Efficient High Dynamic Range Video Using Multi-exposure CNN Flow

Yuchen Guo[1], Zhifeng Xie[1,3(✉)], Wenjun Zhang[1], and Lizhuang Ma[2,3]

[1] Department of Film and Television Engineering, Shanghai University, Shanghai 200072, China
`zhifeng_xie@shu.edu.cn`
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[3] Shanghai Engineering Research Center of Motion Picture Special Effects, Shanghai 200072, China

Abstract. High dynamic range (HDR) imaging has seen a lot of progress in recent years, while an efficient way to capture and generate HDR video is still in need. In this paper, we present a method to generate HDR video from frame sequence of alternating exposures in a fast and concise fashion. It takes advantage of the recent advancement in deep learning to achieve superior efficiency compared to other state-of-art method. By training an end-to-end CNN model to estimate optical flow between frames of different exposures, we are able to achieve dense image registration of them. Using this as a base, we develop an efficient method to reconstruct the aligned LDR frames with different exposure and then merge them to produce the corresponding HDR frame. Our approach shows good performance and time efficiency while still maintain a relatively concise framework.

Keywords: High dynamic range video · Convolutional neural networks
Optical flow

1 Introduction

Due to the limit of sensors used by most image capturing device currently on trade, they lack the capability to capture the wide range of luminance in real world as human eyes can perceive. Thus high dynamic range (HDR) imaging techniques are developed to address this problem. While methods to capture still HDR images have been extensively researched, HDR video is still a comparably less popular subject.

Large portion of HDR video applications up to date have been focused on specialized HDR camera systems [1–4]. These custom hardwares are often either expensive or inconvenient to use, making them hard to be ported for practical use or common consumer market. On the other hand, it's already a common function of digital cameras to capture still HDR image. Utilizing camera's exposure bracketing function, we can take several LDR pictures of same scene with different exposures and merge them to recover larger dynamic range than that of sensors, thus obtaining a HDR image [5, 6].

Similarly, we can also use off-the-shelf cameras to capture a LDR video sequence with alternating exposures. And the aim of HDR video methods is to reconstruct the missing LDR frame of different exposure for each frame in the sequence so they can be merged into a HDR sequence. A sample of the process with our method is shown in Fig. 1. The reconstructed LDR frame should be well-aligned to the original frame of different exposure and temporally coherent to other frames of same exposure, otherwise there will be artifacts like ghosting or jittering in the results. Therefore the process often requires accurate image registration between frames of different exposure due to motions in the sequence. The problems of multi-exposures image registration poses the main challenge for most HDR video application, as traditional motion estimation methods like optical flow often fail in such scenario.
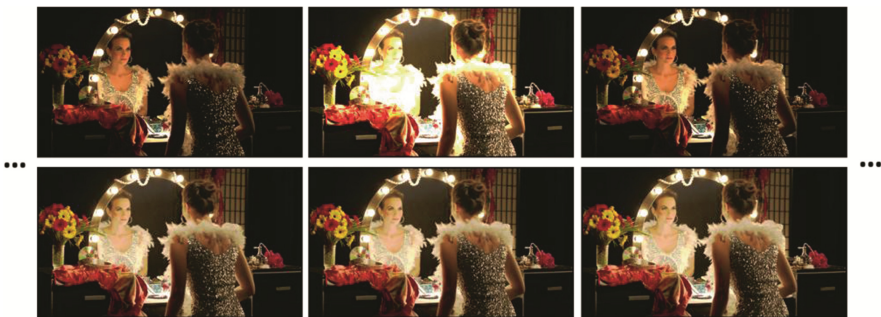
Fig. 1. Sample of HDR reconstruction process. *(Top Row)*: Input sequence of two alternating exposures generated from 'showgirl' sequence of HdM-HDR-2014 dataset [18]. *(Bottom Row)*: HDR sequence (tone-mapped) reconstructed with our method.

On the other hand, recently convolutional neural networks (CNN) have become quite popular in the fields of computer vision after achieve state-of-art performance in problems like object detection, classification, segmentation, etc. Inspired by the research of FlowNet [7] that uses CNN in optical flow estimation, we propose to train an end-to-end CNN model that can handle motion estimation under illumination change using custom-built synthetic dataset.

In this paper, we present a new method to reconstruct HDR video from sequence of alternating exposures using the trained CNN model for motion estimation across different exposure. Leveraging the CNN model's good estimation performance and fast speed, we are able to obtain dense registration between frames of different exposures. A fine registration combined with our occlusion fixing and refinement process, we can achieve good reconstruction results in an efficient way while maintain a relatively simple framework.

In summary, our paper intends to present two main contribution: (1) an end-to-end CNN model trained on custom dataset that can handle multi-exposure motion estimation; (2) an efficient and concise approach to reconstruct HDR video from sequence of alternating exposures by utilizing the above CNN model. We will demonstrate our method and results more specifically in the rest of the paper.

2 Related Work

HDR imaging is a frequently studied subject in this field while there are only a few that are developed specifically for HDR video. And as we mentioned above, a lot of these application are based on custom hardwares like special sensors [1, 2] or devices that register two cameras to capture one scene with different exposures simultaneously [3, 4]. For brevity, in this section we only discuss methods that reconstruct HDR video from LDR sequence of alternating exposures captured by single conventional camera.

Kang et al. [8] propose first practical HDR video approach using sequences of alternating exposures as input. It is a optical flow based method that unidirectionally warp the previous/next frames towards target frame using a variant of the Lucas and Kanade [9] technique in a Laplacian pyramid framework. As for over/under-exposed regions where the optical flow estimation is unreliable, they bidirectionally interpolate the previous/next frames using optical flow between them and further refine the alignment using a hierarchical homography-based registration. Mangiat and Gibson [10] instead choose a block-based motion estimation method in order to overcome the problems of gradient-based method Kang et al. used. They also present a refinement stage that use filtering methods to remove artifacts of mis-registration or block boundary. However, these methods still suffer from the accuracy of motion estimation between multi-exposures frames and often fail when non-rigid or fast motion is present.

The more recent research of Kalantari et al. [11] probably represents the state-of-art result of HDR video reconstruction. They propose a patch-based HDR synthesis method that combines optical flow with a patch-based synthesis approach similar to Sen et al. [12]. Their method enhance temporal coherency using patch-based synthesis and enforce constraints from optical flow estimation to guide patch-based synthesis with a search window map. In this way, they are able to handle more complex motion in the sequence and produce high-quality HDR video output. Although perceptually insignificant, it is still reported that the unstable performance of optical flow estimation may result in artifacts around motion boundaries such as blurring or distorting [13]. Besides, the iteration and optimization process required by the method result in higher running time and complexity compared to other methods.

As the main challenge for HDR video reconstruction is finding correspondences between frames with different exposures, the performance of reconstruction can benefit a lot from an improvement in motion estimation method like optical flow. One of the reasons most variation-based optical flow methods fail when dealing multi-exposures data is the brightness constancy assumption they hold, which was introduced in classical optical flow literature by Horn and Schunck [14]. There were also many attempts to gain robustness against illumination change. Brox et al. [15] added a gradient constancy assumption to the original variational optical flow framework. Mileva et al. [16] tried to make use of photometric invariants in computing an illumination-robust optical flow. Still, the challenge posed by registering frames of different exposure may combine dramatic illumination change, large displacement motion and loss of information due to saturation. It is difficult to design a framework that handles all these issue.

Meanwhile, deep learning techniques, especially CNN, have demonstrated remarkable performance in many computer vision tasks. It is shown to be able to extract features

that otherwise hard to represent in normal ways by learning from large training datasets. Recently, Dosovitskiy et al. [7] first constructed an end-to-end CNN which are capable of solving the optical flow estimation problem as a supervised learning task. However, there is still no learning-based methods developed to overcame the incapability of most motion estimation methods that they can't deal with multi-exposures data.

3   Multi-exposure Optical Flow Based on CNN

Inspired by previous works about CNN-based optical flow, we construct an end-to-end CNN with three main components to predict dense motion vector field between images with different exposures. Besides, in order to supply the networks with sufficient training data to learn from, we build a custom dataset from available flow datasets for multi-exposures motion estimation.

3.1   Network Structure

As shown in Fig. 2, our end-to-end model consists of three main components: low-level feature network, fusion feature network, and motion estimation network.
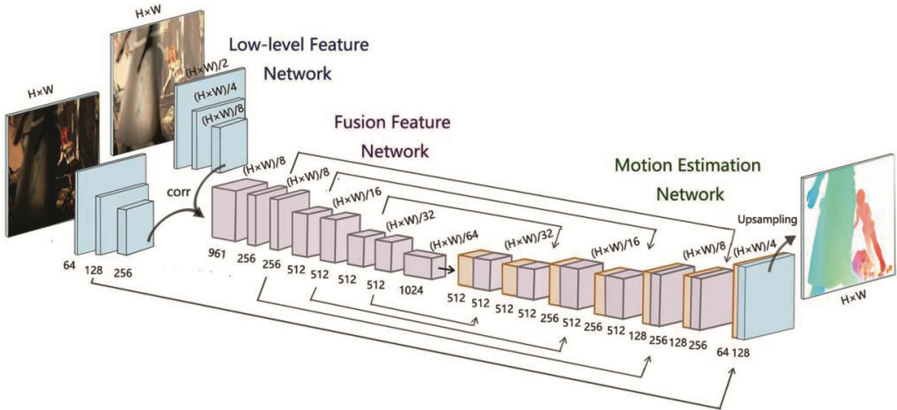
Fig. 2. Our end-to-end CNN consist of three main components: low-level feature network, fusion feature network, and motion estimation network. Given enough training data of multi-exposure image pairs and ground truth flows, our end-to-end model can be trained to predict dense optical flow fields accurately from input images with different exposures.

The low-level feature network contains three convolution layers for each input image. It constructs two separated processing streams for them, which can effectively promote the feature representation and the deep training in the different exposures.

While low-level feature network only focuses on the respective features of the input images rather than their correspondences, we introduce the correlation layer of FlowNet [7] to perform the matching and fusion of two low-level features, and construct a fusion feature network to finally obtain the representation of multi-exposure motion features.

Taking the outputs of the correlation layer from low-level feature network as input, the fusion feature network itself consists of the correlation layer and the convolution layers, which can efficiently handle the matching process between two groups of low-level features and obtain the motion features in the different exposures.

With the entire contracting part completed, we then introduce the motion estimation network in the expanding part, which uses upconvolution layers including unpooling and convolution. It contains seven combination layers, which not only include upconvolution layers but also integrate the outputs of the low-level feature network and the fusion feature network respectively. Each of the combination layers can predict a corresponding coarse flow with 2 outputs, and then upsample the flow as the input of its next layer. In a word, the various features are fused in the motion estimation network, and they are effectively processed by a set of upconvolution and upsampling operations.

3.2   Training Data

In order to effectively train a large-scale CNN, sufficient training data is needed. Besides, neural networks require data with ground truth to learn to perform a prediction task from scratch. These requirements make it difficult to prepare training data for our multi-exposure application as it's quite hard to capture ground truth motion flows from real world scenes.

While there are several public optical flow datasets that contain ground truth flow, most of them are generated from synthetic scenes and, more importantly, maintain the same exposure setting. Therefore we choose to build a custom multi-exposure optical flow dataset using available datasets as bases.

First we choose the public datasets to build on. There are three state-of-art candidates: the Middlebury dataset, the Kitti dataset and MPI-Sintel dataset. The Middlebury dataset is widely used for optical flow evaluation. But it only contains 8 synthetic image pairs with small displacement motions, and thus is too small for learning. The Kitti dataset is a real world scenes dataset captured by automobile platform. Its complexity in lighting and texture makes it a challenging benchmark. Yet due to the limit of capturing device, its ground truth flows are sparse, which makes it unsuitable for our need. MPI-Sintel dataset consists of sequences from an animation movie which include various motion types and scenes. All things considered, we choose the 'final' version of MPI-Sintel dataset with realistic rendering effects such as motion blur and atmospheric effects to get closer to the more complex real world scenes.

After that we need to generate multi-exposure data from the selected dataset. As exposure value (EV) of camera is a number that represents the combination of shutter speed and f-number, with a difference of 1 EV corresponding to a standard power-of-2 exposure step. We utilize gamma correction to synthesize the multi-exposure effect. By increase one frame's exposure while decrease another's, the process create image pairs with drastic brightness change similar to that of exposure difference while maintain same ground truth motion. By comparing results of our post processing with real image with different exposures, it can be observed that our simple simulation can effectively reflect the change between different exposures though not perfectly accurate.

Using the new multi-exposure dataset, we trained our networks on a computer with a CPU of Intel Xeon E5-2620, 16 GB memory and an NVIDIA Tesla K20 GPU. The resulting model converged well and demonstrated good performance on the task of multi-exposure motion estimation, which effectively support our HDR reconstruction application.

4   HDR Video Reconstruction

As mentioned above, the raw data input used for HDR video reconstruction is a LDR video captured with conventional camera that alternates between different exposures for each frame. We will take a two-exposure sequence as example here.

The goal of our method is to reconstruct the missing LDR frame of different exposure for each of the frame in the sequence. As the each frame has different exposure from its neighboring frames, the reconstruction process requires drawing information from its next/previous frame which may be of the same exposure. And that's where an accurate pixel correspondences or motion estimation come into play.

Figure 3 shows a brief structure of our method's process. For certain frame F_n in the alternating exposure sequence, we try to reconstruct the missing LDR image L with a different exposure, shown with dashed red square. Other HDR video methods often use optical flow result as a rough estimation or initiation for the registration of correspondences between frames with different exposures. While by taking advantage of our trained CNN model, we can directly estimate a good motion field as optical flow between F_n and its neighboring frame F_{n-1}/F_{n+1}, which are different in exposure. The improvement
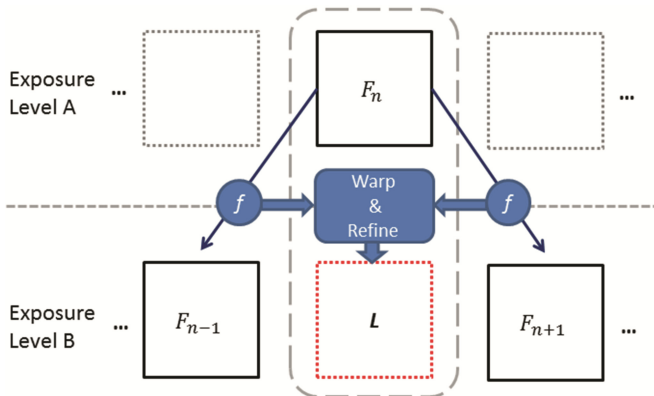
Fig. 3. Reconstruction process of frame n from a sequence alternating between two different exposure levels, which only capture certain exposure at each frame (shown with solid black squares). Our method reconstructs the missing exposure for the current frame F_n using a warp and refine scheme based on optical flow f (shown with solid blue circles) between current frame and its neighboring frame, computed by our CNN model for multi-exposure motion estimation. Once the missing LDR image has been reconstructed, it can be merged together with current frame to produce the HDR frame, which will then form the entire HDR video. (Color figure online)

in quality of motion estimation between frames of different exposure enables us to utilize a more concise and straightforward scheme in the reconstruction of missing LDR frame L. Moreover, in this way we don't need to linearize the image and boost its intensity for better registration like many other methods require, which may involve camera response function (CRF) estimation and therefore limit the application.

To actually reconstruct L with the motion estimation results, we generate two intermediate results by warping the previous/next frame F_{n-1}/F_{n+1} towards current frame F_n to obtain two warped frames W_{n-1}/W_{n+1}. However, we can't directly generate target frame L from the two warped frames, even though the good motion estimation result may yield high quality warped results. Due to occlusion, large-displacement motion or small amount of unreliable flow, it is usually necessary to further refine the results. Therefore we introduce a refinement process to obtain the final reconstructed L with higher quality.

The refinement process uses two main constraints to ensure a satisfactory result. They can be formulated as energy functions below:

$$E = E_c(F_n, L) + E_t(F_n, F_{n-1}, F_{n+1}) \tag{1}$$

In Eq. (1), first term E_c represent the consistency between F_n and L, as they are supposed to be the same frame with different exposures. To measure the consistency in content or structure between two images with different exposures, we employ two metrics. As the two images are supposed to contain the same content and geometry, we assume that there are similar details or gradient where the two images are both well exposed. Besides, to further utilize the performance of our multi-exposure CNN model, we estimate optical flow between the two frames with the model, which can be used to ensure there are no motion between them where the flow are reliable. These two constraints enforce the consistency the original and reconstructed frames and thus help to avoid the ghosting artifacts in HDR merge process. Their formula is shown in the function below:

$$E_c(F_n, L) = \alpha * d(\nabla F_n, \nabla L) + \beta * m(F_n, L) \tag{2}$$

where α resemble the approximation map of how well a pixel is exposed in both image and d(x, y) is L2 distance. While β measures how reliable a motion vector in the optical flow map is, and m(a, b) is the motion distance of each pixel between the two image.

Second term E_t in Eq. (1) maintain the time coherence between reconstructed frame and its previous/next frame with the same exposure. Our refinement procedure approaches this with two main operations. On one hand, we enhance the smoothness of optical flow by comparing all flow fields between the three frames and also the warped results in a bidirectional way to verify the motion's reliability and continuity, which helps to avoid video jittering caused by erroneous motion. On the other hand, sometimes due to large-displacement motion, there are noticeable region of occlusion present which would cause ghosting from previous/next frames in the warped images. To deal with occlusion, we first extract regions of occlusion by comparing motion vectors' origin and destination of flow from neighboring frame to current frame, the difference of which

can be used to extract the occlusion map. Then we fix the region of occlusion in one warped image by drawing information from the other warped image which contains content from another neighboring frame with different occlusion area. The process of handling occlusion is shown in the above Fig. 4. In summary, these operations enforce the constraint of time coherence between reconstructed frame and its neighboring frames with the same exposure, which can be formulated as the function below:

$$E_t\big(F_n, F_{n-1}, F_{n+1}\big) = \sum\nolimits_{i \in pixels} \big(d\big(F_n^i, F_{n-1}^{i+u}\big) + d\big(F_n^i, F_{n+1}^{i+u}\big)\big) \tag{3}$$

where i is a pixel location in F_n, while u represent the motion displacement at i between F_n and its neighboring frame F_{n-1} and F_{n+1}. This ensures similarity and coherence between frames and thus solves the jittering artifacts.

Fig. 4. Example of occlusion fixing process. (*Left Column*): current reference frame *(Left-Top)* and its next frame *(Left-Bottom)*. *(Middle-Top)*: optical flow estimated by our CNN model from current frame to its next frame. *(Middle-Bottom)*: directly reverse warped result using the flow, which shows ghosting at regions of occlusion. *(Right-Top)*: occlusion map extracted from optical flow map. *(Right-Bottom)*: warped LDR result after our occlusion fixing process.

Finally, after the refinement process we combine the two refined warped images to obtain the reconstructed LDR frame of different exposure at current frame time as result. With that, we merge them to achieve the HDR frame and tone-map it for display. Besides, the reconstructed LDR frame can also help to refine the reconstruction process of its neighboring frame.

5   Results and Discussion

We demonstrate and analyze some results of our HDR reconstruction method in this section. All results displayed here are fused and tone-mapped using the exposure fusion method by Raman et al. [17].

In order to obtain sequences with alternating exposures as input data for our method, we make use of the high-quality HDR video sequences dataset by Fröhlich et al. [18]. These sequences are captured using two cameras mounted on a mirror-rig and contained

various scenes with different challenges such as complicate illumination setting, high contrast skin tones and saturated colors, etc. By extracting multiple exposures from original HDR data, synthetic sequences of alternating exposures can be acquired in this
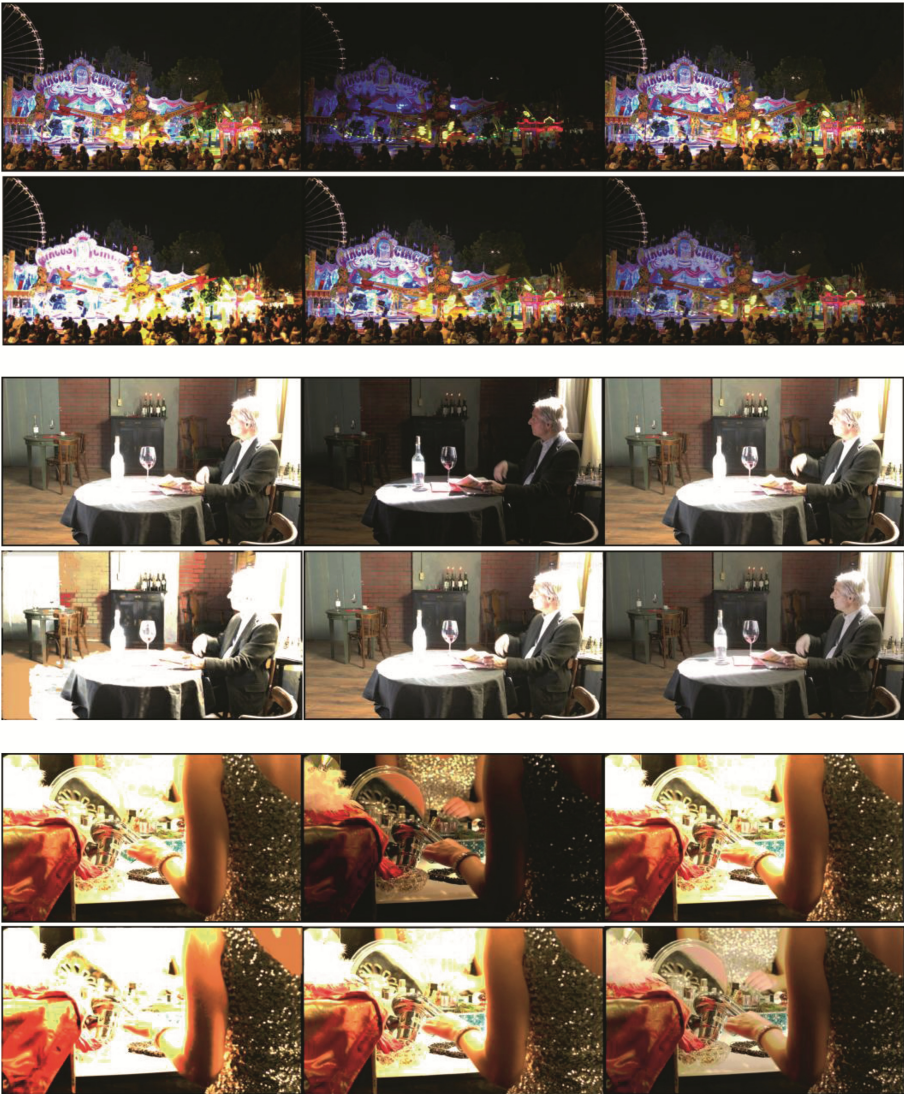
Fig. 5. Examples of test sequences and results. For each group, *(Top Row)*: input triplet of consecutive frames of two alternating exposures, with middle one as current reference frame; *(Bottom Row)*: reconstructed LDR result of *(Bottom-Left)* Kalantari et al. [11] (without corresponding CRF) and *(Bottom-Middle)* ours; *(Bottom-Right)*: our HDR result (tone-mapped). (Color figure online)

way. Moreover, available ground truth data also offer a better evaluation and comparison for the performance of our method.

We test our method using different dynamic scenes from the HdM-HDR-2014 dataset [18], which are extracted into sequences with two alternating exposures of −2EV and +1EV at resolution of 1920*1080 as input. The three scenes in Fig. 5 were chosen to be displayed here due to the unique and representative features they demonstrate. As shown in Fig. 5, the first scene 'carousel fireworks' is filmed at an annual fair where color-saturated highlight and fast moving, self-illuminated objects are present the dark nighttime surroundings. While second scene 'bistro' features a dark bistro-chamber combined with local bright sunlight at the window, creating a high-contrast scene with difficult lighting situation. And the third scene 'showgirl' shows partially illuminated skin and specular highlights on various reflecting props together in a glamorous tone. These scenes can demonstrate the performance of our method when faced with different challenges.

For each frame to be processed, it is combined with its neighboring frames with different exposures to form a triplet of consecutive frames as input, producing the reconstructed LDR frame of missing exposure as output, which is then merged with original frame into the final HDR frame. For brevity, in Fig. 5 we take single triplet from each sequence as example and display the reconstructed results.

Besides, we also run these test cases with the method of Kalantari et al. [11], which is considered one of the state-of-art methods of HDR video reconstruction in regards of reconstruction quality while using conventional camera. Yet it's shown in Fig. 5 that their method fail to reconstruct the correct missing LDR image due to the lack of corresponding camera response function (CRF) for our test data. Though this doesn't affect the good performance of their method when CRF are provided, the comparison demonstrates our method's robustness and wider applicability.

In order to achieve a better evaluation for our method, we compare our reconstructed frame results with the ground truth data generated from original HDR sequence. Using PSNR as main metric, the evaluation results and running of our method for each test sequence are listed in Table 1. From it we can see that our method shows good and stable performance in HDR reconstruction quality as well as high processing speed, much faster than that of Kalantari et al. [11] which may require nearly 10 min to run. It should also be noted that the operation of motion estimation with our multi-exposure CNN model only takes only about one second to run, which implies there is still much room for improvement in time efficiency given better optimization and implementation in refinement stage.

Table 1. Evaluation results

Testing sequence	PSNR/dB (Average)	Running time/s (Per frame)
Carousel fireworks	38.83	103.5
Bistro	42.19	99.7
Showgirl	41.56	84.3
Average (over all scenes)	32.16	98.4

Nevertheless, several limitations are still observed during experiments. When current reference image present large region of glare/saturation due to high exposure time or motion blur caused by fast movement, the optical flow result of our motion estimation may not be accurate because lack of coherence in content between frames, which then leads to a decrease in performance. In addition, sometimes there will be regions of occlusion in current frame that are not present in neighboring frames, causing the algorithm unable to draw information from adjacent frames by using motion as cue, which may require other matching method to fix. Moreover, we also observed that the performance of current CNN model is somehow sensitive to image scale and motion type possibly due to the training data we provided.

To address these problems, our future work will focus on trying different CNN structure design and training scheme in order to solve the current limitations in a more unified framework. And other plans include making better use of similarity between frames in same sequence for achieving better time efficiency.

6   Conclusion

In this paper, we present a new method for HDR video reconstruction from sequence of alternating exposures, which utilize a CNN model with capability of motion estimation across multiple exposures. By training a CNN end-to-end to learn predicting optical flow from image pairs with different exposures, we manage to overcome the problems of image registration between different exposures where many other motion estimation methods failed, and thus use a more concise framework for HDR video reconstruction. With effective refinement process, the results of our method demonstrate competitive performance in both reconstruction quality and efficiency. It also shows the potentials of further application of CNN in the field of HDR synthesis.

References

1. Nayar, S., Branzoi, V.: Adaptive dynamic range imaging: optical control of pixel exposures over space and time. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), vol. 2, pp. 1168–1175. IEEE, Nice (2003)
2. Unger, J., Gustavson, S.: High-dynamic-range video for photometric measurement of illumination. In: Proceedings of the SPIE, vol. 6501 (2007)
3. Tocci, M.D., Kiser, C., Tocci, N., Sen, P.: A versatile HDR video production system. ACM Trans. Graph. **30**(4), 41 (2011)
4. Kronander, J., Gustavson, S., Bonnet, G., Unger, J.: Unified HDR reconstruction from raw CFA data. In: IEEE International Conference on Computational Photography (ICCP 2013), pp. 1–9. IEEE, Cambridge (2013)

5. Mann, S., Picard, R.W.: On being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures. In: IS&T's 48th Annual Conference, pp. 422–428. Society for Imaging Science and Technology, Washington, D. C. (1995)
6. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of SIGGRAPH 1997, pp. 369–378. ACM SIGGRAPH, Los Angeles (1997)
7. Dosovitskiy, A., Fischery, P., Ilg, E., et al.: Flownet: learning optical flow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV 2015), pp. 2758–2766. IEEE, Santiago (2015)
8. Kang, S.B., Uytendaele, M., Winder, S., Szeliski, R.: High dynamic range video. ACM Trans. Graph. **22**(3), 319–325 (2003)
9. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 1981), vol. 2, pp. 674–679. IJCAI, Vancouver (1981)
10. Mangiat, S., Gibson, J.: Spatially adaptive filtering for registration artifact removal in HDR video. In: 18th IEEE International Conference on Image Processing (ICIP 2011), pp. 1317–1320. IEEE, Brussels (2011)
11. Kalantari, N.K., Shechtman, E., Barnes, C., Darabi, S., Goldman, D.B., Sen, P.: Patch-based high dynamic range video. ACM Trans. Graph. **32**(6), 202 (2013)
12. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shecheman, E.: Robust patch-based HDR reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6), 203 (2012)
13. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: The state of the art in HDR deghosting: a survey and evaluation. Comput. Graph. Forum **34**(2), 683–707 (2015)
14. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
15. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24673-2_3
16. Mileva, Y., Bruhn, A., Weickert, J.: Illumination-robust variational optical flow with photometric invariants. In: Hamprecht, Fred A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 152–162. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74936-3_16
17. Raman, S., Chaudhuri, S.: Bilateral filter based compositing for variable exposure photography. In Proceedings of Eurographics 2009. European Association for Computer Graphics, Munich (2009)
18. Froehlich, J., Grandinetti, S., et al.: Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In: Proceedings of the SPIE, vol. 9023 (2014)