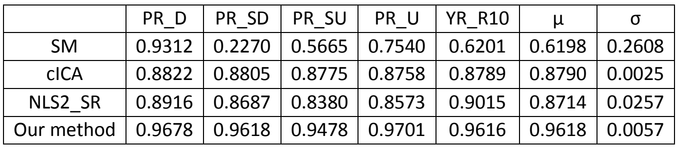
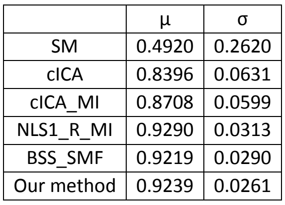


Fig. 21 Comparison of the true depth values and the estimated depth values of the facial feature points of subject 1 in the Bosphorus database.



	PR_D	PR_SD	PR_SU	PR_U	YR_R10	µ	σ
SM	0.9312	0.2270	0.5665	0.7540	0.6201	0.6198	0.2608
cICA	0.8822	0.8805	0.8775	0.8758	0.8789	0.8790	0.0025
NLS2_SR	0.8916	0.8687	0.8380	0.8573	0.9015	0.8714	0.0257
Our method	0.9678	0.9618	0.9478	0.9701	0.9616	0.9618	0.0057

Fig. 22 correlation coefficients for different pose variation.



	µ	σ
SM	0.4920	0.2620
cICA	0.8396	0.0631
cICA_MI	0.8708	0.0599
NLS1_R_MI	0.9290	0.0313
BSS_SMF	0.9219	0.0290
Our method	0.9239	0.0261

Fig. 23 Mean and standard deviation for all subjects in the Bosphorus database.

5 CONCLUSIONS

This paper presents an original attempt of face depth estimation in a passive stereoscopic system. Unlike other general methods used for disparity calculation for general objects, we introduced a dedicated method for face depth estimation that uses the shape characteristics of the human face, obtained by adjusting an ASM, in order to improve the results of the general, local and global, methods. Our method enhances the classical block-matching method for disparity calculation, in terms of depth estimation efficiency, while allowing a very fast processing. The experimental results show that the proposed algorithm produces smooth and dense depth maps of human faces, applicable to a wide range of 3D face reconstruction.

Our approach also opens up many perspectives for improvement and extension. The step of the disparity model reconstruction can also be incorporated in the global methods as graph-cut method, in order to reduce their processing time. The optimization step in the graph-cut methods can be done by identifying an optimal cut in a special graph. In order to construct the graph, this method consider for each pixel, all possible disparities between minimum and maximum values. By integrating our disparity model step, only a small disparity range is selected for each pixel which would reduce considerably the processing time.

The estimation of the disparity model can be improved by using Active Appearance Models [5] instead of ASM, which would give more successful adjustments, because they use the texture information. In order to deal with large pose variations, we aim to use a partial ASM based on the face symmetry in order to deal with profile views because in this case, only one side of the face is required for the fitting step. The 3D Active Appearance Models [30] could also enhance the result to make it robust to large pose variation.

References

1. Berretti, S., Del Bimbo, A., Pala, P.: 3d face recognition using isogeodesic stripes. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(12), 2162 –2177 (2010). DOI 10.1109/TPAMI.2010.43
2. Birchfield, S., Tomasi, C.: Depth discontinuities by pixel-to-pixel stereo. International Journal of Computer Vision **35**(3), 269–293 (1999)
3. Choi, J., Medioni, G., Lin, Y., Silva, L., Regina, O., Pamplona, M., Faltemier, T.: 3d face reconstruction using a single or multiple views. In: Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 3959 –3962 (2010). DOI 10.1109/ICPR.2010.963
4. Chow, C., Yuen, S.: Recovering shape by shading and stereo under lambertian shading model. International journal of computer vision **85**(1), 58–100 (2009)
5. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(6), 681–685 (2001)
6. Cryer, J., Tsai, P., Shah, M.: Integration of shape from shading and stereo. Pattern recognition **28**(7), 1033–1043 (1995)
7. Faltemier, T., Bowyer, K., Flynn, P.: A region ensemble for 3-d face recognition. Information Forensics and Security, IEEE Transactions on **3**(1), 62 –73 (2008). DOI 10.1109/TIFS.2007.916287
8. Fortuna, J., Martinez, A.: Rigid structure from motion from a blind source separation perspective. International Journal of Computer Vision **88**(3), 404–424 (2010). DOI 10.1007/s11263-009-0313-2. URL http://dx.doi.org/10.1007/s11263-009-0313-2
9. Gupta, S., Castleman, K., Markey, M., Bovik, A.: Texas 3d face recognition database. In: Image Analysis & Interpretation (SSIAI), 2010 IEEE Southwest Symposium on, pp. 97–100. IEEE (2010)
10. Hirschmuller, H.: Improvements in real-time correlation-based stereo vision. In: Stereo and Multi-Baseline Vision, 2001.(SMBV 2001). Proceedings. IEEE Workshop on, pp. 141–148. IEEE (2001)
11. Huang, D., Ouji, K., Ardabilian, M., Wang, Y., Chen, L.: 3d face recognition based on local shape patterns and sparse representation classifier. Advances in Multimedia Modeling pp. 206–216 (2011)
12. Huang, Y., Wang, Y., Tan, T.: Combining statistics of geometrical and correlative features for 3d face recognition. In: Proceedings of the British Machine Vision Conference, pp. 879–888 (2006)

13. Kakadiaris, I., Passalis, G., Toderici, G., Murtuza, M., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on **29**(4), 640 –649 (2007). DOI 10.1109/TPAMI.2007.1017
14. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: ECCV (3), pp. 82–96 (2003)
15. Koo, H.S., Lam, K.M.: Recovering the 3d shape and poses of face images based on the similarity transform. Pattern Recognition Letters **29**(6), 712 – 723 (2008). DOI 10.1016/j.patrec.2007.11.018. URL http://www.sciencedirect.com/science/article/pii/S016786550700373X
16. Le, V., Tang, H., Cao, L., Huang, T.: Accurate and efficient reconstruction of 3d faces from stereo images. In: Image Processing (ICIP), 2010 17th IEEE International Conference on, pp. 4265 –4268 (2010). DOI 10.1109/ICIP.2010.5651875
17. Lengagne, R., Fua, P., Monga, O.: 3d stereo reconstruction of human faces driven by differential constraints. Image and Vision Computing **18**(4), 337–343 (2000)
18. Lin, W.Y., Chen, M.Y.: A novel framework for automatic 3d face recognition using quality assessment. Multimedia Tools and Applications pp. 1–17 (2012). URL http://dx.doi.org/10.1007/s11042-012-1092-2
19. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. Computer Vision–ECCV 2008 pp. 504–513 (2008)
20. Park, U., Jain, A.K.: 3d face reconstruction from stereo video. In: Computer and Robot Vision, 2006. The 3rd Canadian Conference on, p. 41 (2006). DOI 10.1109/CRV.2006.1
21. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: Biometrics and Identity Management, pp. 47–56. Springer (2008)
22. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision **47**, 7–42 (2002). DOI 10.1023/A:1014573219977. URL http://portal.acm.org/citation.cfm?id=598429.598475
23. Spreeuwers, L.: Fast and accurate 3d face recognition. International Journal of Computer Vision **93**, 389–414 (2011). URL http://dx.doi.org/10.1007/s11263-011-0426-2
24. Sun, J., Zheng, N.N., Shum, H.Y.: Stereo matching using belief propagation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(7), 787 – 800 (2003). DOI 10.1109/TPAMI.2003.1206509
25. Sun, Z., Lam, K.M.: Depth estimation of face images based on the constrained ica model. Information Forensics and Security, IEEE Transactions on **6**(2), 360–370 (2011). DOI 10.1109/TIFS.2011.2118207
26. Sun, Z.L., Lam, K.M., Gao, Q.: Depth estimation of face images using the nonlinear least-squares model. IEEE Transactions on Image Processing **22**(1), 17–30 (2013)
27. Trucco, E., Verri, A.: Introductory Techniques for 3-D Computer Vision. Prentice Hall PTR, Upper Saddle River, NJ, USA (1998)
28. Wang, S.F., Lai, S.H.: Reconstructing 3d face model with associated expression deformation from a single face image via constructing a low-dimensional expression deformation manifold. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(10), 2115 –2121 (2011). DOI 10.1109/TPAMI.2011.88
29. Wang, Y., Liu, J., Tang, X.: Robust 3d face recognition by local shape difference boosting. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(10), 1858 –1870 (2010). DOI 10.1109/TPAMI.2009.200
30. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2d+3d active appearance models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 535 – 542 (2004)
31. Yan, P., Bowyer, K.W.: A fast algorithm for icp-based 3d shape biometrics. Computer Vision and Image Understanding **107**(3), 195 – 202 (2007). DOI 10.1016/j.cviu.2006.11.001
32. Zheng, Y., Chang, J., Zheng, Z., Wang, Z.: 3d face reconstruction from stereo: A model based approach. In: IEEE International Conference on Image Processing. ICIP., pp. III –65 –III –68 (2007). DOI 10.1109/ICIP.2007.4379247