# A NEW MOTION ESTIMATION METHOD FOR MOTION-COMPENSATED FRAME INTERPOLATION USING A CONVOLUTIONAL NEURAL NETWORK

*Giyong Choi, PyeongGang Heo, Se Ri Oh, and HyunWook Park*

Department of Electrical Engineering, KAIST, Daejeon, Republic of Korea

## ABSTRACT

The motion-compensated frame interpolation (MCFI) methods usually use block matching algorithms (BMAs) for motion estimation (ME). However, the conventional BMAs that are originally developed by minimizing the prediction errors often fail to project the object motion. In this paper, we present a new MCFI method that utilizes a convolutional neural network (CNN) to find the motion vector (MV) with reliability. The CNN model which is used to estimate MVs is trained to track the projected object motion as closely as possible. Experimental results using the standard test video sequences show that our proposed ME method acquired more reliable MVs than conventional ME methods. Furthermore, our proposed MCFI method improves the average peak signal-to-noise ratio (PSNR) of interpolated frames.

***Index Terms***— Motion estimation, frame rate up-conversion, motion-compensated frame interpolation, neural networks

## 1. INTRODUCTION

Hold-type displays, such as liquid crystal display (LCD) televisions, have motion blur artifact when the frame rate is low. Besides, frame rate of transmitting videos is often downsampled to meet bandwidth limitation on communication channels. To overcome the problem, motion-compensated frame interpolation (MCFI) is widely used for frame rate up-conversion (FRUC) which increases the frame rate in the decoder side. As the frame rate of transmitted videos increases in the decoder side, the motion blurriness on LCD can be mitigated.

Intermediate frames between sequential frames are generated by MCFI to increase the frame rate. To harmonize with original frames, object motions in successive frames should be considered in the intermediate frames. Therefore, MCFI firstly calculates motion vectors (MVs) between sequential frames by motion estimation (ME) and produces intermediate frames in accordance with the obtained MVs.

ME has a crucial role in MCFI because MVs estimated by ME strongly affect the interpolated frames. Many MCFI methods have utilized block matching algorithm (BMA) for ME, since it is easy to implement and tracks the movement of objects in a reasonable accuracy [1]-[3]. However, as BMA is originally developed in order to minimize the prediction errors, their performance in tracking and estimating the object motion is not optimal, often finding the MVs that do not agree with the object's true motion.

To find more reliable MVs, many ME methods have been proposed. Kang et al. proposed dual ME method that jointly used the unidirectional and bidirectional ME schemes [3]. Choi et al. proposed the bilateral ME scheme with a side match distortion [4]. In addition, Ha et al. employed overlapped block-based ME (OBME) which applied matching criteria over the overlapped block region to estimate MVs of non-overlapped blocks [2]. However, these methods did not consider variances in shape of the object which disturb finding accurate MVs.

Shape of an object in video sequences often varies in several ways such as rotating and scaling. In these circumstances, it is hard to find reliable MVs by using conventional methods since they determine MVs as the displacement of current block and reference block that minimizes residual errors. This is the reason that conventional BMA often fails to track the projected object motion.

In this paper, we propose a new ME method that utilizes a convolutional neural network (CNN). The CNN is trained to estimate MVs with considering variances in shape of the object. Thus, our CNN model tracks the projected object motion more precisely.

This paper is organized as follows. In the section 2, the new CNN-based ME method is introduced. The experimental results and their analysis are discussed in the sections 3 and 4, respectively.

## 2. PROPOSED MOTION ESTIMATION METHOD

Recently, several methods that utilize CNN to predict how well a pair of image patches match have been proposed [5]-[7]. In these methods, CNN, which is used to predict a similarity of two image patches, has some properties in common. The CNN is trained to classify a pair of image patches into two categories: correct matches and incorrect matches. After training, the CNN produces the probabilities of each category when a pair of image patches is entered to

the CNN as an input. In test phase, a probability of the correct match category is used as a similarity of two image patches.

ME in MCFI is quite similar to comparing image patches because MVs are obtained by finding the most similar image blocks between successive frames. However, the primary difference of the proposed ME method and aforementioned methods is whether the benchmark dataset exists or not. Methods in [5]-[7] used benchmark datasets that had numerous pairs of image patches and its labels, which indicated if the image patch was the correct match or not, to train CNNs to estimate similarity of image patches. Likewise, we need a dataset to train a CNN that can estimate MVs in MCFI, but there is no benchmark dataset which is publicly available for ME in MCFI. Therefore, we generated our own dataset to train a CNN to estimate MVs in MCFI and this is one of the main contributions of this paper.

### 2.1. Construction of the database

To construct the dataset for ME in MCFI, we used several benchmark video sequences: m*other and daughter, bus, coastguard, hall monitor, news, Paris, soccer,* and *tempete*. In conventional MCFI methods, odd frames of each sequence are removed and MVs are estimated between successive even frames. After that, intermediate frames are interpolated in accordance with these obtained MVs and compared with original odd frames. In the basic ME method, MVs are estimated as follows:

$$v = \arg\min_{(dx,dy)\in S} \{ \sum\sum_{(x,y)\in B} | f_{t-1}(x-dx, y-dy) - f_{t+1}(x+dx, y+dx) | \}, \quad (1)$$

where $(\mathrm{dx}, \mathrm{dy})$ denotes the candidate motion vector; B indicates a block; $f_{t-1}$ and $f_{t+1}$ are the previous and following frames; and v and S represent the selected MV of $(\mathrm{dx}, \mathrm{dy})$ and the search range, respectively.

However, MVs which are estimated by this basic ME method do not represent the object's true motion because MVs are decided by minimizing the residual error between blocks of successive frames. To find pseudo-ground truth MVs, we use a modified ME method that considers original odd frames. This method can be written as follows:

$$v^{gt} = \arg\min_{(dx,dy)\in S} \{ \sum\sum_{(x,y)\in B} | f_t(x,y) - \frac{f_{t-1}(x-dx, y-dy) + f_{t+1}(x+dx, y+dy)}{2} | \}, \quad (2)$$

where $v^{gt}$ denotes pseudo-ground truth MV of $(\mathrm{dx}, \mathrm{dy})$ and $f_t$ is the original frame. Because the modified ME estimates MVs that produce the best match with original frames, these estimated MVs can be regarded as pseudo-ground truth MVs.

To train a CNN, we create the dataset which has numerous examples of good and bad matches, which is quite similar way to what Zbontar did in [5]. Two image blocks, one from the previous frame ($f_{t-1}$) and one from the following frame ($f_{t+1}$), are a component of dataset as follows:

$$< B^{P}_{16\times16}(\mathbf{p}), B^{F}_{16\times16}(\mathbf{q}) > \quad (3)$$

where $B^{P}_{16\times16}(\mathbf{p})$ and $B^{F}_{16\times16}(\boldsymbol{q})$ indicate 16×16 image blocks from the previous frame centered at $\mathbf{p}$ and the following frame centered at $\mathbf{q}$, respectively. Good match examples are obtained by using the pseudo-ground truth MV as follows:

$$\mathbf{p} = (x - d^{gt}_x, y - d^{gt}_y), \ \mathbf{q} = (x + d^{gt}_x, y + d^{gt}_y), \quad (4)$$

where $d^{gt}_x$ and $d^{gt}_y$ denote x and y component of the pseudo-ground truth MV. On the other hand, bad match examples are obtained by adding additional offset value to the pseudo-ground truth MV as follows:

$$\mathbf{p} = (x - d^{gt}_x, y - d^{gt}_y), \quad \mathbf{q} = (x + d^{gt}_x + o_x, y + d^{gt}_y + o_y), \quad (5)$$

where $o_x$ and $o_y$ denote offset values which corrupt the match. Each offset value is chosen randomly in range of [4, 12].

### 2.2. Network training

Using the constructed dataset, the CNN is trained to classify examples into two classes: good match and bad match. As our CNN model, 2-channel-based network architecture which shows the best performance among several models in [6] is used. The proposed network architecture is as follows:

C(64,3,1)-C(64,3,1)-C(64,3,1)-F(4096)-F(1024)-F(2), (6)

where C(n, k, s) is a convolutional layer with n filters of spatial size of k×k applied with stride s and F(n) denotes a fully connected linear layer with n output units.

Caffe package [8] is used to implement our proposed model. Training the CNN takes 8 hours with an Nvidia GTX Titan X GPU.

### 2.3. CNN-based motion estimation method

In MCFI, MVs can be estimated by the CNN which is trained to produce similarity of a pair of image patches. When inputs to the CNN are two image blocks from each previous and following frame, the CNN calculates probabilities of the two classes: good match and bad match. The probability of good match class is regarded as the similarity of two blocks, so we use this probability to our proposed ME method as follows:

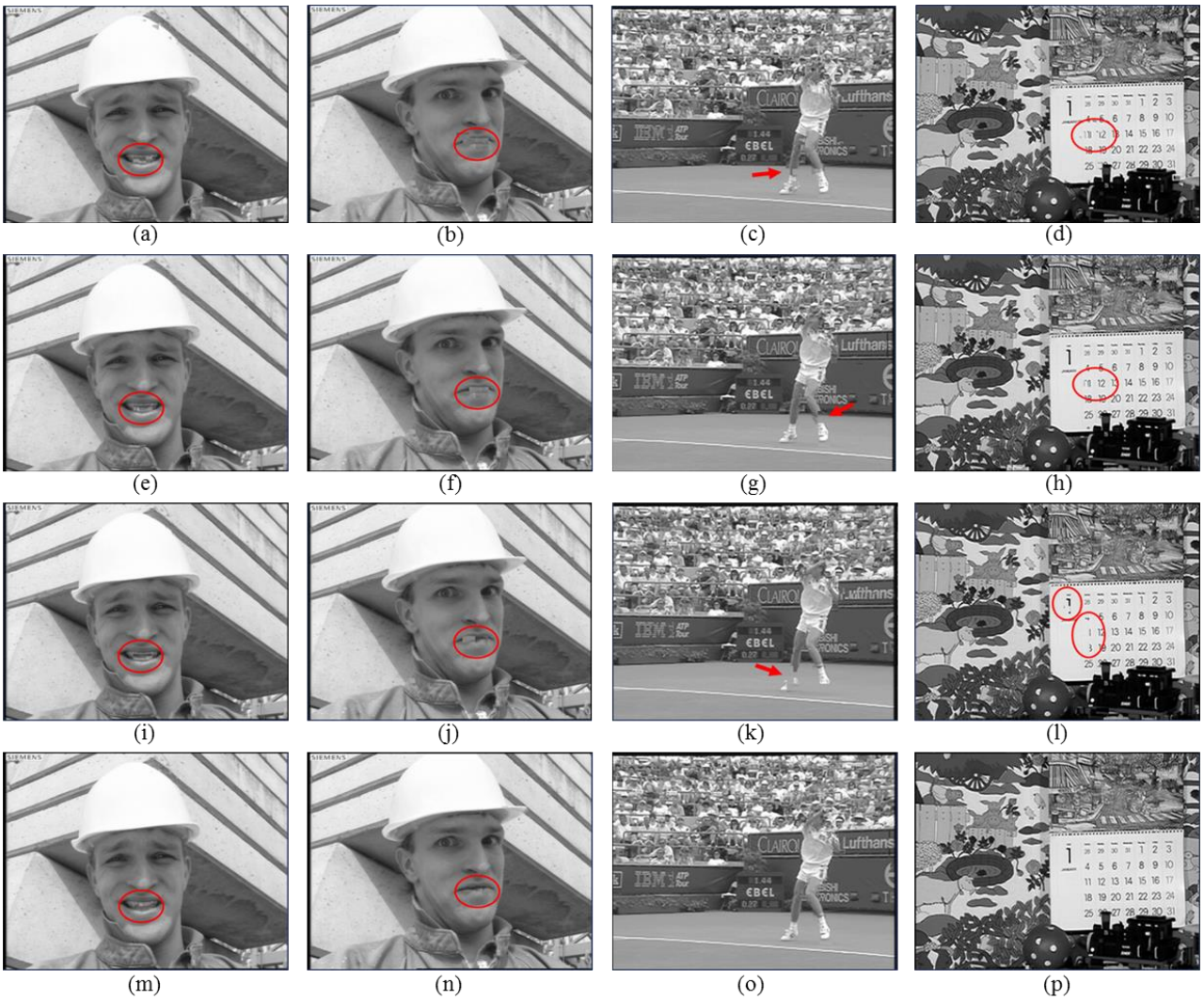

Figure 1. ME results for *foreman*, *Stefan*, and *mobile* sequences. (a)-(d) Results from bidirectional ME in [1]. (e)-(h) Results from OBME in [2]. (i)-(l) Results from dual ME in [3]. (m)-(p) Results from the proposed ME method.

$$v = \arg\max_{\mathbf{p},\mathbf{q}\in S} \{ f_{pos}(B^{P}_{16\times16}(\mathbf{p}), B^{F}_{16\times16}(\mathbf{q})) \}, \qquad (7)$$

where $f_{pos}(B^P, B^F)$ is the probability of good match class when inputs are image blocks of previous and following frames; S denotes search range in each previous and following frame. Through our proposed method, more reliable MVs can be estimated.

## 3. EXPERIMENTAL RESULTS

To evaluate the proposed method, five benchmark video sequences, which are not included in the dataset used to train the network, were utilized: *football, foreman, Stefan, tennis*, and *mobile* in CIF format (352×288). Five video sequences were temporally subsampled by a factor of two, then their original frame rate recovered by generating the intermediate frames using the proposed and other existing methods. After this up-conversion process, PSNR values were measured between the interpolated frame and the original frame in order to measure the objective quality. Then, the average of PSNR over each video sequence was calculated. In addition to objective assessment, the perceptual quality was evaluated subjectively to verify reliability of the estimated MVs.

### 3.1. Subjective assessment of motion estimation method

To test our CNN-based motion estimation method, we compared with three existing ME methods: bidirectional ME in [1], OBME in [2], and dual ME in [3]. To experiment only on ME scheme, not for total MCFI scheme, we generated the interpolated frames with MVFs before motion vector refinement (MVR).

| Test sequence | Average PSNR (dB) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ref [4] | Ref [9] | Ref [10] | Ref [3] | Ref [11] | Ref [12] | Proposed |
| News (90) | 32.884 | 32.969 | 34.671 | 34.946 | 36.25 | 38.214 | **38.259** |
| Stefan (90) | 23.885 | 23.770 | 23.048 | 24.324 | 26.40 | **29.157** | 28.990 |
| Foreman (300) | 27.243 | 28.171 | 28.557 | 29.191 | 32.71 | 33.252 | **33.513** |
| Mother_daughter (300) | 37.405 | 36.607 | 38.105 | 38.726 | 41.45 | 42.708 | **43.103** |
| Mobile (300) | 21.401 | 22.541 | 22.549 | 23.113 | 26.18 | 28.994 | **29.682** |

Table 1. Average PSNRs of test video sequences for the proposed and existing MCFI methods.

The interpolated frames from three existing methods and proposed method are shown in Figure 1. According to the interpolated frames from *foreman* sequence, three existing methods produce noticeable artifacts on lips of the foreman. Because the foreman speaks through this sequence, there are some variances in the shape of his face. Therefore, it is hard to estimate reliable MVs from his lips and three existing methods fail to estimate accurate MVs on his lips. However, we can notice that the interpolated frames from the proposed method have no noticeable artifacts on his lips. Our proposed method uses the CNN which was trained to estimate similarity of a pair of image patches under variances in shape of the object, so it can estimate more reliable MVs on his lips.

According to the interpolated frame from *Stefan* sequence, other ME methods also generate noticeable artifacts on the tennis player's body. Because there are some rotational motions on the tennis player's foot, other ME methods fail to estimate reliable MVs on his foot. These incorrect MVs result in vanishment of his foot in the interpolated frame. In comparison with three existing ME methods, our proposed method successfully estimates MVs on his body. As shown in the interpolated frames from *mobile* sequence, our proposed ME method also estimates more reliable MVs.

### 3.2. Objective assessment of motion-compensated frame interpolation method

To objectively assess our proposed MCFI method, the average of PSNR over each video sequence was calculated and compared with other existing MCFI methods [3], [4], [9]-[12]. In our proposed method, MVR scheme in [13] was used to refine MVFs after ME.

As shown in Table 1, the proposed method outperforms other existing methods for *news*, *foreman*, *mother and daughter*, and *mobile* sequences and ranked 2nd for *Stefan* sequence. These results denote that the proposed method which uses the CNN in estimating MVs produces more reliable intermediate frames.

## 4. DISCUSSION AND CONCLUSION

In this paper, we proposed a new MCFI method that utilized the CNN model to estimate more reliable MVs. The CNN model which was used to estimate MVs was trained to decide whether a pair of image blocks was correct matches or not under the variance in shape of the object. Therefore, our CNN model correctly estimated MVs more accurately than the existing ME methods under the various distorted object environment for the benchmark video sequences. Experimental results showed that the proposed CNN-based ME method outperformed other methods in both objective and subjective results.

In conclusion, as we employ the CNN model to ME, the interpolated frame that has the increased visual quality can be obtained. If we use deeper CNN model to ME scheme, more increased performance of the interpolated frame may be obtained although the complexity becomes higher.

## 5. ACKNOWLEDGEMENT

This work was supported by the Samsung Electronics Company, a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry for Health and Welfare, Korea (HI14C1135), and Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2014M3C7033999).

## 6. REFERENCES

[1] B.-T. Choi, S.-H. Lee, and S.-J. Ko, "New frame rate up-conversion using bi-directional motion estimation," *IEEE Trans. Consumer Electron.*, vol. 46, no. 3, pp. 603–609, Aug. 2000.

[2] T. Ha, S. Lee, and J. Kim, "Motion compensated frame interpolation by new block-based motion estimation algorithm," *IEEE Trans. Consumer Electron.*, vol. 50, no. 2, pp. 752–759, May. 2004.

[3] S.-J. Kang, S. Yoo, and Y. H. Kim, "Dual motion estimation for frame rate up-conversion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1909–1914, Dec. 2010.

[4] B. D. Choi, J. W. Han, C. S. Kim, and S. J. Ko, “Motion compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, no. 4, pp. 407-416, Apr. 2007.

[5] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[6] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[7] X. Han et al., "MatchNet: Unifying feature and metric learning for patch-based matching," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[8] Y. Jia, et al., "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[9] Y. Ling, J. Wang, Y. Liu, and W. Zhang, “A novel spatial and temporal correlation integrated based motion-compensated interpolation for frame rate up-conversion,” *IEEE Trans. Consum. Electron.*, vol. 54, no. 2, pp. 863–869, May 2008.

[10] Y.-L. Lee and T. Nguyen, “Method and architecture design for motion compensated frame interpolation in high-definition video processing,” *in Proc. IEEE Int. Symp. Circuits Syst*., May 2009, pp. 1633–1636.

[11] U. S. Kim and M. H. Sunwoo, "New frame rate up-conversion algorithms with low computational complexity," *IEEE Transactions on Circuits and Systems for Video Technology*, 2014.

[12] S. Dikbas and Y. Altunbasak, "Novel true-motion estimation algorithm and its application to motion-compensated temporal frame interpolation," *IEEE Transactions on Image Processing* 22.8 (2013): 2931-2945.

[13] L. Alparone, et al., "Adaptively weighted vector-median filters for motion-fields smoothing." *Acoustics, Speech, and Signal Processing*, 1996.