the matching cost by performing two-pass aggregation using two orthogonal 1D windows [5], [6], [8]. The two-pass method first aggregates matching costs in the vertical direction, and then computes a weighted sum of the aggregated costs in the horizontal direction. Given that support regions are of size $\omega \times \omega$, the two-pass method reduces the complexity of cost aggregation from $\mathcal{O}(\omega^2)$ to $\mathcal{O}(\omega)$.

B. Temporal cost aggregation

Once aggregated costs $C(p, \bar{p})$ have been computed for all pixels p in the reference image and their respective matching candidates $\bar{p}$ in the target image, a single-pass temporal aggregation routine is exectuted. At each time instance, the algorithm stores an auxiliary cost $C_a(p, \bar{p})$ which holds a weighted summation of costs obtained in the previous frames. During temporal aggregation, the auxiliary cost is merged with the cost obtained from the current frame using

$$C(p,\bar{p}) \leftarrow \frac{(1-\lambda)\cdot C(p,\bar{p}) + \lambda \cdot w_t(p,p_{t\text{-}1})\cdot C_a(p,\bar{p})}{(1-\lambda)+\lambda\cdot w_t(p,p_{t\text{-}1})}, \quad (4)$$

where the feedback coefficient λ controls the amount of cost smoothing and $w_t(p, p_{t\text{-}1})$ enforces color similarity in the temporal domain. The temporal adaptive weight computed between the pixel of interest p in the current frame and pixel $p_{t\text{-}1}$, located at the same spatial coordinate in the prior frame, is given by

$$w_t(p,p_{t\text{-}1}) = \exp\left(-\frac{\Delta_c(p,p_{t\text{-}1})}{\gamma_t}\right), \quad (5)$$

where γ_t regulates the strength of grouping by color similarity in the temporal dimension. The temporal adaptive weight has the effect of preserving edges in the temporal domain, such that when a pixel coordinate transitions from one side of an edge to another in subsequent frames, the auxiliary cost is assigned a small weight and the majority of the cost is derived from the current frame.

C. Disparity Selection and Confidence Assessment

Having performed temporal cost aggregation, matches are determined using the Winner-Takes-All (WTA) match selection criteria. The match for p, denoted as $m(p)$, is the candidate pixel $\bar{p} \in S_p$ characterized by the minimum matching cost, and is given by

$$m(p) = \underset{\bar{p}\in S_p}{\operatorname{argmin}}\, C(p,\bar{p}). \quad (6)$$

To asses the level of confidence associated with selecting minimum cost matches, the algorithm determines another set of matches, this time from the target to reference image, and verifies if the results agree. Given that $\bar{p} = m(p)$, i.e. pixel $\bar{p}$ in the right image is the match for pixel p in the left image, and $p' = m(\bar{p})$, the confidence measure F_p is computed as

$$F_p = \begin{cases} \dfrac{\displaystyle\min_{\bar{p}\in S_p \setminus m(p)} C(p,\bar{p}) - \min_{\bar{p}\in S_p} C(p,\bar{p})}{\displaystyle\min_{\bar{p}\in S_p \setminus m(p)} C(p,\bar{p})}, & |d_p - d_{p'}| \leq 1 \\ 0, & \text{otherwise} \end{cases}. \quad (7)$$

D. Iterative Disparity Refinement

Once the first iteration of stereo matching is complete, disparity estimates D_p^i can be used to guide matching in subsequent iterations. This is done by penalizing disparities that deviate from their expected values. The penalty function is given by

$$\Lambda^i(p,\bar{p}) = \alpha \times \sum_{q\in\Omega_p} w(p,q) F_q^{i\text{-}1} \left|D_q^{i\text{-}1} - d_p\right|, \quad (8)$$

where the value of α is chosen empirically. Next, the penalty values are incorporated into the matching cost as

$$C^i(p,\bar{p}) = C^0(p,\bar{p}) + \Lambda^i(p,\bar{p}), \quad (9)$$

and the matches are reselected using the WTA match selection criteria. The resulting disparity maps are then post-processed using a combination of median filtering and occlusion filling. Finally, the current cost becomes the auxiliary cost for the next pair of frames in the video sequence, i.e., $C_a(p,\bar{p}) \leftarrow C(p,\bar{p})$ for all pixels p in the and their matching candidates $\bar{p}$.

IV. Results

The speed and accuracy of real-time stereo matching algorithms are traditionally demonstrated using still-frame images from the Middlebury stereo benchmark [1], [2]. Still frames, however, are insufficient for evaluating stereo matching algorithms that incorporate frame-to-frame prediction to enhance matching accuracy. An alternative approach is to use a stereo video sequence with a ground truth disparity for each frame. Obtaining the ground truth disparity of real world video sequences is a difficult undertaking due to the high frame rate of video and limitations in depth sensing-technology. To address the need for stereo video with ground truth disparities, five pairs of synthetic stereo video sequences of a computer-generated scene were given in [19]. While these videos incorporate a sufficient amount of movement variation, they were generated from relatively simple models using low-resolution rendering, and they do not provide occlusion or discontinuity maps.

To evaluate the performance of temporal aggregation, a new synthetic stereo video sequence is introduced along with corresponding disparity maps, occlusion maps, and discontinuity maps for evaluating the performance of temporal stereo matching algorithms. To create the video sequence, a complex scene was constructed using Google Sketchup and a pair of animated paths were rendered photorealistically using the Kerkythea rendering software. Realistic material properties were used to give surfaces a natural-looking appearance by adjusting their specularity, reflectance, and diffusion. The video sequence has a resolution of 640×480 pixels, a frame rate of 30 frames per second, and a duration of 4 seconds. In addition to performing photorealistic rendering, depth renders of both video sequences were also generated and converted to ground truth disparity for the stereo video. The video sequences and ground truth data have been made available at `http://mc2.unl.edu/current-research/image-processing/`. Figure 2 shows two sample frames