World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:7, 2013

Video Super-Resolution Using Classification ANN

Ming-Hui Cheng and Jyh-Horng Jeng

***Abstract*—**In this study, a classification-based video super-resolution method using artificial neural network (ANN) is proposed to enhance low-resolution (LR) to high-resolution (HR) frames. The proposed method consists of four main steps: classification, motion-trace volume collection, temporal adjustment, and ANN prediction. A classifier is designed based on the edge properties of a pixel in the LR frame to identify the spatial information. To exploit the spatio-temporal information, a motion-trace volume is collected using motion estimation, which can eliminate unfathomable object motion in the LR frames. In addition, temporal lateral process is employed for volume adjustment to reduce unnecessary temporal features. Finally, ANN is applied to each class to learn the complicated spatio-temporal relationship between LR and HR frames. Simulation results show that the proposed method successfully improves both peak signal-to-noise ratio and perceptual quality.

***Keywords*—**Super-resolution, classification, spatio-temporal information, artificial neural network.

I. Introduction

VIDEO super-resolution methods construct a high-resolution (HR) frame from a set of successive low-resolution (LR) frames in a video sequence by using image/video processing techniques. The application of video super-resolution has gained significant attention in both academia and consumer electronics industries such as video surveillance [1], medical imaging [2], and satellite imaging [3]. In the context, how to explore the useful HR information from limited LR data to generate better super-resolved HR frame plays an important role. In the recent years, many machine learning algorithms have been applied in image/video super-solution [4], [5].

For machine learning based image super-resolution, K. Ni et al. [4] proposed an image super-resolution algorithm using support vector regression (SVR) to learn the relation between HR image and LR image in the transform domain. The input LR image is DCT-transformed and its feature vectors are extracted in the transform domain. To yield better results, a classification method by using Gaussian mixture model is employed to find corresponding parameters. Based on the parameters, the HR super-resolved image is estimated by mixing several SVR prediction results.

In the present paper, our proposed video super-resolution method involves four steps: classification, motion-trace volume collection, temporal adjustment, and artificial neural network (ANN) prediction. For the classifier, three AC coefficients in the frequency domain using discrete cosine transform (DCT) are adopted to group the pixels in the LR frames into five classes. To exploit the spatio-temporal information, we collect a motion-trace volume using motion estimation to track the object motion. Such volume eliminates the unfathomable object motion in the LR frames. The temporal lateral process for the volume is use to adjust the influencing weights to reduce unnecessary temporal features. Finally, a machine learning method is used to learn the complicated spatio-temporal relationship between the LR and HR frames. Simulation results are compared with nearest neighbor (NN) and Bicubic. The proposed method has higher peak signal-to-noise ratio (PSNR) values and shows better visual results.

This paper is organized as follows. In Section II, ANN is briefly described. Section III described the details of the proposed method. The simulation results are presented in Section IV, followed by the conclusions in Section V.

II. Artificial Neural Network

ANN is a biologically motivated learning machine inspired by the biological neuron and nervous system processes [6]. A feed-forward ANN with an input layer of $n+1$ nodes, one hidden layer of $m+1$ nodes with activation function f, and an output layer with p nodes is considered. The network architecture is shown in Fig. 1. Here, $x=[x_1 \;\; x_2 \;\; \cdots \;\; x_n]^T$ is the n -dimensional input vector, and $y=[y_1 \;\; y_2 \;\; \cdots \;\; y_p]^T$ is the p -dimensional output vector. The connection weight matrices are denoted as $\mathbf{v}=[v_{ji}]$ and $\mathbf{w}=[w_{kj}]$, where $1\le i\le n+1$, $1\le j\le m+1$, and $1\le k\le p$, respectively. For the theoretical model of a neuron, each input is multiplied by its weight w_i, and their total goes through the activation function f . The activation function can be an identity, linear, or nonlinear function. Some commonly used nonlinear functions are unipolar logistic, bipolar sigmoidal and hyperbolic tangent functions.

In this study, the back propagation (BP) algorithm is used to train the ANN. The BP algorithm approximates the nonlinear relationship of error between ANN output and the desired output data to adjusting the weight matrices **v** and **w** by using gradient decent algorithm. The combination of weights that minimizes the error is considered as the optimal solution to the learning problem.

M. H. Cheng is with the Computer Science and Information Engineering Department, National Chung Cheng University, Chiayi621Taiwan (e-mail: chengmh@cs.ccu.edu.tw).

J. H. Jeng is with the Information Engineering Department, I-Shou University, Kaohsiung 84001 Taiwan (e-mail: jjeng@isu.edu.tw).

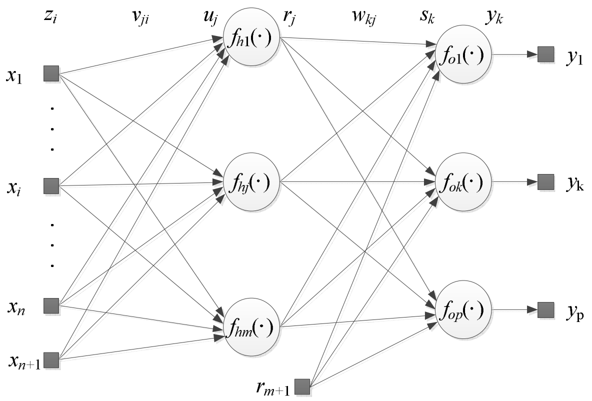

Fig. 1 Feed-forward ANN

III. Proposed Method

The proposed super-resolution method involves off our steps, namely, classification, motion-trace volume collection, temporal adjustment, and ANN training, so that more spatial and temporal data can be exploited and learned. Fig. 2 shows the block diagram of the proposed method. For a given pixel in an LR frame, its class label is first determined by the classifier. Then, the motion-trace volume g for the pixel is collected, obtaining the spatio-temporal information from neighboring LR frames. The volume is further adjusted using temporal lateral information, and it is labeled as x . Meanwhile, the corresponding HR pixels d from the HR frame are also collected to create a pair of training data (x,d) . Based on the class label, the corresponding ANN is trained separately to obtain weight matrices $\mathbf{v}$ and $\mathbf{w}$. Altogether, we have five training sets of ANNs and weight matrices.

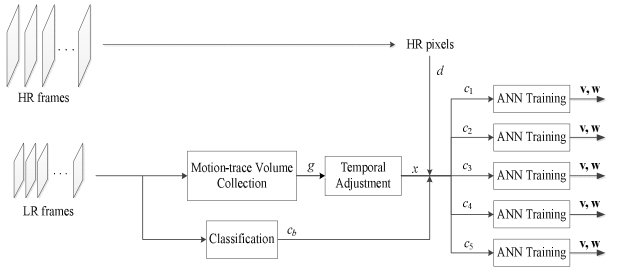

Fig. 2 The proposed super-resolution method in ANN training step

The prediction phase, i.e., super-resolution of LR to HR frame, of the proposed ANN method is shown in Fig. 3. Similarly, for every pixel in the LR frame, the class label is decided, and the motion-trace volume is collected, adjusted, and used as ANN input. The weight matrices are selected according to the class label so that ANN prediction can be performed to obtain the corresponding approximated HR patch, labeled as y . The reconstructed frame is denoted by $\hat{\mathbf{H}}$.

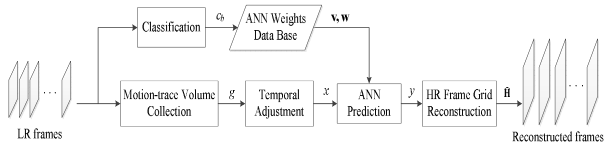

Fig. 3 The proposed super-resolution method in ANN prediction step

A. Classification

In the proposed classification method, the classifier classifies the pixel in the LR frame into one of the five classes. The classification result also shows the motion-trace volume class label, which is collected in next step. Here, the LR frame pixel generates a “ 5×5 ” patch centered at this pixel. Then, patch $g(i,j)$ is transformed using DCT.

The three lowest AC coefficients, namely, $G(1,0)$, $G(0,1)$, and $G(1,1)$ -denoted as V, H, and D, respectively, are used to classify the given patch into one of the five classes. The patch classification results is regarded as the class label of the given center pixel. The five classes consist of smooth and edged patches which include horizontal, vertical, diagonal, and sub-diagonal edges. Patch classification is performed according to V, H, and D together with two thresholds T_s and T_e .

B. Motion-Trace Volume Collection

The purpose of the motion-trace volume collection using motion estimation is to track the object motions. Such volume can remove the unfathomable object motion in the LR frames. For a given pixel (i,j) in LR frame $\mathbf{L}_t$, a 5×5 patch g_t is collected, with this pixel as the center. With g_t as the patch, we then search from the neighboring frames $\mathbf{L}_{t-2}$, $\mathbf{L}_{t-1}$, $\mathbf{L}_{t+1}$,and $\mathbf{L}_{t+2}$ to find the best matches g_{t-2} , g_{t-1} , g_{t+1} , and g_{t+2} , respectively, using motion estimation. Together with the g_t patch, a $5\times5\times5$ motion-trace volume is constructed for pixel (i,j) , denoted as g .After the motion-trace volume g is collected, it will be adjusted according to temporal information first and then serves as the input for ANN training and prediction.

C. Temporal Adjustment

For fast-motion object or complicated scenes, motion estimation cannot capture the correct objects. The $5\times5\times5$ motion-trace volume is temporally adjusted before being fed into the ANN training to exploit the temporal information using the proposed temporal lateral process, derived from the bilateral filter concept [7].Analogous to the bilateral filter, the pixel value $g_\tau(i,j)$ in the motion-trace volume is appropriately adjusted by

$$x_\tau(i,j) = g_\tau(i,j)\cdot\omega_\tau(i,j) \tag{1}$$

where (i,j) and τ are the spatial and temporal coordinates, $-2 \le i,j \le 2$ and $-2 \le \tau \le 2$, respectively. The $x_\tau(i,j)$ term indicates the adjusted pixel value after the temporal lateral process, and the $\omega_\tau(i,j)$ denotes the temporal lateral weights for pixel (i,j) in the L_τ frame. Here, the similarity of each patch in the motion-trace volume is tapped to determine the temporal lateral weight ω . The five patches, namely, g_{t-2} , g_{t-1} , g_t , g_{t+1} ,and g_{t+2} ,calculate their own similarity with center patch g_t to determine the weight, i.e.,

$$\omega_\tau(i,j) = \exp\left(-\frac{\rho(g_{t+\tau}, g_t)}{2\sigma_T^2}\right) \tag{2}$$

where $\rho(\cdot)$ is the similarity function and σ_T^2 is the standard deviation. Note that, $\omega_\tau(i,j)$ is the weight vector along the temporal coordinate, therefore for fixed τ , $\omega_\tau(i,j)$ has the same values for $-2 \le i,j \le 2$. The similarity function is implemented using mean-squared-error (MSE) in this study.

D.ANN Training and Prediction

In the proposed method, five ANNs are set up for the five classes to learn separately the spatio-temporal correlation between LR and HR data. In the training phase, one training set is set up for each ANN, and the BP algorithm is adopted to obtain the optimal weights. In this study, the ANN training is performed in an off-line mode.

The training set for a specific class for its corresponding ANN is constructed from the LR frames in the training video. In each (x,d) pair in the training set, input x has a $5\times5\times5$ dimension with the center pixel belonging to that class. If the magnification factor of the super-resolution is *l*, the desired output d has a dimension of $(l+2)\times(l+2)$, collected from the corresponding patch in the HR frame. During the prediction phase, only an $l\times l$ patch is obtained from the output. In other words, one pixel is expanded to a volume with a dimension of $5\times5\times5$, and ANN produces an $l\times l$ patch output. In the ANN training, the desired output $l\times l$ patch is extended to a $(l+2)\times(l+2)$ patch to reduce the patch artifact caused by discontinuities at the $l\times l$ HR patch boundaries. Considering $l=3$ as an example, the original and the extended patches are shown in Fig. 4.

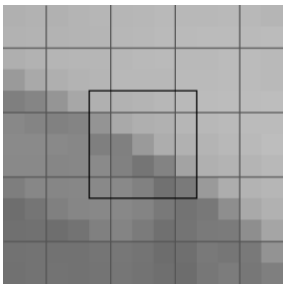

Fig. 4 The 5×5 extended HR patch (marked as black patch) and the centered 3×3 HR patch (marked as black patch) with *l*=3

The five training sets, labeled as S_1, S_2 , S_3 , S_4 , and S_5 , are used for a class of smooth patch, horizontal edge, vertical edge, diagonal edge, and sub-diagonal edge, respectively. Each of these training sets collects K pair training data. After the ANN training, five pairs of weight matrices $\mathbf{v}$ and $\mathbf{w}$ are obtained.

During ANN prediction, every pixel in the LR frame is classified first into one of the five classes. Thereafter, the motion-trace volume, which has a dimension of $5\times5\times5$, is collected and temporally adjusted. The classification result shows that the class label and the corresponding weight matrices $\mathbf{v}$ and $\mathbf{w}$ are obtained for ANN. Finally, the $(l+2)\times(l+2)$-dimensional output is obtained, and the center $l\times l$ HR patch is regarded as the super-resolved result. When all pixels in the LR frame are processed, an HR super-resolved frame is constructed.

IV.Simulation Results

In this study, the experiments are conducted using Borland C++ Builder on an Intel Core i5 CPU 2.5 GHz Microsoft Windows 7 platform. To evaluate the performance of the proposed method versus those of the previous approaches, three video sequences, namely, Akiyo, Miss American, and Foreman are used. For comparison purposes, we implement two interpolation-based methods, i.e., the nearest neighbor (NN) interpolation and the bicubic interpolation (Bicubic).

In the experiment, the magnification factor *l*is set to three ($l=3$). The LR frames are decimated from the original full-resolution clip with a factor of three followed by a corruption of additive Gaussian white noise with zero mean and variance of four. The test video sequences contain 100 frames.

In the ANN training phase, three clips (Akiyo, Foreman, and Carphone) are used as the training video. In the testing phase, three clips (Akiyo, Miss American, and Foreman) are evaluated. The HR frame $\mathbf{H}$ and the down sampled LR frame $\mathbf{L}$ have 348×288 and 116×96, sizes, respectively. The proposed method contains five classes. One ANN is set up for each class, as well as one training set for each ANN. The ANN training aims to obtain the optimal weights of each ANN from the corresponding training set. For each specific class, we randomly choose $K=15000$pixels of that class from the LR frame in the training video, and a pair of training data is collected for each pixel. The input is the temporally adjusted motion-trace volume in which the search range for motion estimation is set to seven. The output is the corresponding HR extended patch of that pixel, which has a size of 5×5 .The number of input nodes n , hidden nodes m , and output nodes p are set to 125, 56, and 25, respectively. Using additional bias nodes as shown in Fig. 1, the dimension of weight matrix $\mathbf{v}$ is 126×56 ,and that of weight matrix $\mathbf{w}$ is 57×25. In the classification process, the thresholds are empirically set as $T_s=10$ and $T_e=3.5$. The temporal lateral parameter σ_t for the motion-trace volume is empirically set to 15.

Fig. 5 shows the super-resolved result of the 35th frame of the Foreman video sequence. The clip, which contains moderate-motion objects, belongs to the training videos. Fig. 6 shows the super-resolved result of the 19th frame of the Miss American video sequence. The clip, which contains small-motion objects, not belongs to the training videos. Figs. 5 and 6 show visual quality of the proposed method are superior to the other two methods. The averaged PSNR of super-resolved frames for the two methods and the proposed method in the three test video sequences are shown in Table I. Again, the proposed method still has better results.

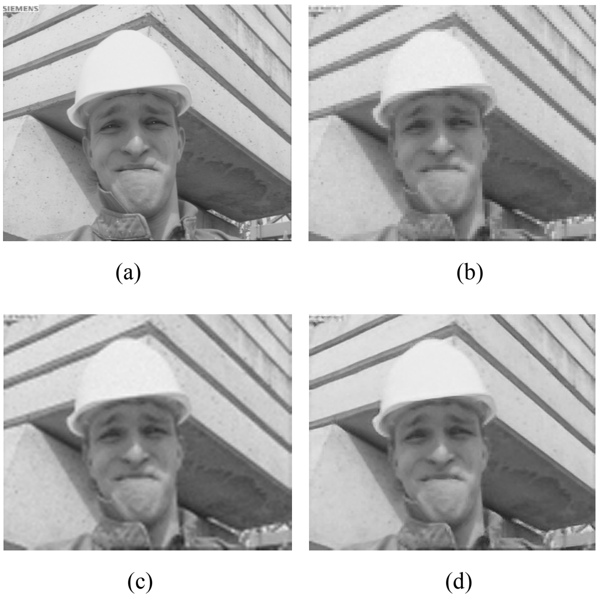

Fig. 5 The 35th frame of the Foreman video sequence with magnification factor of three (a) ground truth (b) NN (c) Bicubic (d) proposed method

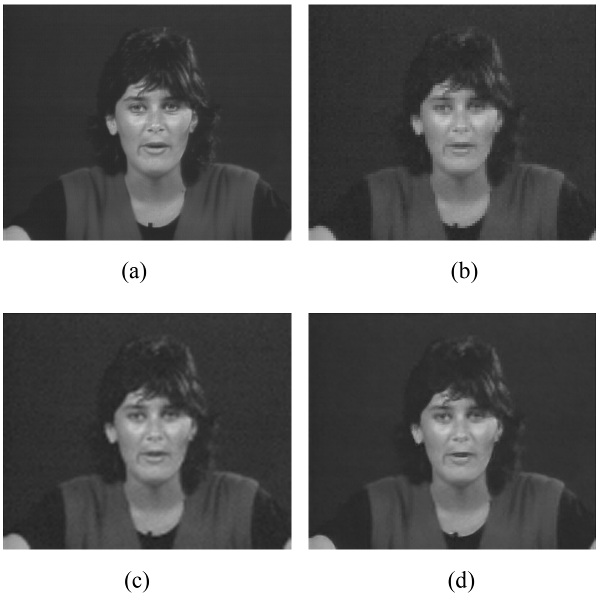

Fig. 6 The 19th frame of the Miss American video sequence with magnification factor of three (a) ground truth (b) NN (c) Bicubic (d) proposed method

TABLE I
THE AVERAGE PSNR (DB) RESULTS FOR THE THREE VIDEO SEQUENCES BETWEEN THE TWO METHODS AND THE PROPOSED METHOD

Sequence	Akiyo	Miss American	Foreman
NN	29.4	35.1	28.1
Bicubic	30.9	37.9	30.4
Proposed	32.6	38.1	31.8

V. CONCLUSIONS

In this study, a learning-based approach for frame resolution enhancement in video super-resolution has been proposed. To improve the ANN training process, the motion-trace volume is captured using motion estimation. To improve further the learning capability, a fast classification method is proposed to classify the motion-trace volume and exploit more spatial information. The temporal adjustment method employing temporal lateral is applied in the motion-trace volume to exploit the temporal information. Then, ANN is used to learn the spatio-temporal relationship between the LR and HR data. Based on the ANN training results, the proposed method indeed improves the average PSNR and the perceptual quality of the super-resolved frame.

ACKNOWLEDGMENTS

The research reported here was supported by the National Science Council, Taiwan, under grant NSC 101-2221-E-214-073.

REFERENCES

[1] G. Caner, A. M. Tekalp, W. Heinzelman, Super resolution recovery for multi-camera surveillance imaging, in Int. Conf. on Multimedia and Expo, Baltimore, USA, 2003, pp. 109-112.

[2] J. A. Kennedy, O. Israel, A. Frenkel, R. BarShalom, H. Azhari, Super-resolution in PET imaging, IEEE Trans. on Medical Imaging 25 (2) (2006) 137-147.

[3] F. Li, X. Jia, D. Fraser, A. Lambert, Super resolution for remote sensing images based on a universal hidden Markov tree model, IEEE Trans. on Geoscience and Remote Sensing 48 (3) (2010) 1270-1278.

[4] K. S. Ni, T. Q. Nguyen, Image super resolution using support vector regression, IEEE Trans. on Image Processing 16 (6) (2007) 1596-1610.

[5] H. Takeda, P. Milanfar, M. Protter, M. Elad, Super-resolution without explicit sub-pixel motion estimation, IEEE Trans. on Image Processing 18 (9) (2009) 1958-1975.

[6] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Prentice Hall, Ontario, 2008.

[7] M. Elad, On the origin of the bilateral filter and ways to improve it, IEEE Trans. on Image Processing 11 (10) (2002) 1141-1151.