HDR Video System Based on Multi-Stage Motion Estimation Using Dual Exposure Camera

Hak-Hun Choi, Jae-Won Kim, Shigeki Hirohata, and Ho-Nam Lee
Samsung Electronics Co. Ltd., Hwasung, Korea
Email: {hakhun.choi, jw9557.kim, s.hirohata, and honam.lee}@samsung.com

Abstract—In this paper, we provide a new HDR video system using a dual exposure camera. In order to prevent ghost artifacts caused by misalignment of motions, multi-stage convolutional neural network (CNN)-based motion estimation is proposed. In the proposed method, true motions and occlusion maps are estimated, and then an HDR video is synthesized with motion compensated images. It provides not only robustness against exposure levels but also flexibility to extend easily search ranges by stacking stages with same CNN coefficients. Experimental results indicate that the proposed method is effective to estimate true motions so that ghost artifacts are eliminated well.

I. INTRODUCTION

An HDR image is typically created by combining a set of different exposures because camera sensors can only capture a limited range of luminance. However, if there exist motions among an input image set, ghost artifacts are produced around motion objects. In order to obtain a ghost-free HDR image, de-ghost HDR techniques have been proposed [1-4]. Input images including motions are motion-compensated to align objects and then synthesized to an HDR image. HDR video processing has been also introduced where an HDR video is created from an image sequence captured while rapidly varying the exposure of a camera [4]. In [4], HDR stitching compensates pixels by global and local estimations. However, since many scenes actually have fast motions, ghost artifacts caused by false motion compensation may still occur.

In order to remove ghost artifacts for fast motions, the architecture to support wide search ranges is needed for HDR video processing. In this paper, we propose a new HDR video system using a dual-exposure camera. In the proposed method, a multi-stage convolutional neural network (CNN) is applied for motion estimation (ME). Search ranges are easily extended by stacking stages that have same coefficients so that fast motions can be estimated precisely.

II. PROPOSED HDR SYSTEM

Fig. 1 depicts the overview of the proposed HDR system. At first, a dual exposure camera captures alternating frames with two different exposure levels (short and long). In the CNN-based ME, optical flows and occlusion maps are estimated. To perform the HDR fusion, an intermediate frame is interpolated by using optical flows and occlusion maps. The interpolated frame and the input frame with a different exposure level are synthesized to an HDR image.

The main purpose of this paper is to eliminate ghost artifact caused by motions in input images. Thus, the most important point is to find accurate object motions which are called true motions. In the proposed CNN-based ME, hierarchical concept is applied to CNN not only estimating true motions well but also extending search ranges easily by stacking each stage having same CNN coefficients. This hierarchical architecture is inspired by PWC-net that improves optical flow estimation by feature pyramids [5].

For the brief explanation of the proposed multi-stage CNN architecture, Fig. 2 shows the example of the hierarchical architecture with two stages. Each stage consists of two parts: a feature extraction part and an ME part.

In the feature extraction part, three input frames with two different exposure levels are analyzed to generate features. This part is organized by CNN and is expected to have a brightness invariant characteristic even though three input frames have different exposures.

The ME part refines up-scaled motions estimated by the ME part of the previous stage with input features delivered by the feature extraction part of the same stage. An ME1 analyzes a correlation among three features that represent input frames with two different exposures. Using the correlation, two ME2s estimates backward motions and forward motions, respectively. In addition, the occlusion map of each motion field is also generated. It is used for the frame interpolation for the HDR fusion. Note that Table I shows the network configurations of three main CNNs used in the proposed ME.

---

Fig. 1 Overview of the proposed HDR system.

Fig. 2 Multi-stage CNN architecture (Number of stages: 2).
Finally, an intermediate frame for the HDR fusion is interpolated with estimated motions and occlusion maps. At this time, occlusion maps help to figure out interpolation methods among cover area, uncover area, bi-directional motion area, and background area. Note that the HDR fusion used in [2] actually generates HDR-like low dynamic range images called wide dynamic range images.

III. EXPERIMENTAL RESULTS AND CONCLUSION

The number of CNN stages of the proposed method for training and test are 4 and 6, respectively. Each input of the network is processed by using a gamma curve to emulate short and long exposure status for training. We have trained the proposed network with 1,332 frames and the size of 160×160.

Total number of coefficients that should be trained of the proposed method and PWC-net are 75,083 and 4,705,064, respectively. For test run with six stages, the numbers of coefficients of the proposed method and PWC-net are 450,498 (75,083×6) and 4,705,064, respectively. Note that the proposed method uses quite less coefficients than PWC-net. As a result, the performance of ME may be degraded relatively as shown in Fig. 3. However, the optical flow estimated by the proposed method is similar as compared with that of PWC-net. Even though input images have different two exposure levels, two optical flows of Fig. 3 (d) and (f) are very similar as well. In addition, if the networks of the proposed method are deeper by stacking stages, bigger motions can be estimated as true motions than PWC-net. On the other hand, PWC-net as fixed networks is difficult to extend search ranges easily.

Fig. 4 is the example of an occlusion map for forward motions. It is seen that occlusions of motion objects are correlated with estimated maps well. In the proposed method, two occlusion maps are employed adaptively to minimize visual side effects such as motion boundary blocking artifacts caused by fast motions.

Finally, in Fig. 5 (a) and (c), ghost artifacts around motion objects make visual quality worse. On the other hand, HDR synthesis with motion compensation shows clear boundaries around motion objects in Fig. 5 (b) and (d). Especially, Fig. 5 (b) shows the better de-ghosted result by the proposed method using occlusion maps and motions as shown in Fig. 3 (f) even though there are fast object motions on slow background motions.

A multi-stage structure helps to find true motions with wide search ranges. Typically, in order to implement wide search ranges, deep neural networks may be needed so that training failures sometimes occur. However, because the proposed method re-uses the trained coefficients for every stage, it is easy to implement the deep network by stacking stages to extend search ranges. This architecture is suitable for applications tracing true and big motions such as HDR video de-ghosting, frame rate-up conversion, and etc. Note that to improve the ME performance of each stage, the network structure of Table I can be reinforced by adding layers or coefficients.

## TABLE I

| Network Configurations for the Proposed Method
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>ME1</td>
<td>ME2</td>
</tr>
<tr>
<td>L0: RGB 10bit</td>
<td>L0: Cost volume (8×1)</td>
<td>L1: 3×3×96, ReLU</td>
</tr>
<tr>
<td>L1: 3×3×8, ReLU</td>
<td>L1: 3×3×4, ReLU</td>
<td>L2: 3×3×48, ReLU</td>
</tr>
<tr>
<td>L2: 3×3×8, ReLU</td>
<td>L2: 2×2×4, ReLU</td>
<td>L3: 3×3×48, ReLU</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>L4: 3×3×10, tanh</td>
</tr>
</tbody>
</table>

![Fig. 3](image.png) Performance of true motions: PWC-net and the proposed method.

![Fig. 4](image.png) (a) an input image with fast motions and (b) the occlusion map for forward motions.

![Fig. 5](image.png) (a) Not compensated (market_5) (b) Proposed (market_5) (c) Not compensated (ambush_7) (d) Proposed (ambush_7) Fig. 5 HDR synthesis results.

---


