21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Towards a Simulation Driven Stereo Vision System

Martin Peris
Cyberdyne Inc., Japan
Email: martin_peris@cyberdyne.jp

Atsuto Maki
Toshiba Research Europe Ltd., UK
Email: atsuto.maki@crl.toshiba.co.uk

Sara Martull
University of Tsukuba, Japan
Email: info@martull.com

Yasuhiro Ohkawa, Kazuhiro Fukui
University of Tsukuba, Japan
Email: {ohkawa,kfukui}@cvlab.cs.tsukuba.ac.jp

Abstract

This paper presents a novel algorithm for estimating stereo disparity which exploits the benefit of learning to the fullest. Given a cost volume of stereo matching, we solve the cost aggregation and disparity computation in one shot by using a classifier; we design a feature called matching cost pattern for the input which we extract from the cost volume while we use simulated stereo patterns for training. To this end, we introduce a highly realistic computer graphics dataset, the new Tsukuba stereo dataset, with ground-truth disparity maps. Through preliminary experiments we show that our algorithm outperforms a simplified AD-Census cost-minimization method, and also that the error ratio decreases as we use a larger number of samples for training.

1. Introduction

Stereo matching is one of the most widely studied research topics in computer vision. A number of new algorithms have been presented every year, a large set of which fall in the taxonomy of Scharstein and Szeliski [11], performing some subset of the following four building blocks: matching cost computation, cost aggregation, disparity computation, and disparity refinement. During cost computation step, each pixel in the reference image is compared against a potentially corresponding pixel in its stereo pair considering a certain disparity. The similarity between these pixels is regarded as a matching cost. In cost aggregation step, the cost is aggregated using either a window of constant size or adaptive size [14]. Then, the disparity computation is performed by selecting the minimal aggregated value, or considering it as an energy minimization problem [13]. Finally, the often noisy disparity map is refined by using post-processing methods such as hole filling [1]. See [11, 13, 2] for some evaluations of different algorithms in terms of their accuracy and computational efficiency.

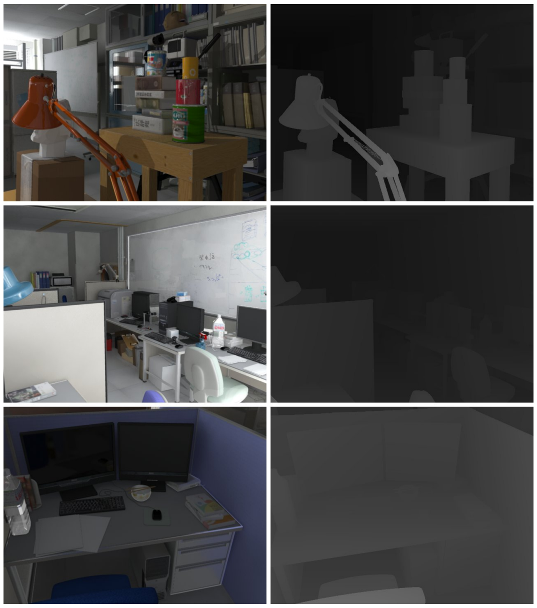

Figure 1: The new Tsukuba CG Stereo Dataset. A few example frames shown with ground-truth disparity maps among four sequences of 1800 stereo image pairs.

In this paper, we focus our attention on learning-based approach to stereo disparity computation, to which relatively little work have been reported despite the generality of learning methods in other tasks of computer vision such as recognition and segmentation; existing stereo methods with few exceptions [10, 5, 4] do not exploit the hidden information that can be learned from ground-truth. The main reason for this is in the deficiency of sufficient training data with ground-truth labels which would be required for a reliable disparity computation. In recent years, however, CG techniques have played inspiring roles for synthesizing training data to fill the void in a few domains of computer vision [12, 9], and a tremendous effect was demonstrated in human pose recognition [12].

In order to overcome the above issue in learning-based stereo, we also introduce a highly realistic CG stereo dataset, the new Tsukuba stereo dataset, containing a large amount of stereo pairs with highly accurate ground-truth disparity maps as a source for learning stereo patterns. Figure 1 shows some of the frames in the dataset with the corresponding ground-truth disparity maps.

The concept of learning can be introduced in different steps of stereo matching [10, 5, 4]. The probabilities of matching errors are learned in [4], and in a global optimization approach [10] free parameters in conditional random field are learned. Armed with a rich CG dataset for training, in our case, we propose to deal with the steps of cost aggregation and disparity computation altogether by learning. For the initial step of computing matching cost, we utilize the state-of-the-art measure of AD-Census [7]. For the final classification of disparity, we choose to use Multiclass Linear Discriminant Analysis (Multiclass LDA)[3]. To the best of our knowledge, this is the first attempt to infer a disparity map based purely on learning with CG data. The goal of the paper is then to study the performance of such a simulation-driven approach using the newly created data set; it can for example learn the peculiarities of stereo patterns in low textured areas and infer an improved disparity value.

To sum, the contributions of this paper are (i) the new learning-based approach to compute stereo disparity, (ii) the new Tsukuba stereo dataset, and also (iii) a method to generate a confidence map for the computed disparities in the framework of classification. We show the benefits of our approach by applying a real stereo image pair to the trained classifier.

2 Simplified AD-Census

In this section we briefly describe the initial matching cost defined by AD-Census [7] which we use for generating the input to our disparity classification.

AD-Census cost initialization. AD-Census combines the Absolute Difference (AD) cost measure and a non-parametric transform called Census [15] which captures the local structure of the image using the relative ordering of the pixel intensities rather than the intensity values themselves.

Following [7], we define at each pixel (u, v) and disparity level d the two measures $C_{census}(u, v; d)$ and $C_{AD}(u, v; d)$. For C_{Census} we use a 9×7 window to encode the local structure in the region centered at each pixel in a 64-bit string, and therewith $C_{census}(u, v; d)$ is defined as the Hamming distance of the two bit strings that stand for pixel (u, v) and its correspondence $(u - d, y)$ in the other image [15]. The second cost value, C_{AD} is defined as the absolute color intensity difference between the pixel (u, v) in the left image and its correspondent $(u - d, y)$ in the right image.

The total AD-Census cost $C(u, v; d)$ is then computed as:

$$C(u, v; d) = f(C_{census}(u, v; d), \alpha_{census}) + f(C_{AD}(u, v; d), \alpha_{AD}), \tag{1}$$

where $f(c, \alpha)$ is a function on variable c defined as $f(c, \alpha) = 1 - exp(-c/\alpha)$ with two goals, first to map the value of the different cost measures to the range $[0, 1]$ and second to control the influence of each cost measure with the parameter α. We used the same α_{census} and α_{AD} as in [7] to calculate the initial cost volume, $\mathcal{C}$, containing $C(u, v; d)$ for all (u, v) and d. For further details on AD-Census cost initialization refer to [7].

Disparity with cost minimization. Once the initial cost volume, $\mathcal{C}$, has been calculated, that information can be used to estimate the disparity map. A simple version of the AD-Census matching is to use a fixed aggregation window of 9×7 pixels around (u, v) to sum the cost of neighboring pixels, instead of an adaptive aggregation window. The cost aggregation C_{Agg} can be performed on each disparity plane, with D being the maximum number of disparities. One standard way to obtain a disparity map, Z, is then to compute

$$Z(u, v) = \operatorname*{argmin}_{d} C_{Agg}(u, v; d), \tag{2}$$

and we will use it for a comparison to our method.

3 Learning disparity

After cost initialization, stereo matching algorithms usually proceeds with three more steps: cost aggregation, disparity computation, and refinement[1]. In this section we introduce our algorithm to compute the stereo disparity using Multiclass LDA, and a method for generating a confidence map for the computed stereo disparity values.

[1] The step could independently be applied to the output disparity map and therefore beyond the scope of this paper.

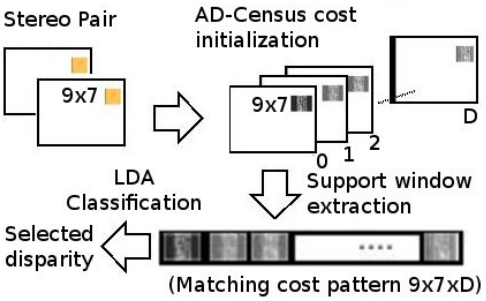
Stereo Pair
AD-Census cost initialization
9x7
9x7
0
1
2
D
LDA Classification
Support window extraction
Selected disparity
....
(Matching cost pattern 9x7xD)

Figure 2: Our new algorithm for computing disparity in a classification framework.

Disparity with Multiclass LDA. Rather than aggregating the cost in the window around a certain pixel, we propose to directly use the costs in the aggregation window of each disparity plane by concatenating them across the depth direction into the features called *matching cost pattern*. See Figure 2. That is, each matching cost pattern will have a size of $9 \times 7 \times D$. Then, those features are used as inputs to a Multiclass LDA classifier that has been previously trained using learning data. Before applying Multiclass LDA, PCA is applied to the matching cost pattern to overcome a singularity, the PCA dimension was set to 33 which has 99.9% cumulative contribution ratio. Hence, we solve a classification problem instead of a minimization problem to compute disparity, i.e. given the input matching cost pattern, $\mathbf{x}(u, v)$, we compute the disparity by

$$Z(u, v) = \operatorname*{argmin}_{d} L(\mathbf{x}(u, v), \mu_d), \tag{3}$$

where $L(\mathbf{x}(u, v), \mu_d)$ is the distance between $\mathbf{x}(u, v)$ and μ_d, the mean vector corresponding to the class of disparity d in discriminant space.

Confidence Map. We also compute a confidence map in order to assess the reliability of the computed disparity. Near depth discontinuities, the window used to extract the feature may well contain information of two disparity values simultaneously, which makes the classification of disparity more difficult. However, the matching cost patterns corresponding to consecutive disparity classes are similar, so if the disparity class with highest score (shortest distance) is d_1 for a non-occluded pattern, then it is very likely that the disparity class with second highest score, d_2, is either of $d_1 \pm 1$. We can use this insight to establish a threshold t and regard each selected disparity value as reliable or unreliable following this rule:

$$\begin{array}{cc} \mathit{reliable} & \text{if } d_2 \in [d_1 - t, d_1 + t] \\ \mathit{unreliable} & \text{otherwise.} \end{array} \tag{4}$$

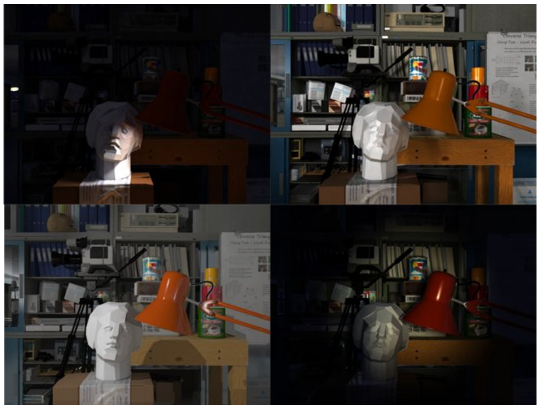

Figure 3: Different illumination conditions of the new CG dataset. Upper-left: Lamps. Upper-right: Fluorescent. Lower-left: Daylight. Lower-right: Flashlight.

4 New Tsukuba CG Stereo Dataset

Learning-based approach to stereo requires a suitable amount of training data although the available number of stereo scenes with ground-truth disparity maps has been quite limited [11]. For this reason we created a highly realistic CG stereo dataset, which will shortly be made available on **http://cvlab.cs.tsukuba.ac.jp**[6]. This dataset was created after the original *head and lamp* scene released by University of Tsukuba [8] and has the following properties.

- 4 Different illumination conditions (see Figure 3).
- 1800 full-color stereo pairs per illumination condition with ground truth disparity maps (1 minute video at 30FPS using an animated stereo camera).
- 256 levels of disparity.
- Non-occluded area mask, near depth-discontinuity mask and 3D camera position and orientation on each frame (see Figure 4).

The objects on the scene have been modeled and photo-realistically rendered using the software Pixologic ZBrush and Autodesk Maya. Each illumination condition offers specific challenges that will be of great use for the improvement of new stereo and tracking algorithms:

- *Fluorescent*: This is considered the default illumination, it features an even illumination on all surfaces.
- *Daylight*: Smooth illumination to the objects in the scene with exception of the areas near the window,

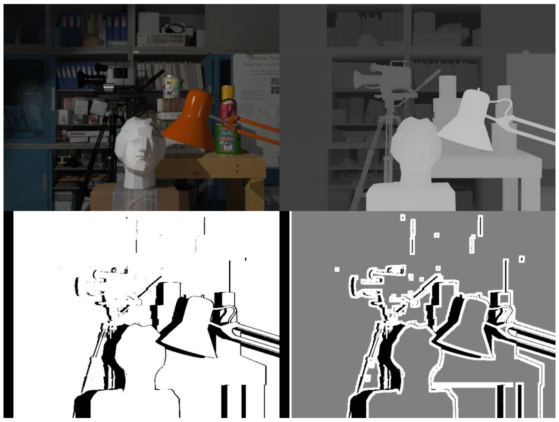

Figure 4: First frame of the new dataset (left camera view). Upper-left: RGB image. Upper-right: disparity map. Lower-left: non-occlusion mask. Lower-right: near depth discontinuity mask.

that appear over exposed due to the intensity of the sun light.

- *Flashlight*: This scene has been rendered only lit by the light of a flashlight attached to the moving Stereo Camera.
- *Lamps*: This is the most challenging of all four illumination conditions, with low and uneven lighting.

The ground truth data was obtained as follows:

- *Disparity map*. Derived from the depth map generated by the render engine by computing

$$d = fT/Z, \quad (5)$$

  where d is the disparity of the pixel, f is the focal length of the camera, T is the baseline distance of the stereo rig and Z is the depth value of the pixel.
- *Non-occlusion mask*. Has been generated by simply cross-checking the left and right disparity maps.
- *Near depth discontinuity mask*. Obtained by detecting the boundaries of the non-occlusion mask.
- *Camera position and orientation*. Extracted by executing a MEL (Maya Embedded Language) script on each frame.

The aim of this dataset is to provide ground truth data that can be useful to evaluate the performance of computer vision algorithms, especially for stereo vision but also for camera tracking and 3D reconstruction algorithms from monocular time-sequential images.

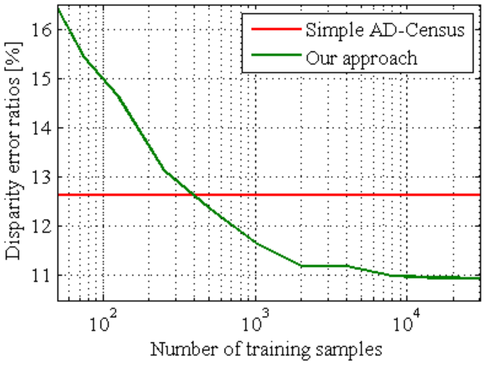

Figure 6: Disparity error ratios [%] of our learning-based approach plotted for varying number of training samples. Larger number of samples reduces misclassification. Also plotted are the error ratio by simplified AD-Census.

5 Experiment

We use the new CG dataset to train our disparity classifier and test it to compute the disparity map of real images. We illustrate the result of using the original *head and lamp* scene [8] as the input. The number of disparities in this scene is 16 so we will consider the same number of classes in our classifier. To be fair and avoid visual similarities between the learning data (CG) and the test data (real images) we discarded the first 261 frames of the CG dataset which contains the *head and lamp* scene. The rest of the dataset was used to uniformly extract $80,000$ samples per disparity value with which we trained the Multiclass LDA classifier.

Figure 5 shows the disparity map generated for the original *head and lamp* scene using the simplified AD-Census with cost-minimization method and our Multiclass LDA-based method. It can be observed that using our method the disparity map generated is smoother than that of simplified AD-Census. In particular, our learning-based approach seems to be more effective in low textured areas (marked with a green square in Figure 5), as our method can extract more information from the cost aggregation window than cost minimization-based methods.

Table 1 shows the results of applying our method and the cost-minimization to the original real image. Our method outperforms the cost-minimization based method in the overall count of pixels (all) and in non-occluded pixels (non-occ). As anticipated, the performance at near depth discontinuity areas (disc) is not as desirable, but we can disregard the resulting disparities in such regions by referring the confidence map; Fig-

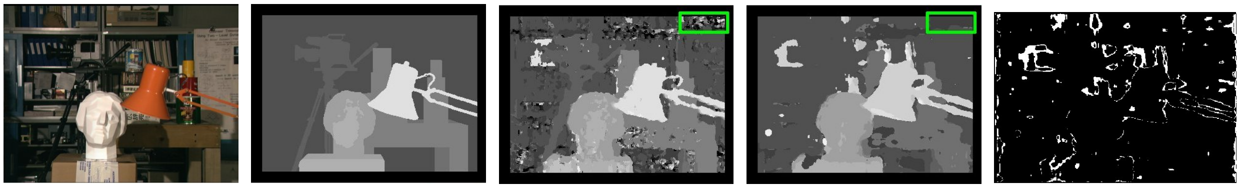

Figure 5: From left to right: Left image from the original *head and lamp* dataset, ground-truth disparity, disparity map with cost-minimization method, disparity map with learning-based approach, and the confidence map.

Method	non-occ.	all	disc.
Simplified AD-Census	10.70%	12.63%	**15.35%**
Our method	**9.16%**	**10.91%**	21.91%

Table 1: The error ratios in all and sub-regions.

ure 5 also shows an example generated by considering a threshold of $t = 3$ in equation 4.

Finally, Figure 6 shows the effect on the error ratios of computed disparity due to varying the number of training samples. The ratio of incorrect matches (an average after performing 100 evaluations) is plotted for each case. The larger the number of samples is, the lower the mis-classification is. It clearly shows the effect of using a large number of training data, supporting our basic idea of learning-based stereo matching.

6 Conclusions

We have introduced a novel learning-based approach to stereo disparity computation as well as the new Tsukuba CG stereo dataset. We have described a feature called *matching cost pattern* which we extract from AD-Census cost volume and classify by Multiclass LDA; we train the classifier exclusively with the synthetic data. To the best of our knowledge, this is the first attempt to infer a disparity map purely based on learning with CG data. In our early experiment with real input data we observed superior performance to a simplified AD-Census in terms of the error ratio. Further evaluations will be made for more variations of input scenes in the future. Finally, the new stereo dataset will be also useful to evaluate any stereo matching algorithms.

7 Acknowledgements

This work was supported by KAKENHI 23650081.

References

[1] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE PAMI*, 30(2), 2008.

[2] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE PAMI*, 31(9), 2009.

[3] A. J. Izenman. *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York, 2008.

[4] D. Kong and H. Tao. A method for learning matching errors for stereo computation. In *BMVC*, 2004.

[5] Y. Li and D. P. Huttenlocher. Learning for stereo vision using the structured support vector machine. *In CVPR*, 0, 2008.

[6] S. Martull, M. Peris, and K. Fukui. Realistic CG stereo image dataset with ground truth disparity maps. In *TrakMark*, 2012.

[7] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. *ICCV Workshops on GPU in Computer Vision Applications*, (9), 2011.

[8] Y. Nakamura, T. Matsuura, K. Satoh, and Y. Ohta. Occlusion detectable stereo – occlusion patterns in camera matrix. In *CVPR*, 1996.

[9] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR*, 2011.

[10] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *CVPR*, 2007.

[11] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3), 2002.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[13] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, A. Agarwala, and C. Rother. A comparative study of energy minimization methods for markov random fields. In *ECCV*, 2006.

[14] F. Tombari, S. Mattoccia, and L. di Stefano. Full-search-equivalent pattern matching with incremental dissimilarity approximations. *IEEE PAMI*, 31(1), 2009.

[15] R. Zabih and J. W. Ll. Non-parametric local transforms for computing visual correspondence. In *ECCV*, 1994.