Increasing the Efficiency of Local Stereo by Leveraging Smoothness Constraints

Yilin Wang, Enrique Dunn and Jan-Michael Frahm
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, USA

***Abstract*—We introduce a novel framework for efficient stereo disparity estimation leveraging the spatial smoothness typically assumed in stereo and formalized by the various smoothness constraints. The smoothness constraint presumes that a neighboring set of pixels shares the same disparity or the disparity varies smoothly. Our key insight is that it hence suffices to evaluate any single one of those pixels at the correct disparity to identify a valid estimate for the entire set. We leverage this insight into the formulation of a complexity reducing mechanism. We distribute the exploration of the disparity search space among neighboring pixels, effectively reducing the set of disparity hypothesis evaluated at each individual pixel. Moreover, we integrate a recently proposed concept to deploy sparsity within this neighborhood of distributed disparities into our novel mechanism, in order to further reduce the computational burden. Our experiments clearly demonstrate the effectiveness of our approach by achieving comparable results to the baseline of exhaustive disparity search. The analysis of the computational complexity of our proposed mechanisms illustrates how, by making moderate assumptions on the smoothness of the observed scene, we can reduce the computational complexity of local stereo disparity search by upwards of two orders of magnitude while maintaining the comparable result quality.**

***Index Terms*—stereo; cost aggregation; sparse distributed disparity sampling;**

I. Introduction

Stereo disparity estimation is an extensively studied topic in computer vision. Historically, there has been a watershed between global and local approaches in terms of both solution quality and computational effort. The qualitative advantage of many global methods can be attributed to the enforcement of smoothness constraints (i.e. as a regularizing term within an energy minimization framework). Nevertheless, such constraint modeling typically comes at a steep computational cost, as it effectively augments the complexity of the disparity search space by introducing global correlations and/or penalties among local disparities estimates. In contrast, we propose the use of smoothness constraints within local stereo disparity search not to explicitly regularize an optimization criterion, but instead, to mitigate computational redundancy.

In its most basic form, stereo disparity estimation exhaustively enumerates and evaluates for each pixel the set of all disparity hypotheses within a predefined range. The accuracy and efficiency of such disparity estimates is contingent upon a variety of factors such as the photo-consistency *measure* used to compare pixel similarity, the scope and form of the *aggregation mechanisms* used to robustify individual pixel similarity measurements, as well as the *search strategy* (if any) used to explore the space of disparity hypotheses. In order to develop an efficient disparity *search strategy*, we leverage the implicit smoothness assumption used in most fronto-parallel photo-consistency measures, where pixels in a local neighborhood are assumed to belong to the same fronto-parallel surface and have the same disparity. Clearly, such a smoothness assumption suggests that there is significant redundancy in an exhaustive search of the disparity search space for each pixel. Hence, our first contribution is to reduce the number of evaluated depth hypothesis by spatially distributing (within a local neighborhood) the sampling of the depth hypothesis space. We introduce this structured approximation as *Distributed Disparity Sampling* (DDS). Our second contribution is to incorporate within this new framework the recently introduced concept of spatial sampling [1], a develop generalization of DDS we denote as *Sparse Distributed Disparity Sampling* (SDDS). The concepts of spatial and disparity sparsity are orthogonal in the sense that they may be used independently or in a combination. In this paper we propose a stereo disparity approach that conjugates these two separate sparsity concepts in a manner consistent with the concept of smoothness in disparity estimation. We note that spatial smoothness is an ubiquitous and implicit assumption throughout stereo methods, whose relevance for algorithm design has been hitherto neglected. Moreover, even though both sparsity mechanisms represent efficiency driven approximations, we combine them into a reduced complexity disparity estimation framework, yielding comparable results to exhaustive search.

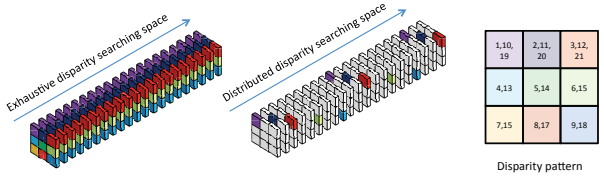

Fig. 1. Relationship between a traditional exhaustive search disparity strategy and our proposed DDS strategy. At left, every pixel within a local neighborhood is evaluated at each possible disparity. At middle, our proposed scheme where each pixel sparsely samples the disparity space, relying on neighboring pixels to infer missing data.

II. Related Work

Reducing the computational requirements of local stereo methods has recently received renewed attention. A standard complexity reducing approach based on changing the *search strategy* is the use of hierarchical stereo methods deploying Coarse-to-Fine (CTF) disparity refinement [2]. The benefit of these approaches is to reduce the total number of disparity comparisons required for estimating a single pixel's disparity. Modifications to the search strategy of CTF approaches, aimed to overcome their biggest limitation of not estimating the depth of thin objects, include the work by Jen *et al.* [3], which actively determines the coarsest scale at which a given pixel may be reliably estimated given its surrounding texture. In this way, the authors achieve a data-driven trade-off among computational efficiency and estimation accuracy. However, computational savings are bounded by maximum efficiency of CTF approaches and depend on the specific scene. In contrast, our proposed method achieves highly efficient disparity search, while performing all computations at the native image resolution.

Another approach is to reduce the computational complexity of pixel *aggregation mechanisms*. In this regard, integral images have been used successfully to reduce area based aggregation to a constant time operation [4]. The more challenging problem of performing variable weighted cost aggregation [5], [6], in constant time has been addressed recently in [7], [8]. While these approaches allow for the use of robust cost functions across arbitrarily large templates they either restrict themselves to the case of asymmetric weighting (notable exception De-Maeztu *et al.* [9]) or involve significant pre-aggregation requirements. Comparatively, our method retains the accuracy benefits of variable cost aggregation while increasing the throughput of the depth estimation process through a lower number of these template comparisons.

Hawe *et al.* [10] explored the use of spatial sub-sampling within the image to reduce the number of pixels whose disparity needs to be explicitly evaluated. The authors used a compressed sensing formulation to reconstruct high quality disparity maps from a reduced set of robust depth estimates (as low as 5% of image pixels). While such an approach eliminates a significant portion of the disparity estimation cost it transfers this cost to the optimization procedure needed for reconstructing the dense disparity map from the sparse samples. Our novel method extends the concept of sparse sampling from the image space to the cost volume space and efficiently propagates robust samples to create dense disparity maps through a localized voting framework.

More closely related to our work is the recent proposal by Min *et al.* [1] for efficient cost aggregation, which combines both a new pixel-wise likelihood histogram aggregation scheme along with sparse image sampling in order to drastically reduce computational complexity. The authors explored the robustness of their cost aggregation scheme across different levels of pixel sampling sparsity and discovered that for their proposal there is not a monotonic correlation between estimation accuracy and the sampling sparseness. In contrast to [1], we favor robustness by deploying full template variable cost aggregation instead of single pixel intensity photo-consistency. To overcome the related performance challenges we couple our template variable cost aggregation with a depth sub-sampling framework, which achieves both lower computational complexity and higher processing speed, while simultaneously improving accuracy.

Most of the top ranked state of the art stereo algorithms are segment-based global stereo approaches (7 out of the top 10 according to the Middlebury benchmark [11]). Some of these methods directly work on segmented images [12], [13], so that the smoothness assumption is implicitly embedded in each of the segments: pixels in the same segment should have the same disparity. More recent work [14] weakened this constraint by using overlapping subsegments. Some local stereo algorithms [15] also use line segmentation to utilize the smoothness assumption. In contrast to global methods using smoothness as an explicit regularizer, we use the smoothness assumption to enable a sparse cost volume computation yielding a speedup of two orders of magnitude.

III. Distributed Disparity Sampling

The smoothness constraint in stereo estimation confers a strong correlation between the spatial proximity of neighboring image pixels and the spatial proximity of their corresponding 3D points in the observed scene. Accordingly, having each pixel evaluate all possible disparities using a photo-consistency measure that implicitly enforces such smoothness constraints, leads unequivocally to the realization that there is significant redundancy in the computation of local stereo methods.

Key Insight. *If we assume a neighboring set of pixels share the same disparity, it suffices to evaluate any single one of those pixels at the correct disparity to identify a valid estimate for the entire set.*

In this paper, we propose to modify the traditional (exhaustive) *search strategy* for disparity estimation into a structured and sparsely spatially distributed search scheme. Namely, lets assume, without loss of generality, a neighborhood of pixels $\aleph = \{p_i | i \in [1, \dots, N]\}$, which satisfies the smoothness constraint. Let us further assign to each p_i a disparity offset value $o_i \in [1, \dots, N]$, such that $\{o_i \neq o_j | \forall i \neq j\}$. In order for the neighborhood of pixels $\aleph$ to *collectively* explore an entire consecutive set of disparities $\mathcal{D}= \{d_k | k \in [1, \dots, D]\}$, where $D \geq N$, it suffices to assign to each p_i a subset of disparities $\{d^i_j = d_k | k = j * N + i \ , \ j \geq 0 \ , \ k \leq D\}$, where the union of the subsets for pixels in $\aleph$ equals the entire disparity range. In this way, each pixel will search a sparse set of disparities where the number of total disparities is bounded by the ratio D/N and the disparity offset between consecutive samples is N.

We denote this new depth sampling scheme *Distributed Disparity Sampling* (DDS). In order to extend DDS across the entire image we may simply tile the image with this sampling pattern, which ensures that in each local stereo window the set of sampled disparities equals $\mathcal{D}$. In this way,

disparity assignment is based on the traditional winner take all (WTA) selection within a vicinity centered on a given pixel. Note that for pixels near the periphery of their sampling pattern, their WTA vicinity encompasses pixels belonging to neighboring sampling patterns. Accordingly, the distribution of offsets within the search patterns, may introduce bias into our estimation. We have found that randomly determining a fixed pattern to be repeated across the entire image generally provides robust results and minimizes the bias. The reduction in computational complexity afforded by DDS is proportional to the size of the disparity distribution neighborhood. Hence, if we define a square neighborhood of side length M=10 for an effective quadratic reduction in computational complexity of two orders of magnitude.

The proposed DDS scheme will provide reliable depth estimates only when the neighborhood used for depth exploration covers a single fronto-parallel surface. Moreover, while DDS indeed provides a remarkable performance to cost ratio (we refer to Section VI for further details), up to this point we have only considered the case where the disparity range is larger than the number of pixels in the disparity distribution neighborhood $\aleph$ (i.e. $D \geq N$). In order to improve upon the efficiency of DDS we have explored the opposite scenario where $D < N$. A straightforward solution is to apply redundant disparity sampling within the DDS neighborhood. Instead, we have incorporated spatial sparsity into our approach and in doing so we have developed a more efficient and accurate generalization of DDS, which is described in the following section.

IV. Sparse Distributed Disparity Sampling

In this section we extend our novel DDS method to be able to leverage the spatial sparsity concept proposed by Min *et al.* [1]. The resulting sparse distributed disparity sampling (SDDS) approach can be summarized as follows.

1) We define a set of neighborhoods (with possible overlap) that cover the entire image. For each of these neighborhoods:
   a) We randomly select individual pixels to be evaluated each at a single specific disparity hypothesis until the entire disparity range $\mathcal{D}$ is sampled without redundancy, i. e. each disparity is sampled once.
   b) Step 1 (a) is performed k times for the neighborhood to obtain a consensus on a reduced set of representative disparities for it.
2) Then we spread a regular sparse pattern of seeds across the entire image and for each seed we evaluate the joint set of representative depths of all the neighborhoods to which the seed belongs.
3) The disparity estimate for each of the seed pixels is the one with minimum cost among those evaluated in Step 2.
4) The disparity estimate for non-seed pixels is obtained through proximity and photo-consistency weighted voting among all seeds in their vicinity.

Our cost aggregation deploys robust symmetric weight aggregation during the windowed matching (steps 1(a) and step 2). The remainder of this section discusses in greater detail the mechanisms and design decisions enabling our SDDS approach.

A. Variable Cost Aggregation

We rely on variable cost aggregation (VCA) in a similar manner to [5], as it has been shown to be a highly accurate and discriminative local photo-consistency aggregation framework. We mitigate the computational burden of using VCA by effectively reducing the number of times such template vs. template evaluations need to be made. Note that this in no way precludes the use of constant time weighted aggregation approaches such as the ones discussed in Section II. For adaptive weight based stereo, the matching cost for pixel p with disparity d is computed as:

$$cost(p,d) = \frac{\sum_{q \in N(p)} w(p,q) w(\bar{p},\bar{q}) e(q,\bar{q})}{\sum_{q \in N(p)} w(p,q) w(\bar{p},\bar{q})} \tag{1}$$

where $\bar{p} = p - d$ and $\bar{q} = q - d$ are the corresponding pixels of p and q on the matching image, while $N(p)$ is the set of p's neighbors. The pixel similarity measure $e(q,\bar{q})$ used in this work is the AD-census cost [16]. The adaptive weight $w(p,q)$ for q with respect to the center p, combines the color distance (in CIELAB space) and the spatial distance:

$$w(p,q) = \exp\left(-\frac{\parallel (I_p - I_q) \parallel_2}{\lambda_{color}} - \frac{\parallel (p - q) \parallel_2}{\lambda_{spatial}} \right) \tag{2}$$

here λ_{color} and $\lambda_{spatial}$ are constants used to adjust the influences of color and spatial differences.

The two most time consuming operations are the pixel-wise comparison $e(q,\bar{q})$ and the weighting $w(p,q)$. Even though very efficient implementations of the AD-Census similarity measure can be achieved [16], we note in general that pixel-wise weight computation $w(p,q)$ is more efficient than similarity computation $e(q,\bar{q})$, due to the overhead associated with the census transform estimation in addition to Census and SAD aggregation. Other operations like disparity sorting could be ignored with respect to these two operations. To analyze the time complexity of VCA we define Q to be the number of pixels in the image, D the number of tested disparity hypotheses, and W the side length of a square matching window. Then the complexity of the exhaustive adaptive weight stereo is $QDW^2(2 * u_w + u_m)$, where u_m and u_w are basic complexity for pixel-wise comparison and weighting. Here, we reduce the window size W by sparse spatial sampling. Let $s \in [1, \ldots, W/2]$ be the sampling step indicating the distance between samples along each image direction, then sub-sampling reduces the total complexity by a factor of $1/s^2$. Note that, such spatial sub-sampling within the photo-consistency measure assumes smoothness between pixels separated by $L = s/2$ pixels. This formulation implies that by performing photo-consistency using every consecutive pixel (i.e. $s = 1$) we are in fact assuming a smoothness level of $L = 0.5$ pixels (i.e. the strong spatial correlations only

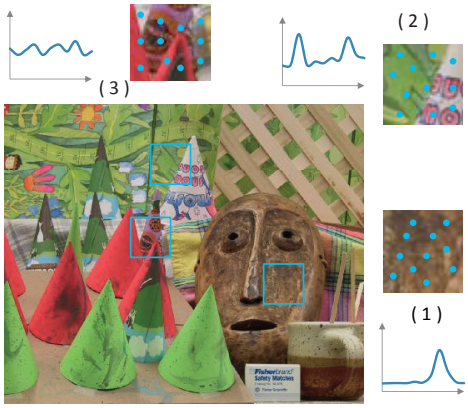

Fig. 2. Disparity sampling analysis. Case 1: only one peak of the score profile for continuous surfaces; case 2: two peaks for simple discontinuous region (only two different surfaces); case 3: no distinguish peaks for patches with complex structures.

extend from the center of the pixel to each edge along both directions). Depending on the smoothness level assumed for a given image scene, the value of s can be reasonably controlled. We empirically determined a value $s = 4$ representing a good trade-off between accuracy and efficiency.

B. Representative Depths of a Neighborhood

For each local patch, we randomly select D (equal to the size of the disparity set) pixels, assign them different test disparities and compute their costs according to Eqn. (1). Next, we sort all the estimated costs and assign to each disparity a score inversely proportional to its order. By repeating this sampling and aggregating the scores for each disparity, we obtain a reliable profile of scores, which approximates the profile of the likelihoods of all disparities within the local patch structure. For example (as illustrated in Figure 2), if we only find one peak in the profile, since weights are computed from randomly selected pixels, it is highly probable that all the pixels in the local patch share the same disparity. Similarly, if there are two peaks, the local patch may be on the boundary of two objects, the absence of distinguishable peaks suggests the complex scene, which may not follow the smoothness assumption.

Let c_i be the cost for the ith disparity after a single sampling, and o_i is the position of the ith disparity in the sorted sequence of the cost of the disparities, then each disparity is scored as $1/o_i$. After sampling k times, disparities with total scores greater than T_s (pre-defined threshold, e.g. $T_s = 1.2$ when sampling 4 times) will be treated as the *representative disparities* for the current local patch. Given that sampled pixels are selected randomly, there is in fact a possibility of failing to find any correct disparities. However, in practice, *SDDS* works well for patches containing three or less structures.

It can be seen that if all all the sampled pixels within a local patch belong to a single fronto-parallel surface (i.e. they all have the same disparity), the probability of identifying the correct disparity is $\mathtt{P}= 1$ given that all disparities of $\mathcal{D}$ are sampled. Accordingly, after the kth sampling, the final score for the correct disparity would be k. Hence for patches from a single surface, we will always find the correct disparity (i.e. one of the samples will be correct and we are certain this sample will have the highest score). This argumentation is predicated on the typical assumptions made in stereo that for a given pixel, the correct disparity will get the smallest cost, or alternatively, that this minimum cost corresponds to the best possible disparity estimate. This implies that when a pixel is evaluated at its correct disparity hypothesis, it will be within the disparities with the smallest cost hence o_i will be small after sorting the results of all disparity values within the patch. It is further assumed that the positions o_i of the other (incorrect) disparity estimates within the patch are randomly distributed, which is true for all regions with a well defined global cost minimum. In all other cases the stereo decision is ambitious. Hence we exclude them from our analysis, although in practice even those cases work robustly. For patches containing multiple surfaces we chose the threshold T_s to ensure a false positive rate (wrong disparities selected into the set of interest disparities) of about 2%. Please, note that any wrong disparity being part of the interest disparities does only slightly influence computational performance but has no quality implications.

C. Sparse Seed Evaluation and Propagation

The distributed attainment of a *representative disparity set* for each neighborhood can be seen as an efficient sampling in disparity space. In the next stage of our pipeline we also perform sampling on the 2D image space to reduce computation even further. Given that neighboring pixels generally have similar matching costs for a given disparity, it is reasonable for the majority of pixels to approximate their corresponding costs from a set of pre-computed cost measurements in their vicinity. This reasoning leads to the sparse sampling of the image by means of pixel "seeds". Seed selection may be color similarity based or spatial proximity based. However, a good color-based sampling usually needs additional processing like clustering or segmentation, which may be prohibitive in cost for local stereo methods. Accordingly, we uniformly distribute m seeds in the image, and for each disparity d belonging to the joint set of *representative disparities* of the seed, we compute their adaptive weight costs $c_{s_1,d}, ..., c_{s_m,d}$. For the ith remaining pixels, their costs $c_{i,d}$ are computed by weighted aggregating from neighboring seeds

$$c_{i,d} = \frac{\sum_{s_j \in N(i)} w(i, s_j) c_{s_j,d}}{\sum_{s_j \in N(i)} w(i, s_j)}. \tag{3}$$

Where $N(i)$ is the set of seeds within the ith pixel's proximity and w_{i,s_j} is the adaptive weight between the ith and s_jth pixels. After cost aggregation, the candidate disparity with minimum cost will be chosen as the final disparity.

D. Computational cost analysis

In this section, we analyze the total computational cost of our SDDS approach. As discussed in Section IV-A, the

complexity for exhaustive adaptive weight based stereo and corresponding sub-sampled aggregation stereo are $C_{EX} = QDW^2(2u_w + u_m)$ and $C_{SP} = QDW^2(2u_w + u_m)/s^2$, where Q is the total number of pixels, D is the size of the disparity set, W is the size (width and height) of matching window, $s$is the spatial sampling ratio inside the matching window, u is the unit computation cost, u_w and u_m are complexity for pixels-wise weighting ($u_w \approx 2u$) and pixels-wise matching cost ($u_m \approx 5u$). In our algorithm, when applying sampled aggregation, the computational cost C_1 to select the representative disparities, the cost C_2 for aggregating the seed costs', and the computational cost C_3 for aggregating the remaining pixels' costs are:

$$C_1 = kQDW^2(2u_w + u_m)/(\mu s^2 B^2) \quad (4)$$

$$C_2 = mQD_cW^2(2u_w + u_m)/(\mu s^2 B^2) \quad (5)$$

$$C_3 = nQD_c u_w/\mu. \quad (6)$$

where B is the size of local patch, D_c is the average size of selected candidate disparities (≈ 3), $1-\mu$ is the overlapping ratio of neighboring patches, k is the number of randomly sampling times, m is the number of seeds per local patch, and n is the number of neighboring seeds for cost aggregation. Accordingly, the total complexity $C_{SDDS} = C_1 + C_2 + C_3$. Substituting the default values used in our experiments (see Table I), and assuming the size of the original disparity set $D = 60$, block size $B = 50$, and sampling $k = 4$ times, we find ratio of computational costs to be $\frac{C_{EX}}{C_{SDDS}} = 1096$ and $\frac{C_{SP}}{C_{SDDS}} = 69$. In fact, even when omitting the use of sub-sampling within the VCA matching (i.e. $s = 1$), we obtain a computational speed-up of $\frac{C_{EX}}{C_{SDDS}} = 131$. We can see our disparity sampling approach reduces the computational cost dramatically. The main reason is that our method reduces the cardinality of the tested disparities per pixel ($\approx$ 3). Additionally, our technique replaces the costly and redundant matching cost computation (patch-wise comparison) by aggregating costs from neighboring seeds, which has only a cost of $2u_w + u_m$ to u_w. Comparing with another fast stereo approach HistoAggr [1], whose computational cost could be approximated by $C_{HA} = QDu_m/(s_{HA}^2) + QD_{fixed}W^2u_w/(s_{HA}^2)$, where $s_{HA} = 3$ is the default spatial ratio and $D_{fixed} = D/10$ is a fixed subset size of disparities. Our experiments clearly demonstrate that our method outperforms [1] by a factor of $2.8 = \left(\frac{C_{HA}}{C_{SDDS}}\right)$.

V. Disparity Refinement

To overcome the typical minor mishaps of local stereo estimation caused for example by occlusions, we propose a voting-based refinement method inspired by the work of [1]. Firstly, pixels P_i^r with reliable disparities d_i^r are found by left-right cross validation. Since the cost c_i^r for each reliable disparity d_i^r is known after the disparity computation, we define the likelihood for d_i^r as $l_i^r = 1 - c_i^r$. Notice that each reliable pixel only needs to keep the likelihood value for its final disparity. When estimating the disparity for an unreliable pixel P_j^u, first we check the ratio of reliable pixels within its neighborhood, if the ratio is greater than some pre-defined threshold T_N, then we build a voting list for all the reliable disparities occurring in its neighborhood, and the voting score is computed by summing weighted likelihoods of the reliable neighbors:

$$v_{j,d_i}^u = \sum_{P_i^r \in N_r(P_j^u)} l_i^r w(P_j^u, P_i^r), \quad (7)$$

where $N_r(P_j^u)$ is the set of reliable pixels within P_j^u's neighborhood N_r, and the weights $w(P_j^u, P_i^r)$ are computed by Equation 2. Also, for each candidate disparity d_i, its confidence f_{j,d_i} is determined by the maximum value of corresponding weight $w(P_j^u, P_i^r)$. Then the disparity with maximum confidence v_{j,d_i}^u greater than threshold T_F (initially $T_F = 0.5$) is chosen as the final disparity d_j^r. If no candidate disparity has good confidence, the minimum disparity will be set as the final disparity (similar to filling holes with the background). Finally, P_j^u becomes a new reliable pixel P_j^r, and its new likelihood is computed as

$$l_j^r = \frac{v_{j,d_j}^u}{\sum_{d_i=d_j} w(P_j^r, P_i^r)} \quad (8)$$

We repeat the above procedure for all unreliable pixels, and gradually decrease the threshold T_F if all the unreliable pixels cannot find enough reliable neighbors. The detailed algorithm is shown in Algorithm 1. Notice that the goal of this paper is to more efficiently perform the local depth estimation without loss of accuracy. We utilized similar post processing to that in [1] in order to better contextualize our results with respect to the Middlebury benchmark and make both approaches comparable. Improved refinement procedures would elevate the ranking of our implementation but that was not our emphasis.

VI. Experiments

We evaluate our methods on the stereo images from the Middlebury benchmark. Table I lists the default parameters used in all experiments, whose corresponding descriptions may be found in Section IV-D.

TABLE I
DEFAULT VALUES FOR EXPERIMENT PARAMETERS.

Parameter	W	s	μ	m	n
default value	31	4	50%	100	20

In the first set of experiments, we investigate the quality of raw disparity images. As an evaluation baseline, we use the results and performance of an exhaustive disparity search with variable cost aggregation. Moreover, we omitted performing our disparity and spatial sub-sampling, yielding a similar approach to that of [7] but using AD-Census as a photoconsistency measure. Results for DDS and SDDS, with varying cardinality of the disparity exploration neighborhood $\aleph$, are compared relatively to the baseline, with all methods being executed on the same machine. For accuracy comparison,

Algorithm 1 Voting Refinement

Parameters:
$\{P^r\}$ and $\{P^u\}$: set of reliable and unreliable pixels
d_i^r: disparity for reliable pixels $P_i^r \in \{P^r\}$
l_i^r: likelihood for reliable pixels P_i^r with disparity d_i^r
$N_r(P_j^u)$: set of neighboring reliable pixels for P_j^u
T_N: Threshold of size of neighboring reliable pixels
T_F: Threshold of confidence

Algorithm:
while $\{P^u\} \neq \emptyset$ **do**
  for $P_j^u \in \{P^u\}$ **do**
    if $|N_r(P_j^u)| < T_N$ **then**
      continue
    end if
    Initialize voting list v^u for candidate set $\{d\}$
    for $d_i \in \{d\}$ **do**
      $v_{j,d_i}^u = \sum_{P_i^r \in N_r(P_j^u)} w(P_j^u, P_i^r) l_i^r$
      $f_{j,d_i} = \max_{P_i^r \in N_r(P_j^u)}(\{w(P_j^u, P_i^r)\})$
    end for
    if $\max(\{f_{j,d_i}\}) < T_F$ **then**
      $d_j = \min(\{d\})$
    else
      $d_j = \text{argmax}_{d_i}(\{v_{j,d_i}^u\})$
    end if
    $l_j^r = v_{j,d_j}^u / \sum_{d_i=d_j} w(P_j^u, P_i^r)$
    $\{P^u\} = \{P^u\} - P_j^u$
    $\{P^r\} = \{P^r\} \cup P_j^u$
  end for
  if $|\{P^u\}|$ not decrease **then**
    $T_N = T_N/2$
  end if
end while

the baseline is the number of pixels with correct disparities (according to the Midlebury benchmark [11]) generated by the baseline. From this value, the **hit ratio** is computed as the number of correct pixels estimated by our methods divided by the same corresponding quantity for the baseline method. Similarly, the **relative processing time** compares the processing times of the exhaustive baseline method and our proposals.

First, we investigate how the performance of DDS changes when the neighborhood size N increases form 3^2 to 9^2, see Figure 3. We find that for images with small disparity set($D < 20$, e.g. Tsukuba and Venus), the hit ratio and processing time do not decrease significantly when $N \geq 5^2$, this is because when $D \leq N$, each pixel only tests a single disparity (the sparsest case for DDS), and no further computation savings can be achieved. For large disparity sets (Teddy and Cones), the relative processing time is about 0.0016 when $N = 9^2$ (approximately 625 times faster than the baseline), while DDS is still able to obtain more than 90% hit ratio. The considerable speed up for DSS is a consequence of both the quadratic

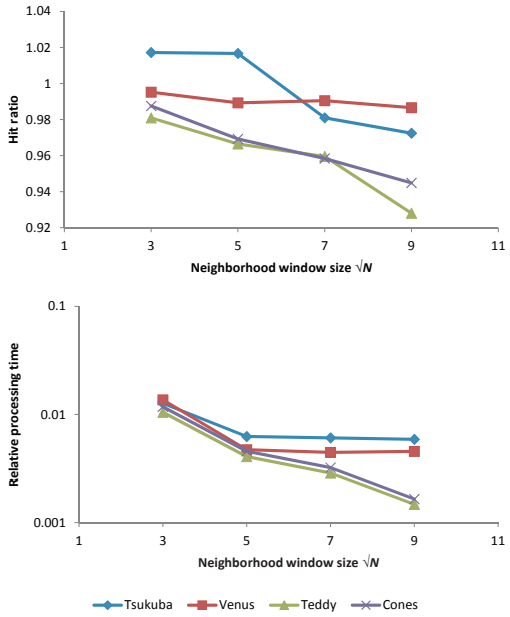

Fig. 3. Raw disparity map evaluation for DDS : Hit ratio and Relative processing time

reduction in depth sampling and the sparse spatial sampling used for variable cost aggregation.

Similar comparisons for SDDS are shown in Figure 4 for various block sizes $11 \leq B \leq 91$ and sampling iterations $2 \leq k \leq 10$. From the hit ratio plot, we can see when $B \leq 51$, SDDS hits more than 95% percent of reliable pixels by sampling 2 times. Even for large block sizes (71 and 91), the hit ratio is greater than 90% after using four samples to build the representative disparities set. Particularly, the hit ratio for Venus images is greater than 1, which means our method could find more reliable pixels than the exhaustive method did. From the processing time comparison, we can see when the block size is larger than 51 and sampling times less than 5, the processing time is improved by 1000 times. Considering that the sparse aggregation scheme speeds up our algorithm by 16 times, the pure benefit of our disparity sampling scheme is more than 60 times.

Comparing DDS and SDDS on the Cones dataset illustrates the relative processing time for DDS ($N = 9^2$) and SDDS($B = 51$ and $k = 4$) are 0.00164($\approx$ 610 times) and 0.00108($\approx$ 926 times) respectively, while the corresponding hit ratios are 0.94 and 1.00. Similar results have been found for other datasets, indicating that SDDS is both faster and more accurate than DDS.

For the second set of experiments we benchmarked our performance against three recently proposed stereo algorithms aimed at improving efficiency: HistoAggr[1] (pixel-wise cost aggregation), ESAW[17] (exponential step cost aggregation), and CSBP[18] (constant-space belief propagation). Our C++ with OpenMP (4 threads) implementation of SDDS was executed on a Quad-core Intel Xeon W3540 @2.93GHz with default parameters shown in Section 4.4. We used our

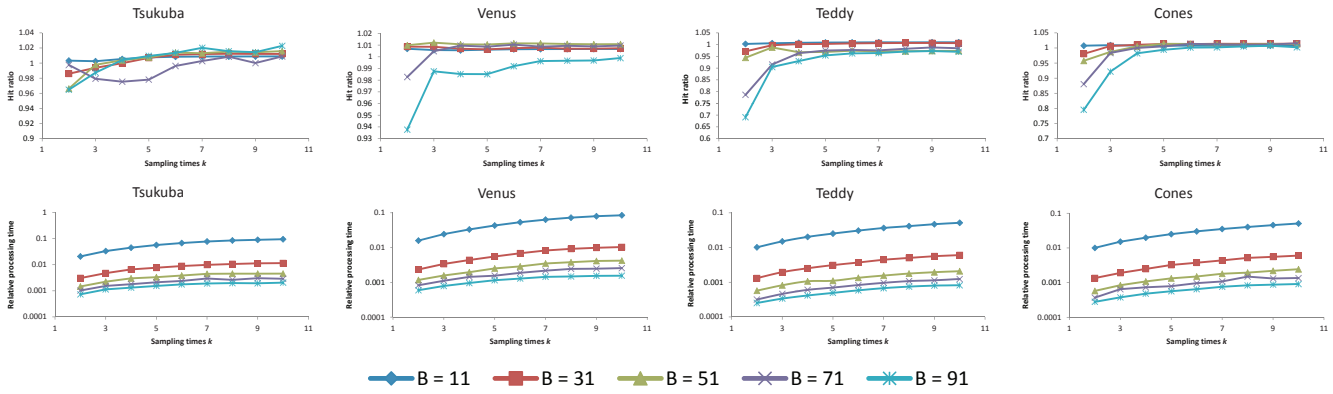

Fig. 4. Raw disparity map evaluation for SDDS: Hit ratio (top) and relative processing time (bottom) for various block size B and sampling times k.

own implementation of HistoAggr in C++ with OpenMP (4 threads). We used the author provided code for CSBP and ESAW. For ESAW we set the number of iterations to six in order to achieve comparable results (longer executions tend to improve quality at the cost of efficiency). Table II shows the Middlebury benchmark results for *raw disparity map* comparison for disparity error (all pixels) and processing time. Notice that our SDDS implementation is CPU based, and is the fastest among all reported CPU methods. Specific comparison with the computational times reported in [1] should reflect on the significance of our proposal. All faster methods on the Middlebury website deploy GPU implementations and there is no aspect of our framework that precludes this.

For the third set of experiments we compared our refined output against those reported on the Middlebury benchmark. Depthmap refinement effectively doubled our execution time given the need to generate both left and right depthmaps and also adds a processing time penalty in the range of 0.3s to 0.5s depending on resolution. Figure 5 shows refined disparity maps for DDS($N = 3$) and SDDS($B = 51$ and $k = 4$), while the corresponding quantitative evaluation results are listed in Table III(results for SDDS have been uploaded to Middlebury website). We find that SDDS consistently outperforms DDS, especially for the discontinuous regions like the lamp arm, which means SDDS is more efficient for sampling disparities than using fixed patterns. Also, we can see SDDS has better performance than other adaptive weight based stereo approaches like HistoAggr [1], FastBilateral[6], and FastAggreg [19]. Our results for Venus and Teddy are even better than the conventional adaptive weight algorithm (AdaptWeight [5]). For large continuous regions with outliers having very low costs for incorrect disparities, our approach usually has better results. One reason is that due to repeated random sampling scheme, scores of incorrect disparities are unlikely to be high enough to be selected as the potential disparities for the whole block, so their influence is limited.

VII. Conclusions

We have presented and efficient and accurate modification to the standard exhaustive search paradigm in local stereo. We have achieved this by incorporating the concept of depth smoothness at the algorithm design level. Our approach incorporates sub-sampling at the disparity as well as the spatial search space. The joint use of these orthogonal performance-oriented optimization mechanisms enables a dramatic reduction in the computational burden of local stereo methods, by reducing the total number of template comparisons. Such an approach enables the efficient implementation of variable cost aggregation and outperforms, in terms of solution quality and speed, recent state of the art work on complexity reduction for local stereo [1].

Future work entails developing a more data adaptive scheme to improve upon our current fixed parameter selection. Additionally, we plan to integrate our approach into to a high-throughput GPU implementation and explore its performance for high resolution images.

Acknowledgment

This work was in parts supported by NSF IIS-0916829 and Navy SPAWAR funding.

References

[1] D. Min, J. Lu, and N. D. Minh, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?" in *ICCV*, 2011.

[2] M. Sizintsev, "Hierarchical stereo with thin structures and transparency," in *Proceedings of the 2008 Canadian Conference on Computer and Robot Vision*, 2008.

[3] Y.-H. Jen, E. Dunn, P. Fite-Georgel, and J.-M. Frahm, "Adaptive scale selection for hierarchical stereo," in *Proc. BMVC*, 2011.

[4] O. Veksler, "Fast variable window for stereo correspondence using integral images," in *CVPR*, 2003.

[5] K. jin Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. PAMI*, vol. 28, pp. 650–656, 2006.

[6] S. Mattoccia, S. Giardino, and A. Gambini, "Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering," in *ACCV*, 2009.

[7] K. Zhang, G. Lafruit, R. Lauwereins, and L. J. Van Gool, "Joint integral histograms and its application in stereo matching," in *ICIP*, 2010, pp. 817–820.

[8] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *CVPR*, 2011.

[9] L. De-Maeztu, S. Mattoccia, A. Villanueva, and R. Cabeza, "Linear stereo matching," in *ICCV*, 2011.

TABLE II
RAW DISPARITY MAP COMPARISON FOR DISPARITY ERROR (ALL REGIONS) AND PROCESSING TIME.

Algorithm	Tsukuba		Venus		Teddy		Cones	
	error	time	error	time	error	time	error	time
SDDS	6.25	0.137s	2.00	0.196s	17.4	0.339s	13.1	0.327s
HistoAggr[1]	6.63	0.256s	2.15	0.392s	17.9	0.507s	13.1	0.490s
ESAW [17]	2.85	0.314s	3.75	1.097s	17.1	2.237s	17.4	2.285s
CSBP[18]	4.17	0.304s	3.11	0.406s	20.2	0.662s	16.5	0.677s

TABLE III
REFINED DISPARITY MAP EVALUATION FOR NON OCCLUSION(NOCC), ALL, DISCONTINUOUS(DISC) REGIONS, AND AVERAGE PERCENT BAD PIXELS (APBP).

Algorithm	Tsukuba			Venus			Teddy			Cones			APBP(%)
	nocc	all	disc	nocc	all	disc	nocc	all	disc	nocc	all	disc	
AdaptWeight[5]	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
SemiGlob [20]	3.26	3.96	12.8	1.00	1.57	11.3	6.02	12.2	16.3	3.06	9.75	8.90	7.50
SDDS	3.31	3.62	10.4	0.39	0.76	2.85	7.65	13.0	19.4	3.99	10.00	10.8	7.19
FastBilateral [6]	2.38	2.80	10.4	0.34	0.92	4.55	9.83	15.3	20.3	3.10	9.31	8.59	7.31
RealTimeABW [21]	1.26	1.67	6.83	0.33	0.65	3.56	10.7	18.3	23.3	4.81	12.6	10.7	7.90
HistoAggr[1]	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
DDS	3.39	3.70	11.6	0.53	0.97	4.14	7.92	14.3	21.1	4.61	10.9	12.1	7.94
FastAggreg [19]	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

Fig. 5. Comparison for refined disparity maps. From top to bottom: color images, refined disparity maps for DDS and SDDS

[10] S. Hawe, M. Kleinsteuber, and K. Diepold, "Dense disparity maps from sparse disparity measurements," in *ICCV*, 2011.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, 2002.

[12] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," 2006.

[13] Z. Wang and Z. Zheng, "A region based stereo matching algorithm using cooperative optimization," in *CVPR*, 2008.

[14] M. Bleyer, C. Rother, and P. Kohli, "Surface stereo with soft segmentation," in *CVPR*, 2010.

[15] X. Sun, X. Mei, S. Jiao, M. Zhou, and H. Wang, "Stereo matching with reliable disparity propagation," in *3DIMPVT*, 2011.

[16] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," in *GPUCV*, 2011.

[17] W. Yu, T. Chen, and J. C. Hoe, "Real time stereo vision using exponential step cost aggregation on gpu." in *ICIP'09*, 2009, pp. 4281–4284.

[18] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching." in *CVPR'10*, 2010, pp. 1458–1465.

[19] F. Tombari, S. Mattoccia, and E. Addimanda, "Near real-time stereo based on effective cost aggregation," in *ICPR*, 2008.

[20] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *CVPR*, 2005.

[21] R. Gupta, S.-Y. Cho, and A. Gambini, "Real-time stereo matching using adaptive binary window," in *3DPVT*, 2010.