Stereo matching with partial information

Y. Cem Sübakan, Ömer Can Gürol and Çağatay Dikici

Boğaziçi University
Electrical and Electronics Engineering
Bebek 34342, Istanbul, Turkey

ABSTRACT

In this paper, we address the stereo matching problem in stereo videos using partially informed Markov Random Fields (MRFs) using the motion information between subsequent frames as a side information. We use the motion vectors within one of the videos to regularize the disparity estimate using this motion field. The proposed scheme enables us to obtain good disparity estimates using faster and simpler disparity finding algorithms in each step.

Keywords: Stereo Matching, Disparity Estimation, Motion Vector, Markov Random Fields, Partial Information

1. INTRODUCTION

With the increasing demand for commercial 3-D technologies in recent years, number of researches on stereoscopy (stereo-matching) has also significantly increased. Among these most of the attention is focused on matching of stereo images. Stereo matching is the problem of finding pixel-wise correspondences in two given spatially coupled images. The interest of this problem for the 3D community is that, after having the correspondences, the disparity field of the stereo images is obtained, which enables reconstruction of the 3D image or estimation of the depth of the scene.

In the literature, there are many approaches to address this problem. Also there are some work, which provide very detailed comparative performance analysis for most of the common stereo matching algorithms.[1,2] Existing stereo matching approaches are defined mostly in two categories, which are local[3–6] and global methods.[7–10] Local methods are also known as correlation based methods, since the matching decision is made according to the correlation between ensembles of pixels from both of the stereo images. On the other hand, global methods take some global energy functions into account. By enforcing some constraints, they try to minimize these functions.

Among the global methods, ones which are based on Markov Random Fields (MRF)[8–10] gained popularity in recent years, since they are more accurate than most of the local approaches and have lesser computational complexity compared to other global methods, such as graph-cuts.[7]

Although there are many successful stereo matching approaches, which yield spatially smooth disparity fields for stationary images, the research on improving the temporal smoothness of the disparity fields for stereo videos has been limited.[6,8] In order to maintain temporal smoothness, the most common method has been extending the approaches, which enforce the spatial smoothness, for temporal domain by assuming small temporal displacement between consecutive images.

In stereo video, there might exist the motion of the scene objects and/or the cameras, and the correspondence problem. In this work, we couple the MRFs used for stereo matching with the side information obtained from the motion estimation within one of the videos. From a probabilistic perspective, we estimate the posterior distribution over all possible matchings given the motion vectors and an initial matching. Intuitively, what we

Further author information: (Send correspondence to Y.C.S.)
Y.C.S.: E-mail: cem.subakan@boun.edu.tr
Ö.C.G.: E-mail: omer.gurol@boun.edu.tr
Ç.D.: E-mail: cagatay.dikici@boun.edu.tr

do is to increase the penalty to the matching combinations which are not consistent with the motion information coming from individual videos.

Note that there are various well-established motion estimation algorithms in the literature.[11,12] What we do is to increase the confidence of stereo matching estimates using these algorithms.

The organization of this paper is as follows. Section 2 gives the classical non informed stereo MRF model definition. The proposed informed model using motion field is introduced in Section 2.1. Finally, experimental results and the conclusion of the paper are given in Sections 3 and 4 respectively.

2. MODEL DEFINITION

The classical non-informed stereo matching model can be defined as follows using MRFs:[13]

$$p(\mathbf{d}) = \frac{1}{Z}\exp(-E_d(\mathbf{d})), \tag{1}$$

$$p(g^*|\mathbf{d}, g) = \prod_{\mathbf{x}} \mathcal{N}(g^*(\mathbf{x}); g(\mathbf{x}+\mathbf{d}(\mathbf{x})), \sigma^2), \tag{2}$$

where, $\mathbf{d} = (d_1, d_2)$ is the displacement vectors that link the two images $g^*(\mathbf{x})$ and $g(\mathbf{x})$. $\mathbf{x}$ is a shorthand for $\mathbf{x} = (x_1, x_2)$ and Z is the normalization constant for the prior distribution of the displacement vectors. So, given the displacement vectors $\mathbf{d} = (d_1, d_2)$ and the image $g(\mathbf{x})$; the other image $g^*(\mathbf{x})$ is normally distributed with mean $g(\mathbf{x}+\mathbf{d}(\mathbf{x}))$, and variance σ^2. E_d is the energy function over displacement vectors which is defined as:

$$E_d(d_1, d_2) = \sum_{i\in image}\sum_{j\in\mathcal{C}_i} F(\mathbf{d}(\mathbf{x}_i), \mathbf{d}(\mathbf{x}_j)). \tag{3}$$

Here, we take the summation over all the sites i. For each site i, we consider the neighborhood $\mathcal{C}_i$, which can be taken as four pixel neighborhood. $F(.)$ is a distance function. It can be for instance taken as Euclidian distance or a binary distance which returns $-\eta$ if inputs are same, or η if they are different, where η is a positive constant. Given this model, we learn the displacement vectors between two given frames. The undirected graph of the classical stereo matching model is given in figure 1.

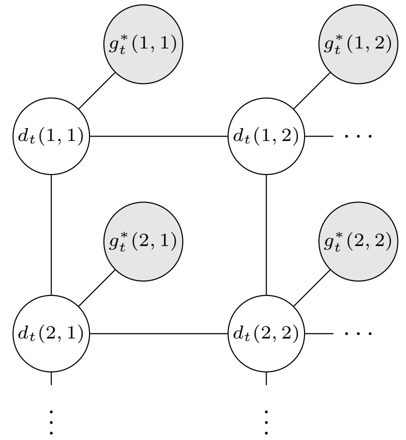


Figure 1: The undirected graph of the classical disparity estimation model.

Note that for disparity estimation in stereo matching problem, we have $\mathbf{d} = (d_1, 0)$. A pictorial description of the model is given in figure 2. In this figure, $g^*(\mathbf{x})$ is modeled as the disparity compensated image $g(\mathbf{x}+\mathbf{d}(\mathbf{x}))$ plus a zero-mean Gaussian noise ζ. Cast as an optimization problem, we would like to solve:

$$\text{maximize}_{\mathbf{d}} \log p(g^*, \mathbf{d}|g) = -E_d(\mathbf{d}) - \frac{1}{2\sigma^2}\sum_{\mathbf{x}}(g^*(\mathbf{x}) - g(\mathbf{x}+\mathbf{d}(\mathbf{x})))^2. \tag{4}$$

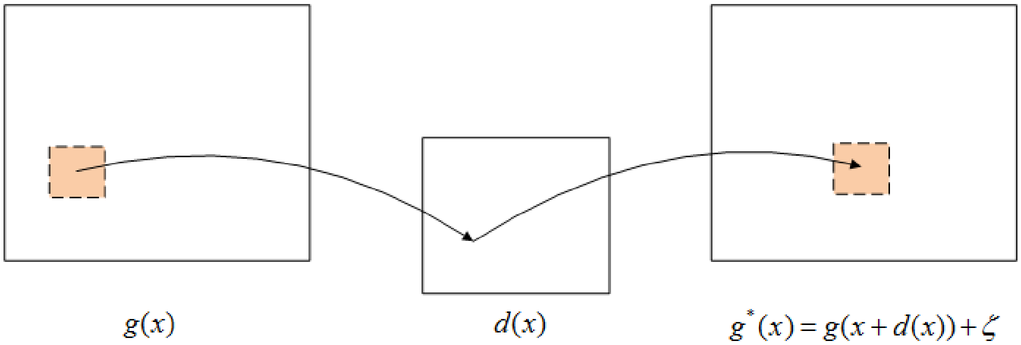


Figure 2: The pictorial description of the classical stereo matching problem. Note that $\zeta \sim \mathcal{N}(0, \sigma^2)$.

This problem is the classical MRF learning problem and it is known that the exact solution can not be computed in polynomial time.[14] Therefore we resort to approximate learning techniques such as simulated annealing or belief propagation.

2.1 Proposed Model

In a video context, the classical disparity estimation problem defined in (4) assumes that all time frames are independent and consequently we do not use any prior knowledge about the disparity image $\mathbf{d}_t$ at time t except the smoothness constraint defined via E_d in eq. (1). However in videos, one can exploit the motion information, in order to incorporate prior knowledge in the disparity estimation. Namely, we can use the motion information to predict the disparity in the current time step, which can in turn be used in the disparity estimation. To achieve this, we introduce an additional Gaussian prior on the disparity, which helps us to incorporate motion information in the disparity estimation (the undirected graph of the proposed model is given in figure 3):

$$p(\mathbf{d}_t|\mathbf{d}_{t-1}, \mathbf{d}_t^m) = \frac{1}{Z} \exp(-E_d(\mathbf{d}_t)) \times \Big[\prod_{\mathbf{x}} \mathcal{N}(d_t(\mathbf{x}); d_{t-1}(\mathbf{x} + d_t^m(\mathbf{x})), \sigma_c^2)\Big], \tag{5}$$

where $d_t^m(\mathbf{x})$ are the motion vectors extracted from $g_t(\mathbf{x})$ and $g_{t-1}(\mathbf{x})$ such that, $g_t(\mathbf{x}) \approx g_{t-1}(\mathbf{x} + d_t^m(\mathbf{x}))$. We basically compensate the disparity estimate, $d_{t-1}(\mathbf{x})$ in time $t-1$ with the motion information $d_t^m(\mathbf{x})$ in order to penalize substantial deviations of $d_t(\mathbf{x})$ from $d_t^c(\mathbf{x}) := d_{t-1}(\mathbf{x} + d_t^m(\mathbf{x}))$. Hence, in a real scenario, in the first frame (at $t = 1$), we compute an accurate disparity estimate $d_1(\mathbf{x})$ with a very reliable (possibly slow) method. Then, in the second frame we use the motion compensated disparity $d_1(\mathbf{x} + d_2^m(\mathbf{x}))$ to estimate $d_2(\mathbf{x})$ (with a faster and possibly less accurate method). Then we use $d_2(\mathbf{x} + d_3^m(\mathbf{x}))$ for $d_3(\mathbf{x})$, and we go on like this until the last frame.

Thus, finding the disparity $d_t(\mathbf{x})$, cast as an optimization problem is as follows:

$$\max_{\mathbf{d}_t} \log p(g^*, \mathbf{d}_t | g, \mathbf{d}_t^c) = -E_d(\mathbf{d}_t) - \frac{1}{2\sigma^2} \sum_{\mathbf{x}} (g_t^*(\mathbf{x}) - g_t(\mathbf{x} + \mathbf{d}_t(\mathbf{x})))^2 - \frac{1}{2\sigma_c^2} \sum_{\mathbf{x}} (\mathbf{d}_t(\mathbf{x}) - \mathbf{d}_t^c(\mathbf{x}))^2, \tag{6}$$

where σ^2 and σ_c^2 are respectively the observation variance and variance of the disparity estimate $d_t(\mathbf{x})$ from $d_t^c(\mathbf{x})$. The key idea in this model is based on the assumption that, the motion vectors between the consecutive frames can be found more accurately than the disparity estimate, since change between the consecutive frames is generally not substantial compared to the change between two stereo images. Therefore, we use this motion information to "steer" the disparity estimates. The model is depicted pictorially in figure 4. To find the motion vectors, we use an off the shelf code in.[15] We also employ the median filtering trick mentioned in that paper.

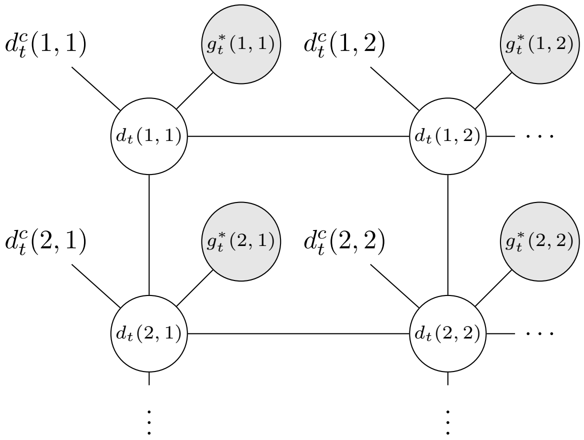


Figure 3: The undirected graph of the informed disparity estimation model, the addition to the first model is the introduction of d_t^c, which "steer" the estimation.

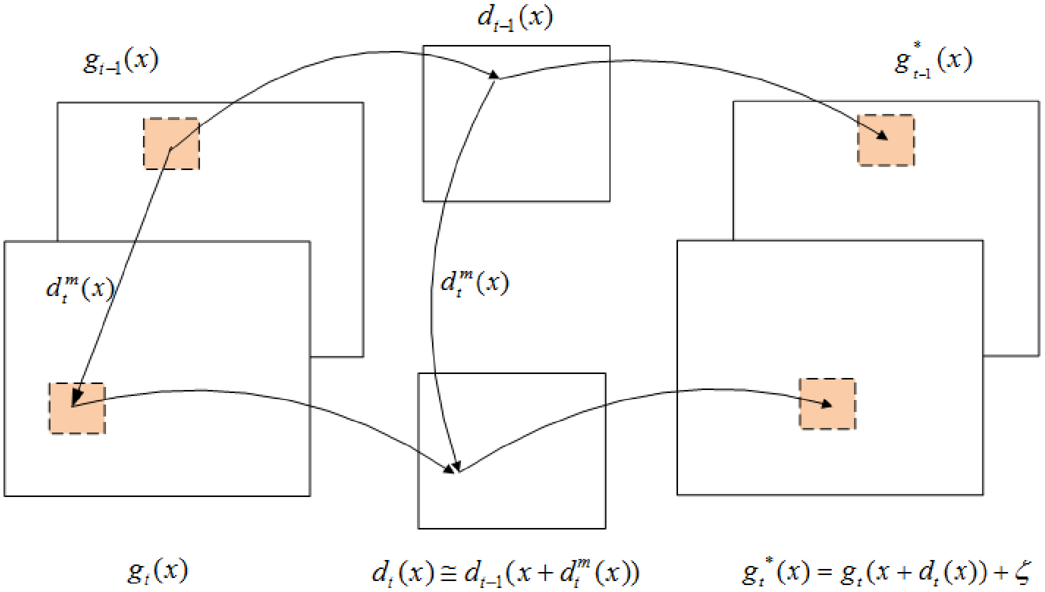

$d_{t-1}(x)$
$g_{t-1}(x)$
$g^*_{t-1}(x)$
$d_t^m(x)$
$d_t^m(x)$
$g_t(x)$
$d_t(x) \cong d_{t-1}(x + d_t^m(x))$
$g_t^*(x) = g_t(x + d_t(x)) + \zeta$

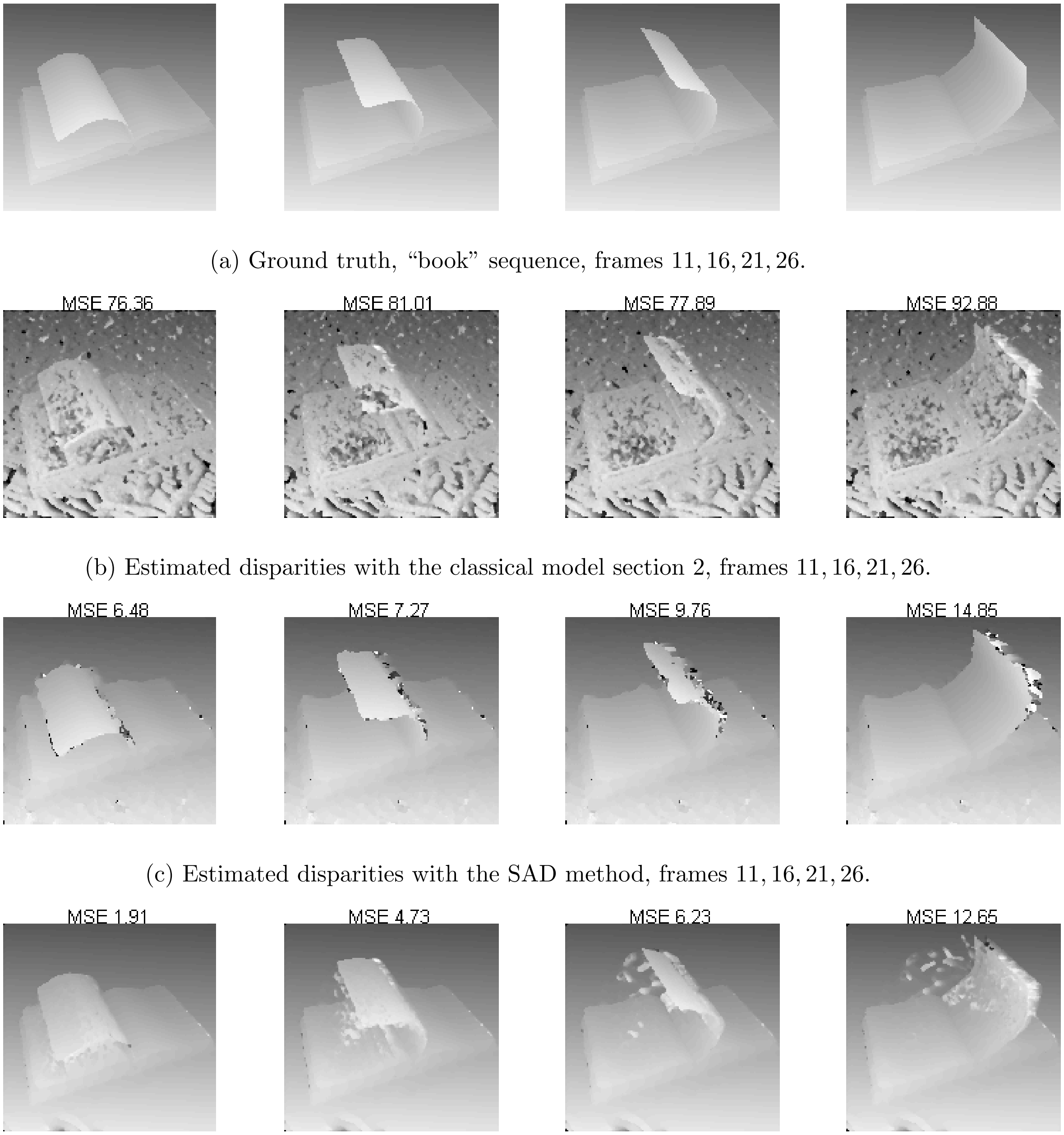
Figure 4: The pictorial description of the classical stereo matching problem. Note that $\zeta \sim \mathcal{N}(0, \sigma^2)$.

3. EXPERIMENTAL RESULTS

We work on the Cambridge stereo image dataset.[6] In the experiments we provide the results obtained on the "book" and "temple" sequences. To incorporate the prior knowledge, we use the ground truth in the first frame and then use the motion information to add in the prior term introduced in our model.

We compare the proposed model with the classical model introduced in section 2 and block matching. In the proposed model and the classical model, we set the Markov Random field to have 4-neighboorhood connection. The $F(.)$ function in eq. (3) is chosen as the l_1 distance. To find the motion vectors we use an off the shelf code in.[15] To estimate the disparities in each time step, we actually use a naive algorithm such as simulated annealing. However, the motion compensation algorithm helps to estimate significantly better results (perceptually and in terms of mean squared error) than the classical uninformed approach as shown in figures 5, 6. In the proposed informed model, at each time step, we initialize $d_t(\mathbf{x}) = d_{t-1}(\mathbf{x} + d_t^m(\mathbf{x}))$.

Block matching is performed by simple sum of absolute differences (SAD) measure, which gives local correlations. The size of the blocks are of size 9×9 pixels. As a preprocessing step a rank filter of size 15×15 is applied on both of the stereo images prior to matching.[2]



(a) Ground truth, “book” sequence, frames $11, 16, 21, 26$.

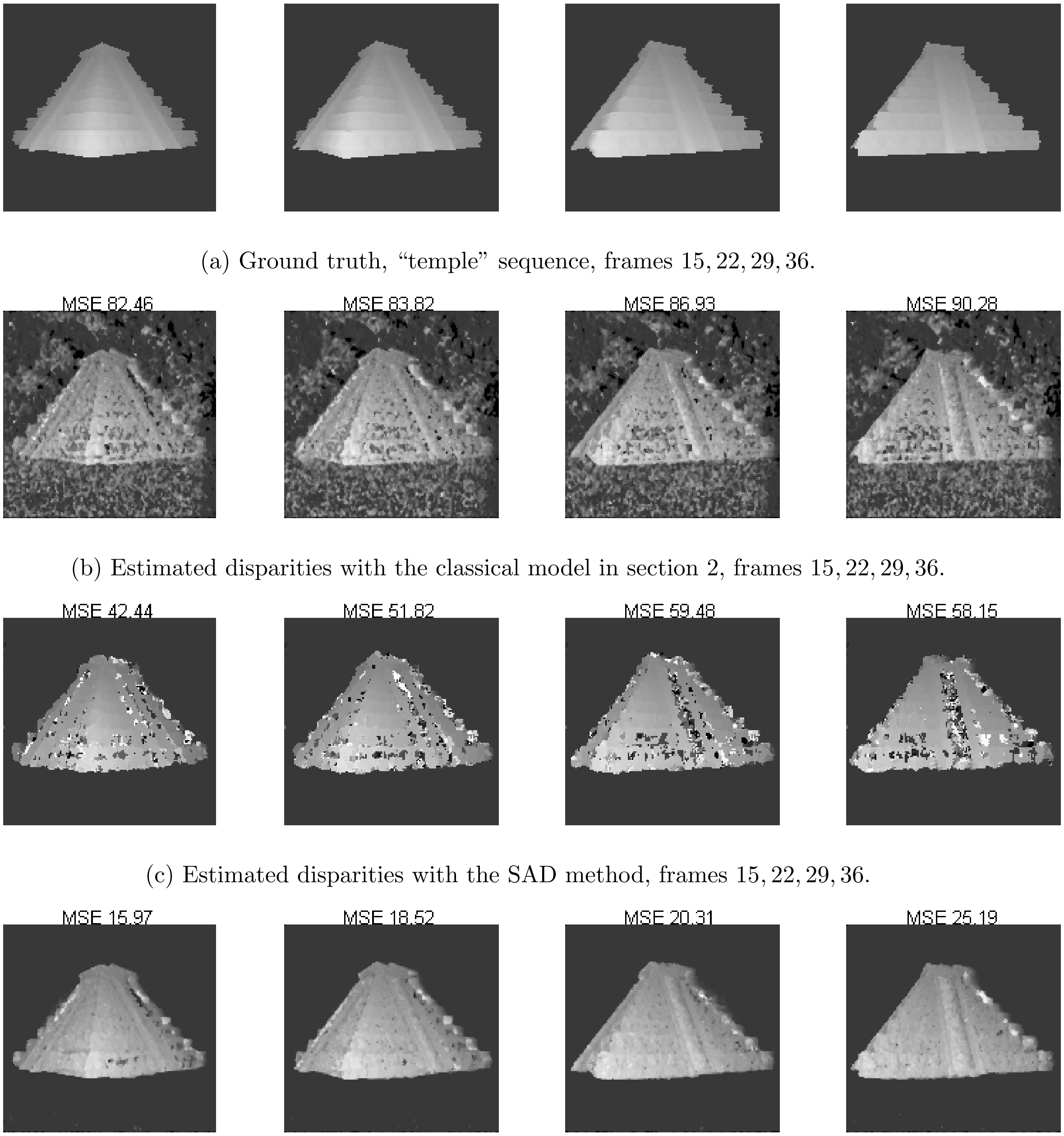
(b) Estimated disparities with the classical model section 2, frames $11, 16, 21, 26$.

(c) Estimated disparities with the SAD method, frames $11, 16, 21, 26$.

(d) Estimated disparities with the proposed model, frames $11, 16, 21, 26$.

Figure 5: Results on the “book” sequence.

These results suggest that, ambiguous regions such as the background of the sequence tend to result in fluctuating disparity estimates in the classical model. Using the motion information enables us to impose a bias towards the previous disparity estimate d_{t-1}. Therefore, in the regions where there is no significant motion, the estimate turns out to be very promising. What we essentially achieve with the proposed algorithm is that, we can obtain competitive results with faster and/or simpler algorithms since, the motion compensation idea significantly shrinks the size of our search space.



(a) Ground truth, "temple" sequence, frames $15, 22, 29, 36$.

(b) Estimated disparities with the classical model in section 2, frames $15, 22, 29, 36$.

(c) Estimated disparities with the SAD method, frames $15, 22, 29, 36$.

(d) Estimated disparities with the proposed model, frames $15, 22, 29, 36$.

Figure 6: Results on the "temple" sequence.

4. CONCLUSIONS AND FUTURE WORK

As the results given in figures 5 and 6 suggest, using the motion information helps us to significantly refine the estimated disparity. In a real-time scenario, one can "initialize" the algorithm with a very accurate disparity estimate in order to have better disparity estimates in the subsequent frames. In this work, we only used a modest learning algorithm such as simulated annealing. Better results can be obtained with more sophisticated approaches such as graph cuts. Also, better spatial priors can be chosen in order to have smoother disparity estimates with less artifacts.

Acknowledgments

Çağatay Dikici has been supported by TUBITAK BIDEB-2232.

REFERENCES

[1] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on*, pp. 131 –140, 2001.

[2] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –8, 2007.

[3] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**(4), pp. 401 –406, 1998.

[4] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, pp. 920–932, Sept. 1994.

[5] K.-J. Yoon and I.-S. Kweon, "Locally adaptive support-weight approach for visual correspondence search," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, **2**, pp. 924 – 931 vol. 2, june 2005.

[6] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. A. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *Proceedings of the 11th European conference on computer vision conference on Computer vision: Part III*, *ECCV'10*, pp. 510–523, Springer-Verlag, (Berlin, Heidelberg), 2010.

[7] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions via graph cuts," in *In International Conference on Computer Vision*, pp. 508–515, 2001.

[8] R. Khoshabeh, S. H. Chan, and T. Q. Nguyen, "Spatio-temporal consistency in video disparity estimation," in *ICASSP'11*, pp. 885–888, 2011.

[9] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –8, june 2007.

[10] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, **1**, pp. I–261 – I–268 Vol.1, june-2 july 2004.

[11] "H.264 : Advanced video coding for generic audiovisual services."

[12] S.-W. Wu and A. Gersho, "Joint estimation of forward and backward motion vectors for interpolative prediction of video," *Image Processing, IEEE Transactions on* **3**, pp. 684 –687, sep 1994.

[13] A.M.Tekalp, *Digital Video Processing*, Prentice Hall, 1995.

[14] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **PAMI-6**, pp. 721 –741, nov. 1984.

[15] D. Sun, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2432 –2439, june 2010.