Image and Vision Computing 38 (2015) 13–23

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A stereo matching approach based on particle filters and scattered control landmarks ☆

Stylianos Ploumpis, Angelos Amanatiadis *, Antonios Gasteratos

School of Engineering, Democritus University of Thrace, 12 Vas. Sofias Str, GR-67100 Xanthi, Greece

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form 20 October 2014
Accepted 1 April 2015
Available online 18 April 2015

Keywords:
Stereo matching
Particle filters
Ground control points
Markov chains
Plane fitting

ABSTRACT

In robot localization, particle filtering can estimate the position of a robot in a known environment with the help of sensor data. In this paper, we present an approach based on particle filtering, for accurate stereo matching. The proposed method consists of three parts. First, we utilize multiple disparity maps in order to acquire a very distinctive set of features called landmarks, and then we use segmentation as a grouping technique. Secondly, we apply scan line particle filtering using the corresponding landmarks as a virtual sensor data to estimate the best disparity value. Lastly, we reduce the computational redundancy of particle filtering in our stereo correspondence with a Markov chain model, given the previous scan line values. More precisely, we assist particle filtering convergence by adding a proportional weight in the predicted disparity value estimated by Markov chains. In addition to this, we optimize our results by applying a plane fitting algorithm along with a histogram technique to refine any outliers. This work provides new insights into stereo matching methodologies by taking advantage of global geometrical and spatial information from distinctive landmarks. Experimental results show that our approach is capable of providing high-quality disparity maps comparable to other well-known contemporary techniques.

1. Introduction

While common digital images provide sufficient 2D information of a scene, several applications including robotics, entertainment and other require full 3D information, such as depth information, which can be accomplished by stereo imaging. Stereo matching is the field of study that aims to determine correspondences in two or more images shot from different viewpoints for obtaining a depth map of the actual scene. This major field of computer vision has been reviewed and categorized accordingly in [1,2] and [3]. Stereo matching techniques can be explicitly classified into two major groups, global and local approaches. The common approach in local algorithms is that the disparity value of a given pixel is calculated only by intensities populating a certain region around that pixel, which is called support window. Several strategies have been introduced to optimize the results of local techniques such as adaptive-weight cost aggregation strategies aiming to reduce implicit assumptions [4]. In global approaches, algorithms determine the disparity values simultaneously based on smoothness assumptions and tend to be iterative for refining their results with energy minimization techniques. The general framework is that local techniques are less computationally expensive than the global techniques and, therefore, they are mostly real time/hardware applicable. Despite this fact, global algorithms can achieve higher accuracy if they deploy cost aggregation techniques in their process. Several applications of belief propagation [5], graph cuts [6] and segmentation combined with plane fitting [7] have been integrated in global and local approaches to boost their performance.

In this paper, we present an approach which tackles the stereo matching problem from a novel perspective compared to known techniques. Our inspiration has been derived from the problem of robot localization in a known environment. This problem can be solved accurately by particle filters, where each particle is a possible state of the robot and each state has its own probability of the robot's actual positioning [8]. The probability is calculated by means of linked series of geometrical and spatial information derived from known landmarks in the environment. We address the stereo matching problem with the same technique, where at first we acquire a set of ground control points (GCPs) [9] by computing multiple disparity maps and subsequently we label them as landmarks. Those features possess an extremely precise disparity value and we utilize them for applying the particle filtering context in a scan line for accurate estimation of disparity values. For improving the effectiveness of particle filtering, we first classify each landmark using a segmentation based technique in each of the paired images. Additionally, we introduce a Markov chain model to increase the accuracy and decrease the computational expensiveness of particle filtering. Finally, we refine our results by applying the RANSAC algorithm [10] combined with a histogram technique

☆ This paper has been recommended for acceptance by Enrique Dunn.

* Corresponding author. Tel.: +30 2541079329.
E-mail address: aamanat@ee.duth.gr (A. Amanatiadis).

http://dx.doi.org/10.1016/j.imavis.2015.04.001
0262-8856/

which, by drawing out the outliers, allows us to obtain high quality disparity maps.

2. Related work

The proposed stereo matching approach is based on particle filtering. Although recent work has introduced particle filtering in 3D optical flow [11], our framework of stereo matching is completely different in every aspect. Moreover, it tends to relate to a small portion of global and local approaches as referred in the stereo related literature.

Global algorithms have lately increased their accuracy due to segmentation-based techniques as introduced by [12,13]. These approaches rely on the assumption of homogeneous color segments which approximate non-overlapping regions in the reference image. Over-segmentation is preferred, since it helps to meet these assumptions in practice, where in every segment the disparity values vary smoothly on a planar surface [7]. Moreover, global approaches have employed belief propagation in their frameworks. Belief propagation was initially established and described in [14] and it has been a popular method for state of the art global algorithms as an energy minimization technique, which is usually constructed by Markov random fields [7,5]. Furthermore, various strategies of graph cuts have been implemented in global approaches to address the same problem of energy minimization [6].

On the contrary, local algorithms examine each pixel independently with the assistance of cost aggregation strategies. Cost aggregation approaches compose a broad chapter in stereo literature which has been evaluated and analyzed extensively in [15]. Several methods of cost aggregation strategies have been introduced over the years based on shiftable windows [16,17], adaptive weights [4], multiple windows [18], and segment based windows [19]. Traditional local algorithms produce less accurate results compared to global ones, but lately this gap has been reduced. A popular method in this category is based on adaptive weights proposed by Yoon and Kweon [4], inspired by the Gestalt principles based on spatial proximity and color similarity. Additional improvements to this method have been proposed by deploying a segmentation based support window [19].

Lastly, there is a category of stereo algorithms which takes advantage of global smoothness assumptions combined with several local constrains. These algorithms fall under the category of semi-global approaches where both global and local techniques are applied for the same correspondence purpose. One of the first semi-global approaches was proposed by Hirschmuller [20], where a different approach of energy minimization was utilized. Semi-global techniques aim at minimizing the global 2D energy function by applying several 1D minimization methods. Generally, semi-global matching algorithms take advantage of scan line energy minimization combined with dynamic programming from several 1D directions. Due to the computational efficiency of semi-global approaches, a wide range of techniques have been introduced to address the stereo correspondence problem, such as discontinuity preserving interpolation in structured environments [21] and segmentation based techniques combined with plane fitting [22].

The overall framework of our approach has been also motivated by recent global approaches which utilize ground control points. These points are described as high confidence matches and their first appearance was in the work of Bobick and Intille [9]. There are two well known methods that can obtain such high confident matches. The first method requires strong feature correspondences, for obtaining high confidence starting points in order to initiate the GCPs calculation [23]. The second technique for acquiring the GCPs relies on the computation of multiple disparity maps based on local approaches and winner-take-all (WTA) strategies as it has been presented in [24]. A similar technique has been presented in [25] where the GCPs are obtained from local matching by oriented spatial filters.

3. Proposed approach

The proposed method has been motivated by the particle filter framework in robot localization. Early particle filter implementations in robot localization can be found in the literature, in which a robot's position has to be recovered from sensor data [8]. In all tracking problems it is essential to decompose the system into three basic models, the *inference* the *state* and the *dynamics* one. In the state model which describes the environment of a mobile robot, a state is usually measured by a two-dimensional Cartesian coordinate system and the orientation of heading. Additionally, in the inference model the sensor data provides the system with an estimation of its position in the environment combining spatial information from the surrounding objects. This spatial sensory information is often called a landmark. Finally, the dynamics model describes the evolution of the system states over time. When the system passes from a certain state $\mathbf{x}_t$ to the next $\mathbf{x}_{t+1}$ in discrete time t, the sensors gather new spatial information from the environment. Furthermore, the current state is related to the previous state $P(\mathbf{x}_{t+1}|\mathbf{x}_t)$ by a Markov chain model. Each state in the Markov chain model is not observable, instead the only observable variable is the set of measurements Z_t acquired from the sensors leading to a stochastic prominence of the true state $\mathbf{x}_t$.

More precisely, particle filters are approximate methods for the calculation of non-linear posterior probabilities in partially observable Markov chain models in discrete time, where an analytic solution to an integral equation is not feasible. In a typical non-linear Bayesian tracking model the list of total measurements up to t is denoted as Z_t and the feature measurements at time t is expressed as $\mathbf{z}_t$, while the set of states over time is expressed as X_t:

$$Z_t = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}, X_t = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}.$$

In order to calculate the posterior density probability $P(\mathbf{x}_t|Z_t)$, conditioned over all given observations until time t, we utilize Bayes' formula:

$$P(\mathbf{x}_t|Z_t) = \frac{p(\mathbf{z}_t|x_t, Z_{t-1})p(\mathbf{x}_t|Z_{t-1})}{p(\mathbf{z}_t|Z_{t-1})}. \quad (1)$$

Furthermore, using the Chapman–Kolmogorov equation for joint probability distributions, and assuming there is independence between observations, Eq. (1) can be rewritten as:

$$P(\mathrm{x}_t|Z_t) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)\int_{x_t-1} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|Z_{t-1})dx_{t-1}}{p(\mathbf{z}_t)}. \quad (2)$$

This solution of the non-linear Bayesian tracking problem is a conceptual solution and it cannot be determined analytically. Several approximations have been proposed over the years in certain sets of cases. In the simplified case of Kalman filters [26] where the dynamics of the system and the observations are not linear, the measurement noise derived from observations forms a distinctive Gaussian distribution. In the case of non-linearities a considerable amount of techniques have been introduced to determine an approximate solution including Monte Carlo approximations [27]. Monte Carlo methods are simulation-based techniques for computing posterior distributions. Particle filters are sequential Monte Carlo models where each particle is a possible state scattered in the known environment as a three dimensional variable containing the orientation and the Cartesian coordinates. In every particle i, there is a certain associated weight $w_t^{(i)}$, which is proportional to the prior probability of importance.

Given a certain amount of support points denoted as particles $X_t = \{\mathbf{x}_1,..., \mathbf{x}_t\}$ up to time t let $\{X_t^{(i)}, w_t^{(i)}\}$, $i = 1 \ldots \nu$ denote a *random measure* that characterizes the posterior density function $P(X_t|Z_t)$, where $\{w_t^{(i)}, i = 1,..., \nu\}$ are the associated weights. The weights are normalized such that $\sum_i w_t^{(i)} = 1$. Then the posterior density at time t can be approximately expressed as:

$$P(\mathbf{x}_t|Z_t) \approx \sum_{i=1}^{\nu} w_t^{(i)} \delta\left(\mathbf{x}_t - \mathbf{x}_t^{(i)}\right) \tag{3}$$

where δ is the Dirac-delta function. This approximation describes the density probability as a discrete weighted function. The weights are chosen recursively using either the method of *sequential importance sampling* (SIS) or in our case the *resampling* method. In the case of SIS a degeneracy problem is observed between particles with negligible weights of insignificant importance. This leads to computational expensiveness due to the continuous update of negligible weights even thought their contribution to the approximate solution is unimportant. On the contrary, the *resampling* approach is capable of eliminating the degeneracy problem by excluding the particles with negligible weights from the update process. Thus, this approach focuses on the particles with important weights for the reducing of the computational complexity.

3.1. Particle filters in stereo correspondence

The proposed approach addresses the stereo correspondence from a different point of view by which we aim to estimate the best disparity value between a range of possible ones, by means of the particle filter framework. The proposed approach is applied to both stereo images combining different sets of characteristics from each one. In our case, the environment for particle filtering corresponds both to the target and the reference image and its possible states correspond to the disparity range of the stereo images. The inference model utilizes the GCPs acquired by the computation of multiple cost efficient disparity maps. Those features serve as landmarks taking advantage of the spatial inference from the global environment and utilize it for the estimation of the posterior probability $P(\mathbf{x}_t|Z_t)$.

3.1.1. The state-dynamics model

The state model can be divided into two categories, the reference state and the set of possible states. The reference state of the model by which we observe the true measurements and compare them to the ones of all possible states, corresponds to a specific pixel in the reference image I_R of size $(m \times n)$ with Cartesian coordinates (x^R, y^R). The reference state at time t can be expressed as:

$$\mathbf{x}_t^R = \left\{x_t^R, y_t^R\right\}.$$

Moreover, the set of possible states correspond to the target image $\boldsymbol{I}_T$. Due to the scan line framework our set of possible states exist in a certain scan line of states that can be determined by the y_t^R coordinate of the reference state $\mathbf{x}_t^R$ and is a subset of the target image:

$$I_T\left(\left\{x_1^T, \ldots, x_n^T\right\}, y_t^R\right) \subset I_T.$$

The particle framework is not applied to the entire scan line but to a small subset defined by the disparity $D_r \in \mathbb{Z}^+$ range of the stereo images. More precisely, the set of possible states in the target image denote the possible disparity states given a certain reference state $\mathbf{x}_t^R$:

$$I_T\left(\left\{x_t^R, \ldots, x_t^R + D_r\right\}, y_t^R\right) \subset I_T\left(\left\{x_1^T, \ldots, x_n^T\right\}, y_t^R\right).$$

The set of possible states $x_t^{(i)}$ in the target image, at time t is defined as:

$$x_t^T = \left\{\left\{x_t^R, \ldots, x_t^R + D_r\right\}, y_t^R\right\}.$$

The current state model has 2 degrees of freedom due to the discrete environment of images and the linear state model. The orientation of states is unnecessary given the fact that the simulated sequences of states are linear and the heading follows only one direction.

The dynamics model describes how the evolution of the system occurs over time in a discrete time-step model. In most cases of particle filter implementations the dynamics of the system has been derived from the system itself and the way the system interacts with the environment. Since we are dealing with rectified stereo image pairs, the dynamics model is forced to follow a scan-line structure. By integrating such model, the particles are propagated by increasing the x coordinate by one pixel per step, allowing us to infer a better converged disparity value for a given reference state. This is achieved by updating the weights according to the likelihood of the new observations. By introducing a hidden state x_t which belongs in the target image in the set of possible states without knowing its parameters the evolution of steps is described as follows:

$$\mathbf{x}_{t+1} = \left\{x_t + 1, y_t^R\right\}.$$

More precisely, the y coordinate of the system remains stationary in a fixed value of the respective scan line, whereas the x coordinate is increased by one pixel. The same dynamics has been applied in each particle in the state model of target image. This dynamics model can be applied only if the following target state of the model takes place under the same segment as the previews state.

The process is repeated over time in a step-wise form, where the weights of particles are updated recursively by the resampling method.

Fig. 1. The extracted GCPs or landmarks of the 'Teddy' target image shown in white color.

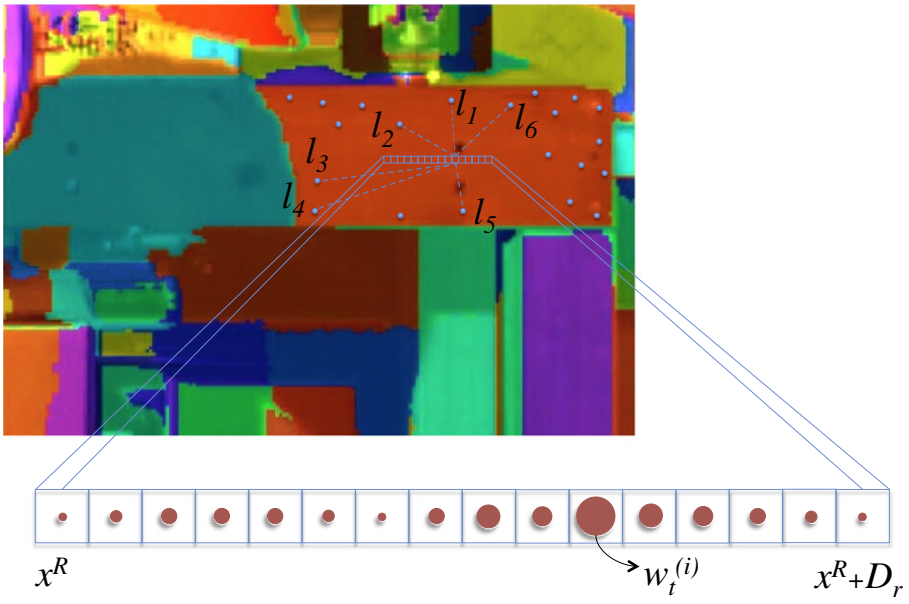

Fig. 2. The scan line particle filtering of the inference-measurement model and the associated weights.

In order to compute the disparity value $d_{\mathbf{x}_t^R}$ for each reference state $\mathbf{x}_t^R$ in the reference image we only utilize the x_t coordinate (parameter) of the hidden state:

$$d_{\mathbf{x}_t^R} = \left| x_t - x^R \right|.$$

After a sufficient amount of resampling periods when the posterior probability is obtained from the measurement model, all the parameters (coordinate x_t) of the hidden state can be estimated by taking expectation of the state with respect to the posterior probability of particle filtering. Then the correct disparity value $\hat{d}_{\mathbf{x}_t^R}$ for each reference state can be computed by:

$$\hat{d}_{\mathbf{x}_t^R} = \left| E\{x_t\} - x^R \right|.$$

Due to the small size of the disparity range the resampling period is decreased proportional to the state model, compared to other implementations of particle filters.

3.1.2. The inference-measurement model

The inference model describes the statistical measurements adopted by the dynamics model to accurately compute the prior likelihood. Given a certain set of measurements in a precise state, the measurement model is denoted as $P(\mathbf{z}_t|\mathbf{x}_t^{(i)})$. In the examined case of stereo correspondence, the measurement model is constructed by the GCPs with known disparity values. The algorithm to obtain the GCPs is applied in both reference and target images and is described in Section 4.

Those GCPs are denoted as landmarks with known Cartesian coordinates in the image environment along with their precise disparity value. Each landmark in the reference image can be expressed as a vector:

$$l_j^R = \left\{ x_j^R, y_j^R, d_j \right\}.$$

The respective landmark's j coordinates in the target image are acquired as follows:

$$l_j^T = \left\{ x_j^R + d_j, y_j^R \right\}.$$

We exploit those landmarks strictly as spatial observations by utilizing the Cartesian coordinates in the two dimensional image space. Since the distinctive set of landmarks possesses various disparity values, we group them under certain planar regions with homogeneous disparity values, in order to use the appropriate landmarks in the inference model under a certain region. The grouping has been achieved by over-segmentation and more precisely by Mean-Shift segmentation [28]. Assuming that the reference state is under a precise segment θ with a certain set of landmarks $L_\theta^R = \{l_1^R, \ldots, l_N^R\}$ the likelihood can be computed by the following procedure.

For each possible state we compute the Euclidean distance between the current state pixel and the landmark coordinates and compare them with the actual Euclidean distance of the reference state pixel and its respective landmark. We only consider landmarks for the computation of likelihood that are closer to the reference state. The number of those landmarks is truncated by a certain variable λ. If the number of landmarks N in a given segment θ is greater than the variable λ, then the closest λ number of landmarks are chosen for the computation of the likelihood:

$$l_j = \begin{cases} closest\ \ l_\lambda & \text{if } \lambda < N \\ l_N & \text{else} \end{cases}.$$

In order to compute the weights in each state, we employ the following equations, where at first we introduce for each landmark the measurement function:

$$z_t^{(j)} = D_j^R - D_{i,j(x_t^{(i)})}^T + u_t^o$$

Table 1
The accuracy and density of GCPs along with the percentage of segments without GCPs.

Stereo pairs	Tsukuba	Venus	Teddy	Cones	Midd1	Wood2	Avg.
Density (%)	33	39.1	26.5	20.8	17.8	31.4	28.68
Accuracy (%)	99.6	99.7	99.2	99.5	98.7	99.8	99.4
Pixels in seg. without GCPs (%)	1.1	0.69	1.3	1.9	0.91	0.82	1.12

where $D^T_{i,j(x_t^{(i)})}$ is the Euclidean distance measured from the landmark l_j^T to a possible state $\mathbf{x}_t^{(i)}$ of the target image and D_j^R is the Euclidean distance measured from the reference state $\mathbf{x}_t^R$ to the respective landmark l_j^R in the reference image, as shown in Fig. 2. The parameter u_t^o is the zero mean Gaussian observation noise. Noisy measurements can be made by measuring the distance between a given pixel and a certain GCP in the image environment. Additionally, for the computation of individual likelihood we employ a conditional probability function based on the previews measurement function:

$$p\left(z_t^{(j)}|\mathbf{x}_t^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp^{-\frac{\left(z_t^{(j)}\right)^2}{\sigma^2/2}}. \tag{4}$$

Given that conditional independence exists between multiple landmarks, a combined likelihood is derived simply by multiplying all the individual likelihoods from the model:

$$p\left(\mathbf{z}_t|\mathbf{x}_t^{(i)}\right) = \prod_{j=1}^{N} p\left(z_t^{(j)}|\mathbf{x}_t^{(i)}\right). \tag{5}$$

In order to calculate the weight $w_t^{(i)}$ for a given state $\mathbf{x}_t^{(i)}$ at time t in the target image, we utilize the following formula and we normalize it accordingly.

$$\hat{w}_t^{(i)} = w_{t-1}^{(i)} p\left(\mathbf{z}_t|\mathbf{x}_t^{(i)}\right) \tag{6}$$

$$w_t^{(i)} = \frac{\hat{w}_t^{(i)}}{\sum_i \hat{w}_t^{(i)}}$$

The more similar these distances are, in a certain state, the bigger the weight of the particular particle is. Since we use the resampling method in our model, particles with bigger weights exhibit higher probability to survive during the resampling process.

Although this structure of inference model has been applied in the entire image, a few segments appear not to be able to support this model due to the lack of landmarks. These limited segments are very small to accommodate even one landmark and they tend to be uniform and not-textured regions. The pixel percentage in each image pair that cannot accommodate the particle filter framework can be seen in Table 1. In some cases as the 'Midd1' stereo pair, large uniform segments appear (the background uniform segment) with a very little portion of landmarks. However the amount of those features is adequate for the application of particle filtering in those segments. Thus, in 'Midd1' stereo pair only the 0.91% of the image is computed solely by the cost aggregation method. Concretely, the pixel percentage that cannot accommodate the particle filter model is proportional not only to the size of the segments but also to the number of landmarks that reside in these segments.

However, in segments that do not accommodate even a single landmark, a typical model of cost aggregation strategy in a support window has been applied in a scan line framework. The method we utilized was the one of adaptive weights [4], which is based on spatial proximity and color similarity. Each pixel in the support window is weighted according

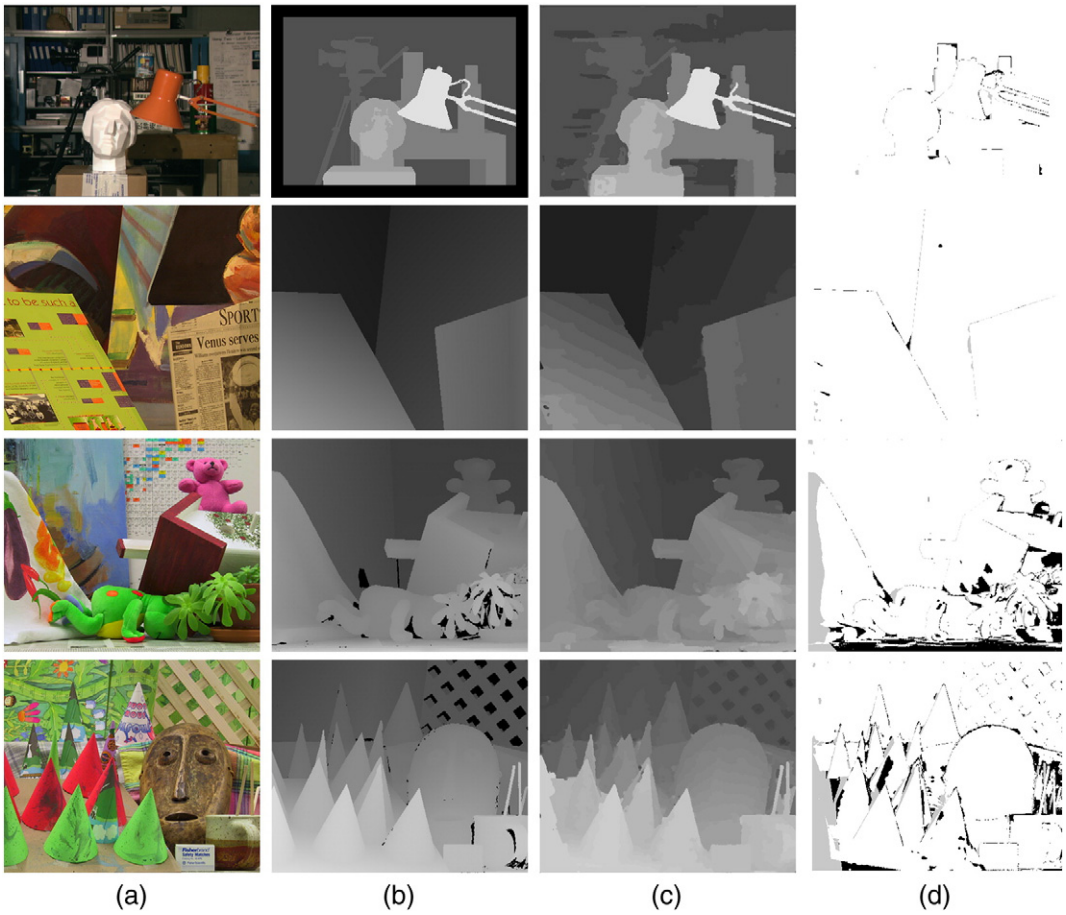

Fig. 3. Results of the stereo image pairs of the Middlebury data set: 'Tsukuba,' 'Venus,' 'Teddy,' and 'Cones' (from top to bottom): (a) original left images of stereo pairs, (b) ground truth maps, (c) results of the proposed method, (d) error maps.

to spatial proximity and color similarity in the CIELAB color space with regards to the central pixel of the window. Let p_k be a random pixel of the support window in reference image and p_c the central pixel, the respective weight can be defined as:

$$w_r(p_k, p_c) = exp\left(-\frac{d_p(p_k, p_c)}{\gamma_p} - \frac{d_c(I_r(p_k), I_r(p_c))}{\gamma_c}\right) \tag{7}$$

where d_p and d_c are the Euclidean distance between coordinated pairs and the Euclidean distance in the CIELAB color space, respectively, according to the central pixel, with γ_p,γ_c the proximity and color parameters. Similarly, we compute the weights $w_t(q_k, q_c)$ for the target image.

The total aggregation cost of correspondence (p_c,q_c) can be expressed as:

$$C(p_c, q_c) = \frac{\sum_{p_k \in W_r, q_k \in W_t} w_r(p_k, p_c) w_r(q_k, q_c) TAD(p_k, q_k)}{\sum_{p_k \in W_r, q_k \in W_t} w_r(p_k, p_c) w_r(q_k, q_c)} \tag{8}$$

where W_t,W_r are respectively the target and the reference support windows of correspondence, while the point score in the support window is evaluated by the Truncated Absolute Difference (TAD).

3.2. Markov chain model

Although we have embedded a belief propagation framework in our method we did not utilize our technique for energy minimization as previously proposed by other authors [7,5]. On the contrary, we use a Markov chain model in a scan line framework so as to reduce the computational complexity of the resampling part and assist the particle filtering structure to converge in the best possible state by adding a correlative weight in the process.

The state model in the Markov chain framework consists of the disparity levels {1,…,D_r}. Each state S_i exhibits a certain probability of occurrence $P(S_i)_{t-1}$ and a transition probability to any other state $P(S_j|S_i)_{t-1}$ at time $t-1$. In order to calculate these transition probabilities, the Markov chain structure requires a sufficient amount of past observations as linked occurrences. The linked observations in our approach are the past disparity values in the respective scan line structure of a certain segment. These observations can be considered satisfactory if and only if all the observations exist under the same segment as the reference state $\mathbf{x}_t^R$ of the particle filter model. Moreover, future probabilities $P(S_i)_t$ can be computed at time t by the sum of conditional probabilities:

$$P(S_i)_t = \sum_{j=1}^{D_r} P\left(S_i|S_j\right)_{t-1} P\left(S_j\right)_{t-1}. \tag{9}$$

In order to avoid overfitting, we apply Laplace smoothing in all of the transition probabilities in our model as follows:

$$P\left(S_i|S_j\right)'_{t-1} = \frac{P\left(S_i|S_j\right)_{t-1} + \alpha}{\omega + \alpha\tau} \tag{10}$$

where $\alpha > 0$ is the smoothing parameter, ω is the number of trials and τ is the number of states in our model, which corresponds to D_r.

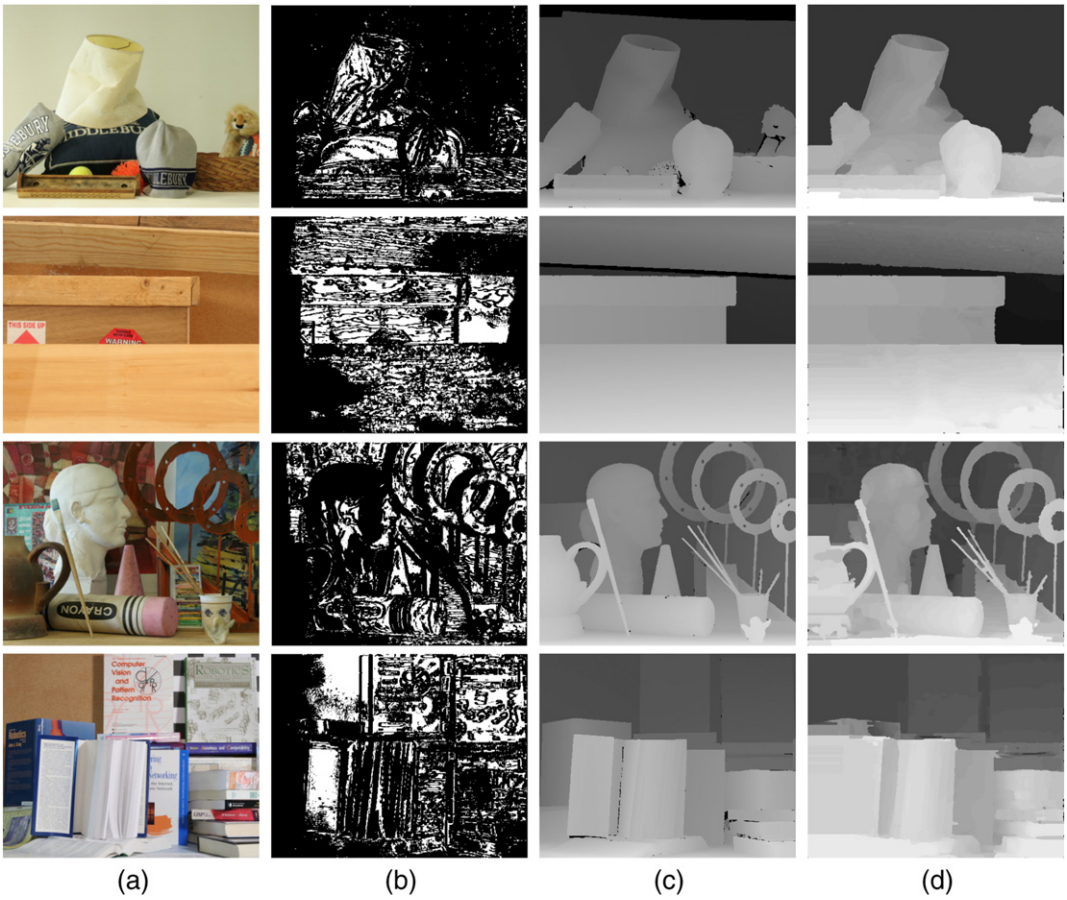

Fig. 4. Results of the stereo image pairs of the Middlebury data set: 'Midd1,' 'Wood2,' 'Art,' and 'Books' (from top to bottom): (a) original left images of stereo pairs, (b) GCPs shown in white color, (c) ground truth maps, (d) results of the proposed method.

Smoothing prevents the assignment of zero probabilities to states that do not occur in a certain sample of past observations. In this way the probability distribution is smoothed across the entire state framework.

Assuming that the next disparity value will be S_i with probability $P(S_i)_t$ for a certain reference state $\mathbf{x}_t^R$ we can acquire the predicted target state in the target image. In order to incorporate the probability $P(S_i)_t$ in our framework an additive weight is assigned to the respective

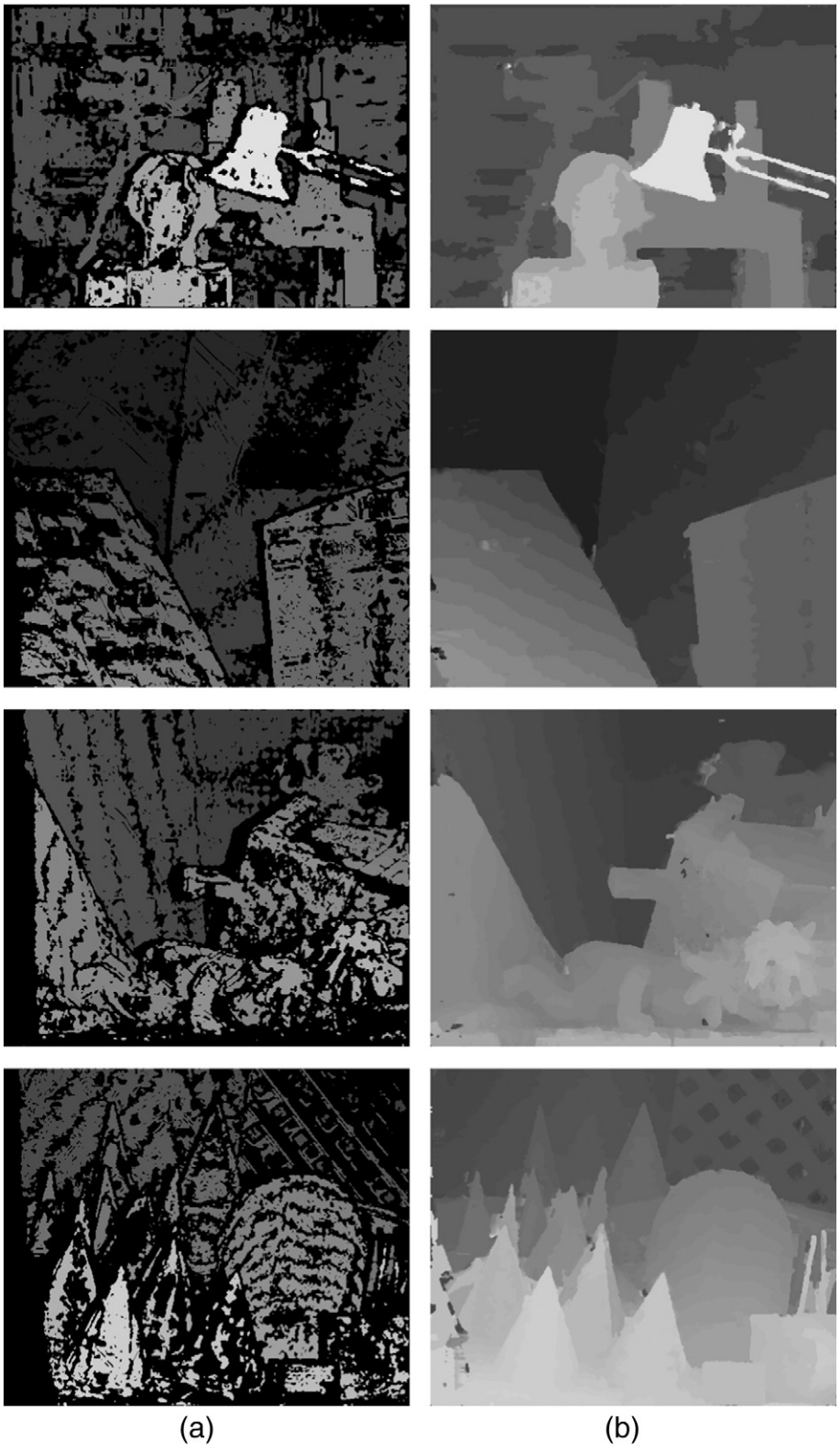

Fig. 5. The initial and converged disparity maps of the stereo image pairs of the Middlebury data set: 'Tsukuba,' 'Venus,' 'Teddy,' and 'Cones' (from top to bottom): (a) initial disparity maps, (b) converged disparity maps.

Table 2
Results of our method compared to other algorithms using the Middlebury test bed.

Algorithm	Tsukuba			Venus			Teddy			Cones			
	Nonocc	All	Disc	Nonocc	All	Disc	Nonocc	All	Disc	Nonocc	All	Disc	APBP (%)
RecursiveBF [31]	1.85	2.51	7.45	0.35	0.88	3.01	6.28	12.1	**14.3**	**2.80**	8.91	**7.79**	5.68
MSWLinRegr [32]	1.46	1.72	7.89	0.57	0.92	6.71	6.11	**11.0**	15.6	3.12	**8.76**	8.52	6.04
Proposed method	**0.98**	**1.53**	**5.31**	**0.25**	**0.69**	**2.60**	9.94	14.8	22.7	6.58	13.3	15.0	7.80
AdaptWeight [4]	1.38	1.85	6.90	0.71	1.19	6.13	7.88	13.3	18.6	3.97	9.79	8.26	6.67
HistoAggr [33]	2.47	2.71	11.1	0.74	0.97	3.28	8.31	13.8	21.0	3.86	9.47	10.4	7.33
PlaneFitSGM [22]	3.13	4.20	14.9	1.08	1.87	14.6	**5.68**	11.6	17.1	3.79	9.26	11.3	8.21
FastAggreg [34]	1.16	2.11	6.06	4.03	4.75	6.43	9.04	15.2	20.2	5.37	12.6	11.9	8.24
CSBP [35]	2.0	4.17	10.5	1.48	3.11	17.7	11.1	20.2	27.5	5.98	16.5	16.0	11.4

Bold values indicate best performance.

target state $\mathbf{x}_t^{(i)}$ of particle filter process. In order to avoid any errors in resampling caused by the Markov chain model, the additive weight is proportional to the original weight $w_t^{(i)}$ of particle filtering in that exact state. The new weight $w_t^{(i)\prime}$ given the probability $P(S_i)_t$ can be described by the following equation:

$$w_t^{(i)\prime} = \left(1 + \frac{1}{D_r(1-P(S_i)_t)}\right) w_t^{(i)}. \tag{11}$$

After the additional weight in the process of particle filtering, normalization is required. This approach of Markov chains reduces the resampling process in the particle filtering model resulting to faster convergence. The resulting weights encompass both the probability of the measurement framework of particle filtering $P(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{z}_t)$ and the highest observation probability of the Markov chain structure $P(S_i)_t$.

From the aspect of accuracy this model assists particle filter structure to converge, particularly in the case of a segmented region which appears to be a planar surface with multiple isometric disparity values. More precisely, by utilizing a separate Markov chain in the process we are able to introduce a smoothness technique in the scan-line framework of particle filtering inside a precise segment. Additionally, by exploiting the past disparity values of a certain scan-line we are able to assist the particle filtering model to converge to the right value of disparity even though some landmarks might not be accurate enough to take measurements from. Only a separate Markov chain that observes only the past disparity values as states and not the measurements from the particle filtering is able to provide us with that information. One disadvantage of that Markov chain model is that the model itself needs a certain amount of past observations in a certain scan-line inside a segment in order to begin its process. Compared to [29] which utilizes multiple neighbor states for the calculation of posterior probability we only utilize neighbor states that are in the respective scan-line of our particle filtering model and we exploit them through the separate Markov chain process.

3.3. Disparity refinement

This section describes how the disparity refinement method has been applied, along with a histogram technique for the optimization of the overall disparity map. The implementation of the refinement includes the RANSAC [10] plane fitting algorithm. The algorithm has been applied in every segment in order to fit a planar surface based on the disparity values in that segment. Plane fitting might, however, produce false results if the parameters of the algorithm are not properly chosen according to the disparity map structure.

Therefore, we apply a histogram technique to resolve this problem and we keep the same parameter values of the RANSAC algorithm in every stereo image pair utilized for this application. Then, each disparity value of the segment is compared with the disparity value before the plane fitting. The disparity error variance V_e of all disparity values is computed for that particular pixel k with the assistance of the cost aggregation technique utilized in the particle filtering section. If the disparity error variance V_d between the disparity value chosen from the plane fitting ($d_{planefitting}$) and the disparity value created from the particle filtering framework ($d_{original}$) is greater than the overall error variance V_e, then the original disparity value is selected as the optimal one, as follows:

$$d_{(k)} = \begin{cases} d_{planefitting} & \text{if } V_d < V_e \\ d_{original} & \text{else} \end{cases}$$

Fig. 6. The extracted landmarks computed by ORB algorithm for the 'Tsukuba' target image.

Fig. 7. The extracted landmarks computed by SIFT algorithm for the 'Tsukuba' target image.

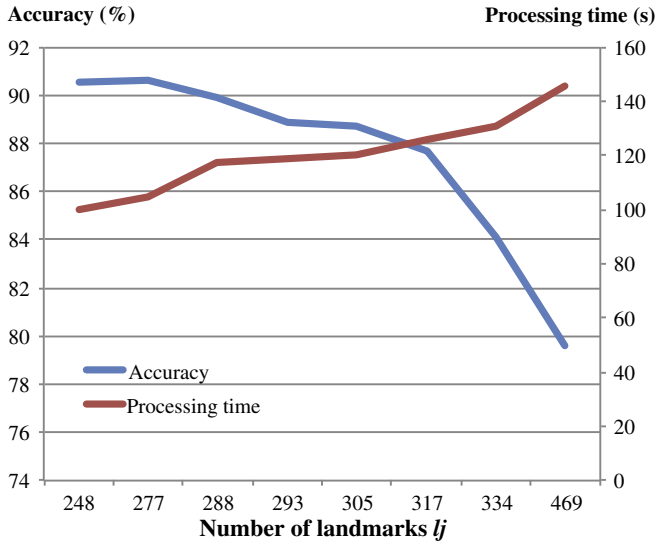

Fig. 8. The accuracy and the processing time of our method with SIFT algorithm compared to the number of landmarks for the 'Tsukuba' stereo pair.

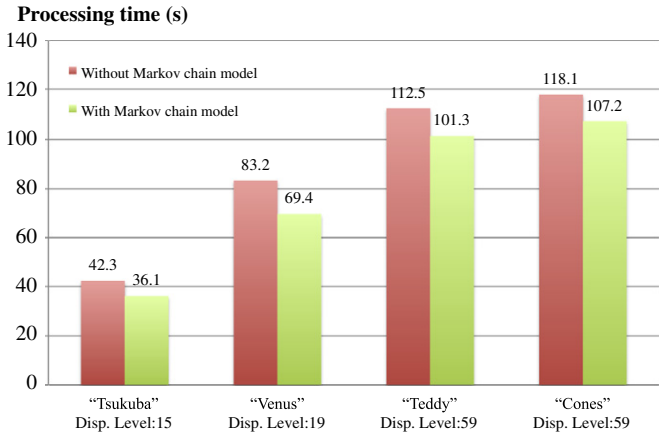

Fig. 10. The overall processing time improvement gained by the application of Markov chain model for various stereo pairs.

4. Experimental results

The proposed method has been evaluated using the Middlebury benchmark [30]. The Middlebury data set include four stereo image pairs: 'Tsukuba', 'Venus,' 'Teddy,' and 'Cones', by which we are able to compare our algorithm with state of the art techniques in stereo vision and evaluate our approach.

Our algorithm utilized the same parameters during the entire process of the disparity computation in all benchmark stereo image pairs without making any adjustments. In the Mean-Shift segmentation part the constant parameter set is: $\sigma_s = 7$ the spatial radius, $\sigma_r = 7$ the range radius and $M_r = 35$ the minimum region size. Moreover, in the inference model the standard deviation of the conditional probability function has been selected to be $\sigma = 0.5$. Additionally, the parameter set utilized for the cost aggregation strategy has the following values: $\gamma_c = 5$ for the color similarity parameter and $\gamma_p = 17.5$ for the spatial proximity parameter; the TAD parameter is $T = 40$ and the size of the support window is 35×35. After the appropriate evaluation of our framework in numerous images, the optimum λ variable in the measurement model was set to 30. Finally, the parameter of the Laplace smoothing equation was assigned to $\alpha = 1$.

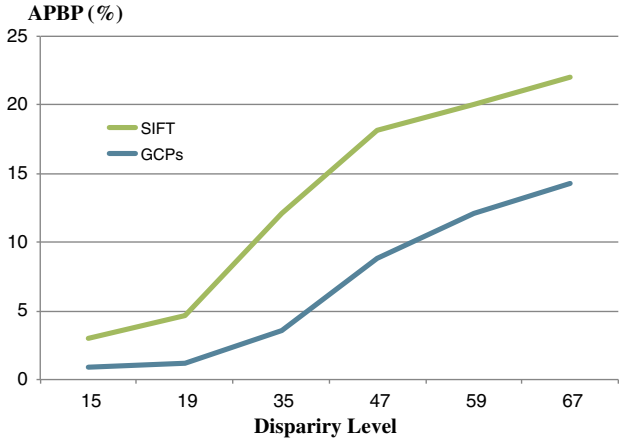

Fig. 9. The average percent of bad pixels compared to the number of disparity levels for the associate methods utilized for the acquirement of landmarks.

Fig. 3, illustrates the estimated disparity maps computed by the proposed method, along with the ground truth maps and the error maps for each stereo pair. The disparity maps are evaluated according to the average percent of bad pixels (APBP) where the absolute disparity error is greater than one. Moreover, Fig. 4 shows additional estimated disparity maps along with the associated GCPs and ground truth maps for every image stereo pair.

In order to acquire the desired landmarks we implemented a voting strategy along with multiple cost efficient disparity maps. More precisely, we compute three disparity maps with winner-take-all strategy (WTA). Firstly, we compute a disparity map based on normalized cross correlation (NCC) [36] with 5×5 window size, followed by a disparity map resulted by the sum of humming distances (SHD) method with the same window size. Lastly, we obtain the third disparity map utilizing the adaptive weights (AW) method proposed by Yoon and Kweon [4] with 39×39 window size. To further reduce any outliers, we apply the same technique in both images and we compute the left-right consistency check [37]. A pixel is labeled as GCP if the disparity level in all of the

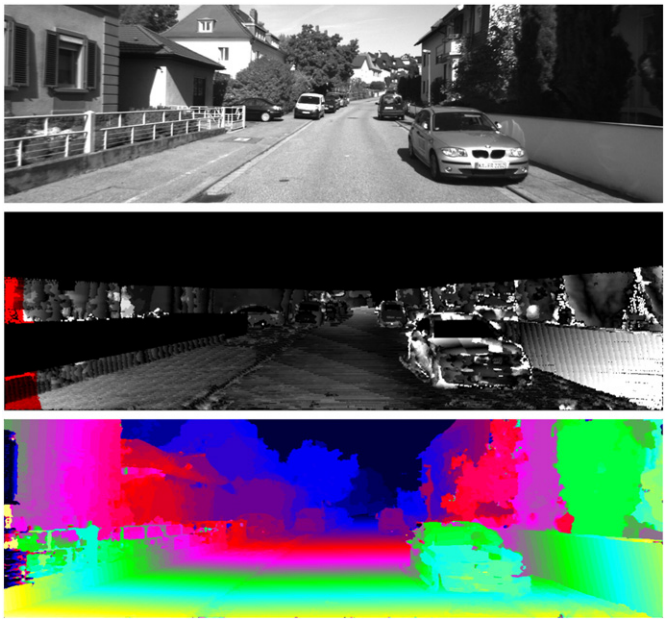

Fig. 11. Results of a stereo image pair of the KITTI data set: reference image, error map, disparity map, (from top to bottom).

three disparity maps is consistent and has survived the left–right consistency. Finally, the pixel should not be at the vicinity of an edge where vast disparity changes might take place and occlusions can be detected; computed by Canny edge detector [38]. The NCC and SHD methods are utilized in order to acquire a sparse set of GCPs whereas the AW method is utilized to choose the most accurate of those features. Additionally, the simplicity of SHD and NCC does not introduce huge computational expensiveness. The density and the accuracy of GCPs in each image pair can be seen in Table 1. The acquired GCPs for the 'teddy' stereo pair are shown in Fig. 1.

Additionally, Fig. 5 depicts the initial disparity maps after one iteration and the converged maps for each stereo pair without any refinement. The areas shown in black color in the initial disparity maps are the areas that have not converged yet. The convergence rate for each stereo pair is different and it depends on the disparity levels of each stereo pair, and on the uniformity of landmarks in each segment. Furthermore, each segment in the image has a different convergence rate and the overall rate is accounted when the last segment converges. The convergence rate for each stereo pair was: 10 for 'Tsukuba', 6 for 'Venus', 14 for 'Teddy', and 15 for 'Cones'.

The evaluation of our method was compared to other similar techniques as shown in Table 2, where APBAP has been divided according to Middlebury benchmark into 3 categories: The bad pixels in the non-occluded region (nonocc), all the pixels (all), and lastly, the pixels near the area of occluded pixels (disc). We compared our approach with similar local and semi global techniques utilizing belief propagation [35] and plane fitting [22], methods utilizing MSW descriptors [32] and cost aggregation strategies [4,33].

For the 'Tsukuba' stereo pair, our algorithm is currently ranked among the best disparity algorithms in the Middlebury benchmark with 0.98 nonocc error, 5.31 disc error and 1.53 all error. Relatively similar results appear also for the 'Venus' stereo pair. As far as the 'Teddy' and 'Cones' stereo pairs are concerned, the results are adequate even if the scattered landmarks are not equally scattered in the image environment. The pixel percentages in the 'Teddy' and 'Cones' synthetic images that do not accommodate any GCPs are 1.3% and 1.9%, respectively. In these relatively extended regions, cost aggregation method was applied instead of particle filtering, thus slightly deteriorated results were reported.

Various methods have been evaluated during our frameworks for the acquirement of landmark, such as feature matching with numerous descriptors. Firstly we implemented our method along with ORB features [39] for a real time applicable scenario. The main disadvantage of this descriptor was that the features were not equally scattered in the image environment and more precisely there were no strong features at all in the main questionable non-textured areas as shown in Fig. 6. Additionally ORB strong features are frequently concatenated in corners and edges where the disparity level changes and occluded regions might appear. Moreover we assessed SIFT algorithm [40] as a feature matching descriptor. Alternative from ORB features the SIFT algorithm extracted more scattered keypoints. Conclusively SIFT algorithm was able to accommodate a few strong features in texture-less areas which they could be exploited as landmarks, as shown in Fig. 7. Nevertheless the quantity of landmarks was considerably less compared to the GCPs in every stereo pair utilized for this research.

Although, we could easily increase the number of features in each stereo pair to overcome these drawbacks, the increment of landmarks does not necessarily imply better accuracy. More precisely this approach of feature matching relies on strong features and it is not possible to enhance the accuracy if we increase the number of landmarks since the additional landmarks would be less strong in terms of accuracy. Fig. 8, depicts the tradeoff between the accuracy of the feature correspondence approach and the extracted number of utilized features.

In terms of accuracy between the two proposed methods we constructed the diagram shown in Fig. 9, where the tradeoff between the disparity levels and the average percent of bad pixels is explained. As the disparity levels increase the accuracy decreases due to the mismatching pixels and the enlargement of ambiguous regions. However, the GCPs approach is always more accurate than the one utilizing the SIFT algorithm.

Table 3
Results of our method using the KITTI test bed.

Error	Out-noc	Out-all	Avg-noc	Avg-all
2 pixels	15.88%	16.64%	1.3 px	1.4 px
3 pixels	8.96%	8.68%	1.3 px	1.4 px
4 pixels	6.25%	6.69%	1.3 px	1.4 px
5 pixels	3.55%	4.01%	1.3 px	1.4 px

All the experiments, as well as the processing time measurements, were executed on a 2.4 GHz Intel Core Duo processor. The overall processing time of our method was: 36.1 s for 'Tsukuba', 69.04 s for 'Venus', 101.3 s for 'Teddy', and 107.2 s for 'Cones'. After the implementation of the Markov chain framework, which aimed at reducing the computational complexity of the resampling part in particle filtering, we observed that the overall time in every stereo pair was reduced approximately by 10% as shown in Fig. 10.

Our method has been evaluated also by the large image data set KITTI [41]. Fig. 11 demonstrates the resulted disparity map for a specific image pair and Table 3 presents the respective results for that stereo pair.

Experimental results reveal that the proposed particle filter framework works better when being moved outside from the laboratory to the real world. The global reasoning of the method along with its local particle filtering provide slightly better results compared to local and semi-global state-of-the-art algorithms. This is mainly because of the fact that the proposed method does not rely on assumptions met in purely local methods, such as single or bi-labeled window assumption [42], or the fronto-parallel segment assumption [43], which are often violated in real world scenes.

5. Conclusion

In this paper, we presented a stereo matching approach motivated by the particle filter framework in robot localization. Strong scale invariant features have been applied in each stereo pair in order to create the inference structure for the particle filter framework. Furthermore, the particle filtering structure has been implemented for the first time to determine the best possible disparity value of each pixel in a stereo pair. Moreover, the Markov chain model has been introduced in the process to reduce the computational complexity of particle filtering, along with a histogram refinement method. The experimental results showed that our method is capable of producing high quality disparity maps by taking advantage of global spatial and geometrical inference from distinctive sets of landmarks.

References

[1] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vision 47 (1) (2002) 7–42.
[2] R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2010.
[3] L. Nalpantidis, G.C. Sirakoulis, A. Gasteratos, Review of stereo vision algorithms: from software to hardware, Int. J. Optomechatronics 2 (4) (2008) 435–462.
[4] K. Yoon, I. Kweon, Adaptive support-weight approach for correspondence search, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 650–656.
[5] Q. Yang, L. Wang, R. Yang, H. Stewénius, D. Nistér, Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2009) 492–504.
[6] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, Proc. IEEE ICCV 2001, pp. 508–515.
[7] A. Klaus, M. Sormann, K. Karner, Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure, Proc. IEEE ICPR 2006, pp. 15–18.
[8] S. Thrun, D. Fox, W. Burgard, F. Dellaert, Robust Monte Carlo localization for mobile robots, Artif. Intell. 128 (1) (2001) 99–141.

[9] A. Bobick, S. Intille, Large occlusion stereo, Int. J. Comput. Vision 33 (3) (1999) 181–200.
[10] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[11] S. Hadfield, R. Bowden, Kinecting the dots: particle based scene flow from depth sensors, Proc. IEEE ICCV 2011, pp. 2290–2295.
[12] S. Birchfield, C. Tomasi, Multiway cut for stereo and motion with slanted surfaces, Proc. IEEE ICCV 1999, pp. 489–495.
[13] H. Tao, H.S. Sawhney, R. Kumar, A global matching framework for stereo computation, Proc. IEEE ICCV 2001, pp. 532–539.
[14] J. Sun, N.-N. Zheng, H.-Y. Shum, Stereo matching using belief propagation, IEEE Trans. Pattern Anal. Mach. Intell. 25 (7) (2003) 787–800.
[15] F. Tombari, S. Mattoccia, L. Di Stefano, E. Addimanda, Classification and evaluation of cost aggregation methods for stereo correspondence, Proc. IEEE CVPR 2008, pp. 1–8.
[16] A. Fusiello, V. Roberto, E. Trucco, Symmetric stereo with multiple windowing, Int. J. Pattern Recognit. Artif. Intell. 14 (08) (2000) 1053–1066.
[17] L. Nalpantidis, A. Amanatiadis, G.C. Sirakoulis, A. Gasteratos, Efficient hierarchical matching algorithm for processing uncalibrated stereo vision images and its hardware architecture, IET Image Process. 5 (5) (2011) 481–492.
[18] H. Hirschmüller, P.R. Innocent, J. Garibaldi, Real-time correlation-based stereo vision with reduced border errors, Int. J. Comput. Vision 47 (1) (2002) 229–246.
[19] F. Tombari, S. Mattoccia, L. Di Stefano, Segmentation-based adaptive support for accurate stereo correspondence, Adv. Image Video Technol. (2007) 427–438.
[20] H. Hirschmuller, Accurate and efficient stereo processing by semi-global matching and mutual information, Proc. IEEE CVPR 2005, pp. 807–814.
[21] H. Hirschmuller, Stereo vision in structured environments by consistent semi-global matching, Proc. IEEE CVPR 2006, pp. 2386–2393.
[22] M. Humenberger, T. Engelke, W. Kubinger, A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality, Proc. IEEE CVPRW 2010, pp. 77–84.
[23] M. Lhuillier, L. Quan, Match propagation for image-based modeling and rendering, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1140–1146.
[24] L. Wang, R. Yang, Global stereo matching leveraged by sparse ground control points, Proc. IEEE CVPR 2011, pp. 3033–3040.
[25] J.C. Kim, K.M. Lee, B.T. Choi, S.U. Lee, A dense stereo matching using two-pass dynamic programming with generalized ground control points, Proc. IEEE CVPR 2005, pp. 1075–1082.
[26] G. Welch, G. Bishop, An introduction to the Kalman filter, 1995.
[27] A. Doucet, N. De Freitas, N. Gordon, et al., Sequential Monte Carlo methods in practice, vol. 1, Springer, 2001.
[28] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[29] S. Yang, Particle filtering based estimation of consistent motion and disparity with reduced search points, IEEE Trans. Circuits Syst. Video Technol. 22 (1) (2012) 91–104.
[30] D. Scharstein, R. Szeliski, Middlebury Stereo Evaluation — Version 2, http://vision.middlebury.edu/stereo/.
[31] Q. Yang, Recursive bilateral filtering, Proc. IEEE ECCV 2012, pp. 399–413.
[32] T. Liu, X. Dai, Z. Huo, X. Zhu, L. Luo, A cost construction via MSW and linear regression for stereo matching, Proc. IEEE ICPR 2012, pp. 914–917.
[33] D. Min, J. Lu, M. Do, A revisit to cost aggregation in stereo matching: how far can we reduce its computational redundancy? Proc. IEEE ICCV 2011, pp. 1567–1574.
[34] F. Tombari, S. Mattoccia, L. Di Stefano, E. Addimanda, Near real-time stereo based on effective cost aggregation, Proc. IEEE ICPR 2008, pp. 1–4.
[35] Q. Yang, L. Wang, N. Ahuja, A constant-space belief propagation algorithm for stereo matching, Proc. IEEE CVPR 2010, pp. 1458–1465.
[36] J. Lewis, Fast normalized cross-correlation, Vision interface, vol. 10 1995, pp. 120–123.
[37] G. Egnal, R.P. Wildes, Detecting binocular half-occlusions: empirical comparisons of five approaches, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1127–1133.
[38] J.F. Canny, Finding edges and lines in images, Massachusetts Inst. of Tech. Report 1 (1983).
[39] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, Orb: an efficient alternative to sift or surf, Proc. IEEE ICCV 2011, pp. 2564–2571.
[40] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[41] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The Kitti Vision Benchmark Suite, Proc. IEEE CVPR, 2012.
[42] M. Agrawal, L.S. Davis, Window-based, discontinuity preserving stereo, Proc. IEEE CVPR, vol. 1 2004, pp. I-66–I-73 http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1315015&abstractAccess=no&userType=inst.
[43] Q. Yang, L. Wang, R. Yang, H. Stewenius, D. Nister, Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2009) 492–504 http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4492788&abstractAccess=no&userType=inst.