World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:8, 2016

A Developmental Survey of Local Stereo Matching Algorithms

André Smith, Amr Abdel-Dayem

***Abstract*—**This paper presents an overview of the history and development of stereo matching algorithms. Details from its inception, up to relatively recent techniques are described, noting challenges that have been surmounted across these past decades. Different components of these are explored, though focus is directed towards the local matching techniques. While global approaches have existed for some time, and demonstrated greater accuracy than their counterparts, they are generally quite slow. Many strides have been made more recently, allowing local methods to catch up in terms of accuracy, without sacrificing the overall performance.

***Keywords*—**Developmental survey, local stereo matching, stereo correspondence.

I. Introduction

IN the field of stereo vision, two-frame matching has become an extensively studied area. The goal of this research is to generate depth information based on a pair of images, typically taken from parallel viewpoints, similar to standard human vision. With the ever-improving technologies in computational ability, both utilizing CPU and GPU capabilities as well as image capturing devices, the research into stereo matching is leading to more efficient and accurate algorithms. Despite the hundreds of research articles pertaining to this (over 160 on the Middlebury website [23] alone), it is still very active. This is in large due to this field approaching its fourth decade. Not only have the changes in technologies greatly affected the direction of research, but also certain breakthroughs have resulted in several shifts in this paradigm. To fully understand this, a review of the progress of this research is necessary.

II. Origins

The basis of stereo matching begins with Marr et al., in 1979 [6], where the first computational theory of stereo correspondence was demonstrated. The initial goal was to break down the human stereo visual system, in order to derive an algorithm to recreate this process. The article goes into great detail about the human visual system, identifying the effects of elements such as the frequency-domain of the images, the distance of elements, and eye motion on a person's ability to determine the distance of visible elements. One problem that is quickly identified when matching elements is that of ambiguity. When similar or identical intensities (i.e. gray scale values in images) are found in multiple places in a particular scene, external factors are required to limit and correctly pair them, left to right image. It should be noted that, while distance is used to describe the separation between the observer and the object, disparity represents the separation between the same object found in left and right views, in this case the distance between pixels. With this, two constraints are put in place to guide and limit the potential matches. The first of these is a uniqueness constraint. This enforces that, for each finite point (e.g. a pixel), only a single disparity value can be assigned to it. This is equivalent to stating that every object can only be found at a particular distance from the viewer, and not at multiple distances. The other constraint is that of smoothness. This, though less restrictive, introduces the supposition that most areas within the scenes are expected to be relatively smooth, or continuous, such as on the surfaces of objects. This suggests that the relative distance between these points and the viewer are very similar, or they follow a predictable trend. One way to perceive this, is that if a flat and smooth object is placed before the viewer, the distance between the viewer and the object is likely to be about the same across its visible surface. From this, a cooperative correspondence algorithm is proposed. To perform matching, a set of four different masks is used, in 12 different orientations. These attempt to match, between both scenes, certain common patterns for edges and borders. The goal is to first assert disparity values for pixels whose matches are of the highest likelihood. These are expected to be near each other (low disparity), and have little to no ambiguity in the match. From this, the missing values, that is to say, ones which cannot be matched with much likelihood, can be determined. Since uniqueness is required, once certain assertions are made, this relieves ambiguity for other matches, as alternatives are eliminated. Correspondence in both directions (left to right and right to left) can further reduce ambiguous matches, since the uncertainty may be in only one of the two cases. The remaining gaps can be filled by propagating the results obtained by previous steps, following the smoothness constraint. To build the disparity map, a dynamic memory structure, entitled a 2 1/2 D sketch is proposed. The resulting algorithm is tested on a set of artificially generated stereograms. Though the results are decent at best, the research created an excellent foundation to future stereo correspondence findings.

During the following decade, a few dozen articles have made their appearance, pertaining to the image

André Smith is with the Department of Mathematics and Computer Science, Laurentian University, Sudbury, Ontario, Canada (e-mail: aw_smith@laurentian.ca).

Amr Abdel-Dayem is with the Department of Mathematics and Computer Science, Laurentian University, Sudbury, Ontario, Canada (phone: +1 705-675-1151 extension (2396); fax: +1–705–673-6591; e-mail: aabdeldayem@lcs.laurentian.ca).

correspondence problem. A survey of these [5], by Dhond et al., came to fruition. This focuses on both image pairs and triplets used for correspondence, for the purpose of 3D structure generation. One noted detail about the techniques illustrated, is that they are passive, meaning they have no information about depth information prior to execution of the algorithms. The survey makes the distinction of two categories of algorithms: old techniques, which perform area-based matching using fixed regions or windows, and new techniques, which perform feature-based matching. The older methods are considered to be a more direct solution to the problem detailed by Marr et al. [6]. These compare the intensities (gray values) of regions of the images in order to match them. These have been further divided into two separate categories: global and local matching. With global techniques, the goal is to maintain an overall consistency. This can be performed by using hierarchical approaches, where large segments of the images are first considered, and after initial correspondence, smaller segments are then used, to improve the results. This process is repeated and results in a disparity map with global consistency. The disadvantage of this is the lack of accuracy for small details. Conversely, local methods focus their matching to specific regions or segments. These are typically rectangular windows with a fixed size. These do present better accuracy in terms of detail, but the global consistency suffers as a result. Despite these, the region matching techniques suffer from the same flaws. Inconsistencies in image contrast, as well as distortions due to changes in viewpoints, result in a reduction of accuracy since matches are pixel-wise.

Unlike area-based methods, ones that use features for comparison overcome some of the previously described issues. These methods identify objects within the scene, and attempt to find correspondence between these as an alternative. Much emphasis is placed on edge detection techniques in order to identify objects. The advantage to these techniques is, since objects are compared instead of pixels directly, errors in terms of contrast and even camera positioning are avoided. Although a parallel axis camera positioning is typically expected, where the cameras are positioned similar to the positioning of human vision, error may be introduced here. Small discrepancies in the calibration, such as the facing angle, distance, and tilt, are another source of inconsistencies between the images. With the feature-based approaches, these can fairly easily be circumvented. The result is a more accurate depth map, with even the possibility of sub-pixel precision.

III. Rectification-Based Approaches

Despite the new direction in methodology, area-based correspondence maintains active interest, and several strides are made to overcome the issues they present. The beginnings of these originate with Grinberg et al. and their presentation of binocular imaging geometry [15]. This demonstrated clear definition of how factors relating to eye and camera positioning can relate, and its effects on the visible scenes. A detailed review of the epipolar geometry expected for stereo rigs, and the configuration used to obtain it is later published by Zhang [22]. While somewhat dated, it still provides an excellent resource for reference. These contributed led to such advancements as the usage of image rectification to calibrate images for stereo matching [8], [20]. The goal of this research is to restructure the geometry of an image pair to conform to that of a parallel axis setup. The idea is that the image planes are rotated and tilted, so that they may conform to this standard. The resulting images are expected to be vertically parallel, where each scan line, or row of pixels, align perfectly left to right, though not perfectly matching due to the change in perspective. Rectification can also correct errors in calibration, where a parallel setup is expected, but has minor discrepancies between the images. As a result, stereo-matching algorithms can overlook this problem entirely, with rectified images. The expectation is thus that all pixels in a particular row in one image can be matched to pixels in the corresponding row in the other image, provided a match exists.

With the simplification of the problem that rectification provides, new details about the correspondence problem are identified. A publishing by Zitnick et al. [13] combine several elements for their matching method. Returning to the roots of the problem, they propose a cooperative matching algorithm, similar to the original publication by Marr et al.. Their method uses an adaptively scaled window to perform matching, which is applied iteratively to refine the results. Efforts are also made to explicitly identify occluded regions (visible in only one of the pair of images), in order to better preserve depth edges. To add to the original constraints, a third restriction of ordering is added. Here, the assumption is that the order in which objects or pixels visible in one image shall be the same in the other. As this is dependent on visible elements, this constraint is not applied to the occlusion detection element. Although ordering does not always hold true, it is another simplification that can reduce the search range for matches.

Other researchers have explored the special case of transparency within elements [16], allowing multiple disparity values per pixel. This may be ideal when analyzing scenes with clear objects like glass. Transparency can also allow for fuzzy object borders. Some researchers have demonstrated the ability to generate disparity maps with sub-pixel precision, using techniques such as interpolation and triangulation. Although the original goal was to match all pixels within both images, this is not always possible as several factors can prevent matching. Because of the change in perspective between the two views, certain reflective surfaces may have a different tint or 'shine' due to the change in viewing angle. There are also segments that are not visible in both images, known as occluded regions. These are typically found on the contours of the images, as well as along depth borders within the scene (e.g. between objects). Newer algorithms are presented at this time in order to work around these details to achieve more accurate results.

Birchfield et al. [10] were able to take advantage of several of the new developments to propose a new algorithm. As they assumed the images to be rectified, they were able to perform

the comparisons of pixel intensity values on their respective scan lines. This drastically reduced computational costs, in comparison with methods that compare windows or areas. A dynamic programming scheme is used for both occlusion detection, and disparity propagation. As the assumption is that along borders or edges, breaks are to appear in the otherwise continuous map, which coincides with the smoothness constraint originally proposed by Marr et al., these are expected to accompany occluded areas. For the regions that are un-textured, or have little texture based on a horizontal differential of the intensities, the smoothness constraint is maintained, meaning the disparity values are propagated along the scan lines. With this, a few adjustments were made to correct certain errors. One problem is the inconsistency in intensity values between images, due to sampling. The initial method calculates a cost value per potential match, where factors such as the difference in intensity and likelihood of occlusion are weighed in. Once certain constraints are applied to limit the potential matches, the minimal cost is used to choose the correct disparity value. To correct errors when taking the difference in intensity, linear interpolation is used between the pixels in question and their direct neighbours, and the minimum possible match between these is used instead. Another problem found in the results are 'streaks' within the disparity map. These are caused when propagation of values across scan lines do not match their neighbours. To correct this, a form of propagation is done between scan lines, as a post-processing step. This identifies small regions, which are considered to be unreliable (sharp unexpected changes in disparity). Once identified, these are adjusted to match the surrounding values. The result is a disparity map with crisp edges, some occlusion detected, and an algorithm capable of obtaining these data in real-time.

IV. Recent Classification

In 2002, a survey of two-frame stereo correspondence algorithms was published [7], which introduced a new categorization for these techniques. The goal was to classify existing methods based on their design decisions. Certain restrictions were put in place to limit the algorithms being considered. These include: not considering feature-based approaches or sparse methods, and assumes only a pair of rectified images is required. In terms of output, a uni-valued disparity map (single value per point) is expected. With this, the article classifies four components in which the algorithms are comprised. It is noted that, although four steps are illustrated, they are not all required. The first element of consideration is the matching cost. This consists of the process used to determine the likelihood of a particular match between points within an image. For example, the calculation of the absolute difference in intensities between two pixels can be considered the matching cost. There are several different methods that can be applied here, such as squared differences, absolute differences, and even linearly interpolated ones. Typically, the higher the cost, the less likely the match is. From here, the second step can be applied, which is the aggregation (or support) of the costs. This utilizes the cost algorithm from the first step to create a cost array or volume, filled with values representative of the likelihood (or unlikelihood) of a correct match. This step may consist of the summation or averaging over a window-based selection. Once the cost information is generated, the final disparity values can be determined in a third step. There are several ways of accomplishing this. The most common local method is by using a winner-take-all approach, where the lowest (or highest) costing values are chosen. A more global approach is to use an energy-minimization technique (e.g. minimum graph cut), where the best fitting cost is applied overall. Although more costly in terms of computation, this can produce more accurate results. Other techniques, such as dynamic programming and cooperation, can also be used here. The last step, which is optional, consists of post-processing refinement procedures. These can include sub-pixel disparity calculations, smoothening procedures, as well as consistency checks to correct errors.

To test published algorithms, two things were required: a rectified pair of images, as well as ground-truth disparity data. The latter of these was considered to be the 'correct' disparity data of the images, and can be used to measure the accuracy of the results generated. To evaluate the results, root-mean-squared statistics were calculated, and the percentage of bad pixels from this was used. Typically, a disparity difference larger than 1 is considered to be incorrect. From this, several conclusions were drawn. The first of these demonstrates that, for window-based techniques, the accuracy is dependent on the level of texture of the regions found in the image. The less there is texture, the less accurate are the results. This is because, with more texture, it is possible to identify a match with higher probability of success, as there is less ambiguity. In contrast to this, the techniques with the best results are those using graph-cut energy-minimization, though at the cost of a significantly increased runtime. Finally, when looking at the refinement step, sub-pixel precision achieves better results when extending the generated output using half-pixel intervals. With these results in place, the article proposed several things for future work. The first is a framework upon which future publications can extend to provide new results. The second is a collection of images along with ground-truth data, available on the Middlebury website [12], freely accessible to the public. These serve as the basis of comparison for most, if not all, future two-frame correspondence algorithms. The main aspirations of future work, to add to this, is the usage of more complex images, as those present at the time of publication lack overall texture and detail commonly found in real-world images.

V. Advancements

The next decade brings several improved approaches to the matching problem, such as the usage of colour pixel data (i.e. RGB) for comparisons, as opposed to solely using intensity values. With such improvements, a look back at previous methods brings new insights as to where improvements can be made. In 2007, Yang et al. [3] introduced a new approach to the usage of sub-pixel precision, meaning that the generated

depth map has a higher resolution than that of the original images. Although the main focus was a software alternative to available hardware for scene depth calculations, which are limited at the time, the applications to stereovision are significant as well. The article proposes a post-processing framework intended to extend available depth map information, increasing its resolution without loss of precision. This consists of a multi-step iterative process. The first requirement is a depth map as an input (can be the result of stereo matching), along with an image (or image pair) to accompany it. A cost volume is then generated from the map, using current depth information along with a search range, to determine costs. A squared differences cost function is used, as this can preserve information used for sub-pixel refinement later. From here, bilateral filtering is applied to each slice of the volume, using the reference image as a guide. This will smooth the surfaces, while preserving depth edges. If stereo images are used, the cost-volume is created using a symmetric correlation approach, presented in other research. A similar symmetric bilateral filtering is applied to the volume slices. Once complete, a winner-take-all minimization is used to reduce the volume, resulting in a final map. This map can then be processed with a polynomial interpolation scheme to obtain a map with a higher resolution. The steps can then be iteratively processed for further refinement. This method was tested on the results of various algorithms found on the Middlebury database [12] to generate the initial disparity data. In all cases, the level of accuracy for the maps had increased, even for methods that already generated maps at sub-pixel resolution. Some results have demonstrated that refinement 100 times maintained accuracy, though the limits to this do not appear to have been tested.

That same year, Hirschmüller et al. submitted an evaluation of functions used to generate the cost information [14]. As a baseline for comparison, the absolute difference between intensity values is measured. This is then compared to several other techniques, including the sampling insensitive method by Birchfield et al., Laplacian of Gaussian (LoG), mean filters, Mutual Information (MI), and others. These were also tested using three different approaches: a local matching method, one semi-local, and a fully global method using graph cut minimization. To test the robustness of different combinations, images taken from the Middlebury database [12] were analyzed on their own, as well as with some modifications, notably adjustments to exposure settings. The goal was to determine the effectiveness with larger inconsistencies between the images, which commonly occur due to lighting and calibration errors. In many cases, a hierarchical MI approach provided the best results, and is effective with inconsistencies between the images.

In 2008, Hirschmüller [11] proposed a matching technique using statistical mutual Information (MI), following up on the previous publishing. Here, the image data are transformed by using a probability distribution of the intensity values, based on the number of correspondences. Typically, this technique would require an existing disparity map to calculate the correspondences. In previous research [17], it had been demonstrated that starting with a randomized disparity map would be sufficient, and using an iterative approach to repeatedly improve upon the map. This one however proposes a more efficient alternative, using a hierarchical approach. Here, a downscaled version of the image is first used. A randomly generated disparity map is used initially, and processed with three iterations. After this, a lesser downscale is then used as a continuation. This is repeated until the full-scale version is used. With this process, a very small overhead is required to process, and it terminates in a finitely determinate amount of time. Furthermore, since the disparity map is being up-scaled, correspondences from previous iterations can be used to estimate minimum and maximum disparity values, allowing for more restricted search ranges in future iterations. A key feature of this article is that it anticipates the needs of future analysis, describing potential hardware limitations when dealing with larger images, and guidelines on how these may be circumvented, though tailored for this algorithm in particular.

To avoid certain difficulties and uncertainties with the problem at hand, Gao et al. [21] propose an auto-rectification and disparity calculation scheme using customized hardware. Since more information about the input is known, assumptions can be made about the camera's initial alignment, as they are positioned in parallel. To calibrate automatically, a single textured background is displayed to the inputs. Since the visible image is at a certain distance, resulting in an effective disparity of zero (too far to be distinguishably different), it is more simple to perform alignment. The right image is kept as-is, and the left is adjusted to minimize the difference between both inputs. There is no guarantee that they will be perfectly identical after alignment, due minute differences in lighting. Once the calibration is done, a simple disparity calculation procedure is introduced, using specialized processing hardware to accelerate the task. While there are certain limitations to this approach, notably the resolution and quality of the input images, it is possible to process the images in real-time (i.e. video processing), and with high accuracy.

A major challenge in local stereo algorithms is the size of windows used for cost evaluation and filtering. Although some solutions have been proposed, such as automatic window size calculations [13], [18], [19], the accuracy of these is very dependent on the contents of the images. In 2012, Yang proposed a non-local approach [4] to solve this problem. Window-based methods are only capable of utilizing data from a limited number of pixels, in a confined region. As an alternative to this, the usage of a tree structure is proposed. This would allow all pixels to be correctly weighed in when processing an individual point. To generate this tree, all the pixels are considered as the nodes, and the edges are relative to the similarity between pixels and the distance that separates them. This effectively creates a complete graph from the image. Edges with the highest distances are then removed, converting the graph into a minimally spanning tree. To avoid the need to re-balance the tree structure, a double processing scheme is proposed. Here, the leaf nodes are first processed towards the roots, or parents. Then the reverse is done to

propagate the data outwards. The algorithm has low cost, as the tree is only generated once, and the processing of each node requires few steps. Because of the elimination of high distance edges, areas with little texture are the most effective for processing, though at the cost of less accuracy for regions with more texture. To correct this, a refinement step is proposed. Here, a consistency check is performed, and pixels that do not correctly correspond between the maps are identified as unstable. The depth information is then re-propagated to these regions. This can correct errors both for high-texture regions, as well as occluded ones. Experimental results suggest that this approach is more accurate than existing local methods.

In 2013, Pham et al. [1] proposed a new technique to improve the performance of local stereo algorithms, without the need to sacrifice quality in the output. This takes the 2D image data and performs a domain transformation, which converts the 5D information (space and colour channels) so that it may be processed using 1D filters, while geodesic information is retained. Performance improvements in terms of processing times come from the ability to process each row and column of the image individually, which can be done in parallel. Filtering is thus performed on both the rows, then the columns of the image as a result. To calculate the cost, a hybrid of common Truncated Absolute Differences (TAD) of the pixels, and the Hierarchical Mutual Information (HMI) techniques is used, where the TAD and an image gradient are combined to generate the cost. A Winner-take-all optimization is performed to generate the map. As with more recent methods, invalid pixels are identified with a consistency check, and these are filled with the minimum disparity found nearby. Experimental results suggest that this algorithm is not only a top competitor for local algorithms, but it is also effective in the presence of noise within the images.

In 2014, Sun et al. proposed a new framework for disparity propagation that is edge-aware [9]. This takes advantage of some former research to create an algorithm that is both accurate and efficient. One difference in particular, over the previously described tree structure from 2012, is that this technique processes each scan line as a separate tree, similarly to the article in 2013 [1]. This allows significant optimization, as processing can take place on a GPU. The first component of this technique is pre-processing. Here the images are analyzed, and an initial cost-volume is generated using techniques similar to previous research. A box filter (non specified) is applied to the results to suppress noise, at each disparity level of the cost volume. Once completed for each image, a disparity map is generated using winner-take-all. A left-right consistency check is then performed to categorize the pixels into stable and unstable. Pixels are only considered stable if their disparities correspond in each map. Once categorized, a new cost volume is generated. This is based on the prior assigned disparity, along with a penalty for diverging from the most probable matches (3 matches in this case). If a pixel is considered unstable, its cost is assigned to zero. Different filtering options are described to propagate the disparities to unstable pixels, while retaining edges. One option listed is the aforementioned bilateral filter. Despite options like these, a new geodesic filter is proposed not only for simplicity and effectiveness, but also its ability to process each scan line independently. The filter performs aggregation in two passes. The first takes the current cost of the current pixel and disparity, and adds the cost of the previous pixel, relative to its edge cost (absolute difference between RGB values). Once passed from left to right, a second pass in the opposite direction is performed, using a similar summation, though with modification to remove directional bias. If desired, the filter may also be applied to the image columns in a similar way. In terms of results, this method outperforms other local methods in both speed and accuracy, averaging at around 9ms runtime for an image pair, with 5% error. As this method does not directly determine sub-pixel disparities, a refinement process can be used, such as quadratic polynomial interpolation, and this maintains a relatively high accuracy in comparison with other sub-pixel methods.

VI. High-Resolution Datasets

In 2014, a collection of high-resolution stereo datasets [2] were published to the Middlebury website [12], as a continuation of the ongoing initiative to bring more complex images for public testing. The technique applied allowed the capture of high-resolution stereo images, along with precise ground-truth depth information for these. The goal is to use these for testing purposes, as they are considered to be minimally flawed. Several techniques were combined to obtain the published results. Firstly, several layers of structured lighting were used. Various sinusoidal patterns of light were projected onto the scene, with varying frequency and ranges of colour. A set of over one hundred images was produced for analysis per scene. This was combined with advanced self-calibration techniques. The rectification process used during analysis not only considered the tilt of the images, but provided corrections to calibrate the projector for the lighting, as well as adjustments for the lenses used, to reduce the effects of warping. The results consist of realistic rectified images, accompanied by ground-truth data with sub-pixel precision. The only limitation presented by this method is the lack of data for occluded areas. This is considered to be a minimal flaw, as stereo-matching alone cannot generate the data for these regions with complete accuracy.

Using the new datasets available on the Middlebury website, Žbontar et al. have created a new matching method [24] that is currently one of the top contenders on its leader board [23]. This technique works by training a convolutional neural network. Unlike many of the methods previously described, this one takes existing datasets along with the ground truth data to identify certain patterns amongst stereo image pairs, in order to estimate attributes best suited for matching. Instead of training the network using the entire images all at once, the goal is to find similarities for small segments, or patches of the images. A binary categorization and classification is done to distinguish similar and dissimilar patches, as part of the training. A patch is segmented by using a cross-based aggregation. “Arms” are extended from a

particular pixel where the difference in intensity is low, in the horizontal and vertical directions. The patch is created by combining these cross segments where vertical "arms" overlap, thus creating the support region. To compare patches, the neural network uses convolution methods based on absolute differences of intensity values. Since the training process can be time consuming, two different approaches are taken into consideration: a fast method and an accurate method. While the raw outputs from the neural network alone are not sufficient to generate a disparity map, as they are not sufficiently accurate particularly in occluded and low-texture regions, they make an excellent basis to initialize the cost values. Once the support regions are used to generate costs, and these are averaged between both images, some optimizations are performed. In order to smoothen the final map, some semi-global methods are applied. Penalties are added to the cost for differences between them and their neighbours. A difference of 1 adds a certain penalty, and differences larger than this add a significantly larger penalty. The penalties are measured in 4 directions, both positive and negative horizontal and vertical. Once this is completed, a winner-take-all minimization is applied to obtain the map. After this, some more refinements are done to correct certain errors. Firstly, a consistency check is performed to identify incorrect matches and occluded regions. Since occluded regions are not always correct as a result of the matching process, some interpolation is performed to fill these places in. After this, a median filter is used to fill in the remaining missing/incorrect values. Next, some quadratic curve fitting is done to perform sub-pixel enhancement. Finally, a median filter followed by a bilateral filter is done to smoothen the image and reduce noise. To test the algorithm, both Middlebury datasets [12] and KITTI datasets [25] are used. For the KITTI set, this approach is a top ranking method, when using the accurate technique. It is also currently the second highest method on the Middlebury leader board. In terms of performance, this method requires (on average) approximately 60 seconds for the accurate method for the KITTI database, and approximately 150 seconds for the Middlebury dataset. The fast method, while less accurate, averages up to 90 times faster. It should also be noted that, for the Middlebury set, half-sized images were used with the results upscaled, due to hardware limitations. While this approach is more costly in terms of execution, the results are highly accurate. It has an average of approximately 4% error on the KITTI set, and 8% for Middlebury. The higher error rate for the latter can be attributed to higher precision of the images and ground truth data, along with loss of quality for the disparity map upscale.

VII. Conclusion

While local stereo techniques have made great strides, particularly over the past decade, there is still a room for improvement. With the recent addition of high-resolution images available to the public, there are more challenges to be faced, particularly terms of performance and accuracy. Currently, the most recent attempts made using these new images average above 7% error, which is higher than attempts with previous datasets (i.e. 2% error). As these newer images can be considered a more accurate portrayal of modern photographs, they are significantly larger and more detailed than older available images, along with more detailed ground truth data. The increased amount of details, not to mention the overall data, push modern local techniques to new limits, though this may expose unknown flaws in the process. While currently, certain paradigms seem to be more prevalent in local matching techniques, such as using WTA minimization, consistency checking, parallel processing optimizations, and data transformation, it's likely some of these may be improved, or even replaced. Despite the research into this problem approaching four decades, one can expect there will be even more strides towards better, faster algorithms in the near future.

References

[1] C. C. Pham, J. W. Jeon, "Domain transformation-based efficient cost aggregation for local stereo matching." Circuits and Systems for Video Technology, IEEE Transactions on 23.7 (2013): 1119-1130.

[2] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth." Pattern Recognition (2014): 31-42.

[3] Q. Yang, R. Yang, J. Davis, D. Nister, "Spatial-depth super resolution for range images." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

[4] Q. Yang, "A non-local cost aggregation method for stereo matching." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[5] U. R. Dhond, and J. K. Aggarwal, "Structure from stereo-a review." IEEE transactions on systems, man, and cybernetics 19.6 (1989): 1489-1510.

[6] D. Marr, and T. Poggio. "A computational theory of human stereo vision." Proceedings of the Royal Society of London B: Biological Sciences 204.1156 (1979): 301-328.

[7] D Scharstein, R Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms." International journal of computer vision 47.1-3 (2002): 7-42.

[8] A. Fusiello, E. Trucco, A. Verri, "Rectification with unconstrained stereo geometry." BMVC. 1997.

[9] X. Sun, X. Mei, S. Jiao, M. Zhou, Z. Liu, H. Wang, "Real-time local stereo via edge-aware disparity propagation." Pattern Recognition Letters 49 (2014): 201-206.

[10] S. Birchfield, C. Tomasi. "Depth discontinuities by pixel-to-pixel stereo." International Journal of Computer Vision 35.3 (1999): 269-293.

[11] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information." Pattern Analysis and Machine Intelligence, IEEE Transactions on 30.2 (2008): 328-341.

[12] http://vision.middlebury.edu/stereo/data

[13] Zitnick, C. Lawrence, and Takeo Kanade. "A cooperative algorithm for stereo matching and occlusion detection." Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.7 (2000): 675-684.

[14] Hirschmüller, Heiko, and Daniel Scharstein. "Evaluation of cost functions for stereo matching." Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

[15] Grinberg, Victor S., Gregg W. Podnar, and Mel Siegel. "Geometry of binocular imaging." IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology. International Society for Optics and Photonics, 1994.

[16] Szeliski, Richard, and Polina Golland. "Stereo matching with transparency and matting." Computer Vision, 1998. Sixth International Conference on. IEEE, 1998.

[17] Kim, Junhwan, Vladimir Kolmogorov, and Ramin Zabih. "Visual correspondence using energy minimization and mutual information." Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003.

[18] Kanade, Takeo, and Masatoshi Okutomi. "A stereo matching algorithm with an adaptive window: Theory and experiment." Robotics and

Automation, 1991. Proceedings., 1991 IEEE International Conference on. IEEE, 1991.
[19] Veksler, Olga. "Fast variable window for stereo correspondence using integral images." Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Vol. 1. IEEE, 2003.
[20] Fusiello, Andrea, Emanuele Trucco, and Alessandro Verri. "A compact algorithm for rectification of stereo pairs." Machine Vision and Applications 12.1 (2000): 16-22.
[21] Gao, Xinting, Richard Kleihorst, and Ben Schueler. "Implementation of auto-rectification and depth estimation of stereo video in a real-time smart camera system." Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on. IEEE, 2008.
[22] Zhang, Zhengyou. "Determining the epipolar geometry and its uncertainty: A review." International journal of computer vision 27.2 (1998): 161-195.
[23] http://vision.middlebury.edu/stereo/eval, Middlebury Stereo Evaluation, Version 2, 2015.
http://vision.middlebury.edu/stereo/eval3, Middlebury Stereo Evaluation, Version 3, 2015.
[24] Žbontar, Jure, and Yann LeCun. "Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches." arXiv preprint arXiv:1510.05970 (2015).
[25] http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo, The KITTI Vision Benchmark Suite, Stereo Evaluation, 2015.