Temporally Coherent Superresolution of Textured Video via Dynamic Texture Synthesis

Chih-Chung Hsu, Li-Wei Kang, *Member, IEEE*, and Chia-Wen Lin, *Senior Member, IEEE*

***Abstract*— This paper addresses the problem of hallucinating the missing high-resolution (HR) details of a low-resolution (LR) video while maintaining the temporal coherence of the reconstructed HR details using dynamic texture synthesis (DTS). Most existing multiframe-based video superresolution (SR) methods suffer from the problem of limited reconstructed visual quality due to inaccurate subpixel motion estimation between frames in an LR video. To achieve high-quality reconstruction of HR details for an LR video, we propose a texture-synthesis (TS)-based video SR method, in which a novel DTS scheme is proposed to render the reconstructed HR details in a temporally coherent way, which effectively addresses the temporal incoherence problem caused by traditional TS-based image SR methods. To further reduce the complexity of the proposed method, our method only performs the TS-based SR on a set of key frames, while the HR details of the remaining nonkey frames are simply predicted using the bidirectional overlapped block motion compensation. After all frames are upscaled, the proposed DTS-SR is applied to maintain the temporal coherence in the HR video. Experimental results demonstrate that the proposed method achieves significant subjective and objective visual quality improvement over state-of-the-art video SR methods.**

***Index Terms*— Video super-resolution, video hallucination, dynamic texture synthesis, video upscaling, motion-compensated interpolation.**

I. Introduction

With the rapid development of multimedia and network technologies, delivering and sharing multimedia contents through the Internet and heterogeneous devices has been more and more popular. However, limited by the storage capability, channel bandwidth, and source resolution, videos distributed over the Internet may exist in low-resolution (LR) versions degraded from the sources. Moreover, consumer multimedia devices with high-definition and ultra-high-definition displays have also been very popular. Nevertheless, the resolutions of most existing videos are still not such high. In this paper, we focus on investigating an efficient video super-resolution (SR) approach for resolution enhancement of a dynamic-texture video captured by a resource-limited device (e.g., low-cost surveillance camera) or stored in a lower resolution than the capability of a display device. Enhancement of video resolutions would be beneficial for other extended applications, such as face, action, or object recognition, behavior analysis, and video retrieval.

Manuscript received May 30, 2014; revised October 4, 2014; accepted December 16, 2014. Date of publication January 5, 2015; date of current version January 26, 2015. This work was supported by the National Science Council of Taiwan under Grant MOST 101-2221-E-007-121-MY3, Grant MOST 100-2218-E-224-017-MY3, and Grant MOST 103-2221-E-224-034-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Frakes.

C.-C. Hsu is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: m121754@gmail.com).

L.-W. Kang is with the Department of Computer Science and Information Engineering, Graduate School of Engineering Science and Technology-Doctoral Program, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan (e-mail: lwkang@yuntech.edu.tw).

C.-W. Lin is with the Department of Electrical Engineering, Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, and also with the Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2387416

A. Image Super-Resolution

Most SR methods in the literature were mainly designed for image SR. The goal of image SR is to recover a high-resolution (HR) image from one or multiple LR input images, which is essentially an ill-posed inverse problem [1]. There are mainly two categories of approaches for image SR: (i) traditional approaches and (ii) exemplar/learning-based approaches. In the traditional approaches, one sub-category is reconstruction-based methods, where a set of LR images of the same scene are aligned with sub-pixel accuracy to generate a HR image [2]. The main disadvantage of such kind of approaches is that they require multiple input LR images with accurate image registration. The other sub-category of the traditional approaches is frame interpolation [3], which usually generate over-smoothing images with ringing and jagged artifacts. Furthermore, it has also been shown that the limitation on magnification factor achieved by traditional approaches is not easy to break [4].

The exemplar/learning-based methods [5]–[14] hallucinate the high frequency details of a LR image based on the co-occurrence prior between LR and HR image patches in a training set, which has proven to provide much finer details compared to traditional approaches. More specifically, for a LR input, exemplar-based methods [5]–[8] search for similar image patches from a pre-collected training LR image dataset or the same image itself based on self-examples, and use their corresponding HR versions to produce the final SR output. Nevertheless, the HR details hallucinated by such kind of approaches may not provide the true HR details. Hence, the performance of this approach relies highly on the similarity between the training set and test set or the self-similarity in the image itself.

Moreover, learning-based SR approaches [9]–[14] focus on modeling the relationship between different resolutions

of images. For example, Yang *et al.* [9] proposed to apply sparse coding techniques to learn a compact representation for HR/LR patch pairs for SR based on pre-collected HR/LR image pairs (SC-SR). Then, a coupled dictionary training approach was proposed in [10] for SR based on patch-wise sparse recovery, where the learned couple dictionaries relate the HR/LR image patch spaces via sparse representations. In [11], a sparse representation based framework was proposed for image deblurring and SR based on adaptive sparse domain selection and adaptive regularization, where two adaptive regularization terms are introduced. In addition, Ren *et al.* [12] proposed to utilize context-aware sparsity prior to enhance the performance of sparsity-based restoration approach for image denoising and SR. In addition, self-learning frameworks based on self-similarity of an image were introduced for SR in [13] and [14].

B. Video Super-Resolution

Image SR techniques can be extended to video SR by incorporating temporal information. Most video SR methods rely mainly on motion estimation for interpolating LR frames between two key-frames (usually assumed to be of high resolution) in a video [15]–[17]. In [15], a video SR method based on key-frames and motion estimation was proposed. In [16], an energy-based algorithm for motion-compensated video SR was proposed for up-scaling a standard definition video to high-definition video via optical flow. In addition, a video SR algorithm was proposed in [17] to interpolate an arbitrary frame in a LR video from sparsely sampled HR key-frames which are assumed to be always available for a LR video input.

On the other hand, exemplar/learning-based techniques have been proposed for video SR. In [17], in the case that motion-compensated error is large, an input LR patch is spatially upscaled using the dictionary learned from the LR/HR key-frame pair. In [18], adaptive regularization and learning-based SR were integrated for web video SR by learning a set of LR/HR patch pairs. An exemplar-based video SR based on the codebooks derived from key-frames was also proposed in [19]. Moreover, the property of nonlocal-means was adopted for video SR in [20], which upscales each input LR patch by linearly fusing multiple similar LR patches based on self-similarity with no explicit motion estimation.

C. Texture Image/Video Super-Resolution

A main challenging problem in video SR is SR for dynamic textural information [21]–[23], [34] which was rarely investigated in the literature. In [24] and [25], texture synthesis (TS) techniques were proposed for image/video synthesis, but not for SR. For image SR with texture synthesis, a TS-based SR (TS-SR) scheme that upscales an image via texture hallucination was proposed in [26]. This method interprets a LR image as a tiling of distinct textures and each of which is matched to an exemplar patch in a database of relevant textures, extended from the exemplar-based approach in [5]. Although TS-SR can reconstruct fine HR textural details, the exemplar-based TS is time consuming, making the SR of whole video via TS-SR computationally very expensive. Furthermore, individually hallucinating the HR textural details of successive video frames usually renders the HR textural details in a time incoherent manner, which leads to visually annoying artifacts. Although such temporal incoherence artifacts can be mitigated by imposing temporal smoothness constraints in the optimization formulation of TS-SR, its significantly increased computational cost would make this method impractical. Therefore how to use TS-SR to hallucinate fine HR textural details of a LR input video in a computationally efficient way while maintaining the temporal coherence of hallucinated HR textures still remains a challenging problem.

D. Contribution of Proposed Method

To address the above problem, we propose a video SR framework via dynamic texture synthesis (DTS) to effectively and efficiently enhance the resolution of a LR video with dynamic textures while maintaining the temporal coherence of the reconstructed HR details. The proposed method divides the input LR video frames into key-frames and non-key-frames. We first apply the texture-synthesis-based SR (TS-SR) method to hallucinate the HR textural details of each key-frame, followed by employing a low-complexity bi-directional overlapped block motion compensation (BOBMC) method to interpolate the HR details of the non-key-frames between two successive key-frames. To solve the problem of temporal incoherence, we propose an exemplar-based DTS method to refine the HR details of the super-resolved video based on the temporal dynamics of the input LR video.

As shown in Fig. 1, our scheme first divides the input LR video frames into key-frames and non-key-frames, with a fixed (or dynamic) interval length between two successive key-frames. Each LR key-frame is upscaled using patch-based TS-SR [26]. Then, individual non-key-frames between two successive key-frames are first upscaled by bicubic interpolation, followed by BOBMC [17] to further interpolate their HR details from the two anchor key-frames. After all frames are upscaled, the proposed DTS-SR is applied to refine the HR details so as to maintain the temporal consistency between neighboring frames in the HR video. Similar to [26], we collect a set of exemplar textures in advance to form a multi-scale textural image database for texture synthesis.

The main contribution of this paper is two-fold: (i) we propose an efficient framework which can hallucinate visually fine and pleasing HR textural details of a LR video in a cost-efficient manner; and (ii) our novel DTS-based SR (DTS-SR) method can well maintain the temporal coherence in the hallucinated HR video by learning the texture dynamics from the input LR video. This problem, to the best of our knowledge, was not well studied before.

Compared with the preliminary conference version [35] of this paper, besides the significantly more detailed descriptions about the proposed method, this paper has been significantly extended in the following aspects: (i) This paper provides comprehensive and in-depth analyses and interpretations about the experimental results to offer good insights about the

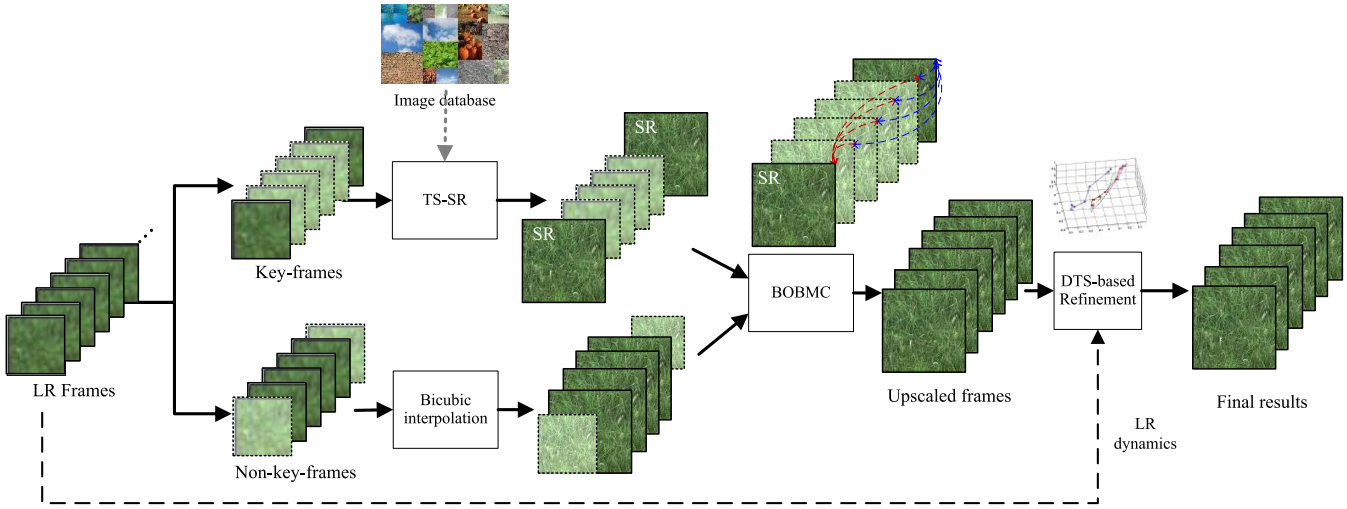


Fig. 1. Block diagram of the proposed DTS-SR framework. The key-frames are first uniformly sampled from the LR input video with an interval of K frames. These LR key-frames are upscaled using TS-SR [26]. Bicubic interpolation followed by BOBMC [19] are then used to upscale and reconstruct the HR details of individual non-key-frames in between two successive key-frames. Finally, these upscaled HR frames are rendered using the proposed DTS-SR to maintain the temporal coherence.

proposed method. (ii) We have provided subjective evaluation results in terms of "visual quality," "temporal coherency," and "details reconstruction" based on subjective paired comparisons. (iii) We have added a run-time complexity comparison of various methods.

The rest of this paper is organized as follows. Sec. II presents the proposed hybrid TS/BOMC video SR scheme. Sec. III describes the proposed DTS-based refinement scheme for maintaining the temporal coherence of reconstructed HR video. In Sec. IV, experimental results are demonstrated. Finally, Sec. V concludes this paper.

II. Hybrid Texture-Synthesis/ Interpolation-Based Video Super-Resolution

A. Texture-Synthesis-Based SR for Key-Frames

Based on [26], each input LR key-frame can be divided into different segments according to texture descriptors and then each segment can be classified into an exemplar texture from the pre-collected multi-scale textural image database. For each segment in a LR key-frame I_t^{LR}, where t denotes the frame index, the best matched patch $\mathbf{z}_{p,t}^{\mathrm{LR}}$ (p denotes the patch index) for each patch $\mathbf{x}_{p,t}^{\mathrm{LR}}$ in I_t^{LR} is searched in the exemplar texture set T by

$$\mathbf{z}_{p,t}^{\mathrm{LR}} = \arg\min_{\mathbf{z}\in T} d\left(\mathbf{x}_{p,t}^{\mathrm{LR}}, \mathbf{z}\right), \tag{1}$$

where d denotes the distance between two textural patches, P_1 and P_2, defined by [26]

$$d\left(P_1, P_2\right) = \left| \frac{P_1 - \mu_1}{\sigma_1} - \frac{P_2 - \mu_2}{\sigma_2} \right|, \tag{2}$$

where the symbols, μ_i and σ_i, respectively, denote the mean and standard deviation of the pixel values in each patch P_i, $i = 1, 2$.

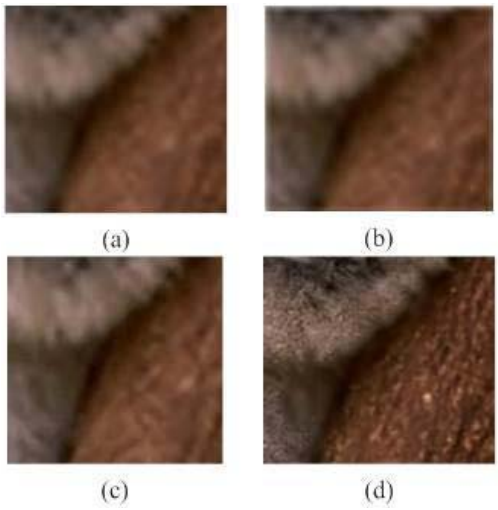


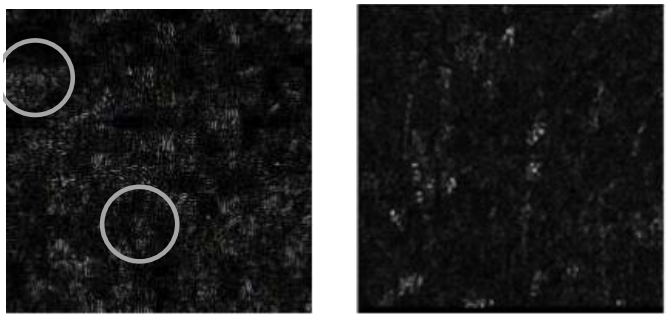
Fig. 2. Texture synthesis results obtained by (a) Bicubic; (b) NLM-based SR [20]; (c) SC-SR [9]; and (d) TS-SR [26] (employed in our method).

Then, the SR version $\mathbf{x}_{p,t}^{\mathrm{SR}}$ of $\mathbf{x}_{p,t}^{\mathrm{LR}}$ can be hallucinated based on the HR version $\mathbf{z}_{p,t}^{\mathrm{HR}}$ of $\mathbf{z}_{p,t}^{\mathrm{LR}}$ from the database and the texture synthesis by [26]

$$\mathbf{x}_{p,t}^{\mathrm{TS}} = \left(\mathbf{z}_{p,t}^{\mathrm{HR}} - \mu\left(\mathbf{z}_{p,t}^{\mathrm{HR}}\right)\right) \frac{\sigma\left(M_I\left(\mathbf{x}_{p,t}^{\mathrm{LR}}\right)\right)}{\sigma\left(M_I\left(\mathbf{z}_{p,t}^{\mathrm{HR}}\right)\right)} + \mu\left(\mathbf{x}_{p,t}^{\mathrm{LR}}\right), \tag{3}$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ respectively represent the mean and standard deviation of patch $\mathbf{x}$, and $M_I(\mathbf{x})$ denotes the middle-frequency component of $\mathbf{x}$.

Based on (3), the high-frequency components of the reconstructed HR patch can be synthesized directly from the textural image database, as exemplified in Fig. 2, which shows that TS-SR [26] provides significantly finer details compared to the other SR methods. Nevertheless, it is expected that individually applying TS-SR to each LR frame will cause the



(a) (b)

Fig. 3. (a) The difference image between two neighboring SR frames synthesized using TS-SR; and (b) the difference image between the two corresponding HR ground truths. Comparing (a) with (b), false motions due to incoherent texture synthesis can be observed as indicated by the grey circles.

temporal incoherence artifacts in the reconstructed SR video. For example, as illustrated in Fig. 3(a), the difference image between two neighboring SR frames synthesized using TS-SR shows much motion in textures. However, comparing Fig. 3(a) with the difference image of the two corresponding HR ground-truths [see Fig. 3(b)], we can observe that some regions [e.g., those indicated by the gray circles in Fig. 3(a)] actually are false motions due to incoherent texture synthesis since they cannot find corresponding motions in the ground-truth difference image in Fig. 3(b). Hence, in the proposed method, we only apply TS-SR to upscale a set of LR key-frames. Then, we apply BOBMC [17] to interpolate the HR details of non-key-frames between two successive key-frames, followed by the proposed DTS-SR to refine the HR details so as to maintain the temporal coherence of the resulted HR video, as described in Sec. II-B and Sec. III, respectively.

B. Bidirectional Interpolation for Non-Key-Frames

After upscaling the key-frames via TS-SR for an input LR video, we apply BOBMC to interpolate the HR details of non-key-frames between two successive key-frames. Considering a pair of upscaled HR key-frames I_t^{TS} and I_{t+K}^{TS} where K denotes the distance between the two successive key-frames, each LR non-key-frame I_{t+n}^{LR} in between them is initially up-scaled to the desired HR size via bicubic interpolation (denoted by I_{t+n}^{U}), their HR details are then reconstructed using BOBMC. To perform forward motion estimation (ME) for each non-overlapped patch in I_{t+n}^{U} with respect to I_t^{TS}, the patch together with its surrounding pixels are extracted for overlapped block matching to estimate the motion vector (MV) $\mathbf{v} = (v_x, v_y)$ by

$$\mathbf{v}^* = \arg\min_{\mathbf{v}\in\Omega} \mathrm{SAD}\,(\mathbf{v}), \tag{4}$$

where $\mathbf{v}^*$ denotes the estimated best-match MV for this patch, Ω denotes the search window, and SAD (sum of absolute differences), which is the most commonly used metric for block matching-based motion estimation due to its simplicity and promising performance [36], is defined as

$$\mathrm{SAD}\,(\mathbf{v}) = \sum_{i,j} \left| I_{t+n}^{U}\left(o_x + i, o_y + j\right) - I_t^{\mathrm{TS}}\left(o_x + i + v_x, o_y + j + v_y\right) \right|, \tag{5}$$

where $\left(o_x, o_y\right)$ denotes the location of the current patch. Similarly, backward ME between I_{t+n}^{U} and I_{t+K}^{TS} can be performed by the same technique in the reverse motion direction.

After performing the bidirectional ME for I_{t+n}^{U} with respect to I_t^{TS} and I_{t+K}^{TS}, each patch in I_{t+n}^{U} can be upscaled using the motion-compensated blocks with the estimated MVs (forward and/or backward MVs according to the SADs) and the pre-specified weight matrices as

$$\begin{aligned}\mathbf{x}_{p,t+n}^{\mathrm{BOBMC}}(i,j) &= W^C(i,j)\cdot\mathbf{x}_{t+n}^{C}(i,j) + W^T(i,j)\cdot\mathbf{x}_{t+n}^{T}(i,j) \\ &\quad + W^B(i,j)\cdot\mathbf{x}_{t+n}^{B}(i,j) + W^L(i,j)\cdot\mathbf{x}_{t+n}^{L}(i,j) \\ &\quad + W^R(i,j)\cdot\mathbf{x}_{t+n}^{R}(i,j),\end{aligned} \tag{6}$$

where $\mathbf{x}_{p,t+n}^{\mathrm{BOBMC}}$ denotes the p-th upscaled patch via BOBMC in I_{t+n}^{U}, $\mathbf{x}_{t+n}^{C}(i,j)$, $\mathbf{x}_{t+n}^{T}(i,j)$, $\mathbf{x}_{t+n}^{B}(i,j)$, $\mathbf{x}_{t+n}^{L}(i,j)$, and $\mathbf{x}_{t+n}^{R}(i,j)$ respectively represent the pixel values at (i,j) in the motion compensated patches corresponding to the MVs of the current patch, the top neighboring patch, bottom neighboring patch, left neighboring patch, and righting neighbor patch, respectively, and the weight matrices, W^C, W^T, W^B, W^L, and W^R, are shown in Fig. 4 [17]. Note that the pixel values of overlapped regions are normalized by their summed weight value.

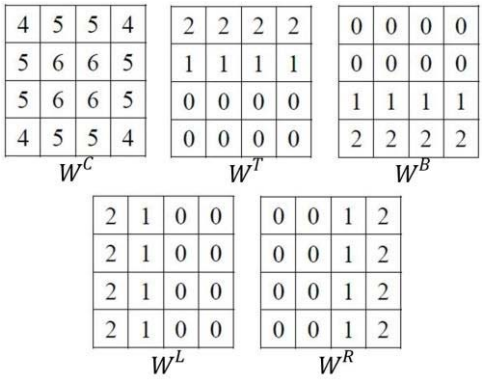


Fig. 4. The weight matrices for 4×4 block-based BOBMC [17].

After applying TS-SR and BOBMC to upscale an input LR video, the resulting HR video still suffers from the problem of temporal inconsistency. We then apply the proposed DTS-SR to solve the problem, as explained below.

III. Dynamic Texture Synthesis-Based Refinement

Although the hybrid TS-SR/BOBMC SR scheme can generate fine HR details, it usually leads to significantly different temporal texture dynamics from that of the original HR video because the HR details of key-frames are separately hallucinated making them temporally incoherent and those of non-key-frames are interpolated from the temporally

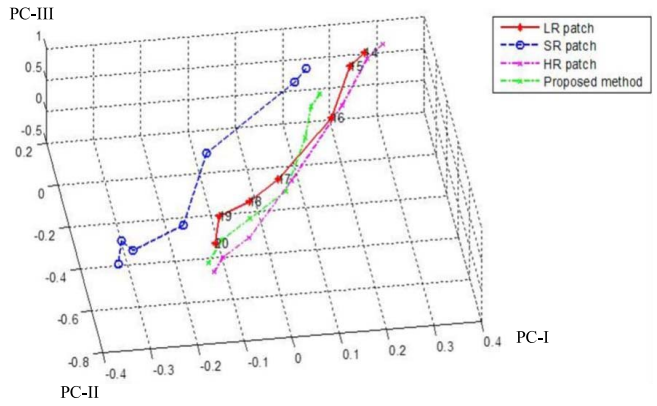


Fig. 5. Examples of low-dimensional trajectories projected from the ground-truth HR textural patches (red), its downscaled LR patches (pink), the HR patches upscaled from the LR patches via TS-SR + BOBMC (blue), and the SR patches via the proposed DTS-SR method (green), respectively. The three axes, PC-I, PC-II, and PC-III indicate the first, second, and third principal components of a video patch projected using PCA.

incoherent key-frames. Such inconsistent texture dynamics yield temporal incoherence artifacts in the SR video. To verify this, we crop a dynamic-texture video patch from a LR or HR video and project it to a low dimensional subspace using principal component analysis (PCA). As a result, in the subspace of the first few principal components, we can visualize and compare the low-dimensional trajectories of HR, LR, and SR dynamic-texture video patches over time. For example, as illustrated in Fig. 5, the low-dimensional trajectory of a ground-truth HR video patch in the subspace of the first three principal components is similar to that of its downscaled LR patch, but is significantly different from the trajectory of the SR patch upscaled from the LR patch using hybrid TS-SR/BOBMC. This motivates to learn the temporal texture dynamics of the HR video from its LR version and use the learnt texture dynamics to refine the HR details based on dynamic texture synthesis so as to effectively mitigate the temporal incoherence artifacts.

Fig. 6 depicts the proposed DTS-based refinement scheme, in which both the input LR frames and the reconstructed HR frames obtained via hybrid TS-SR/BOBMC are used to derive temporally coherent HR video frames. Our method is based on the assumption that the content of a textural patch varies along time and the transition between the textures can be modeled as a linear or nonlinear system [21]–[23], [34]. For sake of simplicity, we adopt the following linear model [22]:

$$\mathbf{x}_{p,t+1} = \mathbf{A}\mathbf{x}_{p,t}, \tag{7}$$

where $\mathbf{x}_{p,t}$ and $\mathbf{x}_{p,t+1}$ denote two patches at the same position p of the two successive frames with indices t and $t+1$, respectively, and $\mathbf{A}$ denotes the transition matrix of dynamic texture for $\mathbf{x}_{p,t}$. Based on [22], dynamic textures can be formulated as a linear autoregressive (AR) system as:

$$\mathbf{y}_{p,t+1} = \mathbf{A}\mathbf{y}_{p,t} + N(0, \Sigma_t), \tag{8}$$

where $\mathbf{y}_{p,t} = \mathbf{C}^T\mathbf{x}_{p,t}$, $\mathbf{C}$ is an orthogonal projection matrix used to reduce the dimensionality of the dynamic textures so that the dimensionality is smaller than or equal to the number of frames to avoid the underdetermined problem, and $N(0, \Sigma_t)$ denotes the zero-mean normally distributed additive noise that captures the uncertainty with covariance matrix Σ_t.

Since the AR model in (8) is established for the low-dimensional projected data $\mathbf{y}_{p,t}$, prior to estimating the model parameters $\mathbf{A}$ and Σ_t in (8), we need to estimate the projection matrix $\mathbf{C}$ first. As shown in Fig. 6, the HR projection matrix $\mathbf{C}^{\mathrm{HR}}$ is estimated from the SR frames obtained via hybrid TS-SR/BOBMC by using principal component analysis (PCA) as follows:

$$\mathbf{C}^{\mathrm{HR}} = \arg\min_{\mathbf{C}} \left\| \mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}} - \mathbf{C}\mathbf{C}^T\mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}} \right\|^2 \tag{9}$$
$$\text{s.t.} \quad \mathrm{rank}(\mathbf{C}) = d \text{ and } \mathbf{C}\mathbf{C}^T = \mathbf{I},$$

where $\mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}}$ denotes a HR patch upscaled via hybrid TS-SR/BOBMC and d is the number of principal components.

Based on our experiments exemplified in Fig. 5, since the HR and LR versions of a video have similar temporal texture dynamics, the transition matrix $\mathbf{A}^{\mathrm{LR}}$ for characterizing the dynamic textures of a downscaled LR video is similar to that of its HR version. Therefore, we can estimate $\mathbf{A}^{\mathrm{LR}}$ based on the least squares approximation in (10) and use it to approximate its HR counterpart $\mathbf{A}^{\mathrm{HR}}$ to maintain the temporal coherence in the reconstructed SR frames.

$$\mathbf{A}^{\mathrm{LR}} \approx \mathbf{Y}_{2:N}^{\mathrm{LR}} \left(\mathbf{Y}_{1:N-1}^{\mathrm{LR}}\right)^{\mathrm{T}} \left(\mathbf{Y}_{1:N-1}^{\mathrm{LR}} \left(\mathbf{Y}_{1:N-1}^{\mathrm{LR}}\right)^{\mathrm{T}}\right)^{-1}, \tag{10}$$

where matrix $\mathbf{Y}_{1:N}^{\mathrm{LR}} = \left[\mathbf{y}_1^{\mathrm{LR}} \mathbf{y}_2^{\mathrm{LR}} \cdots \mathbf{y}_N^{\mathrm{LR}}\right]$ represents the projected data from the first frame to the N-th frame, and $\mathbf{y}_t^{\mathrm{LR}} = \left(\mathbf{C}^{\mathrm{LR}}\right)^T \mathbf{x}_t^{\mathrm{LR}}$, where the LR projection matrix $\mathbf{C}^{\mathrm{LR}}$ is estimated from the LR frames via PCA similar to (9).

By replacing $\mathbf{A}^{\mathrm{HR}}$ with $\mathbf{A}^{\mathrm{LR}}$ in the AR model in (8), the projected low-dimensional data $\mathbf{y}_{p,t}^{\mathrm{SR}}$ can be rendered in a temporally coherent way using dynamic texture synthesis as follows:

$$\mathbf{y}_{p,t+1}^{\mathrm{SR}} \approx \mathbf{A}^{\mathrm{LR}}\mathbf{y}_{p,t}^{\mathrm{SR}} + N(0, \Sigma_t^{\mathrm{HR}}), \tag{11}$$

where $\mathbf{y}_{p,t}^{\mathrm{SR}} = (\mathbf{C}^{\mathrm{HR}})^T \mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}}$, and $N(0, \Sigma_t^{\mathrm{HR}})$ is defined in (8) which can be estimated from the projected data $\mathbf{y}_{p,t}^{\mathrm{SR}}$ [22].

Finally, each HR patch $\mathbf{x}_{p,t}^{\mathrm{SR}}$ can be reconstructed from the projected low-dimensional version $\mathbf{y}_{p,t}^{\mathrm{SR}}$ using the projection matrix as $\mathbf{x}_{p,t}^{\mathrm{SR}} = \mathbf{C}^{\mathrm{HR}}\mathbf{y}_{p,t}^{\mathrm{SR}}$. After each patch is processed using the proposed DTS-based refinement scheme, the temporal incoherence artifacts can be effectively mitigated.

To evaluate the performance of the proposed scheme, Fig. 5 shows the four projected trajectories of the ground-truth HR patches, the downscaled LR patches, the SR patches via hybrid TS-SR/BOBMC, and the SR patches via the proposed DTS-SR method (i.e., hybrid TS-SR/BOBMC followed by DTS-based refinement), respectively. We can observe from Fig. 5 that the trajectory of the SR patches obtained via DST-SR is much closer to the ground-truth trajectory compared to that of the SR patches obtained via hybrid TS-SR/BOBMC. As a result, the proposed DTS-SR method can well address the temporal incoherence problem in video SR which can also be observed from the SR videos available in [28]. The whole proposed DTS-SR algorithm is summarized in TABLE I.

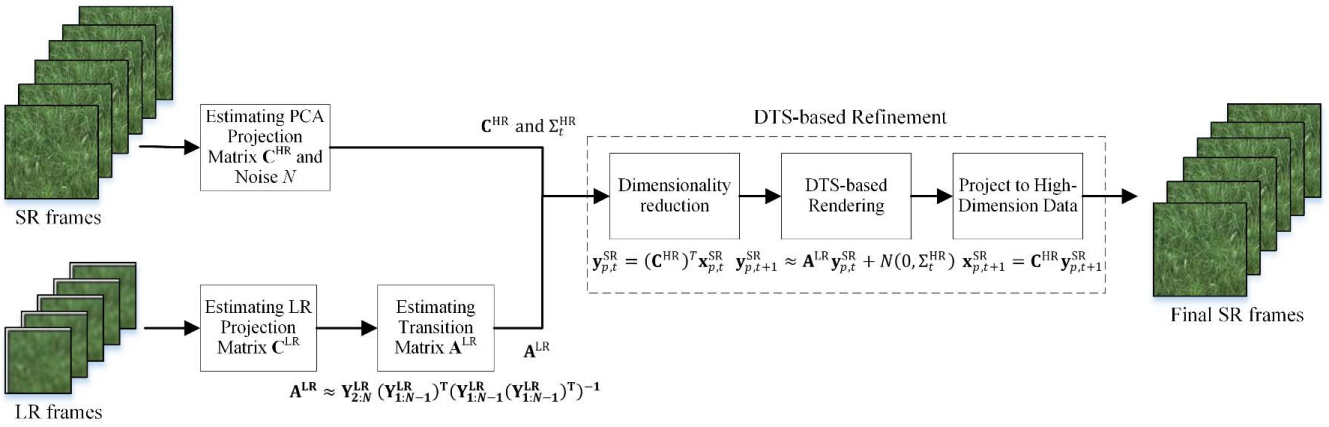


Fig. 6. Block diagram of proposed DTS-based refinement.

TABLE I
PROPOSED DTS-SR ALGORITHM

Input: a LR video with dynamic textures.
Output: a SR version of the input video.

1. **TS-SR for key-frames:**
   Hallucinate the HR details of each patch $\mathbf{x}_{p,t}^{\mathrm{LR}}$ in each key frame I_t^{LR} via
   $$\mathbf{x}_{p,t}^{\mathrm{TS}} = \left(\mathbf{z}_{p,t}^{\mathrm{HR}} - \mu\left(\mathbf{z}_{p,t}^{\mathrm{HR}}\right)\right) \frac{\sigma\left(M_I\left(\mathbf{x}_{p,t}^{\mathrm{LR}}\right)\right)}{\sigma\left(M_I\left(\mathbf{z}_{p,t}^{\mathrm{HR}}\right)\right)} + \mu\left(\mathbf{x}_{p,t}^{\mathrm{LR}}\right),$$
   where $\mathbf{z}_{p,t}^{\mathrm{LR}} = \arg\min_{\mathbf{z} \in T} \mathbf{d}\left(\mathbf{x}_{p,t}^{\mathrm{LR}}, \mathbf{z}\right)$.

2. **BOBMC interpolation for non-key-frames:**
   Reconstruct the HR details of each patch of non-key-frame I_{t+n}^{U} (bicubic-interpolated version of LR non-key-frame) between two successive key-frames I_t^{TS} and I_{t+K}^{TS} via BOBMC as
   $$\mathbf{x}_{p,t+n}^{\mathrm{BOBMC}}(i,j) = W^C(i,j) \cdot \mathbf{x}_{t+n}^{C}(i,j) + W^T(i,j) \cdot \mathbf{x}_{t+n}^{T}(i,j) + W^B(i,j) \cdot \mathbf{x}_{t+n}^{B}(i,j) + W^L(i,j) \cdot \mathbf{x}_{t+n}^{L}(i,j) + W^R(i,j) \cdot \mathbf{x}_{t+n}^{R}(i,j).$$

3. **DTS-based refinement:**
   Estimate $\mathbf{C}^{\mathrm{HR}}$ from $\mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}}$ using PCA and estimate $\mathbf{A}^{\mathrm{LR}}$ using (10)
   Apply the following AR model to obtain temporally coherent low-dimensional sequence by
   $$\mathbf{y}_{p,t+1}^{\mathrm{SR}} \approx \mathbf{A}^{\mathrm{LR}} \mathbf{y}_{p,t}^{\mathrm{SR}} + N(0, \Sigma_t^{\mathrm{HR}})$$
   where $\mathbf{y}_{p,t}^{\mathrm{SR}} = (\mathbf{C}^{\mathrm{HR}})^T \mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}}$, where $\mathbf{x}_{p,t}^{\mathrm{TS\_BOBMC}}$ is an upscaled patch obtained by hybrid TS-SR/BOBMC

   Back-project the low-dimensional sequence to the original high-dimensional space to reconstruct the HR video
   $$\mathbf{x}_{p,t+1}^{\mathrm{SR}} = \mathbf{C}^{\mathrm{HR}} \mathbf{y}_{p,t+1}^{\mathrm{SR}}.$$

IV. EXPERIMENTAL RESULTS

A. Performance Evaluation of Super-Resolution

To evaluate the performance of the proposed video SR method, we collected a training dataset of 52 HR textural images. Five 57-frame 720×480 videos (videos #1–#4 and #6) with fine and dynamic textures and three 57-frame 1280×720 videos (videos #5, #7–#8) with mixed textures and non-textures, as shown in Fig. 7, are used as HR ground-truths in our experiments. These videos are first downscaled by a factor of three both horizontally and vertically as the LR test videos, which are then upscaled back to their original resolution using various SR schemes. The experimental settings are described as follows. For each LR test video, the number of non-key-frames between two successive key-frames is set to $K = 8$. The patch sizes used in TS-SR and BOBMC are both 16×16, and the patch size used in DTS-SR is 60×60. We compare the performance of the proposed DTS-SR method with that of the following approaches: (i) bicubic interpolation (denoted by Bicubic) [3], (ii) SR via sparse coding (denoted by SC-SR) [9], (iii) SR via non-local iterative back-projection (denoted by NLIBP-SR) [6], (iv) SR via adaptive sparse domain selection and adaptive regularization (denoted by ASDS-SR) [11], (v) SR via texture hallucination (denoted by TS-SR) [26], and (vi) video SR via BOBMC [17]. The complete test results can be found in our project website [28].

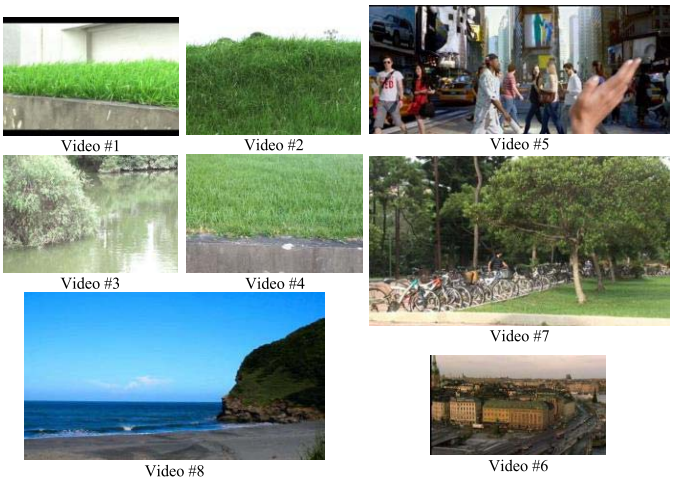


Fig. 7. Example frames of the eight test videos with dynamic textures, where the resolution of video #5, #7, and #8 is 1280×720, and the resolution of the rest videos is 720×480.

1) Objective Quality Evaluation: To quantitatively evaluate the performances of various SR schemes, we use the motion-based video integrity evaluation (MOVIE) metric proposed in [27] for video quality assessment. The MOVIE metric is a full-reference quality assessment metric which utilizes a general, spatio-spectrally localized multi-scale framework for evaluating dynamic video fidelity that integrates both spatial and temporal (and spatio-temporal) aspects of distortion assessment. The smaller the MOVIE index of an evaluated video is, the higher the visual quality of this video will be. MOVIE has proven to be fairly consistent with human subjective judgments. Since it takes into account the temporal distortion, the MOVIE metric is much more

Fig. 8. Video SR results: a) the original consecutive HR frames; and the SR results of the corresponding LR frames obtained by (b) the proposed method (MOVIE = **0.183**); (c) SC-SR [9] (MOVIE = 0.36); (d) ASDS-SR [11] (MOVIE = 0.24); (e) NLIBP-SR [6] (MOVIE = 0.29); and (f) Bicubic [3] (MOVIE = 0.95).

suitable for evaluating the fidelity of an upscaled video with dynamic textures compared to other spatial quality assessment metrics which do not consider temporal information [e.g., the peak signal-to-noise ratio (PSNR) metric and the structure similarity (SSIM) metric, and their variants].

Fig. 8 shows the SR results cropped from a set of five successive upscaled frames for Video #2 using the proposed method, SC-SR [9], ASDS-SR [11], NLIBP-SR [6], and Bicubic [3] along with their respective MOVIE scores. It shows the visual qualities of the HR frames obtained by the

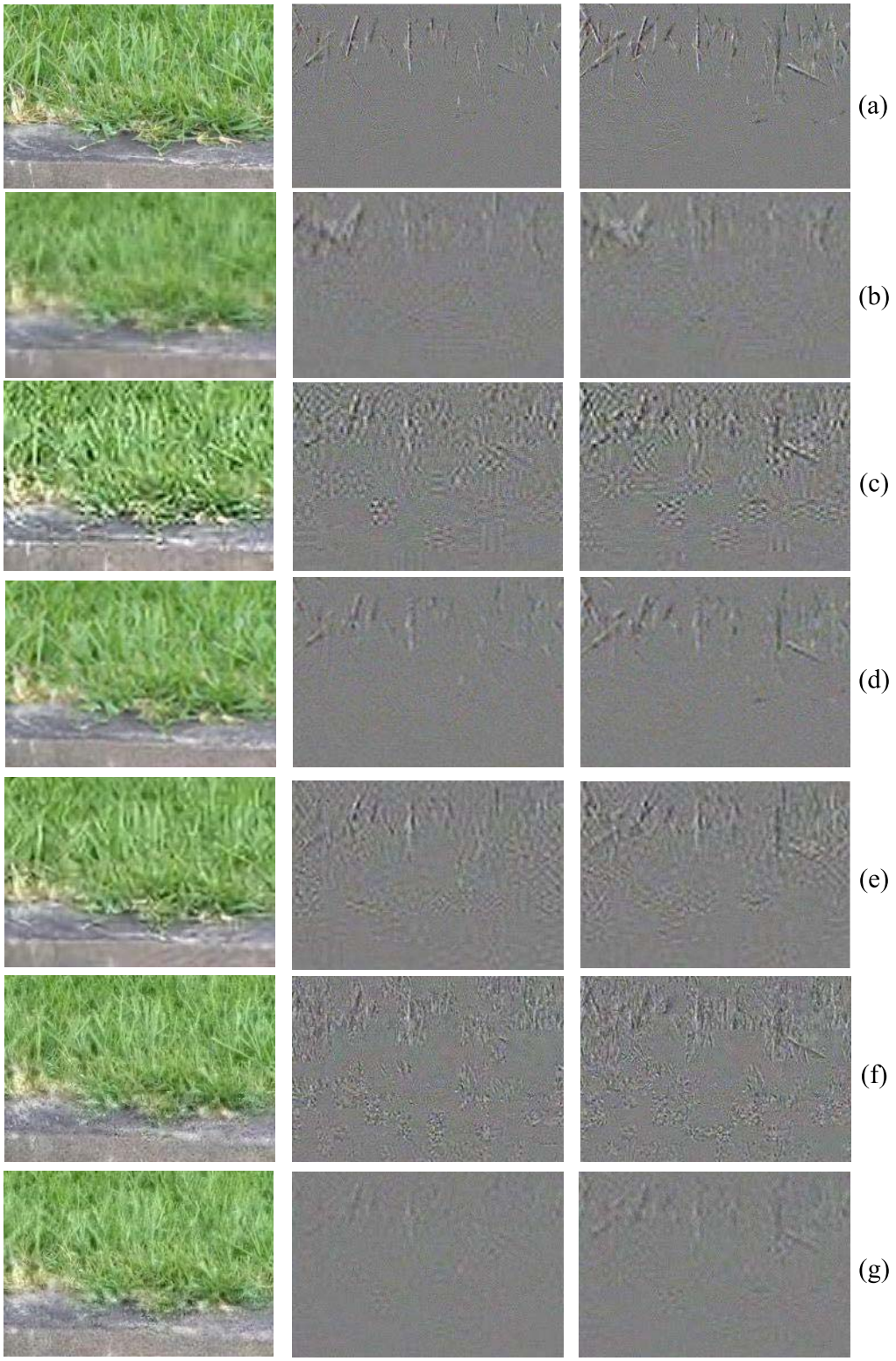


Fig. 9. Video SR results: (a) the original HR frame; and its SR results obtained by (b) Bicubic [3]; (c) NLIBP-SR [6]; (d) ASDS-SR [11]; (e) SC-SR [9]; (f) TS-SR [26]; and (g) the proposed method. The first, second, and third columns show the respective SR results, the difference between the frame in the first column and its immediately next frame, and the difference between the frame in the first column and its next frame with a two-frame distance between them.

proposed method are significantly better than those obtained by the four compared methods. More specifically, more high-frequency components (i.e., textures) in the reconstructed frames can be well reconstructed using our method, whereas the other SR methods usually cannot provide sufficiently fine details.

In addition, Fig. 9 shows the highlighted SR results and the corresponding temporal consistency for the input LR video

TABLE II
OBJECTIVE EVALUATION BY MOVIE FOR THE RECONSTRUCTED SR VIDEOS OBTAINED USING THE BICUBIC [3], SC-SR [9], NLIBP-SR [6], ASDS-SR [11], TS-SR [26], BOBMC [17], AND THE PROPOSED METHOD (SMALLER MOVIE VALUE INDICATES HIGHER VISUAL QUALITY)

Method	Bicubic	SC-SR	NLIBP-SR	ASDS-SR	TS-SR	BOBMC	Proposed
Video #1	1.12	0.36	0.24	0.29	0.22	0.24	**0.18**
Video #2	0.93	0.50	0.29	0.21	0.43	0.44	**0.20**
Video #3	0.98	0.41	0.34	0.22	0.29	0.32	**0.20**
Video #4	1.51	0.51	0.17	0.11	0.17	0.21	**0.11**
Video #5	1.14	0.68	0.42	0.35	0.36	0.27	**0.21**
Video #6	0.97	0.45	0.33	0.31	0.29	0.23	**0.19**
Video #7	1.31	0.41	0.43	0.45	0.37	0.38	**0.29**
Video #8	0.88	0.57	0.39	0.31	0.24	0.31	**0.22**

using Bicubic [3], NLIBP-SR [6], ASDS-SR [11], SC-SR [9], TS-SR [26], and the proposed method. It can also be observed from Fig. 9 that the visual qualities of the SR results obtained by the proposed method are significantly better than those obtained by the five methods used for comparisons. Moreover, compared to the other methods, our method also achieves better temporal coherence in the reconstructed SR video as illustrated by the lower differences between neighboring frames.

TABLE II compares the MOVIE indices between the HR ground-truths of the eight test videos and the corresponding SR results using Bicubic, SC-SR, NLIBP-SR, ASDS-SR, TS-SR, BOBMC [17], and our method. To fairly compare our method with BOBMC, we apply TS-SR to upscale each LR key-frame, and then use BOBMC to upscale non-key-frames because in [17], each HR key-frame was assumed to be already available. TABLE II shows that the proposed method outperforms these compared methods in terms of MOVIE based on the fact that our method can well maintain the temporal consistency for consecutive upscaled video frames. More experimental results are provided in our project website [28].

The proposed method was implemented without specific code-level optimization in MATLAB 64-bit version with Windows 8 operation system on a personal computer equipped with Intel i7 processor and 16 GB memory. To evaluate the computational complexity of the proposed algorithm, the runtime of each compared method is listed in TABLE III, which shows that the proposed method is significantly faster than ASDS-SR and TS-SR and comparable with BOBMC.

2) Subjective Quality Evaluation: In order to subjectively evaluate the performance of the proposed SR method, we conduct a paired comparison-based subjective user study [29]. We invited 20 subjects to join the experiments, where each subject was given two side-by-side SR videos obtained by two different evaluated SR methods (in a random order) at a time, and was asked to choose their preference from the two SR videos in terms of visual quality, temporal coherence, and details reconstruction, respectively.

The visual quality is based on subjective user preference. Moreover, the temporal coherence is to evaluate the temporal consistency between neighboring frames of the SR videos, whereas the details restoration is to evaluate the performance of the ability of the HR details recovery from LR videos. The 20 subjects include 13 males and 7 females, whose ages ranging from 21 to 31, without prior knowledge about the evaluated SR methods. The device used to display these SR videos was a full-HD 23-inch LCD display with color temperature 4300K.

TABLE III
RUM-TIME (IN SECONDS) COMPARISONS AMONG THE EVALUATED METHODS USED FOR COMPARISONS AND THE PROPOSED METHOD

Method	Bicubic	SC-SR	NLIBP-SR	ASDS-SR	TS-SR	BOBMC	Proposed
Video #1	29	1121	557	6679	11791	1710	1751
Video #2	22	1981	609	6857	8875	1337	1379
Video #3	30	1363	542	6513	7347	1086	1124
Video #4	29	1216	701	5491	9104	1375	1419
Video #5	22	4099	2299	18221	14174	5233	5323
Video #6	33	1364	713	6350	7671	2231	2327
Video #7	51	3121	1438	11735	14567	6364	6455
Video #8	47	2998	1515	10124	15181	6412	6671

TABLE IV
SUBJECTIVE "VISUAL QUALITY" EVALUATION BY PAIRED COMPARISONS (IN RELATIVE WINNING PERCENTAGE) FOR THE EIGHT RECONSTRUCTED HR VIDEOS OBTAINED USING OUR METHOD, ASDS-SR [11], NLIBP-SR [6], SC-SR [9], BICUBIC [3], AND TS-SR [26]

Method	Proposed	ASDS-SR	NLIBP-SR	SC-SR	Bicubic	TS-SR	Average
Proposed	–	**80.63%**	**83.13%**	**88.75%**	**93.13%**	**88.13%**	**86.75%**
ASDS- SR	19.38%	–	56.88%	63.13%	86.88%	68.75%	59.00%
NLIBP-SR	16.88%	43.13%	–	61.88%	83.13%	73.75%	55.75%
SC-SR	11.25%	36.88%	38.13%	–	85.63%	56.88%	45.75%
Bicubic	6.88%	13.13%	16.88%	14.38%	–	43.75%	19.00%
TS-SR	11.88%	31.25%	26.25%	43.13%	56.25%	–	33.75%

In our subjective experiments, we compare the proposed method with Bicubic, SC-SR, NLIBP-SR, ASDS-SR, and TS-SR for the eight test videos. Each SR method is pairwise compared with the others by totally 5 (methods) $\times$ 8 (test videos) $\times$ 20 (subjects) $=$ 800 times, implying that 160 comparisons are made between every two methods for the eight test videos. To quantify the subjective evaluation results, we calculate the winning frequency matrix $[w_{ij}]$, $i, j = 1, 2, \ldots, 6$ proposed in [29], where the (i, j)-th entry w_{ij} indicates the number of times that the i-th method outperforms the j-th method determined by the subjects in the paired comparisons. For each categorization of performance evaluation (visual quality, temporal coherence, and details reconstruction), as respectively shown in TABLES IV–VI, we calculate the relative winning percentage w_{ij}/N_{ij}, between the i-th and j-th methods, where $N_{ij} = 160$ is the number of comparisons made between every two methods.

TABLES IV–VI show that the proposed method performs the best subjectively in visual quality and details reconstruction, and the second best in temporal coherence based on the subjective quality evaluation criterion proposed in [29]. Note, TABLE V shows that the Bicubic method outperforms the

TABLE V
SUBJECTIVE "TEMPORAL COHERENCE" EVALUATION BY PAIRED COMPARISONS (IN RELATIVE WINNING PERCENTAGE) FOR THE EIGHT RECONSTRUCTED HR VIDEOS OBTAINED USING OUR METHOD, ASDS-SR, NLIBP-SR, SC-SR, BICUBIC, AND TS-SR

Method	Proposed	ASDS-SR	NLIBP-SR	SC-SR	Bicubic	TS-SR	Average
Proposed	–	61.88%	**60.63%**	**69.38%**	43.13%	87.50%	64.50%
ASDS- SR	38.13%	–	54.38%	53.75%	25.63%	86.25%	51.63%
NLIBP-SR	39.38%	45.63%	–	55.63%	41.88%	86.88%	53.88%
SC-SR	30.63%	46.25%	44.38%	–	30.63%	79.38%	46.25%
Bicubic	**56.88%**	**74.38%**	58.13%	69.38%	–	**91.88%**	**70.13%**
TS-SR	12.50%	13.75%	13.13%	20.63%	8.13%	–	13.63%

TABLE VI
SUBJECTIVE "DETAILS RECONSTRUCTION" EVALUATION BY PAIRED COMPARISONS (IN RELATIVE WINNING PERCENTAGE) FOR THE EIGHT RECONSTRUCTED HR VIDEOS OBTAINED USING OUR METHOD, ASDS-SR, NLIBP-SR, SC-SR, BICUBIC, AND TS-SR

Method	Proposed	ASDS-SR	NLIBP-SR	SC-SR	Bicubic	TS-SR	Average
Proposed	–	**83.75%**	**76.25%**	**87.50%**	**92.50%**	**53.75%**	**78.75%**
ASDS- SR	16.25%	–	43.75%	61.88%	75.00%	30.00%	45.38%
NLIBP-SR	23.75%	56.25%	–	64.38%	83.13%	36.88%	52.88%
SC-SR	12.50%	38.13%	35.63%	–	86.25%	27.50%	34.88%
Bicubic	7.50%	25.00%	16.88%	39.38%	–	24.38%	22.63%
TS-SR	46.25%	70.00%	63.13%	72.50%	75.63%	–	65.50%

others in temporal coherence, which is the only item in which our method does not perform the best. The main reason is that the bicubic method is simply based on interpolation, where only the pixel values within the LR version of an image itself are used for upscaling and the interpolation scheme is temporally coherent, thereby resulting in better temporal coherency while leading to poor performance in both visual quality and details reconstruction. In contrast, in our method, the missing HR details are reconstructed by hallucination, which somehow unavoidably results in some temporal inconsistency even if the proposed novel DTS scheme has addressed this problem to some extent.

B. Analyses and Discussions

1) Selection of Dimensionality Reduction Techniques: As shown in Sec. III, in our implementation, the orthogonal projection matrix, generated via PCA, is used to project each image patch into the subspaces. The feasibility of using PCA for dimensionality reduction has been verified by visualizing the trajectory of the dynamics for each patch as illustrated in Fig. 5. To further verify the benefit coming from PCA, we implement two additional popular dimensionality reduction techniques, namely, the locality preserving projection (LPP) [30] and the orthogonal LPP (OLPP) [31] on top of our framework, as well as evaluate their performances based on the MOVIE index. TABLE VII shows the MOVIE values of the reconstructed HR videos obtained by the three dimensionality reduction techniques implemented in our SR framework for the four test videos. TABLE VII shows that PCA outperforms or is comparable with LPP and OLPP techniques in terms of objective visual quality as PCA achieves the best energy compaction performance which is usually helpful in preserving the dynamics in (8). Moreover, PCA consumes lower or similar computational complexity compared to LPP and OLPP. Hence, we employ PCA in our SR framework considering both visual quality and computational complexity.

TABLE VII
OBJECTIVE EVALUATION BY MOVIE INDEX FOR THE RECONSTRUCTED HR VIDEOS OBTAINED USING PCA, LPP, AND OLPP FOR DIMENSIONALITY REDUCTION BASED ON OUR SR FRAMEWORK (A SMALLER MOVIE VALUE INDICATES HIGHER VISUAL QUALITY)

Method	PCA	LPP	OLPP
Video 1	0.18	0.18	**0.17**
Video 2	**0.20**	0.26	0.23
Video 3	**0.20**	0.22	0.21
Video 4	0.11	0.13	**0.10**
Video 5	**0.21**	0.26	0.24
Video 6	**0.19**	0.27	0.26
Video 7	**0.29**	0.33	0.31
Video 8	**0.22**	0.24	0.22

2) Visual Quality Versus Computational Complexity: In our SR framework, instead of performing TS-SR [26] to all video frames, the HR details of non-key-frames are first reconstructed using BOBMC [17] and then refined by he proposed DTS-based scheme to maintain the temporal coherence. The main concern about using BOBMC is to achieve good tradeoff between visual quality and computational complexity. Even if individually performing TS-SR to each frame (i.e., $K = 0$), followed by performing the proposed DTS-SR can result in good SR quality, the computational cost will be high. Hence, in our SR framework, TS-SR is only performed on key-frames, whereas BOBMC is performed on non-key-frames, which can significantly reduce computation (see TABLE III).

Besides, to investigate the impact of DTS models on SR performance, we also implement two state-of-the-art nonlinear DTS models, High-order DTS (HO-DTS) [23] and high-order-SVD-DTS (HOSVD-DTS) [34], to replace the linear model in (8) used in the proposed DTS-SR. Fig. 10 illustrates three reconstructed SR frames for Video #2 using linear DTS method [22], HO-DTS, and HOSVD-DTS. The complexities of the DTS methods in [23], [34] are significantly higher than that of the linear model in [22], whereas the visual qualities of the reconstructed HR videos using these three DTS are almost visually indistinguishable due to the short distance between two neighboring key-frames and small patch size used for SR (see the demo videos provided in our project website [28]). Therefore, in this work we choose the linear model for the sake of its low complexity. But one can easily replace the DTS model used in the proposed framework.

Note, the selection of the interval K of key-frames will influence both the visual quality and computational complexity. Fig. 11 compares the visual qualities of reconstructed HR videos using our method with different values of K for the eight test videos. Fig. 9 shows that,

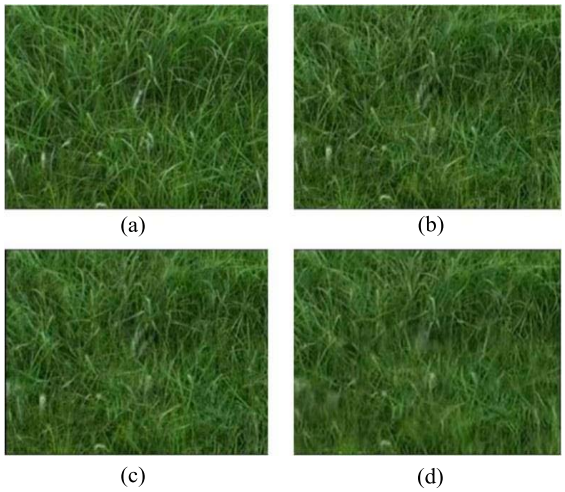


Fig. 10. The (a) ground truth frame and the synthesized frames using (b) linear DTS model in [22], (c) HO-DTS model in [23], and (d) HOSVD-DTS in [34].

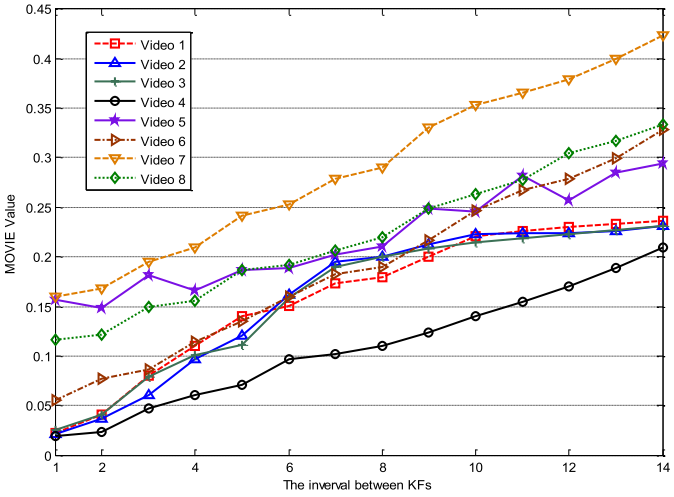


Fig. 11. Objective visual qualities evaluated by MOVIE for the four reconstructed SR Videos with different interval lengths between two successive key-frames (KFs).

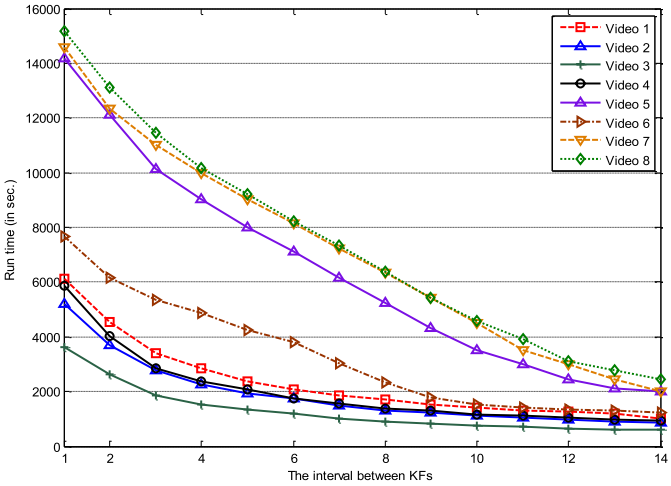


Fig. 12. Run-time complexity (in seconds) comparison of the proposed SR method for the four test videos with different interval lengths between two successive key-frames (KFs).

as K decreases, the MOVIE index will also decrease (i.e., better visual quality is achieved), because more frames (key-frames) will be upscaled by TS-SR, making the motion estimate/compensation process more accurate due to shorter distances between a non-key-frame and its two neighboring anchor key-frames. Nevertheless, the computational complexity will increase due to the increased number of key-frames as depicted in Fig. 12. In contrast, the MOVIE index increases with the value of K, leading to lower visual quality but also lower computational complexity. As a result, in our method, the parameter K can be adjusted to achieve a good tradeoff between visual quality and complexity.

3) SR for a Video With Mixed Dynamic-Textures and Non-Dynamic-Textures: Since the proposed DTS-SR scheme is designed for synthesizing HR dynamic textures, it may be inefficient for dealing with still-texture regions and may not be as effective as other SR schemes for upscaling non-textural regions. To address this problem, we can first separate dynamic-texture regions, static-texture regions, and non-textural regions using existing dynamic texture recognition techniques (e.g., in [37] and [38]). After the region classification, an adaptive SR manner can be used. For dynamic texture regions, the proposed DTS-SR scheme can do a good job, whereas for static-texture regions, the hybrid TS-SR/BOBMC (i.e., skipping the DTS-based refinement step in the proposed DTS-SR to reduce computational complexity) can be used. For the remaining non-textural regions, we can apply traditional interpolation or SR methods (e.g., SC-SR and NLIBP-SR). In our experiments, we adopt NLIBP-SR to upscale the non-textural regions. Finally, the three types of regions are combined to obtain HR frames. Several of our test videos (e.g., videos #1, #5–#8) are with mixed dynamic-textures and non-dynamic-textures and the results can be found in the project website [28].

V. Conclusion

In this paper, we proposed a video SR framework via dynamic texture synthesis to effectively enhance the resolution of a LR video while maintaining the temporal coherence of the reconstructed HR video. The proposed method divides the input LR video frames into key-frames and non-key-frames. We first apply the texture synthesis-based SR method to upscale each key-frame, followed by a low-complexity bi-directional overlapped block motion compensation method to reconstruct the HR details of each non-key-frame between two successive anchor key-frames. To address the problem of temporal incoherence artifacts, we have proposed a self-learning-based DTS-based refinement scheme to render the upscaled video based on the temporal dynamics learned from the input LR video. Our experimental results demonstrate that the proposed method outperforms the state-of-the-art super-resolution methods in terms of visual quality of reconstructed video both subjectively and objectively with reasonable computational complexity.

REFERENCES

[1] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[2] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[3] H. S. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 26, no. 6, pp. 508–517, Dec. 1978.

[4] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[6] W. Dong, L. Zhang, G. Shi, and X. Wu, "Nonlocal back-projection for adaptive image enlargement," in *Proc. 16th IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 349–352.

[7] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 349–356.

[8] C.-C. Hsu and C.-W. Lin, "Image super-resolution via feature-based affine transform," in *Proc. IEEE 13th Int. Workshop Multimedia Signal Process. (MMSP)*, HangZhou, China, Oct. 2011, pp. 1–5.

[9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[10] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. S. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[11] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[12] J. Ren, J. Liu, and Z. Guo, "Context-aware sparse decomposition for image denoising and super-resolution," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1456–1469, Apr. 2013.

[13] L.-W. Kang, B.-C. Chuang, C.-C. Hsu, C.-W. Lin, and C.-H. Yeh, "Self-learning-based single image super-resolution of a highly compressed image," in *Proc. IEEE 15th Workshop Multimedia Signal Process.*, Sardinia, Italy, Sep./Oct. 2013, pp. 224–229.

[14] M.-C. Yang and Y.-C. F. Wang, "A self-learning approach to single image super-resolution," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 498–508, Apr. 2013.

[15] F. Brandi, R. de Queiroz, and D. Mukherjee, "Super-resolution of video using key frames and motion estimation," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 321–324.

[16] S. H. Keller, F. Lauze, and M. Nielsen, "Video super-resolution using simultaneous motion and intensity calculations," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1870–1884, Jul. 2011.

[17] B. C. Song, S.-C. Jeong, and Y. Choi, "Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 274–285, Mar. 2011.

[18] Z. Xiong, X. Sun, and F. Wu, "Robust web image/video super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2017–2028, Aug. 2010.

[19] E. M. Hung, R. L. de Queiroz, F. Brandi, K. F. de Oliveira, and D. Mukherjee, "Video super-resolution using codebooks derived from key-frames," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1321–1331, Sep. 2012.

[20] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan. 2009.

[21] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa, "Video textures," in *Proc. ACM SIGGRAPH*, New Orleans, LA, USA, Jul. 2000, pp. 489–498.

[22] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *Int. J. Comput. Vis.*, vol. 51, no. 2, pp. 91–109, Feb. 2003.

[23] M. Hyndman, A. D. Jepson, and D. J. Fleet, "Higher-order autoregressive models for dynamic textures," in *Proc. Brit. Mach. Vis. Conf.*, Warwick, U.K., Sep. 2007, pp. 1–10.

[24] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: Image and video synthesis using graph cuts," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 277–286, Jul. 2003.

[25] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 795–802, Jul. 2005.

[26] Y. HaCohen, R. Fattal, and D. Lischinski, "Image upsampling via texture hallucination," in *Proc. IEEE Int. Conf. Comput. Photography*, Cambridge, MA, USA, Mar. 2010, pp. 1–8.

[27] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[28] *NTHU Video Super-Resolution Project*. [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/videoSR/index.html, accessed May 2014.

[29] J.-S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 564–571, Feb. 2014.

[30] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 16. Vancouver, BC, Canada, Dec. 2003.

[31] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.

[32] L. Shao and R. Jin, "Subspace learning for silhouette based human action recognition," *Proc. SPIE*, vol. 7744, p. 77441S, Jul. 2010.

[33] U. I. Bajwa, I. A. Taj, M. W. Anwar, and X. Wang, "A multifaceted independent performance analysis of facial subspace recognition algorithms," *PLoS ONE*, vol. 8, no. 2, p. e56510, 2013.

[34] R. Costantini, L. Sbaiz, and S. Süsstrunk, "Higher order SVD analysis for dynamic texture synthesis," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 42–52, Jan. 2008.

[35] C.-C. Hsu, L.-W. Kang, and C.-W. Lin, "Video super-resolution via dynamic texture synthesis," in *Proc. IEEE 16th Int. Workshop Multimedia Signal Process.*, Jakarta, Indonesia, Sep. 2014, pp. 1–6.

[36] M. Ghanbari, *Standard Codecs: Image Compression to Advanced Video Coding*, 3rd ed. London, U.K.: IET, 2011.

[37] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[38] B. Ghanem and N. Ahuja, "Maximum margin distance learning for dynamic texture recognition," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 223–236.

Chih-Chung Hsu received the B.S. degree in information management from the Ling-Tung University of Science and Technology, Taichung, Taiwan, in 2004, the M.S. degree in electrical engineering from the National Yunlin University of Science and Technology, Yunlin, Taiwan, in 2007, and the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2014.

He is currently a Post-Doctoral Researcher with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. His research interests mainly lie in computer vision, image/video processing, and image protection/watermarking.

Dr. Hsu received the top 10% Paper Award from the IEEE International Workshop on Multimedia Signal Processing in 2013.

Li-Wei Kang (S'05–M'06) received the B.S., M.S., and Ph.D. degrees in computer science from National Chung Cheng University, Chiayi, Taiwan, in 1997, 1999, and 2005, respectively.

He has been with the Graduate School of Engineering Science and Technology-Doctoral Program, and the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan, as an Assistant Professor, since 2013. He was with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as an Assistant Research Scholar, from 2010 to 2013, and a Post-Doctoral Research Fellow from 2005 to 2010. His research interests include multimedia content analysis and multimedia communications.

Dr. Kang has served as an Editorial Board Member of the *International Journal of Distributed Sensor Networks* and the Editor-in-Chief of the *Gate to Multimedia Processing* (Science Gate Publishing). He served as an Editorial Advisory Board Member of a book entitled *Visual Information Processing in Wireless Sensor Networks: Technology, Trends and Applications* (IGI Global), a Guest Editor of the *International Journal of Electrical Engineering* (Taiwan), the Special Session Co-Chair of APSIPA ASC 2012, and the Registration Co-Chair of APSIPA ASC 2013. He serves as the Demo/Exhibition Co-Chair of the IEEE ICCE-TW 2015. He received a top 10% Paper Award from the IEEE MMSP 2013.

Chia-Wen Lin (S'94–M'00–SM'04) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is an Adjunct Professor with the Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University (CCU), Chiayi, Taiwan, from 2000 to 2007. Prior to joining academia, he was with Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. His research interests include image and video processing and video networking.

Dr. Lin has served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the *IEEE Multimedia*, and the *Journal of Visual Communication and Image Representation*. He is an Area Editor of the *EURASIP Signal Processing: Image Communication*. He is also the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He served as the Technical Program Co-Chair of the IEEE International Conference on Multimedia and Expo (ICME) in 2010, and the Special Session Co-Chair of the IEEE ICME in 2009. His paper received the top 10% Paper Award by the IEEE MMSP 2013, and the Young Investigator Award by VCIP 2005. He was a recipient of the Young Faculty Awards by CCU in 2005 and the Young Investigator Awards by the National Science Council, Taiwan, in 2006.