Improving Global Motion Compensation for Frame Interpolation with High-Resolution and High-Frame-Rate Video

Keita Ukihashi†1 Takashi Imagawa†2 Hiroshi Tsutsui†3 Yoshikazu Miyanaga†3 Hiroyuki Ochi†1,2

†1 Graduate School of Information Science and Engineering, Ritsumeikan University
†2 College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, 525-8577 Japan
is0248he@ed.ritsumei.ac.jp, takac-i@fc.ritsumei.ac.jp, ochi@cs.ritsumei.ac.jp

†3 Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814 Japan

Abstract— In this paper, we propose a novel global motion compensation method to be used in frame interpolation from input video that consists of high-resolution less-frequent frames (keyframes) and low-resolution high-frame-rate (LR-HF) frames. To generate better-interpolated background from two keyframes using homography transformation, we improve the accuracy of global motion estimaion by eliminating and interpolating feature points (FPs) and by detecting erroneous homography matrix. We also introduce an adaptive weight model for superimposing transformed keyframes. The experimental results show that the proposed method achieves interpolated frames with better quality than the conventional one.

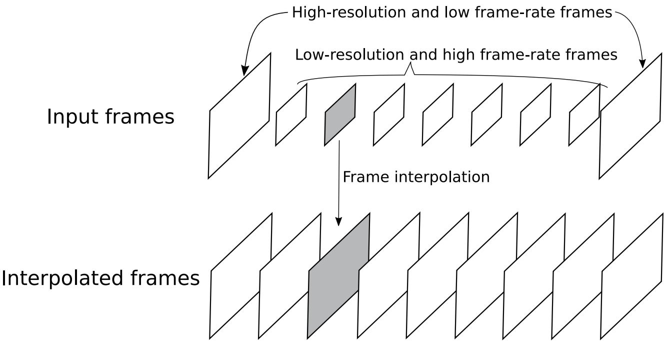
High-resolution and low frame-rate frames
Low-resolution and high frame-rate frames
Input frames
Frame interpolation
Interpolated frames

Fig. 1.: Frame interpolation for videos with two different resolutions and frame-rates [1]

I. Introduction

To realize a safe and secure society, the realization of a high-quality real-time surveillance camera network covering a wide area is desired. Collection of high-quality videos from multiple viewpoints using wireless communication is becoming realistic since technologies for camera devices and wireless communication devices have been matured and commoditized.

A large number of high-resolution high-frame-rate cameras, however, would overuse the limited bandwidth of the wireless network, and video compression at each camera node is a must. The advanced video compression such as H.265 reduces the amount of data while maintaining the quality of videos, but requires a high-performance and high-power processor, which is undesirable to be embedded in the sensor nodes. Therefore, high-quality and low-power video compression method should be introduced.

One simple method for video data compression is to reduce the resolution of the frames on the sensor node and to enlarge the resolution on the server. Another one is to reduce the frame rate of the video on the sensor node and to interpolate the frames on the server. Both methods require only a small processing effort on the sensor node to reduce the amount of data. Instead, a very large processing effort for restoring resolution or frame-rate is required on the server to maintain the video quality. This transfer of computational load from the sensor node to the server is compatible with the IoT system that assumes high-performance computing resource on the "cloud."

A recent paper [1] has proposed a hybrid video that consists of high-resolution low-frame-rate frames (keyframes) and low-resolution high-frame-rate (LR-HF) frames as shown in Fig. 1, and has developed a frame interpolation method for it. This method generates global images (or background) of interpolated frames by applying homography transformation to keyframes, while low-resolution frames are used to calculate the homography matrix and to compensate local motion for the foreground objects. The evaluation results suggested that the proposed frame interpolation method generates higher quality interpolated frames from the video of the same amount of data, as compared to the case when only the frame rate is reduced, or only the frame resolution is reduced.

To generate better-interpolated background from two keyframes using homography transformation, this paper improves the accuracy of global motion estimation by eliminating and interpolating feature points (FPs) and by detecting erroneous homography matrix. This paper also introduces an adaptive weight model for superimposing transformed keyframes.

The rest of this paper is organized as follows. Section 2 introduces the conventional frame interpolation method [1] and its problem. Section 3 describes proposed techniques to improve the accuracy of homography transformation and superimposing model of transformed frames. Section 4 evaluates the quality of the proposed methods and compares them to a conventional method. Section 5 concludes this paper.

II. Preliminary

A. Related Work

Our approach illustrated in Fig. 1 can be considered as a process of increasing (1) resolution of low-resolution high-frame-rate (LR-HF) frames, or (2) frame-rate of high-resolution low-frame-rate frames (keyframes).

Technologies for (1) is known as image super-resolution (SR) that recovers a high-resolution image from a low-resolution one. Recently, various deep-learning-based SR methods have been proposed and outperform conventional ones. There are also SR methods for videos that exploit information from the target frame as well as its neighboring frames. However, these methods often produce excessively smoothed and hallucinated images because of pixel loss and lack of high-resolution information [2].

Technologies for (2) is known as frame interpolation that generates new frames between given frames. Frame interpolation has also been evolved by deep-learning-based methods. However, these methods typically produce only one image between frames or assume a small time interval between frames [3] compared to our approach.

In contrast, since our approach utilizes both frames (LR-HF frames and keyframes), higher quality can be expected. Even better quality can be expected if we introduce the deep-learning-based methods to our approach, but the research on them is left for future work.

B. The conventional frame interpolation method for video with two different resolutions and frame-rates

There are two types of motion in videos. One is global motion, which is the apparent motion of the background caused by the motion of the camera. The other is local motion, which is the actual motion of the target objects (foreground). To improve the quality of interpolated frames, each of them need to be compensated by an optimal method according to its characteristics.

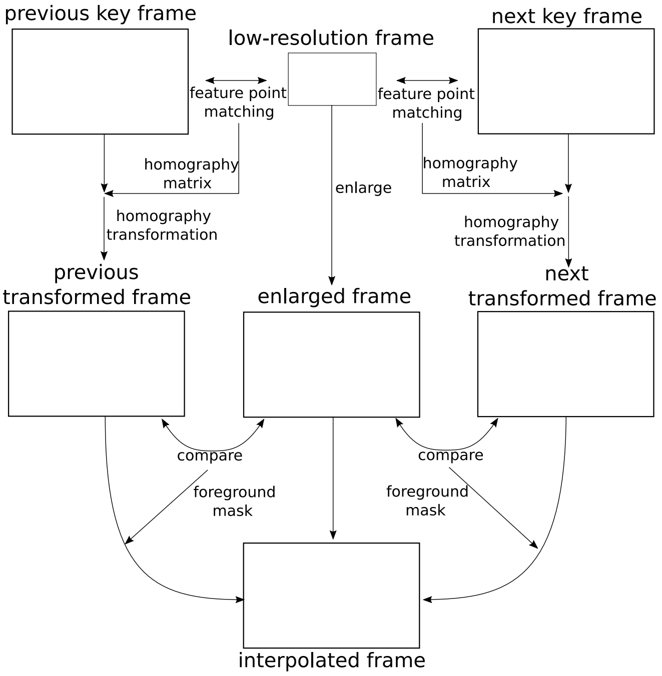

Fig. 2.: The frame interpolation method where local and global motion are compensated.

The input video of [1] consists of high-resolution low-frame-rate frames (keyframes) and low-resolution high-frame-rate (LR-HF) frames (Fig. 1). Global motion is compensated by homography transformation of keyframes. The homography matrix required for this transformation is calculated by feature point (FP) matching among an enlarged LR-HF frame of the target time and neighboring keyframes. The neighboring two keyframes are transformed by this procedure and superimposed to generate a high-resolution frame with global motion compensation. On the other hand, the local motion compensation uses the pixels of the enlarged LR-HF frame of the target time. To find the region where the local motion should be compensated, [1] calcurates the difference between the transformed keyframes and the enlarged LR-HF frame. Areas where the difference is large from both transformed keyframes are considered to require local motion compensation, and the pixel values are overwritten by those of the enlarged LR-HF frame. This procedure proposed in [1] is illustrated in Fig. 2.

C. Problems of conventinal interpolation method

C.1. Accuarcy of foreground detection

In [1], the foreground region where local motion compensation should be applied is extracted based on the difference among the transformed keyframes and the enlarged LR-HF frame. If the noise elimination is insufficient or the homography transformation is inaccurate, an unnecessarily large area is considered as the foreground, and lower-quality image from enlarged LR-HF frame is used for the

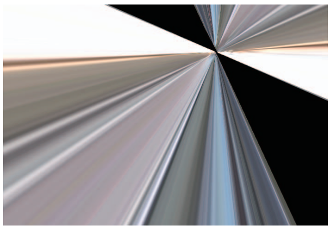

Fig. 3.: A collapsed keyframe transformed by an erroneous homography matrix.

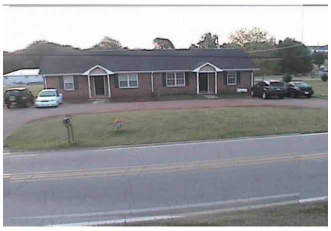

Fig. 4.: An interpolated frame made from misaligned keyframes. Two white lines are seen at the bottom.

interpolated frame instead of that from the transformed keyframes. Another previous work [4] introduced contour extraction by a watershed algorithm in order to improve the accuracy of foreground region extraction. In this paper, we use FgSegNet [5], which is a neural-network-based foreground region extraction method, to improve the accuracy of foreground region extraction.

C.2. Incorrect homography transformation

When there are many FPs on a large foreground, or when the number of background FPs is not large enough, an erroneous homography matrix is occasionally derived. In the former case, keyframes transformed by the erroneous homography matrix become collapsed, as shown in Fig. 3. In the latter case, the transformed frame is slightly misaligned from the original position, resulting in a double-exposure-like interpolated image, as shown in Fig. 4. These erroneous transformations generate background frame that is inconsistent to the enlarged LR-HF frame in a large portion. As a result, a large portion of the frame is treated as foreground, causing excessive use of the enlarged LR-HF frame, and a low-quality interpolated frame is generated.

Figure 5 plots the result of a preliminary evaluation of the quality of the generated interpolated frame when the foreground extraction method in the conventional method is simply replaced with FgSegNet. The red plot is the result when using FgSegNet while the green plot is that of the conventional method. While some frames achieve improvement in PSNR using FgSegNet, we can observe PSNR degradation in many frames. In particular, the PSNR degradation is remarkable in the frames with indexes around 950. These degradations are caused by the increased contribution of the pixels of the erroneously homomorphically transformed keyframe.

D. *A superimposing model for transformed keyframes*

In the conventional method, when superimposing transformed keyframes for global motion compensation, the operation is based on the following model.

$$I(x,y) = (1-\alpha)T_0(x,y) + \alpha T_1(x,y),$$

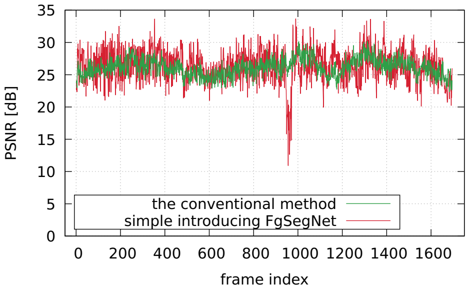

Fig. 5.: PSNR when simple introducing FgSegNet is applied.

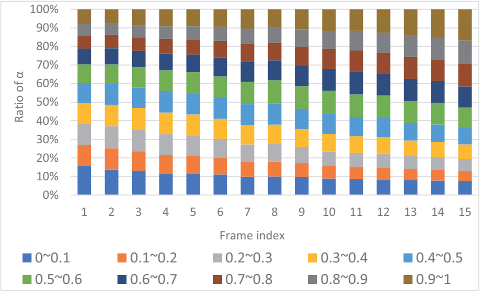

Fig. 6.: Distribution of the value of α in each interpolation.

where $T_0(x,y)$, $T_1(x,y)$ and $I(x,y)$ are pixel values at coordinates (x,y) in previous and next transformed keyframes, and interpolated frame, respectively. α is a weight value that takes a value between 0 and 1, and is determined by $\alpha = t/T$, where t is a relative index of target frame from previous keyframe and T is a total number of frames between two keyframes. In the conventional methods, α is constant in one interpolated frame.

We experimentally calculated the pixel-by-pixel optimal α using the two transformed keyframes and the original frame at each target time. In this experiment, the video consists of a repetition of one keyframe followed by 15 LR-HF frames. The distribution of α calculated for each LR-HF frame is shown in Fig. 6. From the results, we can observe that the model $\alpha = t/T$ is not appropriate. For example, by using $\alpha = t/T$, α in the 15th frame is $15/16 = 0.9375$. However, from Fig. 6, $\alpha > 0.9$ is at most 20%, and $\alpha = t/T = 0.9375$ is far from the experimentally calculated average, $\alpha = 0.5881$. Consequently, it is evident that the model $\alpha = t/T$ is inappropriate and constant α in one frame is also inappropriate.

III. Proposed interpolation methods

A. *Elimination of FPs from foreground*

In the conventional method, the foreground region is detected after transforming keyframes by a homography

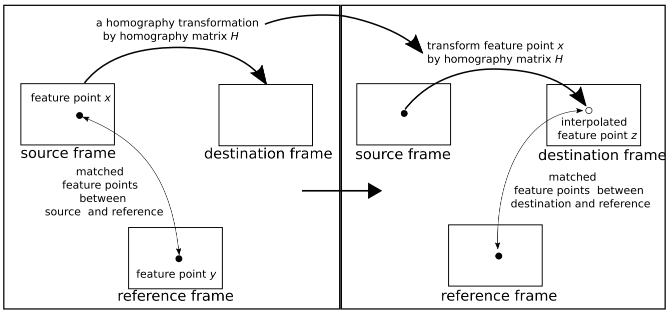

Fig. 7.: The concept of interpolation of FPs.

matrix. To calculate the homography matrix, all FPs including those in the foreground are used. Therefore, the calculated homography matrix is affected by local motion, degrading the accuracy of the homography transformation. In contrast, FgSegNet detects foreground region before calculating the homography matrix. This is beneficial to prevent local motion from affecting the homography matrix.

B. Detection of obviously erroneous transformation

To filter out an obviously erroneous homography transformation, we propose to use eigenvalues of a homography matrix as a key metric. Since the frame rate of LR-HF frames (e.g., 60fps) is high, we can assume that camera device's relative motion is very small, and we can consider that a frame is very close to a slight rotation and parallel shift of neighboring frames. In general, eigenvalues of a homography matrix corresponding to such transformation are close to 1.0. Therefore, if the eigenvalues are far from 1.0, we can consider that the induced transformation is erroneous. For simplicity, instead of using eigenvalues we use a trace value and a determinant value which are a sum and a product of eigenvalues, respectively. When an erroneous transformation is detected, we only use regions of the transformed frame where the difference to the enlarged LR-HF frame is small, as in [1].

C. Interpolation of FPs

To improve the accuracy of homography transformation, we propose to increase FPs by interpolation. Here, interpolation of FPs refers to transforming FPs in one frame to another frame by homography transformation. As Fig. 7 shows, we consider three frames, a source frame, a destination frame, and a reference frame. Our goal is to derive an accurate homography matrix between the reference frame and the destination frame when the destination frame has only a small number of FPs.

Assume that there are enough matched FPs between the source frame and the destination frame, an accurate homography matrix $\boldsymbol{H}$ can be calculated. Also, assume that there is an FP $\boldsymbol{y}$ in the reference frame that matches an FP $\boldsymbol{x}$ in the source frame and there is no FP in the destination frame that matches $\boldsymbol{x}$. Under these assumptions, there are not enough FPs matched between the destination frame and the reference frame, resulting homograhpy transformation between them will be inaccurate. In such a cese, we transform $\boldsymbol{x}$ from the source frame to the destination frame using $\boldsymbol{H}$ and generate a new FP $\boldsymbol{z}$ in the destination frame. By this operation, an FP in the destination frame which matches an FP in the reference frame is generated. Hence, using these FPs more accurate homography matrix can be calculated.

In this paper, we attempt to interpolate FPs between any two frames, out of two neighboring keyframes and all the LR-HF frames between them. Note that inappropriate interpolation of FPs can degrade the accuracy of homography transformation. We calculate the sum of absolute difference (SAD) of the transformed frame and the enlarged frame, and the interpolated FPs are accepted only when the SAD value is smaller.

D. A superimposing model of keyframes with adaptive weights

In a region where two transformed keyframes can be used for global motion compensation, pixel values in the interpolated frame are determined by a weighted average of the two frames. Using a constant α for a whole frame to generate an interpolated frame is inappropriate as mentioned in Section 2. The proposed model uses pixel-by-pixel α based on the difference of the transformed frame from the enlarged frame. In detail, α is calculated according to the following equation.

$$\begin{aligned} d_0 &= |E(x,y) - K_0(x,y)|\,,\ d_1 = |E(x,y) - K_1(x,y)|\,, \\ \alpha &= \frac{d_0}{d_0 + d_1}, \\ I(x,y) &= (1-\alpha)K_0(x,y) + \alpha K_1(x,y), \end{aligned} \tag{1}$$

where $E(x, y)$ is a pixel value in the enlarged frame at the coordination (x, y). In this model, the pixel value that is closer to the value in the enlarged frame is reflected in the interpolated frame more strongly.

IV. Evaluation

This section evaluates the quality of interpolated frames generated by the proposed methods. The video used for the evaluation is ContinuousPan in CDnet2014 Dataset [6]. From this video, we generate the sequence of frames consisting of keyframes and LR-HF frames. Specifically, it consists of a repetition of one keyframe followed by 15 LR-HF frames. The LR-HF frames are generated by discarding 3/4 of rows and 3/4 of columns so that both their height and width are reduced to 1/4.

We evaluate each method proposed in Section 3 as well as the method where all of them are applied, by means of PSNR. Average PSNR, maximum PSNR, minimum PSNR and runtime of each method are summarized in Table I. Each method is implemented with C++ and OpenCV 3.3.0.

A. *Elimination of FPs from foreground*

As row (b) and (c) of Table I shows, the average PSNR is slightly improved by eliminating FPs in the foreground region. This result suggests that the accuracy of homography transformation can be improved as expected when there are enough FPs in the background region. In contrast, the minimum PSNR is degraded because the FPs in the foreground is inadequately eliminated and/or there are not enough FPs in the background region. Although minimum PSNR is slightly degraded, we found that elimination of FPs from foreground can be more effective by combining with interpolation of FPs.

B. *Detection of obviously erroneous transformation*

The comparison of PSNR between (b) interpolated frames with only introducing FgSegNet and (d) those additionally with detection of erroneous transformation is shown in Table I and Fig. 8. From Fig. 8, the degradation of PSNR caused by obviously erroneous transformation at frame index around 950 is successfully eliminated. When applying the detection of erroneous transformation, the minimum PSNR in Table I is significantly improved and the average PSNR is improved as well.

The determinants of homography matrices for previous and next keyframes are shown in Figs. 9 and 10, respectively, in comparison with PSNR of interpolated frames with FgSegNet. According to these figures, the determinants fluctuate considerably when PSNRs are degraded. Consequently, the detection of erroneous transformation by the determinant works as expected. Since the trace have a similar tendency, we omitted its plot.

However, even with this detection method, the minimum PSNR is still lower than that of the conventional one. It is because this detection method is limited to detecting completely incorrect case and partially incorrect transformation cannot be detected. With this detection method, the minimum PSNR is observed at an interpolated frame that has a transformed frame with partially incorrect transformation.

C. *Interpolation of FPs*

The comparisons of PSNR between (b) interpolated frames with only introducing FgSegNet and (e) those additionally with interpolation of FPs are shown in Table I and Fig. 11. The PSNR values of many frames are successfully improved. From Table I, the average PSNR is also better than that of the conventional method. The minimum PSNR is observed at the frame that induces erroneous transformation. So the detection of erroneous transformation and interpolation of FPs are expected to work complementarily.

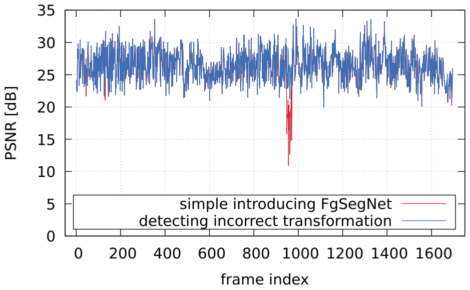

Fig. 8.: PSNR when detecting incorrect transformation is applied.

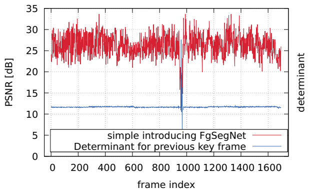

Fig. 9.: Determinants for previous key frames.

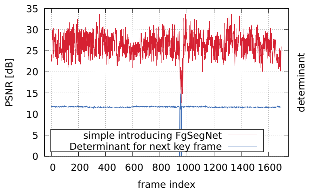

Fig. 10.: Determinants for next key frames.

D. *A superimposing model of key frames with adaptive weights*

The comparison of PSNR between (b) interpolated frames with only introducing FgSegNet and (f) those additionally with an adaptive superimposing weight α is shown in Table I and Fig. 12. From Table I, improvement of the average and the minimum PSNR are observed, and from Fig. 12, many frames get better PSNR. The frames which are especially improved are those where erroneous transformation occurs. This is because when one of the keyframes is correctly transformed and the other is erroneously transformed, the difference of pixel values between the correctly transformed frame and the enlarged frame is not significant, but that between incorrectly transformed frame and the enlarged frame are large. Hence, pixel values of the correctly transformed

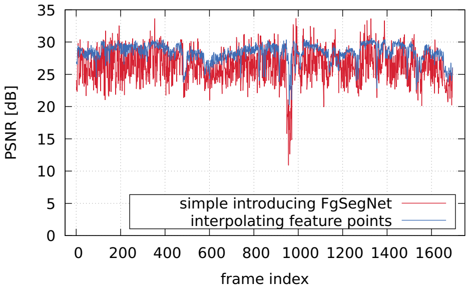

Fig. 11.: PSNR when interpolating FPs.

TABLE I

: PSNRs of each method.

	PSNR (dB)			runtime (s)
	avg	max	min	
(a) conventional method	26.10	30.21	22.33	432
(b) (a) + FgSegNet	26.30	33.68	10.91	621
(c) (b) + eliminating FPs from foreground	26.33	33.68	10.41	621
(d) (b) + detecting erroneous transformation	26.44	33.68	19.93	621
(e) (b) + interpolating FPs	28.35	30.60	20.20	3117
(f) (b) + using adaptive superimposing weights	26.67	32.80	14.68	627
(g) applying all the proposed methods	28.40	30.79	22.86	3137

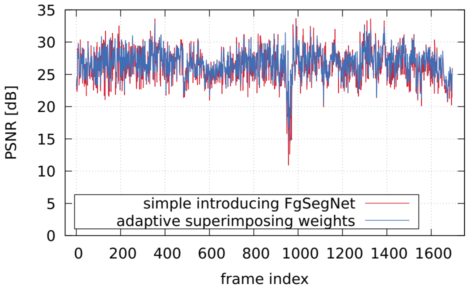

Fig. 12.: PSNR when adaptive superimposing weights are applied.

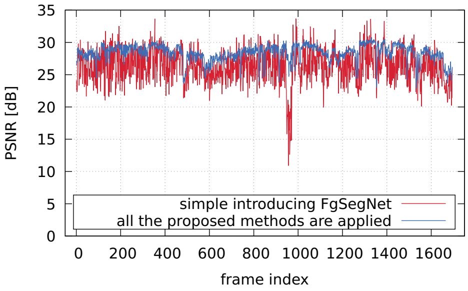

Fig. 13.: PSNR when all the proposed methods are applied.

frame are strongly reflected to the interpolated frame.

E. All the proposed methods are applied

The comparison of PSNR between (b) interpolated frames with only introducing FgSegNet and (g) those with all the proposed techniques is shown in Table I and Fig. 13. According to Table I, the average PSNR and the minimum PSNR are the best of the entries in Table I. The influence of the erroneous transformation is eliminated and many frames are improved. However, the maximum PSNR is degraded compared to (b). This is because incorrect interpolation of FPs induces inappropriate homography transformation.

V. Conclusion

In this paper, we have proposed four methods to improve the accuracy of global motion compensation in frame interpolation whose input video consists of keyframes and LR-HF frames. The elimination of FPs on the foreground region, detection of inappropriate homography matrix, and interpolation of FPs on the background region improves the accuracy of homography transformation of high-resolution key frame. We also introduce an adaptive weight model for superimposing of transformed key frames while a constant weight is used for each composition. The experimental results show the PSNR values of interpolated frames are improved when all the proposed methods are applied.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 26870009.

References

[1] H. Uesaka, T. Imagawa, H. Tsutsui, and Y. Miyanaga, "An Accuracy Evaluation of Motion-Compensated Frame Interpolation Using High-Resolution Video and High-Frame-Rate Video," in *In Proc. of International Workshop on Smart Info-Media Systems in Asia (SISA)*, Sep. 2016, pp. 131–135.

[2] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *ArXiv*, vol. abs/1902.06068, 2019.

[3] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. G. Learned-Miller, and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," in *In Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

[4] H. Ihara, T. Imagawa, H. Tsutsui, Y. Miyanaga, and H. Ochi, "A Study on Quality Improvement of Frame Interpolation Method with High-Resolution and High-Frame Rate Video Using Foreground Elimination and Contour Extraction," *VLD2017-98*, vol. 117, no. 455, pp. 55–60, Feb. 2018.

[5] L. A. Lim and H. Y. Keles, "Learning Multi-scale Features for Foreground Segmentation," *arXiv:1808.01477*, pp. 1–11, 2018.

[6] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An Expanded Change Detection Benchmark Dataset," in *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 393–400.