MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Temporally Consistent Stereo Matching Using Coherence Function

Dongbo Min, Sehoon Yea, Anthony Vetro

TR2010-040 July 2010

Abstract

This paper proposes a novel method for estimating temporally consistent disparity maps. The proposed approach utilizes a coherence function that represents the temporal consistency between the disparity maps of consecutive frames. The coherence function is computed according to a motion probability between consecutive frames based on the assumption that a disparity varies smoothly over time. The temporal consistency between consecutive disparity maps is enforced based on the coherence function. In order to cope with scene changes, the similarity of consecutive frames is also considered in the coherence function. Experimental results show that the proposed method yields satisfactory disparity maps and synthesized views on several challenging stereo videos, including clips with fast object and camera motions as well as scene change.

3DTV Conference 2010

TEMPORALLY CONSISTENT STEREO MATCHING USING COHERENCE FUNCTION

Dongbo Min, Sehoon Yea and Anthony Vetro

Mitsubishi Electric Research Laboratories (MERL), Cambridge, USA

ABSTRACT

This paper proposes a novel method for estimating temporally consistent disparity maps. The proposed approach utilizes a coherence function that represents the temporal consistency between the disparity maps of consecutive frames. The coherence function is computed according to a motion probability between consecutive frames based on the assumption that a disparity varies smoothly over time. The temporal consistency between consecutive disparity maps is enforced based on the coherence function. In order to cope with scene changes, the similarity of consecutive frames is also considered in the coherence function. Experimental results show that the proposed method yields satisfactory disparity maps and synthesized views on several challenging stereo videos, including clips with fast object and camera motions as well as scene change.

Index Terms— stereo matching, disparity, temporal consistency, coherence, view synthesis, 3D video

1. INTRODUCTION

Stereo matching methods estimate a depth map based on spatial correspondence with two or more images taken from different viewpoints. A number of methods have been proposed in the literature to improve the performance of stereo matching methods, including methods based on global optimization and color segmentation [1, 2, 3, 4, 5].

Most of these methods have mainly focused on improving the performance of the disparity map, especially in the textureless, occluded and discontinuities region. However, temporal aspects have not been extensively studied or considered. Zitnick et al. showed that it is possible to provide fluctuation-free disparity maps without considering the temporal aspects if the estimated disparity maps are accurate enough [5]. However, it is often difficult to get accurate disparity maps for general scenes with complex structure and with significant local or global motion. The fluctuation in depth maps over time may cause a serious problem, especially when used for view synthesis since the human visual system is very sensitive to high-frequency visual artifacts. It is noted that depth estimation and view synthesis are essential components in the 3D Video framework of MPEG, which is an active area of exploration [6].

Several methods have been proposed to enforce the temporal consistency as part of the stereo matching process. Space-time stereo used a set of spatial windows in the temporally neighboring frames [7, 8], but assumes that the scene and camera are static or the motion is very small compared to the capturing frequency. Tao [9] considered the temporal consistency by assuming that the pixels in one color segment have the same motion. In this way, motion smoothness could be maintained over time by using spatial and temporal homography without any optical flow information.

Leung [10] proposed to enforce the temporal consistency by minimizing the disparity value differences between consecutive frames, but it does not work well when the movements of object or cameras are large. There are also joint estimation methods that calculate both the motion and disparity maps to obtain temporally consistent disparity maps [11, 12, 13]. For instance, Gong [13] defined disparity flow, which helps to enforce the temporal consistency between the consecutive frames using only the cross-validated disparity and disparity flow values. In [14], Bartczak et al. proposed a method that provides denser prediction maps by reducing uncertainty due to the discrete hypotheses in [13].

In this paper, we propose a novel method for enforcing temporal consistency in stereo video. This work aims to address the fluctuation problem that is caused by the estimation of incoherent disparity maps. We define a coherence function that represents the temporal consistency between the disparity maps of consecutive frames. We assume that disparity values vary smoothly over time, therefore the coherence function can be computed with the disparity of the previous frame and a motion probability between consecutive frames. The coherence function is incorporated into a cost function to compute the disparity map with temporal consistency. Furthermore, in order to address scene changes, a weight that represents the similarity of depth distributions between frames is also used.

The reminder of this paper is organized as follows. We introduce the estimation models for enforcing temporal consistency in Section 2, and then explain the proposed method in Section 3. We present the experimental results and conclusion in Section 4 and 5, respectively.

2. ESTIMATION MODEL FOR TEMPORAL COHERENCE

In order to make disparity maps consistent over time, we consider additional information such as the motion in the scene as part of our estimation model. Consider the case that two corresponding points in the stereo video have unique motion vectors. Given the disparity and motion information, we can calculate the disparity value of the current frame with a joint estimation constraint as shown in Fig. 1(a):

$$d_t = m_l + d_{t-1} - m_r \tag{1}$$

where m_l and m_r denote the motions between previous and current frames for the left and right views, respectively, and d_t represent the disparity vector at frame t. The joint estimation modeling can be divided into categories with different assumptions on the level of information that is available. We outline three joint models:

- explicit motion and occlusion (both motion and disparity)
- motion probability and disparity occlusion
- motion probability only

Fig. 1 shows examples of these estimation models. Note that in all of these cases the disparity map of the previous frame is fixed (a solid line). The first model assumes explicit motion and occlusion information is available. In this case, the disparity vector of the current frame can be calculated using eqn. (1), except at the occluded pixels (namely for both motion and disparity occlusions).

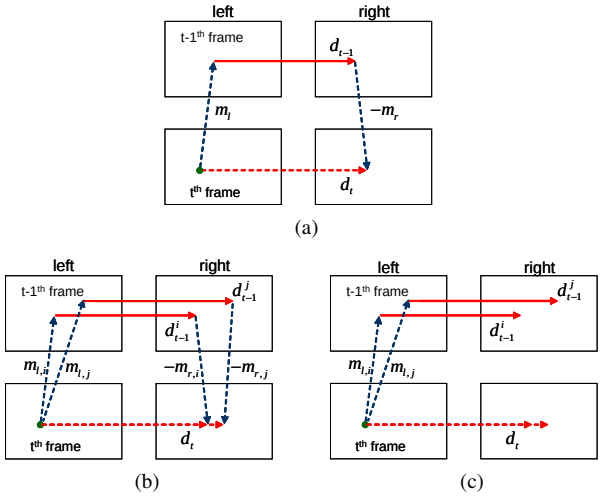

Fig. 1. Joint estimation models: (a) with explicit motion and occlusion, (b) with motion probability and disparity occlusion, (c) with motion probability

The second model uses the probability of left and right motions and explicit disparity occlusion of the previous frame. Multiple disparity candidates are considered according to the motion probability. This is similar in spirit to non-local means (NLM) filtering methods used in video denoising [15]. NLM filtering calculates weighted sums of all pixels in a motion search window using weights determined by the block similarity with reference pixels; these weights can be referred to as probabilities after a normalization step. According to experimental results in [15], this process makes the denoising method robust to errors in the motion estimation.

The third estimation model is based on the assumption that the disparity varies smoothly over time and uses only a motion probability. The coherence function is computed based on the assumption that the corresponding points with motion vector m_i between frames $t-1$ and t have similar disparity value. In this work, we use this third estimation model since it has the advantage that one motion map is used.

3. PROPOSED METHOD

Given disparity map of the previous frame, the coherence function for the current frame is computed based on the motion probability. In this work, only the disparity map of the previous frame is used for computing the coherence function, and it does not change to enable a causal implementation. In other words, the disparity map of the current frame does not have any influence on that of the previous frame, i.e., the proposed scheme is non-iterative. Our work aims to obtain fluctuation-free disparity maps which can be used for virtual view synthesis. Since the human visual system is generally sensitive to temporal artifacts caused by the disparity fluctuation problem, we address this problem by enforcing the temporal consistency in the stereo matching process. This section defines the coherence function proposed in this work and describes its use for temporally consistent disparity estimation.

3.1. Coherence function

The coherence function $C(p_t, d)$ represents the likelihood that the disparity p_t of a pixel p of the current frame t is equal to d. This function can be computed using the disparity map d_{t-1} of the previous frame and the motion probability map as follows:

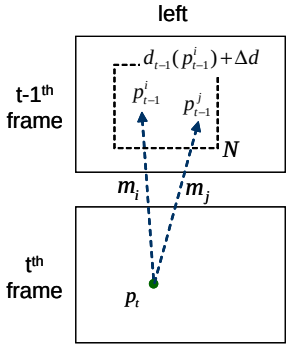

Fig. 2. Computation of coherence in the motion search window.

$$C_t(p_t, d) = \frac{1}{|\Delta d|} \sum_{i \in N(p_t)} \sum_{\Delta d} P(p_t, m_i) f(d, d_{t-1}(p_t + m_i) + \Delta d) \tag{2}$$

where P represents the probability of a motion candidate $m_i = p^i_{t-1} - p_t$ in a motion search window $N(p_t)$. Also, $f(a, b)$ is set to 1 if $a = b$ and is 0 otherwise. Δd represents the disparity variation across the frames and function f is normalized by using $|\Delta d|$ which indicates the magnitude of the disparity variation. Note that the motion probabilities and function f should each sum to 1, as follows.

$$\begin{aligned} &\sum_{i \in N(p_t)} P(p_t, m_i) = 1 \\ &\frac{1}{|\Delta d|} \sum_{d} \sum_{\Delta d} f(d, d_{t-1}(p_t + m_i) + \Delta d) = 1 \end{aligned} \tag{3}$$

Fig. 2 shows the process that calculates the coherence at pixel p_t. The probability $P(p_t, m_i)$ of the motion vector m_i determines the weighting that represents the disparity similarity between temporally-corresponding points (p_t and p^i_{t-1}). In other words, the disparity of a pixel at the frame $t-1$ whose motion vector has higher probability is more coherent to the disparity of the current frame. Δd controls the temporal smoothness of disparity map over time. In this work, Δd is set to $(-2 \sim 2)$.

The motion probability function $P(p_t, m_i)$ can be computed using block similarity, optical flow or global optimization. In this work, a block matching method is used, similar to non-local means filtering:

$$\begin{aligned} P(p_t, m_i) &= \frac{1}{/Z(p_t, m_i)} e^{-BM(p_t, m_i)/\sigma_M} \\ BM(p_t, m_i) &= \sum_{n \in BW} |I_t(p_t + n) - I_{t-1}(p_t + m_i + n)| \end{aligned} \tag{4}$$

where BW is a block matching window and Z is a normalization factor. σ_M is a weighting constant for block matching distances and I_t represents intensity values at frame t. A hierarchical scheme and integral images are used to reduce the complexity of the implementation.

3.2. Incorporating coherence function

In this work, the occluded pixels on the disparity map of the previous frame are handled before processing, so that all pixels including the

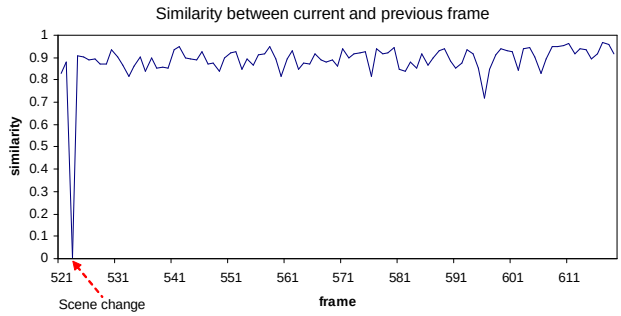

Fig. 3. Scene similarity α_t between consecutive frames.

occluded pixels of the previous frame are used for calculating the coherence function. The final temporally-consistent cost function $E_t^C(p_t, d)$ is computed by incorporating the coherence function into the cost function $E_t(p_t, d)$ on the current frame. The cost function used in our current design is defined as follows:

$$E_t^C(p_t, d) = \frac{E_t(p_t, d)}{1 + 3\alpha_t C_t(p_t, d)} \tag{5}$$

where α_t is a time-varying weighting factor that controls the influence of the coherence function C and represents the scene similarity between consecutive frames. The depth distribution of a scene is a distinct characteristic of the scene and can be represented by a disparity histogram of feature points. Therefore, the similarity of disparity histograms can be used to identify scene change. We use the SURF (Speeded Up Robust Features) [16] tracker, which is a scale and rotation invariant interest point detector and descriptor, to determine the feature points and generate the disparity histogram. Specifically, using pairs of matched feature points, the disparity histogram is computed as follows:

$$h(i) = \sum_{j=1}^{N} f\left(D(i), \left\lfloor \frac{d_j}{B} \pm 0.5 \right\rfloor \cdot B\right) \qquad i = 1, 2, \cdots M \tag{6}$$

where $h(i)$ is the histogram count for the i^{th} bin, M is the total number of bins, and $f(a, b)$ is set to 1 if $a = b$ and is 0 otherwise. By quantizing each disparity value d_j of a matching-points pair (out of the N total pairs) with the bin-size B, a histogram bin count with the closest representative value $D(\cdot)$ will be incremented by one. The disparity histogram in eqn. (6) has also been used to estimate a temporally-consistent disparity search range in stereo video [17]. The similarity of disparity histograms can be represented with the weighting factor α_t between frames t and $t-1$ as follows:

$$\alpha_t = \exp\left(-\sum_{i}^{M} |h_t^{nor}(i) - h_{t-1}^{nor}(i)|/\sigma_S\right) \tag{7}$$

where σ_S is a weighting constant for distance between histograms. A normalized histogram h^{nor} should be used in eqn. (7) since the total number of matching points vary among consecutive frames. The distance ranges ranges from 0 to 1 since it is calculated with the normalized histograms. Fig. 3 shows the scene similarity α_t between consecutive frames. α_t is usually close to 1, except at 523^{th} frame where a scene change occurs.

3.3. Implementation

Fig. 4 shows a block diagram of the proposed temporally-consistent stereo matching scheme, where the proposed methods are shaded.

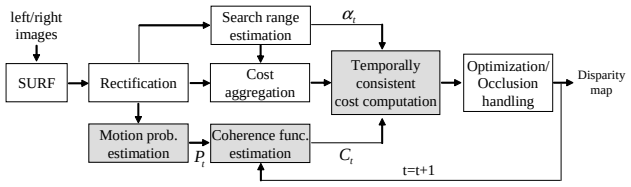

Fig. 4. Block diagram of main processes in the proposed scheme.

The disparity search range is calculated using temporally-consistent range estimation method proposed in [17]. The cost function in eqn. (5) is obtained using the simplified method of [18] and WTA (Winner-Takes-All) is used as optimization method, but other methods such as belief propagation or graph cut can be also used. The temporally consistent cost function is then obtained using the coherence function C_t and scene similarity α_t calculated for the rectified input stereo sequences. The occluded pixels that are detected by a cross-checking method are handled using a support-and-decision process [19], and the final disparity map for frame t is then used to compute the coherence function for frame $t+1$. Note that the proposed system is causal, i.e., the disparity map of the previous frame is not influenced by the updated disparity map of the current frame.

4. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed method using a variety of stereo videos including clips from the 'Heidelberg' (1280×720) and 'Rhinevalley' (720×576) sequences, which are available online [20]. The proposed method is tested using the same parameters for all images. The block matching window BW and motion search window for computing a motion probability are both set to 11×11. The bin size B is 7, and weighting constants σ_S and σ_M are set to 0.4 and 1.0, respectively.

Fig. 5 shows the results of the proposed method for the 'Heidelberg' sequence. The estimated search ranges are $(-4, 24)$ and $(-10, 20)$ using the technique in [17] and novel views are synthesized at 30% of the original baseline distance. We find that the disparity maps generated without temporal coherence cause visual artifacts in the synthesized results. For example, see the iron window bar in Fig. 5(a) and the right window in Fig. 5(b). In this work, $2D$ disparity maps which were derectified using an estimated homography are used to synthesize the virtual views, but only the left x-disparity maps are shown in the figures.

The results for 'Rhinevalley' are shown in Fig. 6. The estimated search ranges for the three 3 sample frames are $(-12, 24)$, $(-24, 24)$ and $(-24, 19)$ [17]. We can see that the 2^{nd} disparity map and synthesized view have serious visual artifacts even though the consecutive frames have very similar depth distributions. It is evident from these results that the temporal consistency in the stereo matching method ensures more reliable disparity maps, which improve the quality of synthesized views.

5. CONCLUSIONS

In this paper, we presented a novel method for enforcing the temporal consistency in the stereo matching. The coherence function is obtained using a motion probability and the disparity map on the previous frame based on the assumption that a disparity varies smoothly over time. The scene similarity is also considered in order to cope with scene changes. The experimental results show that the proposed method work well for several stereo video sequences. In further research, we plan to analyze how the coherence function works in the several parts such as the motion and disparity occlusions.

(a) (b) (c) (d)

Fig. 5. Disparity maps and synthesized views for 'Heidelberg': (a)(b) without temporal consistency, (c)(d) with temporal consistency

(a) (b) (c) (d) (e) (f)

Fig. 6. Disparity maps and synthesized views for 'Rhinevalley': (a)(b)(c) without temporal consistency, (d)(e)(f) with temporal consistency

6. REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7-42, Apr. 2002.

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

[3] H. Tao and H. S. Sawhney, "Global matching criterion and color segmentation based stereo," *Proc. IEEE Workshop on the Application of Computer Vision*, pp. 246-253, 2000.

[4] Q. Yang, L. Wang, R. Yang, H. Stewenius and D. Nister, "Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation, and Occlusion Handling," *IEEE Trans. on PAMI*, vol. 31, no. 3, pp. 1-13, 2009.

[5] C. Zitnick, S. Kang, M. Uyttendaele, S.Winder, and R. Szeliski, 'High-quality video view interpolation using a layered representation,' *SIGGRAPH*, pp. 600-608, 2004.

[6] MPEG Video and Requirements, "Vision on 3D Video," ISO/IEC JTC1/SC29/WG11 Doc. N10357, Feb. 2007.

[7] L. Zhang, B. Curless and S. Seitz, "Spacetime stereo: shape recovery for dynamic scenes," *Proc. IEEE CVPR*, 2003.

[8] J. Davis, D. Nehab, R. Ramamoothi and S. Rusinkiewicz, "Spacetime stereo: a unifying framework for depth from triangulation," *IEEE Trans on PAMI*, 2005.

[9] H. Tao, H. Sawhney and R. Kumar, "Dynamic Depth Recovery from Multiple Synchronized Video Streams," *Proc. IEEE CVPR*, 2001.

[10] C. Leung, B. Appleton, B. Lovell, and C. Sun, "An energy minimisation approach to stereo-temporal dense reconstruction," *Proc. IEEE ICPR*, pp. 72-75, 2004.

[11] M. Isard and J. MacCormick, "Dense Motion and Disparity Estimation Via Loopy Belief Propagation," *Proc. ACCV*, pp. 32-41, 2006.

[12] D. Min, H. Kim, K. Sohn, "Edge-preserving joint motion disparity estimation in stereo image sequences," *Signal Processing: Image Communication*, vol. 21, no. 3, pp. 252-271, 2006.

[13] M. Gong, "Real-time joint disparity and disparity flow estimation on programmable graphics hardware," *Computer Vision and Image Understanding*, pp. 90-100, 2009.

[14] B. Bartczak, D. Jung and R. Koch, "Real-Time Neighborhood Based Disparity Estimation Incorporating Temporal Evidence," *Proc. DAGM*, pp. 153-162, 2008.

[15] A. Buades, B. Coll, and J. Morel, "Nonlocal Image and Movie Denoising," *International Journal of Computer Vision*, vol. 76, no. 2, pp. 123-139, 2008.

[16] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding* (CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

[17] D. Min, S. Yea, Z. Arican and A. Vetro, "Disparity search range estimation: enforcing temporal consistency," *Proc. IEEE ICASSP*, 2010.

[18] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Trans. Image Processing*, vol. 17, no. 8, pp. 1431-1442, Aug. 2008.

[19] D. Min, S. Yea and A. Vetro, "Occlusion handling based on support and decision," *Proc. IEEE ICIP*, 2010. (submitted)

[20] http://www.3dtv.at