Signal Processing 93 (2013) 2612–2625

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Classification-based video super-resolution using artificial neural networks

Ming-Hui Cheng [a], Kao-Shing Hwang [b,d], Jyh-Horng Jeng [c,*], Nai-Wei Lin [a]

[a] Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621, Taiwan
[b] Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[c] Department of Information Engineering, I-Shou University, Kaohsiung 84001, Taiwan
[d] Department of Electrical Engineering, National Chung Cheng University, Chiayi 621, Taiwan

ARTICLE INFO

Article history:
Received 12 September 2012
Received in revised form
16 February 2013
Accepted 18 February 2013
Available online 26 February 2013

Keywords:
Artificial neural network (ANN)
Bilateral filter
Classification
Motion estimation
Super-resolution

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.

1. Introduction

Image or video super-resolution is an important issue in contemporary applications. Two main reasons prompt the demand for such technology, namely, low-resolution (LR) capturing and low bandwidth communication, in which high-resolution (HR) display is required at the user end. Super-resolution applications, such as video surveillance [1,2], medical imaging [3,4], and satellite imaging [5–7], have gained significant interests in consumer electronics, academia, and industries.

Image super-resolution methods construct an HR image by exploring the spatial correlations from a set of LR images. Freeman et al. [8] proposed an example-based super-resolution algorithm. In their method, training mechanism was used to learn the relation between LR and HR patches. The training set consists of a large number of LR patches and their corresponding HR patches which are collected from a set of training images. For each LR patch, a best match is searched in that training set, for which the corresponding HR patch is the predicted super-resolved image. Later, Li et al. [9] adopted the classification approach to reduce computation complexity and further improve the prediction accuracy.

In video super-resolution, construction of one HR frame can be performed from a set of successive LR frames, instead of from just one LR frame. Three main steps are involved to obtain video super-resolution, namely, registration (motion estimation), interpolation, and restoration, which are implemented either sequentially or simultaneously [10]. Traditional sequential techniques largely rely on the availability of accurate motion estimation. Narayanan et al. [11] employed the partition-based weighted sum filters, which utilized the iterative gradient-based registration techniques, to estimate accurate local motions prior to interpolation and restoration. Su et al. [12] proposed a

* Corresponding author. Tel.: +886 7 6577207; fax: +886 7 6578944.
E-mail addresses: jjeng@isu.edu.tw (J.-H. Jeng).

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.02.013

framework to combine rule-based and learning-based classification, single-image enhancement, and feature extraction. They divided the HR frame into adaptive-size blocks to overcome the registration error. In the meanwhile, the method makes the best use of the advantages of different conventional super-resolution algorithms. Encouraging super-resolution improvements are achieved for real-life videos.

Non-motion estimation-based sequential methods have recently been proposed [13,14]. Protter et al. [13] generalized the non-local means (NLM) algorithm (a denoising algorithm) without explicit motion estimation in which mobile search strategy and adaptive patch size method were proposed to improve the efficiency of the NLM algorithm in [14]. However, the collected volume did not correctly track the object motion and many artifacts appeared. In addition, much processing time was required because many parameters had to be calculated online.

Simultaneous video super-resolution approaches include frequency domain-based [15,16], spatial domain-based [17–22] and machine learning-based methods [23–25]. Tsai and Huang [15] exploited the shift property of the Fourier transform and developed a set of system equations to construct the relationship between HR frames and LR frames in the frequency domain. Kim et al. [16] extended Tsai and Huang's approach and proposed a recursive least-square solution in which motion blurring and additive noise are also included. It assumes that all LR frames have the same motion blurring and noise property.

In the spatial domain approaches, Costa and Bermudez [17] proposed a strategy to reduce outlier effects for video super-resolution. Li et al. [18] constructed two new regularization items, locally adaptive bilateral total variation and gradient consistency, to preserve edges and flat regions in the LR frames, respectively. Farsiu et al. [19] proposed bilateral total variation operator as a regularization term to form a new observation model. This combination makes the algorithm robust to motion outliers.

Shen et al. [20] used joint maximum a posteriori (MAP) to formulate motion estimation, segmentation, and super-resolution together. They adopted the cyclic coordinate decent process to solve the formulation, in which the motion fields, the segmentation fields, and the HR image are treated as unknowns and estimated jointly using the available data. Yuan et al. [21] applied the U-curve method on the MAP-based super-resolution model with Laplacian prior regularization to select an optimal parameter adaptively. Keller et al. [22] proposed an energy-based algorithm to replace MAP to formulate motion-compensated video super-resolution model. In their formulation, the estimation of super-resolution sequence and its flow field are jointly considered. Then, the calculation of variations leads to a coupled system of partial differential equations for video sequence and motion estimation. Compared to the methods in frequency domain, the regularization method may produce better results whereas it is computationally expensive.

In machine learning-based methods, Ni and Nguyen [23] proposed an algorithm that uses support vector regression to acquire the relationship between HR and LR images in the transformed domain. Their method was only applied to images, rather than videos. For video super-resolution, object motion has a critical function in modeling the video signals. Previous approaches usually depend on accurate (pixel and sub-pixel) motion estimation to explore spatio-temporal information. Takeda et al. [24] proposed a framework that used motion estimation without sub-pixel accuracy. To obtain better results, multi-dimensional kernel regression was employed at high computation costs. Yang et al. [25] proposed a sparse-coding method in which LR and HR patch pairs share the same sparse representation in terms of coupled dictionaries jointly trained. The sparse representation of an LR patch can be applied to the HR dictionary to generate an HR patch.

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. Combination of these coefficients reflects the edge properties near the pixel in the LR frame; thus, spatial information can be obtained. 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. Failure of motion estimation to capture the correct objects due to fast motions or complicated scenes has been known. Hence, we adopt the temporal lateral process for the volume by adjusting 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. This paper uses five ANNs for five corresponding classes. The learning process is time consuming. However, it can be done off-line, and only a little computation is required during the prediction phase. Simulation results are compared with nearest neighbor (NN), Bicubic, Lanczos, NLM, and 3D ISKR methods. The proposed method has higher peak signal-to-noise ratio (PSNR) values and shows better visual results. Moreover, it is faster than the NLM method.

This paper is organized as follows. In Section 2, ANN is briefly described. Section 3 described the details of the proposed method. The simulation results are presented in Section 4, followed by the conclusions in Section 5.

2. Artificial neural network

ANN is a biologically motivated learning machine inspired by the biological neuron and nervous system processes [26,27]. ANN functions as a powerful computational tool for nonlinear prediction problems in various applications, including image coding, pattern recognition, and medical imaging [28–30].

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,