Prediction of Disparity and Its Application to Stereo Matching

Intae Na\textsuperscript{1} and Hong Jeong\textsuperscript{2}

Abstract—In this paper, we present a novel matching framework for stereo video sequence based on disparity prediction by stereo-motion fusion. Fusion-based methods tackles the problem by integrating different depth modalities in cooperative way, so as to resolve matching ambiguities from each modality and acquire more accurate disparity map. Unfortunately, only few works consider processing with video sequences, which appears commonly in practical applications of stereo vision. The proposed method attempts to achieve high-quality disparity map, while maintaining temporal consistency by disparity prediction process which tracks features of disparity map. Experimental results show that our fusion algorithm quantitatively and qualitatively outperforms previous fusion-based matching methods.

I. INTRODUCTION

Reconstructing the 3D scene from the images is emerging as a promising area in the future due to the fast advancing displays, mobile computing, and the need of smart multimedia. In early stage researches, algorithms so called as Shape from X, where X notes certain modality of observation, were popular. Various kinds of depth cues and corresponding algorithms were suggested, such as stereo vision\textsuperscript{1}, structure from motion\textsuperscript{2} (SfM), photometric stereo\textsuperscript{3}, and shape from shade\textsuperscript{4}. Since modeling and computational efficiency for the reconstruction were their major concerns, there were remarkable progress in terms of processing time and accuracy of resulting depth map.

Nowadays, needs for real-time and high-quality 3D processing are continuously increasing, and it becomes harder to satisfy their requirements on performance by utilizing only single depth cue. Since each cue has characteristic pros and cons, we can consider cooperative framework which utilizes different depth cues\textsuperscript{5},\textsuperscript{6},\textsuperscript{7} to achieve target performance. Most of practical applications deals with stereo video data, where traditional stereo matching problem basically estimates depth map frame-by-frame. These stereo-only systems works well both in controlled and general scenarios, however, they frequently suffer from temporal inconsistency in depth estimates. Therefore, one needs to incorporate temporal continuity of depth data with stereo correspondence\textsuperscript{8},\textsuperscript{9},\textsuperscript{10} to enhance performance of the system dealing with stereo video data.

\textsuperscript{1}I. Na is with the Department of Electrical Engineering, Pohang University of Science and Technology, Pohang, Korea bluewing@postech.ac.kr
\textsuperscript{2}H. Jeong is with Faculty of the Department of Electrical Engineering, Pohang University of Science and Technology, Pohang, Korea hjeong@postech.ac.kr

A. Related Works

Several methods have been suggested to provide precise and consistent depth maps by utilizing both stereo correspondences and temporal information. Some authors introduced extended matching cost which considers pixels from different time instants for the same location\textsuperscript{11},\textsuperscript{12}, other authors extended overall cost volume instead\textsuperscript{13},\textsuperscript{14}. Usually their cost volume is 3-D and these methods are called as spacetime stereo. Spacetime stereo works well under static environment-consecutive scene with no significant relative motion-so performance improvement is limited under more general scenes such as mobile robot vision or surveillance systems.

Tracking correspondences along temporal axis with stereo images have been also popular in this area. Variety of algorithms have been suggested that simultaneously estimates depth and motion using iterative relaxation\textsuperscript{15}, loopy belief propagation\textsuperscript{16}, or variational method\textsuperscript{7}. They often called as stereo-motion fusion, or scene flow estimation in various contexts. In general, these stereo-motion fusion methods deals with complex energy function which includes both stereo disparity and optical flow vectors as variables. Complexity of algorithm became the most important issue during estimation process because of increased dimension of target variable. Recently, parallel processing on dedicated hardware, or GPU became majority in real-time processing area.

The main contributions of our work is twofold. First, we suggest a novel depth tracking framework based on invariant feature from the prior depth estimates. Under this framework, one can easily plug-in temporal consistency into any existing stereo matching methods. Second, the stereo-motion optimization algorithm based on dynamic programming (DP)\textsuperscript{17},\textsuperscript{18},\textsuperscript{19}. DP is popular in real-time stereo matching because of moderate performance and high efficiency in computation. In this paper, we modified DP algorithm suggested in\textsuperscript{19} to incorporate temporal continuity and occlusion information. The two additional information are modeled as additional constraint to the original energy minimization problem. The performance and reliability of proposed method is verified in both quantitative and qualitative ways.

II. PROBLEM STATEMENT

Stereo matching problem can be defined as searching corresponding pixels between two or more images from distinct cameras. We assume that the binocular stereo image streams \( I_t = \{(I_t^l(p), I_t^r(p)) | p \in \Omega, t \in [0, T]\} \), defined on 2D image support \( \Omega \), are well calibrated and rectified so
their epipolar lines are parallel to the scanline. Given \( \mathcal{I} \), we need to find an optimal sequence of disparity maps \( \mathcal{D} = \{ D^t(p) | p \in \Omega, t \in [0, T] \} \).

Assuming Markov property holds for temporal neighbors of \( p \), we only need to consider information from the previous frame \( \mathcal{T}_{t-1} \), \( D_{t-1} \) to estimate disparity at certain time instant \( D^t \). Hence, the optimal disparity estimation of stereo video frames can be reduced to four-frames (\( t \in [t-1, t] \)) correspondence instead of filtering (\( t \in [0, T] \)) or smoothing (\( t \in [0, T] \)), which means that we estimate current \( D^t \) given \( \mathcal{T}_{t-1}, \mathcal{T}^t, \) and \( D_{t-1} \). For further description, we define an augmented spatial domain \( \Omega' = \{((p), d)) | (p) \in \Omega, d \in [d_{min}, d_{max}] \} \) which is 3D space with additional disparity axis, and consider \( \Omega' \) as standard image domain. For example, intensity of reference image \( I^t_{l-1}(p') = I^t_{r-1}(p) \) as its disparity equals to zero. In this configuration, motion of a point between two consecutive time instants \( t-1 \) and \( t \) is represented as a vector \( \mathbf{v} = [u, v, w]^T \), which comes from the projective transform of 3D motion vector \( \mathbf{V} = [U, V, W]^T \), as shown in Fig. 1.

Let right images are reference images, then we can make our basic comparison by considering motion vectors as,

\[
I^t_l(p + v^{t-1, t}) = I^t_{r-1}(p),
\]

where \( p \in \Omega' \). Eq. (1) says that if we know the projected motion vector and previous disparity map, we can compute current disparity map \( D^t \). Actually, \( \mathbf{v} \) is called as scene flow and several preceding works have tried to estimate this [7], [20].

So far, we considered important definitions about the consecutive stereo problem. They can briefly represented as follows.

- Given four consecutive stereo frames \( \mathcal{T}_{t-1}, \mathcal{T}^t \) and past disparity map \( D_{t-1} \), compute \( D^t \).

We divided the problem into following two subproblems:

- Prediction problem: Given \( \mathcal{T}_{t-1}, D_{t-1} \), determine \( D^{t|t-1} \).

- Updation problem: Given the prediction \( D^{t|t-1} \) and current measurement \( \mathcal{T}^t \), determine \( D^{t|t} \triangleq D^t \).

Based on our problem definition, the proposed algorithm would likely to have two-step form with feedback, where the predicted solution is utilized in the updation process and vice versa.

The prediction process may include explicit flow estimation between two consecutive time instants, like optical flow estimation. Approaches in stereo-motion contexts which adopts explicit optical flow estimation often suffer from propagated errors in flow estimation; errors in estimated flows affect the disparity prediction and updation. Moreover, previous disparity has nothing to do with the prediction process except for the warping, that is, the prediction is made by warping \( D^{t-1} \) with estimated 2D flow \( (u, v) \), which has no connection with \( D^{t-1} \) at all. In this paper, we use rather previous disparity map \( D^{t-1} \) for the prediction process than the pixel intensities to avoid noisy prediction and to realize closer fusion between stereo-motion cues.

III. PROPERTY OF DISPARITY

As mentioned in previous section, disparity prediction based on explicit flow estimation is jeopardized by error propagation problem. One way to minimize effect of the problem is to extract features for flow estimation. In order to search robust and invariant features on consecutive disparity images, we need to analyze the properties of time-varying disparity map.

In stereo configuration, a motion vector in 3-D space \([U, V, W]^T \) corresponds to 2-D vectors \([u, v]^T, [u, v]^T \) for respective images, which is actually equivalent to scene flow representation \([u, v, w]^T = [u_r, v_r, u_r - u_t]^T \) in image domain.

Our objective is to implement two-step approach which predicts next disparity \( D^t \) by warping \( D^{t-1} \) using explicit flow estimation based solely on the disparity image data. In general stereo videos, however, disparity of certain position varies between different time instants. This fact makes it impossible to apply constant intensity assumption to our flow estimation process.

To overcome this limitation, we first begin with simple quasi-orthogonal assumption between two consecutive time instants. Time-varying disparity of a point can be represented by differentiation in time as,
\[
\frac{d}{dt} D(p) = -\frac{f_B}{Z^2} \frac{dZ}{dt}, \\
= -\frac{f_B}{Z^2} W.
\]

In quasi-orthographic configuration, \( Z \gg W \) so we can further assume \( \frac{dZ}{dt} \approx 0 \) and finally,

\[
\frac{d}{dt} D(p) = 0.
\]  

This assumption gets rid of the effect of varying disparity in corresponding locations.

### IV. Prediction of Disparity Signals

The disparity flow problem is inherently a motion estimation between \( D(t-1) \) and \( D(t) \). One may rely upon the many optical flow algorithms. However, the nature of disparity is different from the image intensities. In optical flow, the basic assumption is, \( df(t) = 0 \): the brightness of object point under examination is not changing between successive frames, which does not hold for disparity. However, we still can find sparse invariants of disparity map by utilizing \( \nabla D \). Using \( \nabla D \) as our feature map, we applied a simple Lucas-Kanade algorithm to find sparse optical flow.

Once the optical flow, \( (u, v) \), is available, disparity map for the next time instant can be estimated as:

\[
d(x, y, t|t-1) \approx d(x - u, y - v, t - 1).
\]  

The first two terms are the ordinary matching cost and smoothness cost. The third term is newly introduced to direct the search direction around the predicted disparity, \( d(t|t-1) \). The far the path deviates from the recommended track, the more penalty is expected. Also, \( \mu(x) \) is an indicator function that exists only where \( d(x, t|t-1) \) is available. In this manner, the cases when the prediction is sparse or the confidence is variable can be managed by \( \mu(x) \).

This energy equation can be solved by the DP. Let \( \delta(i, j) \) denote the accumulated cost at the node \( (i, j) = (x, d) \). Also, \( \psi(i, j) \) denotes the parent node of the node \( (i, j) \). For each, \( i = 0, \ldots, N - 1 \) and \( j = 0, \ldots, D_{max} \), the following is computed.

\[
\begin{align*}
\delta(i, j) &= \min_{k \in \mathcal{N}}(\delta(i - 1, k) + \lambda(j, k)) + (I^l(i) - I^r(i + j))^2 + \mu(i)(d(i, t|t-1) - j)^2, \\
\psi(i, j) &= \arg \min_{k \in \mathcal{N}}(\delta(i - 1, k) + \lambda(j, k)).
\end{align*}
\]
Here, $\lambda(i,j)$ is a balancing factor between data and smoothness terms, and $\mu(i)$ is a weighting factor for the recommended disparity.

The optimization by dynamic programming (DP) consists of forward and backward processing. Most of preceding DP algorithms adopt uniqueness constraint which allows only one-to-one correspondences between two views. This results in restricted disparity transition between neighbors $\Delta d \in [-1,0,+1]$, which is not expected in occluded regions, where abrupt transitions may occur. To overcome this problem, we omitted the uniqueness constraint in our algorithm, which results in wide range of disparity transition $\Delta d \in [d_{min}, d_{max}]$.

Increased disparity transition levels may results in heavy computational burden. To prevent large amount of computations, we adopted distance-transform based technique [19] in forward processing. Moreover, original matching cost and smoothness cost are truncated to reduce the effect of matching outliers,

$$E_{data}(d) = \min(|f'(i) - f'(i+j)|, C_d),$$
$$E_{smooth}(d) = \min(|d(i,j) - d(i-1,j')|, C_v). \tag{8}$$

Algorithm 1 depicts our modified forward processing in pseudocode. Backward DP processing is actually same as normal DP algorithms, such that assigns the sub-optimal path for neighboring transitions.

**Algorithm 1** Modified forward DP processing

```plaintext
for x = 0 \rightarrow width - 1 do
  if x == 0 then
    B(x, d) \leftarrow D(x, d).
  else
    M(x, d) \leftarrow D(x - 1, d).
    for d = 1 \rightarrow d_{max} do
      M(x, d) \leftarrow \min(M(x, d), M(x, d - 1) + 1).
    end for
    M(x, d) \leftarrow \min(M(x, d), M(x, d + 1) + 1).
  end for
  B_n \leftarrow \min(B(x - 1, d)).
  B(x, d) \leftarrow \lambda D(x, d) + \mu P(x, d) + \min(M(x, d), B_n + C_v).
end if
```

**B. Considering Outliers**

As we use predicted disparity from previous frames, error propagation may occurs if we don’t consider outlying disparity or flow estimate during the processing. To check reliability of predicted flow and current matching, two methods were examined; the significance of best match and the entropy computation. The former checks for matching scores of each pixel’s best and second best matches, $C_1(p)$ and $C_2(p)$, respectively. Then we have computed the significance $C_p(p)$ as,

$$C_p(p) = \frac{C_2(p) - C_1(p)}{C_2(p)}. \tag{9}$$

We applied resulting value directly as additional weight in Eq. (7) as $C_p \in [0, 1]$. Then the original algorithm is modified as,

$$\begin{align*}
\delta'(i,j) &= \min_{k \in \mathbb{N}}(\delta'(i-1,k) + \lambda(j,k)) \nonumber \\
&\quad + (1 - C_p(p))(f'(i) - f'(i+j))^2 \\
&\quad + \mu(i)(d(i,t|t-1) - j)^2, \\
\psi'(i,j) &= \arg \min_{k \in \mathbb{N}}(\delta'(i-1,k) + \lambda(j,k)).
\end{align*} \tag{10}$$

The latter normalizes each matching cost vector $C(p)$ for a pixel to make probability distribution $P(I|D)$. Then the entropy of the distribution is computed as,

$$\epsilon(p) = - \sum_{d=|d_{min}|}^{d_{max}} p(I|d) \log p(I|d). \tag{11}$$

Since the values of $\epsilon(p) \notin [0, 1]$, we normalized resulting entropy and scaled $\epsilon'(p) = \kappa \epsilon(p)/|\epsilon(p)|$. Its application is similar to former case, except for additional fixed scale parameter $\kappa$.

**VI. Experiments**

So far, we have presented our predict-and-matching algorithm based on optical flow estimation by disparity map tracking. In this section, we aim to verify our algorithm both qualitatively and quantitatively by utilizing sequential data set with ground truth [21]. The data include five different scenarios with 100 frames per each set. Every sequences are created in 400×300 resolution, and ground truth data have 64 discrete levels of disparity scaled by 4.

**TABLE I: Implemented parameters**

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\kappa$</th>
<th>$C_d$</th>
<th>$C_v$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>13</td>
<td>70</td>
<td>10</td>
</tr>
</tbody>
</table>

The algorithm is configured to use the same parameters for every data sets. The search range of optical flow is restricted to $[-5, 5]$ to avoid outlying tracking results. The remaining algorithm parameters are determined experimentally, which are described in Table. I.

Fig. 6 shows the percentage of erroneous pixels that change in time (tPER) for four data sets. Each row includes the result from different data set. This offers a basic quantitative analysis for our work. As we can see from the figure, error rates are dramatically decreased, especially for pyramid and building sequences. Overall error reduction rate is around 5 15 percents, which shows the effectiveness of our approach. More interesting aspect of our result is that oscillation, or jittering in tPER is significantly smoothed, as we can see from the figure without statistical analysis. This improvement is more significant in well-textured sequences as the maze data set, which means that our algorithm not
only makes consecutive estimates consistent, but also reduce the effect of occlusion, pattern-less regions, or image noise in individual frame by help of temporal integration of matching information. Predict-and-update framework works well with soft constraints on energy minimization, hence it is prone to error propagation problem, and this fact is also shown in the figure as the errors are not increasing monotonically.

In Fig. 5, an example of quantitative result is represented. The marked region shows bad estimate of disparity because of the ambiguous patterns on image data. The regions marked in red show where pixels are not confident enough so they have high probability of estimation error. As we can see from the figure, applying our method successfully reduces the effect of outlying matching candidates and jittering of intensity during time. Since our algorithm is based on simple LS optical flow estimator, several false matchings may appear in still regions (regions without any motion). Applying advanced optical flow algorithm incurs a trade-off problem between efficiency and accuracy of overall system.

VII. Conclusion

In this paper, we have derived a predict-and-update framework for disparity estimation of stereo video. Sparse invariant features of previous depth estimate are extracted and tracked to make sparse prediction of disparity map. An efficient modified DP algorithm optimizes the energy function, which includes both current observation and temporal continuation terms. Proposed algorithm is verified using depth sequences with ground truth data. Quantitative and qualitative analysis show that our algorithm shows superior results than frame-by-frame algorithm and is hardly affected by error propagation.

The proposed framework is applicable to almost all kinds of existing stereo and motion estimation methods. Our future interest is the improvement of algorithm by relaxation and the implementation of real-time system.

ACKNOWLEDGMENT

This work was supported by the Ministry of Knowledge Economy, Korea, under the Core Technology Development for Breakthrough of Robot Vision Research support program supervised by the National IT Industry Promotion Agency.

REFERENCES

Fig. 6: tPER plot for experimental data (Rows: Tank, Maze, Pyramid, Building). First column shows the results without confidence measure, second and third columns consist of the results with significance and entropy measures, respectively.