Texton-Based Super-Resolution for Achieving High Spatiotemporal Resolution in Hybrid Camera System

Kenji Kamimura*, Norimichi Tsumura[1], Toshiya Nakaguchi[1], and Yoichi Miyake[2]

Graduate School of Science and Technology, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

[1]*Graduate School of Advanced Integration Science, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan*

[2]*Research Center for Frontier Medical Engineering, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan*

Many super-resolution methods have been proposed to enhance the spatial resolution of images by using iteration and multiple input images. In a previous paper, we proposed the example-based super-resolution method to enhance an image through pixel-based texton substitution to reduce the computational cost. In this method, however, we only considered the enhancement of a texture image. In this study, we modified this texton substitution method for a hybrid camera to reduce the required bandwidth of a high-resolution video camera. We applied our algorithm to pairs of high- and low-spatiotemporal-resolution videos, which were synthesized to simulate a hybrid camera. The result showed that the fine detail of the low-resolution video can be reproduced compared with bicubic interpolation and the required bandwidth could be reduced to about 1/5 in a video camera. It was also shown that the peak signal-to-noise ratios (PSNRs) of the images improved by about 6 dB in a trained frame and by 1.0–1.5 dB in a test frame, as determined by comparison with the processed image using bicubic interpolation, and the average PSNRs were higher than those obtained by the well-known Freeman's patch-based super-resolution method. Compared with that of the Freeman's patch-based super-resolution method, the computational time of our method was reduced to almost 1/10.

KEYWORDS: super-resolution, texton, video, wavelet transform

*E-mail adress: kamimura@auone.jp

1. Introduction

High-resolution videos are desired for image processing application, photometry, surveillance, TV broadcast, and also personal use. For example, in Japan, the TV broadcast will completely change from the analog NTSC TV to the digital high-definition TV (HDTV; 1440×1080) in 2011. The increase in image resolution brings outstanding experience to us; however, it requires a high bit rate of a reading pixel to record a vast amount of information. We consider that the high bit rate of a reading pixel is difficult for both network use and the next-generation HD camera. For example, a $4K \times 2K$ resolution camera requires a bit rate that is 26 times higher than that of an standard-definition (SD) camera. Therefore, the development of a resolution enhancement technology is required to solve these problems. Image interpolation techniques, such as bicubic interpolation, are the most common and simple methods for resolution enhancement. However, they cannot recover high-frequency (detail) information on certain structures, such as edges and textures. Super-resolution is the term generally applied to the problem of transcending the limitation of an imaging system with image processing.[1] If we can recover high-frequency information by super-resolution, we can reduce the amount of data (also bandwidth) required. Since super-resolution is an ill-posed problem, we need regularizations to solve the problem (to limit the answer space).[2] There are generally two manners to construct a regularization for super-resolution; (1) use a high-accuracy registration of many input images[3] and (2) use (example-based) training about a subject and a single input image.[4–6]

Elad and Datsenko showed that the example-based regularization is more effective for super-resolution.[2] Furthermore, for a video sequence, the use of the example-based method is preferred to maintain the frame rate.[7,8] Since the video sequence consists of many frames, lower computational cost is also necessary for practical use. However, most of the existing example-based super-resolution methods use patch-based feature matching that requires a high computational cost to determine the appropriate patch from the database and to maintain consistency between patches.

To reduce computational cost, we have already proposed a simple pixel-based super-resolution method named "texton substitution" for a texture image.[9] Our method is

based on the pixel-based feature "textons" for texture recognition proposed by Julesz[10] and Leung and Malik.[11] We modified and simplified Leung and Malik's concept for super-resolution. The texton is simply substituted in the learned manner to obtain a high-resolution output texture image, and this super-resolution process does not have iteration steps. This could reduce computational cost compared with a conventional method.

However, since our target was only the texture image in a previous paper, there are still problems to resolve to achieve super-resolution for a video sequence; e.g., how to obtain the training data, how to process a multiobject scene (since conventional texton substitution did not consider multiple textures), and how to utilize temporal information.

Thus, we improved texton substitution in our SIGGRAPH posters.[12–14] These posters did not include details of the algorithm and evaluation. Furthermore, these posters still leave problems; (1) we did not address the method for obtaining the training data and (2) the use of temporal information was not suitable for our texton substitution. Therefore, in this study, we combine texton substitution with a hybrid camera system, which can obtain training data from the target scene itself. The texton substitution method is improved to make it suitable for a hybrid camera using a modified texton, which involves the utilization of temporal information from the look-up table (LUT). In §2, we briefly explain conventional texton substitution. In §3, we describe the hybrid camera system and the modification of the texton, and we introduce temporal information to the texton substitution. In §4, we describe the experiment using the simulated camera, and finally, in §5, we present our conclusions.

2. Texton Substitution

Texton substitution is a training-based method; thus, it consists of two phases, namely, training and inference.[9] The overall texton substitution is illustrated in Fig. 1. In the texton substitution, we utilize a pixel-based feature named "texton", which comes from Leung and Malik's "texton" concept with a small modification.

2.1 Texton

Leung and Malik analyzed an image using a filter bank, which is a set of orientation and spatial-frequency selective linear filters.[11] The responses yield a feature vector at each pixel, which is useful in many image processing applications; however, the vector is generally overly redundant. Therefore, Leung and Malik applied clustering to the vectors to obtain a small set of prototype response vectors. They called the resultant prototype vectors "textons".[11] For our super-resolution application, the vector is still redundant and requires a high computational cost. Thus, we generate a texton with frequency information originating from stationary wavelet transform (SWT)[15] instead of a filter bank.

A diagram of our texton generation is shown in Fig. 2. In this paper, since it is well known that the resolution of the human eye is less sensitive in the chrominance channel than in the luminance channel, we consider only the luminance channel (Y) in texton generation. We transform RGB color values to luminance (Y) - color difference signal (Cb, Cr) values with the following equation:

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{pmatrix} 0.2989 & 0.5866 & 0.1145 \\ -0.1687 & -0.3313 & 0.5000 \\ 0.5000 & -0.4187 & -0.0813 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{2.1}$$

To store and utilize the frequency information of an image, we analyze the Y signal using 2D SWT. We decompose an image into three resolution layers using the Haar kernel, and each layer has three orientation properties, namely, horizontal, vertical, and diagonal details for spatial frequency. The residual approximation (low-frequency) component in SWT is eliminated, since low-frequency information is already contained in the low-resolution image. Thus, we obtain a nine-dimensional coefficient vector at each pixel. The vectors are clustered using the K-means technique to create a small set of prototype feature vectors (the number is K) called textons.

2.2 Training phase

First, we prepared a pair of high- and (degraded) low-resolution images for training. The high-resolution image yields high-resolution textons by the method described above,

and then, every pixel in the high-resolution image is approximated by a texton. In other words, each pixel is represented by the label of the texton, and we call it the texton image. The low-resolution image is represented in the same manner. As a result, we can obtain a pair of high- and low-resolution texton images.

Since the low-resolution image is generated from the high-resolution image, both images are completely aligned. Thus, texton relationships between the high- and low-resolution images can be obtained using the spatial distribution of the pixel and are stored in the LUT as the "substitution table" (i.e., the upper part of Fig. 1). If one low-resolution texton is related to a certain high-resolution texton, the substitution table records all of these relationships with their frequencies of occurrence.

2.3 Inference phase

Since texton substitution is a pixel-based method and we utilize the pixel-to-pixel correspondence in the substitution table, we increase the pixel number of the input image to a target resolution by preprocessing. Then, the input image is decomposed in the same manner as in the training phase. The obtained coefficient vectors of the input image are classified into the trained low-resolution textons and produce a low-resolution texton image. Each pixel of the low-resolution texton image is substituted on the basis of the trained substitution table. Each pixel of the substituted image has the label of a high-resolution texton. From the training phase, we have correspondences between labels and mean vectors (prototype vectors of a high-resolution texture), and thus, the label image is transformed into a wavelet coefficient vector image by substituting labels to center vectors. The obtained coefficient vector image is processed by inverse stationary wavelet transform and the inverse of eq. (2.1) to produce a super-resolution output.

If we found multiple candidates in the substitution table, we simply choose the mode value of candidates. If an appropriate texton is not found in the trained database, we cannot perform any processing to the pixel since texton substitution is a training-based method. In this case, the input texton is not substituted, and thus, the input pixel value is transferred into an output pixel value without any change. In other words, the resultant and input pixels have the same values.

3. Texton Substitution for Video

3.1 Camera system

To obtain a training image pair in a video sequence, we consider a hybrid camera system proposed as a dual-sensor camera.[16–18] The basic idea is utilizing a multicamera to capture two sequences with different spatiotemporal resolutions.[19] Figure 3 shows our assumed camera system. The incident light is divided by a beam splitter to supply light simultaneously to two video cameras with different characteristics. Camera H has a high-spatial-resolution image sensor with a low frame rate (i.e., low-temporal resolution). On the other hand, camera L has a high frame rate with a low spatial resolution. Since these two cameras have the same optical axis, the obtained two video sequences can be registered with a high accuracy. The registered two sequences have a common capturing timing with a uniform interval, and the frames captured with the same timing act as training frames in our method.

Super-resolution processing is applied to a low-spatial-resolution sequence captured by camera L using the information from camera H. In this case, the bit rate of a reading pixel is simply the sum of the two cameras and is lower than that of a true high-resolution camera. For instance, if we use a 720×480, 30 frames per second (fps) sequence and a 1920×1080, 3 fps sequence, we reduce the reading rate by about 1/3.4.

3.2 Improved texton

In a video sequence, the lighting condition depends on spatiotemporal factors, such as shadow, weather, and color temperature. A low light intensity brings about a reduction in image contrast, resulting in different wavelet coefficients in feature vectors. In our previous method, we did not consider these contrast effects since our previous method was developed for the super-resolution of a simple texture image. Thus, if we process a complex scene with the previous method, we have to train all possible contrast patterns and have to maintain a huge database of textons specified for the scene. To eliminate the effect of image contrast from feature vectors, orientation frequency components are normalized using the low-frequency component (approximation component) in SWT, resulting in normalized textons.

Since texton substitution was designed for a texture image, the training and inference should be achieved with respect to each texture (object). Thus, for a video frame, it is preferred to divide the image into objects as a preprocessing step. There are many existing object recognition methods; however, they are very complex and difficult to implement for our super-resolution. Therefore, in this paper, we consider the color information of the image. As mentioned above, the color image is transformed into a luminance image and two chrominance images in texton substitution. Although these chrominance images are wasted in our conventional method, here we put them into feature vectors, resulting in eleven-dimensional feature vectors (nine normalized frequency components and two chrominance components).

In the substitution table, since we have many correspondences from a single low-resolution texton to multiple high-resolution textons due to the lack of a high-frequency component in the low-resolution image [Fig. 4(a)], the substitution from low- to high-resolution textons cannot give a unique solution. To obtain a unique solution, we adopted the mode value for substitution in the previous method. However, this results in spatial inconsistency in the output image (noisy image), since it obviously brings inappropriate substitutions even in the case of a training frame. The direct approach for increasing the number of unique solutions is to use either a larger filter kernel or more resolution layers in the low-resolution image for wavelet decomposition. However, these methods increase the calculation cost and come close to the original Leung's texon. Therefore, in this paper, we use the spatial connections of the textons in the low-resolution image to obtain a unique solution, as illustrated in Fig. 4(b). A similar idea is found in texture synthesis to enforce spatial consistency.[20] The texton of each pixel is connected to four adjacent pixels to construct a crossed texton. These connections increase the number of apparent varieties of the texton. The substitution table is constructed using the relationship between the low-resolution crossed and high-resolution textons. While straightforward feature vector connections become 55-D real value vectors, we address these texton connections as texton label connections that become 5-D integer vectors. The recognition of these 5-D vectors can be implemented using a five-stage (or 5-D) LUT and the calculation cost of texton connection becomes relatively lower than that of straightforward implementation. This

spatial connection increases the number of LUT varieties by a factor of 20–30.

3.3 Temporal treatment

In the case of video super-resolution, as successive frames have similar data, we can use this temporal similarity to improve output quality. In our substitution method, temporal information is considered as LUT with respect to temporal changes of textons. However, in our assumed camera system described in §3.1, the image pairs for training are obtained with specific intervals, not obtained in two successive frames. Thus, we have to train the temporal changes from a training frame pair. To achieve this training, we assume that temporal changes of textons are identical to spatial changes. Under this assumption, we consider a neighborhood region of size $m \times m$ pixels at each pixel location on the training pair. Figure 5(a) shows the training for the temporal LUT with $m = 5$. The center-to-neighbor relations of the low-resolution neighborhood region (indicated by a dotted circle (center) and a dotted triangle/rectangle (neighbor)) and their corresponding high-resolution relations (indicated by a continuous line) are stored in the temporal LUT.

At the inference phase, if we cannot determine the output from the substitution table, we check the previous frame and obtain a high- and low-resolution texton pair at the same spatial location. By considering this pair as the center pair and the low-resolution texton of the current frame as the low-resolution neighbor, we search the appropriate entry from the temporal LUT. If the entry is found, we set the high-resolution neighbor texton to the corresponding pixel of the output texton image [Fig. 5(b)]. In other words, we use the intraframe training of texton transition to achieve interframe substitution. Note that we do not use any motion vector; thus, we need not estimate the motion vector in the super-resolution process.

4. Experimental Methods

To confirm the effectiveness of the proposed method, we performed super-resolution processing using the simulated camera. As mentioned above, our method requires a hybrid camera system for input; thus, we simulated that condition by synthesizing two sequences from the original 720×480, 30 fps video sequence. For a low-resolution (high-frame-rate) camera, we downsized each frame to 180×120 pixels by 4×4 binning operation

(averaging 16 adjacent pixels). The binning operation is common and available for many cameras. For a high-resolution (low-frame-rate) camera, we obtained a high-resolution sequence with 3 fps by decimating frames from the original sequence.

We trained for every 10 frames using the obtained high- and low-resolution pairs to achieve super-resolution processing to a low-resolution sequence. Since texton substitution utilized pixel-based correspondence, we enlarged the low-resolution input to the size of the high-resolution sequence by bicubic interpolation as preprocessing.

Under this condition, the increase ratio of pixels is calculated to be approximately 5.5 $[720 \times 480 \times (10-1)]/[180 \times 120 \times 10 + 720 \times 480]$. The increase ratio of SDTV to HDTV is approximately 5.1 ($[1440 \times 1080]/[640 \times 480]$).

For stationary wavelet transform, we used the Haar wavelet as a mother wavelet. The number of textons was 5000 at low and high resolutions. The final size of the spatial LUT was increased to about 100,000 - 300,000 by spatial connections (depends on the image). The temporal LUT has 30,000 - 100,000 entries (also depends on the image).

For better understanding, we compare our super-resolution method with well-known bicubic interpolation and Freeman's super-resolution method.[5] The features and differences among our method and these methods are briefly shown in Table 1.

Figure 6 shows the results of super-resolution of the training frame; (a) shows the result of bicubic interpolation, (b) that of the true high-resolution frame, (c) that of Freeman's method, and (d) that of our texton substitution method. In this frame, we have both high- and low-resolution input images; thus, the results show only the quality of the texton representation (patch representation for Freeman's method). Since we represented the image by 5000 textons, the output has a small degradation compared with the high-resolution image. However, the output has a much higher quality than the result of bicubic interpolation and has a similar quality to the result of Freeman's method. The PSNRs are calculated against the true high-resolution image and shown with each image. Since we apply super-resolution to the luminance (Y) image only, the PSNRs are calculated on the luminance channel.

Figure 7 shows the results for the test frame (3 frames after the training frame). These results show that our method can significantly reproduce fine details such as the woman's

eye, compared with bicubic interpolation and that the sharpness of the entire image is also improved. Freeman's method reproduced fine details; however, it produced a large artifact in the results. Although our texton substitution could not improve small features compared with Freeman's method, texton substitution did not produce such a large artifact and the PSNR was higher than the result of Freeman's method.

Both texton substitution and Freeman's method cannot improve the details of a fast moving object (the cognac bottle in this experiment) compared with a slow object (woman's face). Since we trained the relationships between the high- and low-resolution images, we can handle only blur caused by resolution degradation. The bottle not only exhibits resolution degradation but also motion blur, indicating that the training-based super-resolution fails in such regions.

Figure 8 shows PSNRs of the entire image for 30 successive frames. The PSNRs improved by about 6 dB on training frames and by about 1 dB on test frames. Such improvement reduces with respect to temporal distance to the training frames, since the contents of these frames are different from the trained contents. Freeman's method has higher PSNRs in the training frames; however, in the test frames, it has a lower PSNRs than ours, and in some frames, it has a lower PSNRs than bicubic interpolation.

Since the results have large PSNR gaps between the training frame and the target frames, the observers could detect a temporal flicker artifact. Our method (and also Freeman's method) is mainly an intraframe approach, and we currently cannot handle such a temporal flicker artifact. To measure and reduce such an artifact, we will introduce human visual characteristics in the spatiotemporal domain[21] and structural similarity measurements[22] in our future work. These methods could be used to reduce small artifacts (noise) mainly caused by substitution error.

Figure 9 shows the results of another sequence, which captures a yachting scene. The quality of the training frame is sufficient, and here we showed the result of the test frame only. In this sequence, the camera tracked the yacht; thus, the motions in the sequence are mainly translations. Under such conditions, Freeman's method markedly improves the image (similar to a high-resolution image), while texton substitution slightly improves the image compared with the case of bicubic interpolation. The PSNRs of Freeman's method

are not markedly improved owing to small shifts in spatial position and luminance.

Figures 10-12 show results for other sequences and Fig. 13 shows PSNRs of yachting, driving, a Japanese room, and square sequences. Since they have the same trend in all the training frames (10-frame interval), we first showed 10 frames only. All sequences have the same trend as the two previous examples. Freeman's method is strong for translation motion and has a higher quality than texton substitution in a successful case. However, this method causes a large artifact in a failure case and thus in the average PSNR. Texton substitution is better than Freeman's method except for the yachting sequence (which has translational motion).

Since our method uses an 11-dimensional vector as a texton, its computational cost is much lower than that of Freeman's method, which uses a 174-dimensional vector when using the parameter described by Freeman and coworkers in a previous paper. In our brute force implementation, the computational time is almost 1/10 that of Freeman's method. In addition, texton substitution has no dependence on other pixels and thus can be accelerated by parallel processing, while Freeman's method requires scan line order processing to maintain consistency between patches (which cannot be paralleled). Once input feature vectors are classified into low-resolution textons, high-resolution feature vectors are immediately obtained by the trained LUT. This indicates that the computational time of inference is determined by classification to textons. Thus, if we approximate classification by LUT, a real-time output will be obtained using the trained substitution table.

5. Conclusions

In this paper, we proposed a simple and efficient training-based method to obtain a super-resolution image. Experimental results show that our method has a higher quality than bicubic interpolation. Our method is ten times faster than Freeman's super-resolution method and has higher PSNRs except for the yachting sequence with translational motion (Freeman's method is suitable for the translational motion). However, the image quality seems visibly slightly lower than that of Freeman's method since Freeman's method has a much higher quality in a successful case. On the other hand, in a failure case, Freeman's method causes a huge artifact; thus, the total PSNRs are higher in our

method. In our proposed method, the temporal changes of a video sequence were estimated using the spatial information of the training frame. Therefore, the inconsistency of the texton change along the temporal domain can cause a flicker. In our experiment, we used a binning operation to downsize the frame. As is well known, the binning operation induces an aliasing effect, which seems to negatively affect our algorithm. However, the effect of aliasing would be suppressed by the following two mechanisms: (1) If one pixel is classified to an incorrect texton by aliasing, the spatial connections of textons act to keep the consistency among neighboring pixels. (2) The aliasing effect is enhanced only in the low-resolution image, not in the high-resolution image and texton, and thus not in the resulting image. If the camera has the function of prefiltered down sampling, one should use it to eliminate aliasing. To improve image quality, it is necessary to consider the consistency of the texton and to introduce the object-based approach based on the segmentation of the object or create a substitution table that depends on spatial position. Furthermore, it will also be important to develop physical parameters that can be used to measure the spatiotemporal quality of a video image.

Acknowledgements

This work was partially supported by a Grant-in-Aid for JSPS Fellows. Norimichi Tsumura is partially supported by a Grant-in-Aid for Scientific Research (No. 19360026) from the Japan Society for the Promotion of Science. The authors received generous support from Hideto Motomura (Panasonic Corporation).

References

1) S. Farsiu, D. Robinson, M. Elad, and P. Milanfar: Int. J. Imaging Syst. Tech. **14** (2004) No. 2, 47.

2) M. Elad and D. Datsenko: Comput. J. **50** (2007) No. 4, 1.

3) D. Capel and A. Zisserman: IEEE Signal Process. Mag. **20** (2003) 75.

4) W. T. Freeman, E. C. Pasztor, and O. T. Carmichael: Int. J. Comput. Vision **40** (2000) No. 1, 25.

5) W. T. Freeman, T. R. Jones, and E. C. Pasztor: IEEE Comput. Graphics Appl. **22** (2002) No. 2, 56.

6) S. Baker and T. Kanade: IEEE Trans. Pattern Anal. Mach. Intell. **24** (2002) No. 9, 1167.

7) M. C. Bishop, A. Blake, and B. Marthi: Proc. 9th Conf. Artificial Intelligence and Statistics, 2003, published on CD-ROM and online.

8) D. Kong, M. Han, W. Xu, H. Tao, and Y. Gong: Proc. British Machine Vision and Applications, 2003, p. 197.

9) K. Kamimura, N. Tsumura, T. Nakaguchi, T. Sugaya, and Y. Miyake: Eizo Joho Media Gakkaishi **60** (2006) No. 10, 1655 [in Japanese].

10) B. Julesz: Nature **290** (1981) No. 12, 91.

11) T. Leung and J. Malik: Int. J. Comput. Vision **43** (2001) No. 1, 29.

12) K. Kamimura, T. Nakaguchi, N. Tsumura, H. Motomura, K. Kanamori, and Y. Miyake: presented at ACM SIGGRAPH Posters, 2005.

13) K. Kamimura, N. Tsumura, T. Nakaguchi, Y. Miyake, H. Motomura, and K. Kanamori: presented at ACM SIGGRAPH Research posters, 2006.

14) K. Kamimura, N. Tsumura, T. Nakaguchi, Y. Miyake, and Hideto Motomura: presented at ACM SIGGRAPH Posters, 2007.

15) G. P. Nason and B. W. Silverman: *Wavelets and Statistics* (Springer, New York, 1995) Vol. 103, p. 281.

16) H. Nagahara, A. Hoshikawa, T. Shigemoto, Y. Iwai, M. Yachida, and H. Tanaka: Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance, 2005, p. 450.

17) T. Shigemoto, A. Hoshikawa, H. Nagahara, Y. Iwai, M. Yachida, and T. Suzuki: IPSJ Trans. Comput. Vision Image Media **47** (2006) No. SIG5, 35.

18) K. Watanabe, Y. Iwai, H. Nagahara, M. Yachida, and T. Suzuki: IEICE Trans. Inf. Syst. **E89-D** (2006) No. 7, 2186.

19) T. Q. Pham and L. J. van Vliet: Proc. SPIE **6077** (2006) 607708.

20) M. Ashikhmin: I3D'01: Proc. 2001 Symp. Interactive 3D Graphics (ACM, New York, 2001) p. 217.

21) D. H. Kelly: J. Opt. Soc. Am. **69** (1979) No. 10, 1340.

22) Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli: IEEE Trans. Image Process. **13** (2004) No. 4, 600.

Figure captions

Fig. 1. (Color online) Overview of texton substitution. We modified the Leung and Malik's "texton" to make it suitable for our super-resolution application. Low-resolution textons are substituted using the trained substitution table to create a high-resolution output.

Fig. 2. (Color online) Diagram of texton generation.

Fig. 3. Hybrid camera system.

Fig. 4. (Color online) Connection of textons.

Fig. 5. Temporal treatment for texton substitution.

Fig. 6. (Color online) Results of the super-resolution for training frame (closed data).

Fig. 7. (Color online) Results of the super-resolution for test frame (three frames after training frame)

Fig. 8. PSNR comparison of the super-resolution.

Fig. 9. (Color online) Results for "yachting" sequence (three frames after training frame).

Fig. 10. (Color online) Results for "driving" sequence (three frames after training frame).

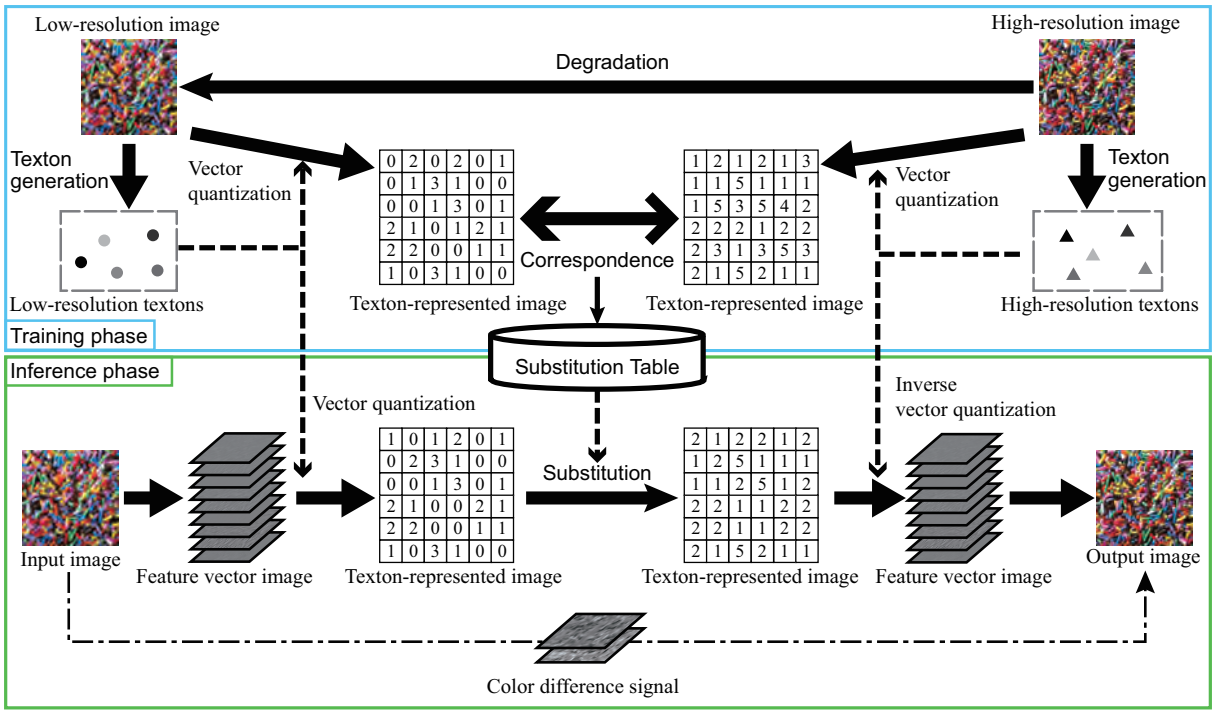
Fig. 11. (Color online) Results for "Japanese room" sequence (three frames after training frame).

Fig. 12. (Color online) Results for "square" sequence (three frames after training frame).

Fig. 13. PSNR comparisons for various sequences.

Table 1. Comparison of resolution enhancement methods.

	Our method	Freeman's method	Bicubic interpolation
Feature vector	Pixel-based	Patch-based	Intensity itself
Dimensions of feature vector	11	174	1
Paralell processing	Possible	Impossible	Possible
Recovery of high-freq. detail	Possible	Possible	Impossible
Implementation	Search, LUT	Search	Calculation



*The number in a texton-represented image is the label of the texton.

Fig. 1

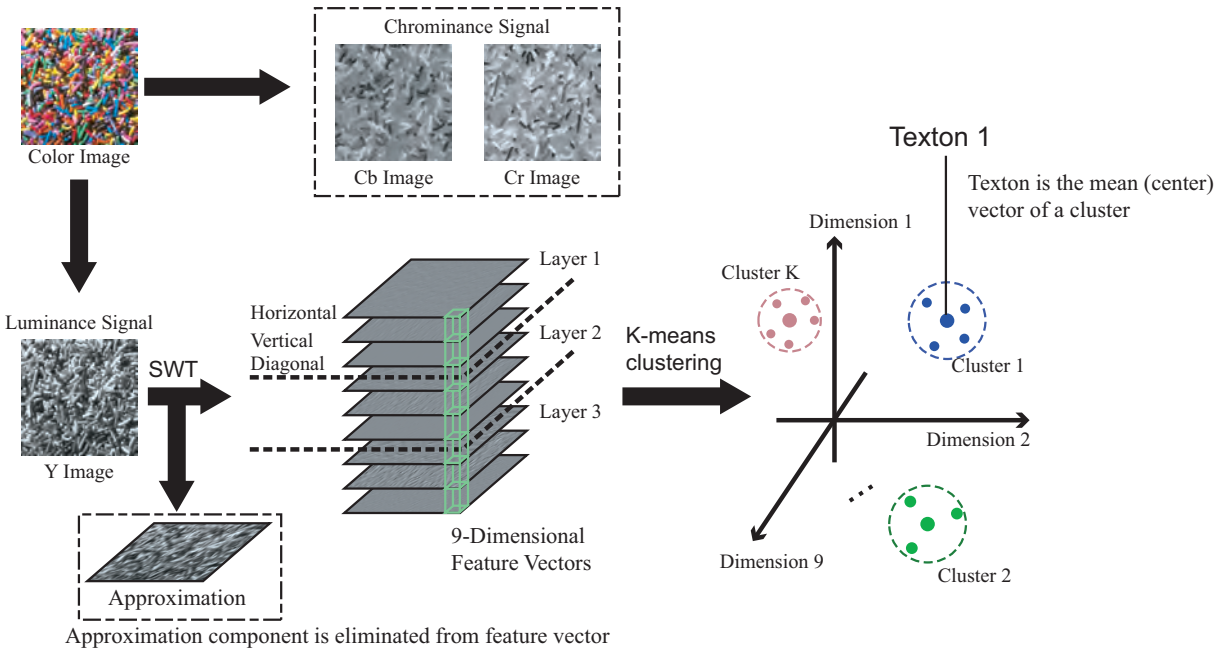

Color Image
Chrominance Signal
Cb Image
Cr Image
Luminance Signal
Y Image
SWT
Horizontal
Vertical
Diagonal
Layer 1
Layer 2
Layer 3
9-Dimensional
Feature Vectors
Approximation
Approximation component is eliminated from feature vector
K-means
clustering
Texton 1
Texton is the mean (center)
vector of a cluster
Dimension 1
Cluster K
Cluster 1
Dimension 2
Dimension 9
Cluster 2

Fig. 2

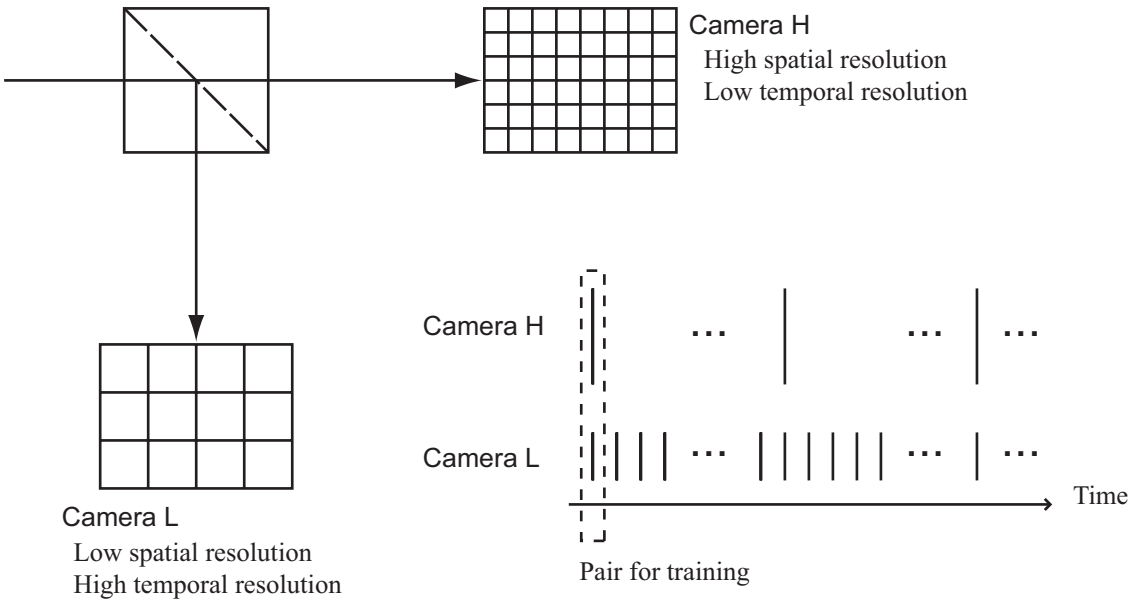
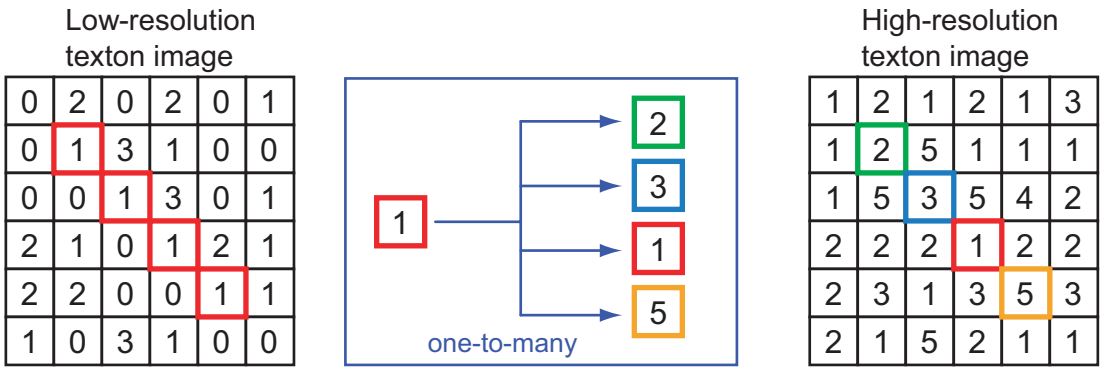
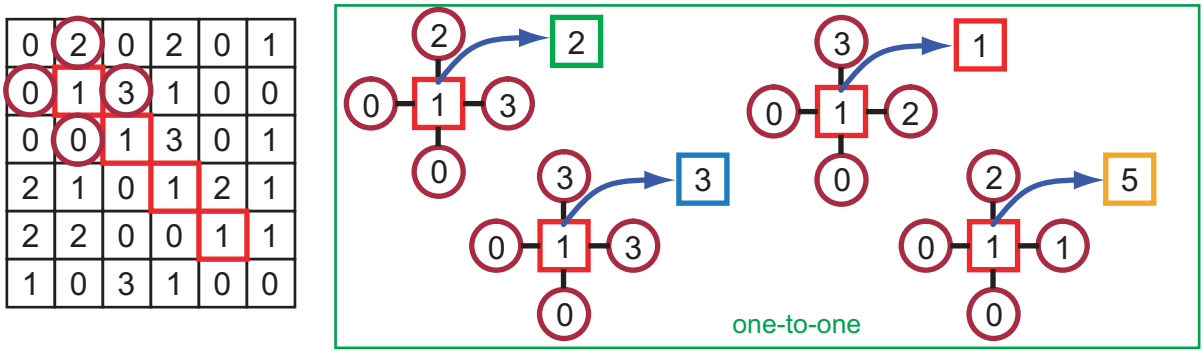

Camera H
High spatial resolution
Low temporal resolution
Camera L
Low spatial resolution
High temporal resolution
Camera H
Camera L
Time
Pair for training

Fig. 3

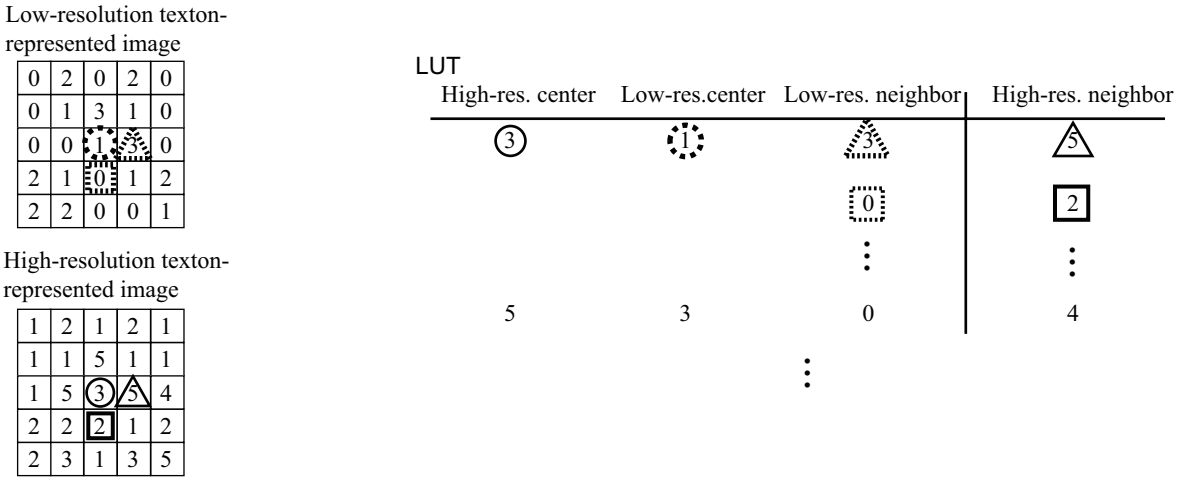


(a) Conventional texton substitution

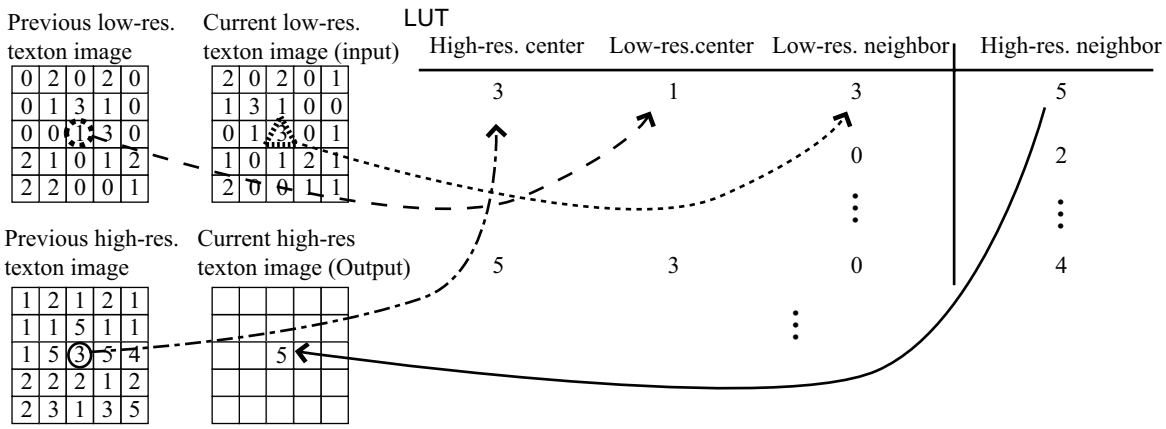


(b) Texton connection with 4-adjacent textons

Fig. 4



(a) Training of temporal LUT from spatial information



(b) Temporal substitution using temporal LUT

Fig. 5

PSNR = 26.7 dB PSNR = 31.8 dB

PSNR = 29.6 dB

(a) Bicubic interpolation

(b) High resolution

PSNR = Infinity PSNR = 56.0 dB

PSNR = 33.9 dB PSNR = 33.6 dB

PSNR = 38.4 dB

(c) Freeman's method

PSNR = 37.4 dB

(d) Texton substitution

Fig. 6

PSNR = 19.2 dB PSNR = 32.0 dB

PSNR = 29.8 dB

(a) Bicubic interpolation

(b) High resolution

PSNR = 26.3 dB PSNR = 32.2 dB

PSNR = 27.1 dB PSNR = 33.0 dB

PSNR = 30.5 dB

(c) Freeman's method

PSNR = 30.7 dB

(d) Texton substitution

Fig. 7

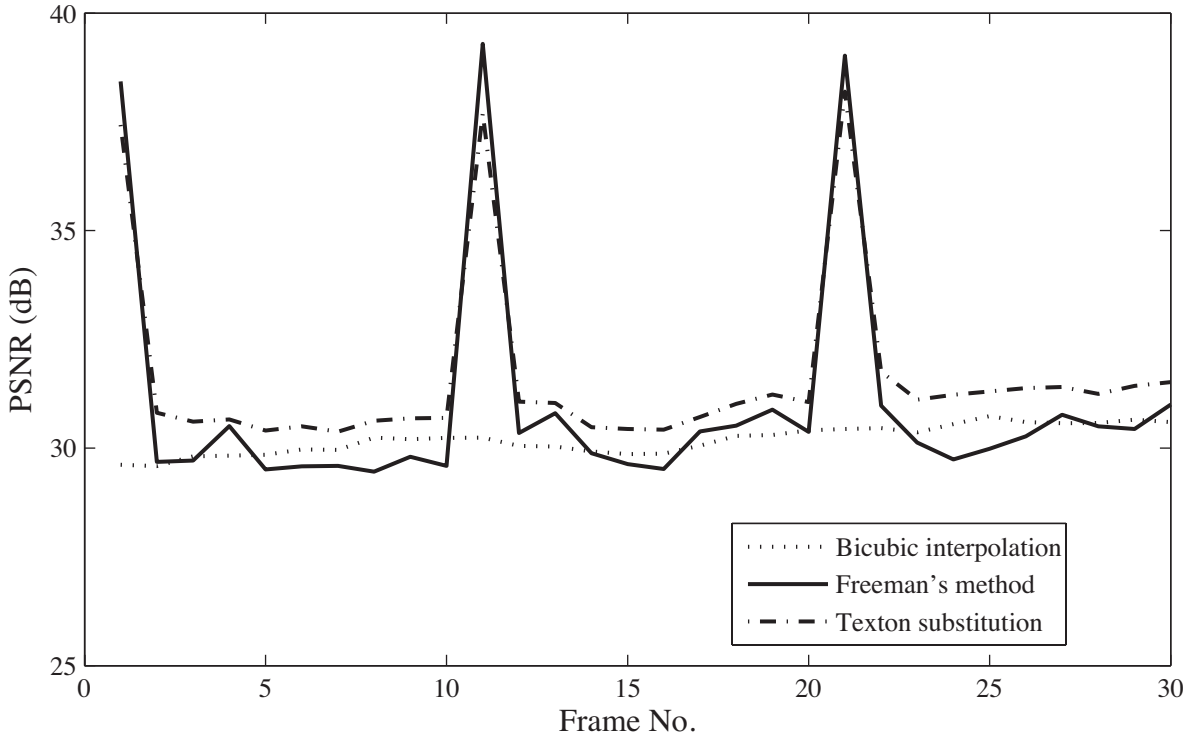
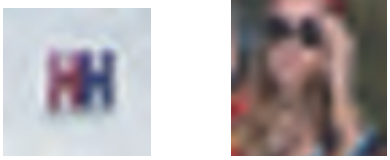

40
35
30
25
0
5
10
15
20
25
30
PSNR (dB)
Frame No.
Bicubic interpolation
Freeman's method
Texton substitution

Fig. 8



PSNR = 22.6 dB

PSNR = 22.1 dB

PSNR = 25.7 dB

(a) Bicubic interpolation

(b) High resolution

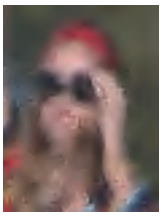
PSNR = 20.5 dB

PSNR = 21.7 dB

PSNR = 28.2 dB

(c) Freeman's method

PSNR = 22.5 dB



PSNR = 22.2 dB

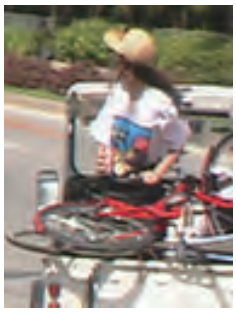
PSNR = 26.7 dB

(d) Texton substitution

Fig. 9

PSNR = 25.1 dB

PSNR = 19.3 dB



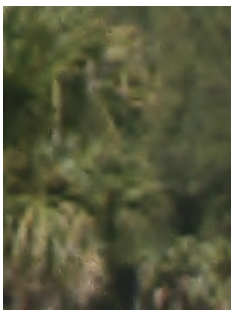
PSNR = 25.3 dB

(a) Bicubic interpolation

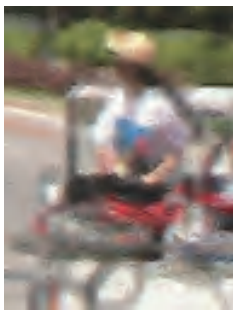
(b) High resolution

PSNR = 24.0 dB

PSNR = 18.9 dB



PSNR = 25.1 dB



PSNR = 19.4 dB

PSNR = 25.4 dB

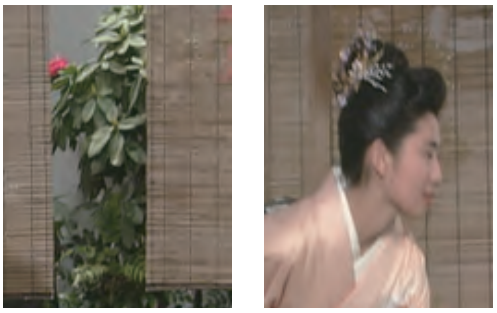
(c) Freeman's method

PSNR = 25.7 dB

(d) Texton substitution

Fig. 10

PSNR = 25.4 dB PSNR = 28.8 dB



PSNR = 25.7 dB

(a) Bicubic interpolation

(b) High resolution

PSNR = 24.1 dB PSNR = 28.1 dB

PSNR = 25.3 dB PSNR = 28.6 dB

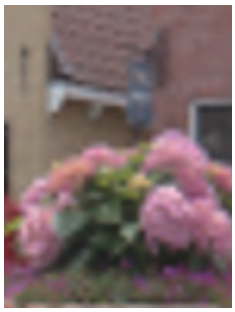
PSNR = 25.3 dB

(c) Freeman's method

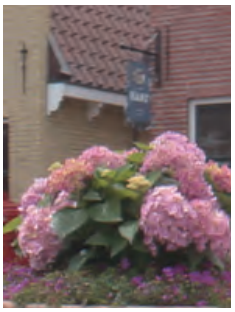
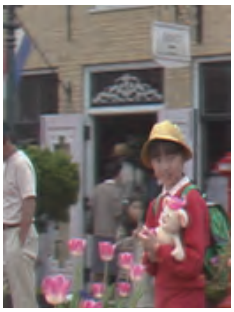
PSNR = 26.4 dB

(d) Texton substitution

Fig. 11



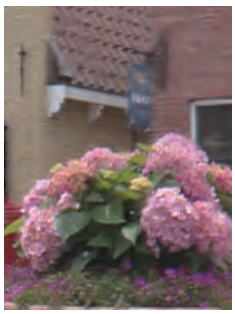

PSNR = 25.3 dB PSNR = 24.1 dB

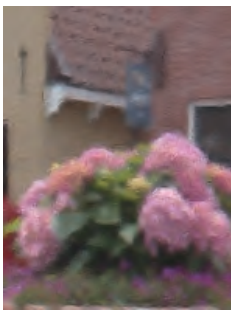
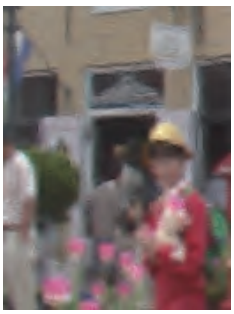

PSNR = 25.4 dB

(a) Bicubic interpolation

(b) High resolution




PSNR = 25.9 dB PSNR = 23.9 dB

PSNR = 25.9 dB PSNR = 24.8 dB

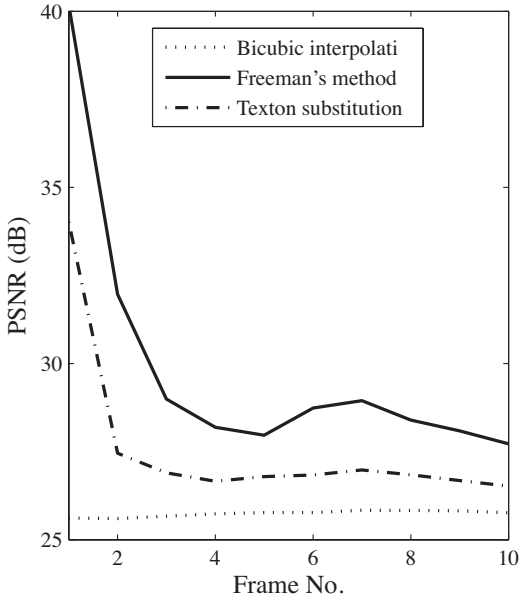

PSNR = 25.6 dB

(c) Freeman's method

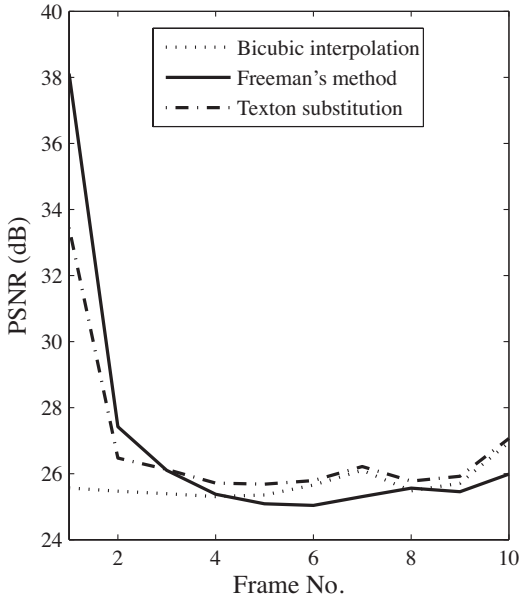
PSNR = 26.1 dB

(d) Texton substitution

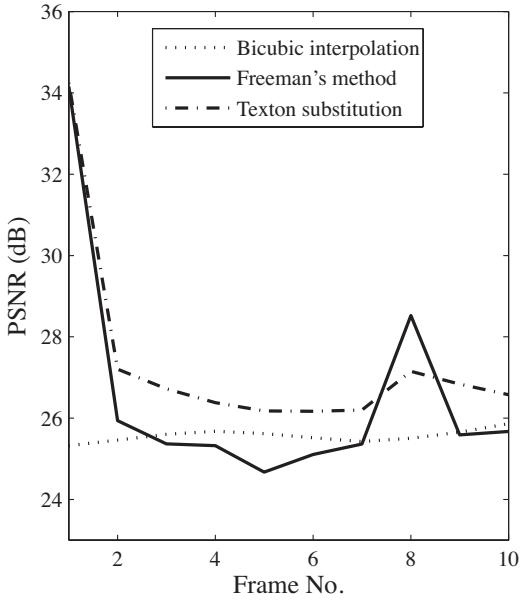
Fig. 12



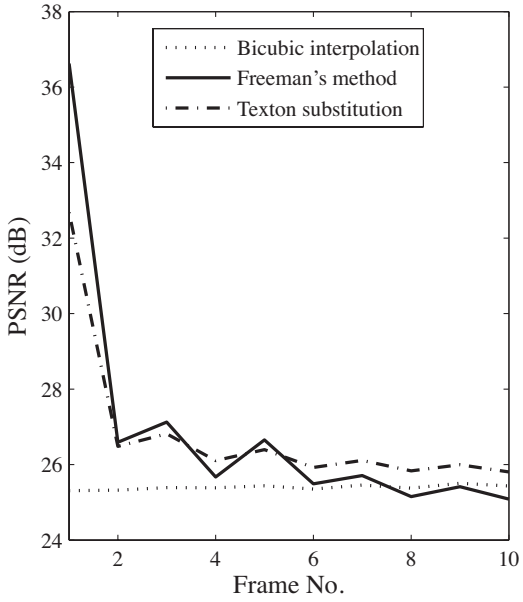
(a) Yachting



(b) Driving



(c) Japanese room



(d) Square

Fig. 13