CORNER VIEW DISPARITY ESTIMATION FOR LOSSLESS LIGHT FIELD COMPRESSION

Emre Can Kaya, Ioan Tabus

Tampere University
Computing Sciences Unit
Tampere, Finland

ABSTRACT

In this paper we study disparity estimation for high density camera arrays using convolution neural networks (CNN) with the goal of achieving an overall description of all disparity images needed in the warping process for light field compression. Furthermore, we present a lossless compression scheme that makes use of the obtained disparity estimates. The scheme provides random access to individual views, as opposed to most published lossless light field compression, which require the simultaneous decoding of all views. Following similar steps as in EPINET, which is a recently published CNN based estimator of the disparity at the center view, we design a CNN for estimating the disparity of the views in corner positions. EPINET uses several data augmentation techniques through rotation and flipping of light field and we study similar transformations of the light field that provide data augmentation when estimating corner views disparities. The performance of the estimated disparities is evaluated first in terms of the traditional mean square error and percentage of pixels above a certain threshold. Additionally, we validate the quality of the disparity estimates in terms of their successful usage for the warping stage in a lossless color view compression scheme. The lossless compression achieved by the estimated disparities is shown to be close to the lossless compression achieved when using ground truth disparities. Codes are available at *https://github.com/marmus12/CornerView*.

I. INTRODUCTION

Disparity estimation is an active area of research that has several applications in computer vision. Disparity maps contain the essential information about how the consecutive views in a certain angular direction of the camera array are related, making them very useful for light field compression.

Recently, CNN based methods proved to be successful in several light field processing tasks [1], [2]. EPINET [3] is a CNN model for estimating a disparity map of a light field image. The disparity map generated by EPINET corresponds to the center view which is the disparity between the center view and its closest right horizontal neighbour view. In the following, we use the term corner view disparity to refer to the disparity between the corner view and its closest right horizontal neighbor, or equivalently the negative disparities to the left neighbor. EPINET has a multi-stream architecture which efficiently combines the information coming from different view stacks. In the network, several stacks of views are considered, extracting information from the epipolar plane images at different angular directions in the camera array. The four input view stacks are intersecting at the center view. For estimating the center view disparity with EPINET, it is needed that input views from around the center view are available. Here, we present a different scheme, in which the disparities can be estimated for the most extreme locations in a camera array, namely for the corner views. It should be noted that once the corner and center view estimates are available, reliable estimates can be obtained for any desired view, since reliable information exists for displacements on various angular directions.

II. CORNER VIEW DISPARITY ESTIMATION

The proposed CNN architecture for estimating a corner view disparity map is depicted on Fig. 1. We call this architecture CEPINET, short for Corner EPINET. This architecture is derived from the original EPINET architecture by replacing the 45° and 135° diagonal subnetworks with a single diagonal subnetwork, since in CEPINET there is a single diagonal stack. Moreover, the number of filters in the merge network is reduced to 210 from 280 to remain consistent with the number of input stacks. The CEPINET network encompasses 2.96 million trainable parameters, while the number for EPINET is 5.12 millions.

In EPINET, data augmentation of the light field through rotation is performed by rotating the light field image around the center view. Therefore, the center view remains at the center after rotation. Hence, the resulting light field is still a valid input for the network that estimates the center disparity. In CEPINET, the situation is different in the sense that whenever the light field is rotated about the center of view array, the position of every corner changes. In order to train a single network, which by proper preprocessing can estimate any of the four corners, we introduce rules for forming the input stacks, with the convention that the disparity in the upper left corner has to be estimated. According to this convention, when estimating the other three corners, the input light field is rotated by multiples of 90 degrees to bring the corner view that is being estimated to the upper left position. On the other hand, we also apply transposition of the light field as an augmentation which also yields a valid light field. Upper left corner view remains in the same place after transposing the light field, thus not violating our convention. In summary, eight different training samples are obtained by rotating and transposing a light field image: two for each corner (original and transpose). The other augmentation strategies such as color scaling are kept the same as in EPINET.

III. LOSSLESS COMPRESSION USING CORNER AND CENTER VIEW DISPARITIES

Once the disparity maps for the corner views and the center view are obtained, any target view can be reconstructed with small error by warping reference images and combining the warped images. The proposed lossless light field compressor, dubbed here LLFC, is depicted in Fig. 2. The scheme is close in spirit to the disparity based and region based plenoptic image compression from [4], but has a more refined combining of warping from several references, based on a region based best performance switch, as described next. Unlike the previous work in [4], LLFC makes use of the corner view disparity estimations in addition to the center view for predicting one general position view.

For each target view, a disparity image is generated by warping the closest reference's disparity to the target location. The target disparity is quantized and divided into connected regions. For each region, one marks in an image, called *best reference labels image*, the index of that warped image which yields the smallest MSE over the region. The best reference labels image is constructed for each target and it is used in conjunction with the warped references to predict the target view. Best reference labels image has all elements as integers 1 to 5 (indexing the winner, out of 4 corners and center). First the side information is encoded: 5 reference disparity maps, 5 reference color images, and 1 best reference labels image for each target. Furthermore, since we want to perform lossless compression, residual images for each target are also transmitted. Best reference labels images are very sparse allowing them to be compressed efficiently. We compress reference color, disparity images and residual images with lossless JPEG 2000 [5] prior to transmission. The best reference labels image can be encoded either with JPEG 2000 or CERV [6]. We present results for both cases.

IV. EXPERIMENTAL WORK

In this section, we present experimental results for corner view disparity estimation and lossless compression. Our training and validation set consists of 13 samples from the HCI Benchmark [7]. Three samples that contain reflective surfaces (*vinyl*, *kitchen* and *museum*) are chosen as test samples. Since reflective surface disparities are hard to estimate, MSEs for these samples are much higher than the validation samples in Tables I-II.

We train 2 CEPINET and 2 corresponding baseline EPINET models, each time leaving out 1 sample (first *greek*, then *town*) for validation so that the training set consists of 12 samples. From each light field image, 8 different training samples are obtained by rotating the light field with multiples of 90 degrees and taking the transpose. These 8 different cases are randomly sampled during training. One training batch consists of 48 randomly chosen multi-scale patches with size 25x25. Learning rate is initially set as 10^{-4} and it is dropped by a factor of 0.5 whenever training loss reaches a plateau. We present MSE and Bad Pixel Ratio results for 5 different views, 4 being corners and 1 being center on Tables I-IV. Qualitative results obtained with 2 EPINET and 2 CEPINET models with the corresponding validation samples are presented on Figure 3.

Table I. MSE*10^3 with Town as Validation

View	Train Mean	Vinyl	Kitchen	Museum	Town
NW	**7.17**	97.13	141.73	65.34	3.67
NE	7.55	**79.11**	**138.28**	**60.66**	**2.95**
SW	7.69	124.50	157.88	83.01	6.03
SE	9.21	156.59	153.87	107.58	5.14
Center	15.13	112.09	164.57	116.68	3.30

Table II. MSE*10^3 with Greek as Validation

View	Train Mean	Vinyl	Kitchen	Museum	Greek
NW	22.74	114.53	148.17	66.88	**95.12**
NE	20.65	**90.33**	**138.69**	82.28	206.86
SW	21.43	115.13	175.99	**52.73**	250.77
SE	23.79	126.98	152.06	65.29	272.86
Center	**12.00**	117.70	154.84	76.05	118.61

We report the compressed size, as total compressed file size over the total number of pixels $(9 \times 9 \times 512 \times 512)$ of the light field, expressed in bits per pixel (bpp), for 16 test samples out of the training set averaged over all target views with LLFC and JPEG2000 are presented on Tables V-VI. According to Tables V-VI LLFC compression method yields superior results to JPEG2000 for all samples. Using CERV [6] for compressing best reference labels, instead of JPEG2000, yields slightly better results as evident on Tables V-VI. CERV provides a specialized framework for encoding constant value regions in an image, therefore it is expected to yield better results when compared to JPEG 2000 which is a generic image compression scheme. On the other hand, JPEG 2000 has the advantage of providing a less complicated encoder and decoder architecture while not sacrificing a lot from accuracy.

The compressed size obtained with ground truth (GT) corner and center disparities are also presented as an ideal reference on Table V. GT disparity results provide an

Table III. Bad Pixel Ratio with Town as Validation (Threshold =0.07)

View	Train Mean	Vinyl	Kitchen	Museum	Town
NW	**0.037**	**0.229**	0.251	0.113	0.049
NE	**0.037**	0.237	**0.238**	**0.107**	**0.043**
SW	0.043	0.254	0.246	0.130	0.047
SE	0.043	0.230	0.244	0.132	0.052
Center	0.043	0.260	0.259	0.132	0.047

Table IV. Bad Pixel Ratio with Greek as Validation (Threshold = 0.07)

View	Train Mean	Vinyl	Kitchen	Museum	Greek
NW	0.108	0.272	0.279	0.133	0.309
NE	0.108	0.267	0.278	**0.125**	0.295
SW	0.103	0.294	0.275	0.139	0.313
SE	0.109	0.273	0.271	0.143	0.337
Center	**0.075**	**0.245**	**0.253**	0.132	**0.279**

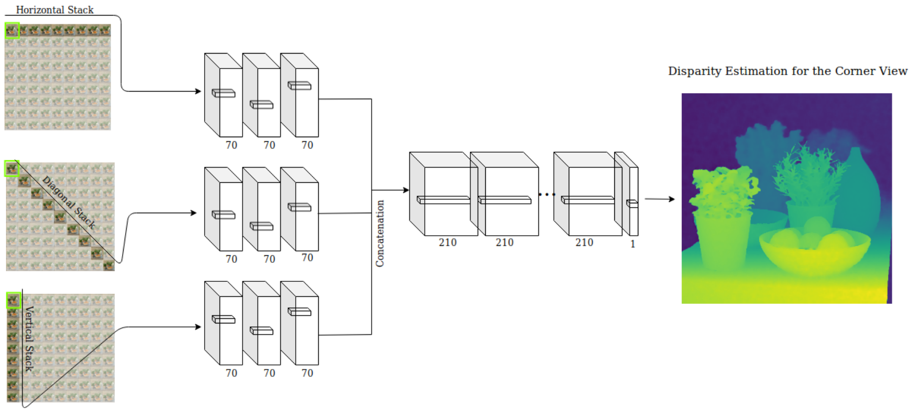

Figure 1. CEPINET: the angular directions for estimating the disparity map of the upper left corner view (green square) are depicted as black arrows.

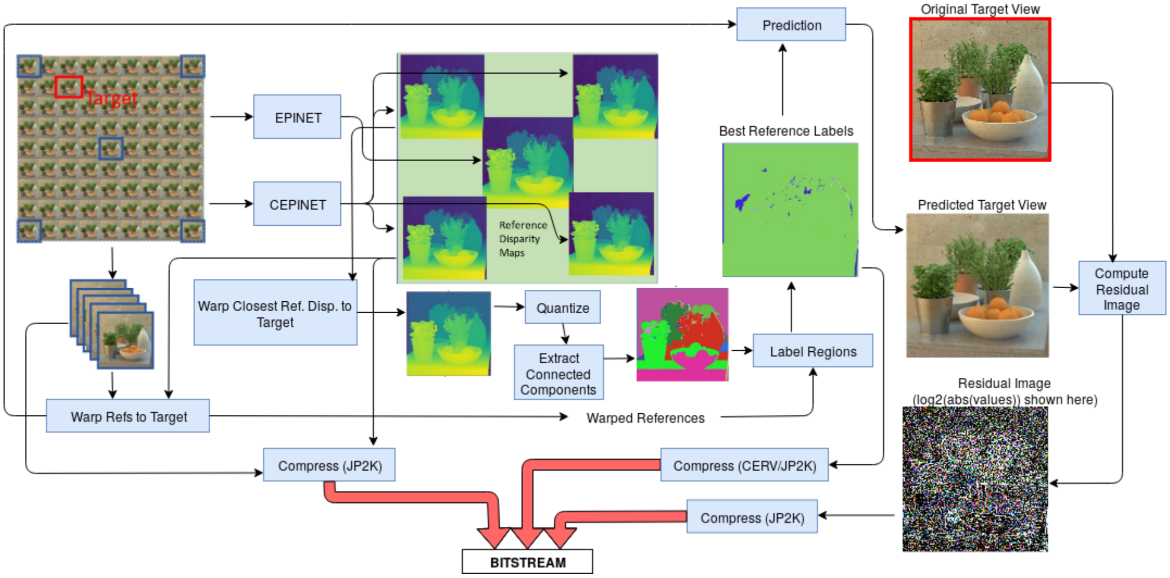

Figure 2. Proposed encoder architecture for lossless compression.

Table V. The compressed size for samples with GT disparity available

Sample	JP2K	LLFC(JP2K/CERV)	LLFC(GT Ds)
vinyl	7.38	4.41/4.30	**4.08**
kitchen	9.07	6.27/6.15	**5.78**
museum	10.97	7.01/6.92	**6.67**
greek	8.15	5.22/5.10	**4.94**

Table VI. The compressed size for samples without GT disparity

Sample	JP2K	LLFC(JP2K)	LLFC(CERV)
dino	9.65	5.84	**5.79**
dots	24.16	20.43	**20.06**
bedroom	10.13	6.94	**6.90**
pyramids	19.88	13.40	**13.32**
stripes	3.44	2.01	**1.92**
bicycle	12.85	8.80	**8.65**
b.gammon	16.62	11.40	**11.30**
origami	10.53	6.97	**6.83**
boxes	11.34	8.12	**7.95**
cotton	6.96	3.26	**3.21**
sideboard	13.93	9.56	**9.42**
herbs	11.93	8.15	**8.01**

Table VII. The compressed size at different Views (LLFC(CERV))

Ref	7.81	7.96	8.01	8.10	7.95	7.91	7.80	**Ref**
7.92	8.70	8.77	8.82	8.23	8.83	8.76	8.70	7.93
8.02	8.73	8.80	8.77	8.14	8.76	8.83	8.75	8.02
8.04	8.75	8.73	8.69	7.94	8.68	8.75	8.79	8.06
7.95	7.89	7.86	7.72	**Ref**	7.73	7.89	7.94	8.00
8.21	8.72	8.69	8.66	7.88	8.64	8.72	8.75	8.19
8.12	8.81	8.71	8.67	7.97	8.67	8.73	8.81	8.13
7.96	8.73	8.80	8.71	7.99	8.68	8.77	8.73	7.97
Ref	7.79	7.94	7.98	7.92	7.93	7.89	7.77	**Ref**

upper limit for the performance of our framework that can be attained by improving the disparity estimation stage. Best reference labels are encoded using CERV.

The compressed size results of our lossless compression scheme at every target view averaged over 16 samples (listed on Tables V-VI) are presented on Table VII. These values are computed by averaging the common costs due to transmission of reference color and disparities over all targets and adding each target's own cost for best reference labels and residuals. It should be noted that this computation corresponds to the scenario when whole light field image is being transmitted. Lossless JPEG2000 yields an average bpp of 11.7 for the same data for all views. Hence, corner view based compression scheme yields better results when compared to JPEG2000 at all camera array locations as evident from Table VII.

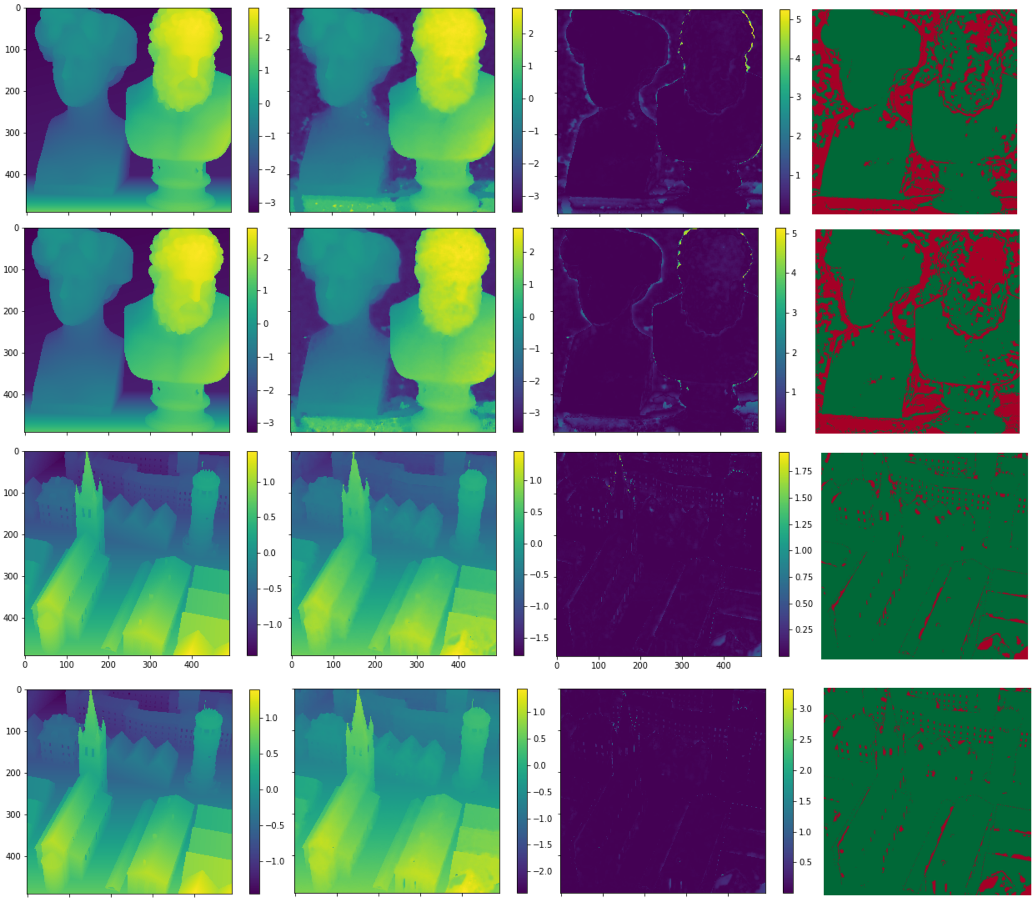

Figure 3. Columns from left to right: Ground Truth, Estimation, Absolute Error, Bad Pixel Mask (0.07). Rows from top to bottom: Center, NW Corner, Center, NW Corner.

V. CONCLUSION

In this work, we constructed the CEPINET for estimating corner view disparity maps of a light field image. It is observed that this variant is able to generate corner view disparities at a similar precision or even better than the center view estimates by EPINET. The proposed lossless compression method provides random access to individual targets, at the cost of transmitting first only the references, and its compression ratio is expected to be lower than the methods that don't possess the random access feature. On the other hand, the compression ratio of LLFC is better than the independent encoding of views by JPEG 2000, which provides instantaneous random access.

VI. REFERENCES

[1] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 193, 2016.

[2] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 24–32.

[3] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4748–4757.

[4] Ioan Tabus, Petri Helin, and Pekka Astola, "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and jpeg 2000," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4567–4571.

[5] Charilaos Christopoulos, Athanassios Skodras, and Touradj Ebrahimi, "The jpeg2000 still image coding system: an overview," *IEEE*, vol. 46, no. 4, pp. 1103–1127, 2000.

[6] Ioan Tabus, Ionut Schiopu, and Jaakko Astola, "Context coding of depth map images under the piecewise-constant image model representation," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4195–4210, 2013.

[7] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 19–34.