Neurocomputing 94 (2012) 140–151

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

# Video super-resolution with 3D adaptive normalized convolution

Kaibing Zhang [a], Guangwu Mu [a], Yuan Yuan [b,*], Xinbo Gao [a], Dacheng Tao [c]

[a] School of Electronic Engineering, Xidian University, Xi'an 710071, China
[b] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
[c] Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering & Information Technology, University of Technology, Sydney, NSW 2007, Australia

ARTICLE INFO

*Article history:*
Received 9 May 2011
Received in revised form
10 March 2012
Accepted 11 March 2012
Communicated by Tao Mei
Available online 22 May 2012

*Keywords:*
Normalized convolution (NC)
Motion estimation
Video super-resolution (SR)

ABSTRACT

The classic multi-image-based super-resolution (SR) methods typically take global motion pattern to produce one or multiple high-resolution (HR) versions from a set of low-resolution (LR) images. However, due to the influence of aliasing and noise, it is difficult to obtain highly accurate registration with sub-pixel accuracy. Moreover, in practical applications, the global motion pattern is rarely found in the real LR inputs. In this paper, to surmount or at least reduce the aforementioned problems, we develop a novel SR framework for video sequence by extending the traditional 2-dimentional (2D) normalized convolution (NC) to 3-dimentional (3D) case. In the proposed framework, to bypass explicit motion estimation, we estimate a target pixel by taking a weighted average of pixels from its neighborhood. We further up-scale the input video sequence in temporal dimension based on the extended 3D NC and hence more video frames can be generated. Fundamental experiments demonstrate the effectiveness of the proposed SR framework both quantitatively and perceptually.

## 1. Introduction

The objective of multi-image-based super resolution (SR) technique is to produce one or multiple high-resolution (HR) images from a set of low-resolution (LR) inputs. By super-resolving the LR images, it becomes possible to break through the resolution limitation of image acquisition devices and obtain one or more HR images that the traditional digital cameras cannot capture from a real scene.

The classical image SR reconstruction technique consists of three fundamental steps: registration, fusion, and post-processing. Before fusion, it is necessary to know how the LR inputs are generated from an underlying scene. Fig. 1 illustrates a common degradation model showing how an HR image is distorted and how the LR images are formed. From Fig. 1 we can see that, due to atmospheric, lens and optical fuzzy or other equipment problems, the real sequences captured by a CCD sensor are significantly degraded. Therefore, the SR reconstruction aims to recover the original HR image from a set of observed LR images, solving the inverse problem of the image formation.

The SR technique was first proposed by Tsai and Huang in frequency domain [1]. Since then the SR reconstruction has received intensive attentions in image processing communities and a variety of SR algorithms have been proposed. Roughly, the existing multi-image-based SR approaches can be classified into two categories: frequency domain based and spatial domain based methods. Rhee et al. [2] proposed a DCT-based method. Bose et al. [3] presented a recursive total least squares algorithm. Ur and Gross [4] presented a generalized sampling theorem in the frequency domain for SR reconstruction. The frequency-domain methods are simple and easy to implement. However, the SR performance of this family of methods is limited by the presumed imaging model and the preclusion of any prior knowledge from the reconstruction process. Since SR problem is inherently ill-posed, it is crucial to incorporate some prior knowledge to make SR estimate well-posed. Consequently, for years, the spatial domain based methods have become popular. The representative methods include iterative back projection (IBP) [5], probability methods (e.g., ML and MAP) [6], [7], the projection onto-convex-sets approaches [8–11]. Compared with the frequency domain methods, the major advantage of spatial domain methods lies in more flexibility in motion and degradation models. Particularly, some prior knowledge can be incorporated into the reconstruction process to obtain more stable SR estimate. Nonetheless, the weaknesses of these algorithms are relatively complex and computationally intensive.

For the classic multi-image-based SR methods, a key step is to perform an accurate registration between the LR images before fusion. Unfortunately, most registration algorithms are limited to the global motion model. If a local motion pattern, such as appearance or disappearance of some new objects among images happens, the classic SR reconstruction approaches cannot work well. Recent progresses have focused on a strategy without explicit motion estimation. Takeda et al. [20] proposed a 3-D ISKR for SR reconstruction by introducing a steerable kernel

[*] Corresponding author.
*E-mail addresses:* yuany@opt.ac.cn, yuan369@hotmail.com (Y. Yuan).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.03.012

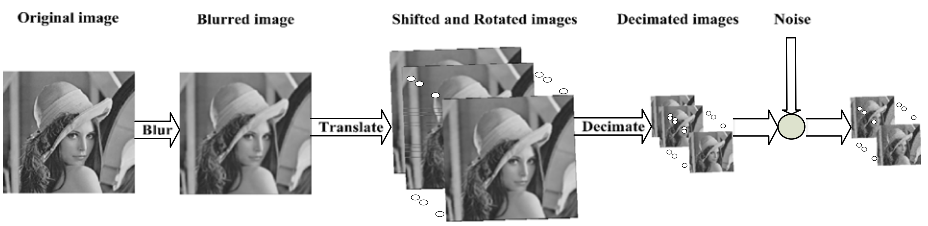

**Fig. 1.** A common image degradation model.

method [12]. Protter et al. [13] presented a generalized non-local means algorithm for SR reconstruction with no explicit motion estimate. Danielyan et al. [14] presented a Video-BM3D SR by extending block-matching 3-D filter [15]. Besides, a probabilistic motion estimation SR approach was proposed in [16]. The major advantage of these methods lies in that they do not require any explicit motion estimation and are capable of handling video sequence that contains complex motion patterns. The weakness lies in that: the Video-BM3D method is prone to blurring edges; the 3-D ISKR method suffers from intensively computational complexity and requires multiple iterations; the non-local means-based method tends to generate unwanted artifacts.

In this paper, we propose a new video SR framework, called as 3D adaptive NC, without explicit motion estimation. The proposed SR algorithm aims to maintain sharp edges, to suppress annoying artifact, and to take a moderate computational complexity. Our idea is inspired from the traditional normalized convolution [18] that assumes that a target pixel can be estimated by taking a weighted average of its neighborhood pixels with similar characteristics. Based upon this assumption, a novel video SR framework is developed by utilizing the local structure of video sequence during reconstruction. By analyzing the local structure and similar characteristics, we distribute different weighs to different pixels around the target pixel and therefore achieve SR reconstruction without explicit motion estimation. Furthermore, the extended 3D adaptive NC can up-scaling the video sequence along temporal dimension, obtaining more HR frames.

The rest of the paper is organized as follows. Section 2 shows a brief overview of the traditional as well as the adaptive normalized convolution algorithm, including its application in SR reconstruction [17]. Section 3 shows the proposed 3D adaptive NC algorithm and its application in video SR reconstruction. Experimental results are presented in Section 4 and Section 5 concludes.

## 2. 2D robust and adaptive normalized convolution

In this section, we will briefly review the normalized convolution (NC) [18] and its application to SR. The NC is a technique that performs a general convolution operation on tensor signals by modeling or analyzing the local structure of an image. The traditional NC separates both data and operator into a signal part and a certainty part. In the case of uncertain data, the certainty of the data must be estimated simultaneously. For a missing pixel, the certainty is just simply set to zero. Besides, an applicability function is applied in the localization operator. Similarly, this paper sets the applicability function to zero when a defined domain is outside of a given window.

### 2.1. Notations

For convenience, some important notations are described briefly. $\{s, s_0\}$ and $\{x, y, t\}$ represent the global spatial coordinate and the local spatial one, respectively. $f(s)$ stands for a tensor representation of an input signal (e.g., image gray value). The certainty function and applicability function are denoted as $c(s, s_0)$ and $a(s, s_0)$, respectively. $B(x)$ represents a filter basis of operator. Generally, a polynomial basis is chosen for normalized convolution. $P$ represents a projection coefficient vector. In next subsection, we will introduce the polynomial-based NC. We introduce the subscripts to distinguish the 2D and 3D NC.

### 2.2. Polynomial-based 2D adaptive NC

Given a polynomial basis, $B_{2D}(x)$ can be constructed from $N$ input data within their local coordinates. We set $B_{2D}(x)$ to $\{1, x, y, x^2, y^2, xy, ...\}$, where $1=\begin{bmatrix}1 & 1 & \ldots & 1\end{bmatrix}^T$ ($N$ entries contained), $x=\begin{bmatrix}x_1 & x_2 & \ldots & x_N\end{bmatrix}^T, x^2=\begin{bmatrix}x_1^2 & x_2^2 & \ldots & x_N^2\end{bmatrix}^T$ and so on. By using the polynomial basis, the NC can be denoted as a Taylor series expansion within a local neighborhood centered at $s_0$=$\{x_0, y_0\}$, i.e.,

$$\hat{f}_{2D}(s,s_0)=p_0(s_0)+p_1(s_0)x+p_2(s_0)y+p_3(s_0)x^2+p_4(s_0)xy+p_5(s_0)y^2+\cdots, \tag{1}$$

where $s$ is the global coordinate and $\{x, y\}$ represents the local coordinate with respect to the center pixel. $\hat{f}_{2D}(s, s_0)$ is an estimated intensity value at $s$ when expanded at the centered analysis pixel $s_0$. $P_{2D}$ is denoted as a vector $\begin{bmatrix}p_0, p_1, \cdots, p_m\end{bmatrix}^T$to represent the projection coefficients.

For the 2D robust and adaptive NC, the certainty function is defined as:

$$c_{2D}(s, s_0)=\exp\left(-\frac{\left|f(s)-\hat{f}_{2D}(s,s_0)\right|}{2\sigma_r^2}\right), \tag{2}$$

where $c_{2D}(s,s_0)$ is the robust certainty depending not only on the global coordination $s$, but also the analysis center coordination $s_0$. The parameter $\sigma_r$ defines an acceptable range of the residual error.

The family of applicability functions that defines the localization of the convolution operator in the traditional NC is given as

$$a=\begin{cases} r^{-\alpha}\cos^{\beta}\left(\frac{\pi r}{2r_{max}}\right), & r<r_{max} \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $r$ denotes the space distance to the analysis center. Both $\alpha$ and $\beta$ are positive integers. Different values of $\alpha$ and $\beta$ affect the sharpness of the application function. Fig. 2 shows an example of an applicability function, where the values of $\alpha$ and $\beta$ are 0 and 2 respectively, and $r_{max}$ is equal to 8.

When a structure-adaptive applicability function is formed, the initial estimation of the output intensity $I$ and the gradient images $I_x$ and $I_y$ are employed to estimate the gradient structure tensor (GST) [19], i.e.,

$$GST_{2D}=\overline{\nabla I_{2D}\nabla I_{2D}^T}=\begin{bmatrix}\overline{I_x^2} & \overline{I_xI_y} \\ \overline{I_xI_y} & \overline{I_y^2}\end{bmatrix}=\lambda_{\mu}\mu\mu^T+\lambda_{\nu}\nu\nu^T, \tag{4}$$