J. Vis. Commun. Image R. 31 (2015) 225–230

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Short Communication

Dictionary based surveillance image compression ☆

Jing-Ya Zhu, Zhong-Yuan Wang *, Rui Zhong, Shen-Ming Qu

NERCMS, School of Computer, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 20 March 2014
Accepted 1 July 2015
Available online 6 July 2015

Keywords:
Video surveillance
Image compression
Dictionary learning
Sparse representation
DCT transform
Still images
JPEG
JPEG2000

ABSTRACT

Common image compression techniques suitable for general purpose may be less effective for such specific applications as video surveillance. Since a stationed surveillance camera always targets at a fixed scene, its captured images exhibit high consistency in content or structure. In this paper, we propose a surveillance image compression technique via dictionary learning to fully exploit the constant characteristics of a target scene. This method transforms images over sparsely tailored over-complete dictionaries learned directly from image samples rather than a fixed one, and thus can approximate an image with fewer coefficients. A set of dictionaries trained off-line is applied for sparse representation. An adaptive image blocking method is developed so that the encoder can represent an image in a texture-aware way. Experimental results show that the proposed algorithm significantly outperforms JPEG and JPEG 2000 in terms of both quality of reconstructed images and compression ratio as well.

1. Introduction

Compression of images relies on the ability to capture and exploit temporal, spatial and perceptual redundancies in images. The most common image compression approach is based on a framework equipped with transform coding, quantization and entropy coding. Transform coding is a widely used image compression technique, where entropy reduction can be achieved by decomposing the image over a dictionary which provides compaction. In particular, a Fourier-related transform such as the Discrete Cosine Transform (DCT) is widely used. The more recently developed wavelet transform is also used extensively. Existing algorithms, such as JPEG [1] and JPEG2000 [2], utilize fixed DCT and wavelet dictionaries for pixel-domain transform.

In contrast to fixed dictionaries, the idea of image compression was recently extended from approaches based on linear transform towards sparsity promoting coding schemes with the help of learned over-complete dictionaries. A learned over-complete dictionary is optimized for a specific class of images and thus provides a more flexible and faithful representation of images as compared to a fixed dictionary. Bryt and Elad [3] are the first to employ a learnt over-complete dictionary successfully in image compression, who proposed an algorithm for facial image compression based on a K-SVD dictionary. OMP (Orthogonal Matching Pursuit) [4] is employed to select the most appropriate dictionary elements in sparse representation (SR). Their method is shown to achieve an improvement over JPEG2000 for facial images owing to the optimized dictionaries. However, it uses an image alignment procedure as an essential pre-process stage, which is hard to do in practical applications.

Following the work of [3], Ref. [5] used dictionaries trained by recursive least squares dictionary learning algorithm (RLS-DLA) to compress images, but its compression efficiency is still lower than JPEG 2000s. In an iteration-tuned dictionary (IDT) scheme [6], a single hierarchical IDT is pre-trained for a specific class of images and is used to compress the input image patches. The results tested with facial images showed that it can convincingly outperform JPEG and JPEG2000 for this class of images. In [7], the authors used a dictionary specifically trained over the image content. Because this approach requires the dictionary as part of the compressed stream, its performance substantially outperforms JPEG but does not approach JPEG2000. A coding scheme incorporating dictionaries learned by Independent Component Analysis (ICA) was proposed in [8], which also gave better results than discrete cosine transform.

However, most of the above mentioned research is related to facial images while there is a lack of research on other specific classes of images. In recent years, there has been a growing interest in the study of video surveillance. A major aspect of any video surveillance system is to efficiently compress the huge volumes of recorded video to facilitate subsequent processing. Meanwhile, still image compression is also an indispensible task for almost all of the video surveillance systems. In a monitoring and alarm system, the video encoder is always accompanied with a still image

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.

E-mail address: wzy_hope@163.com (Z.-Y. Wang).

http://dx.doi.org/10.1016/j.jvcir.2015.07.002
1047-3203/© 2015 Elsevier Inc. All rights reserved.

encoder to deliver snapshot images once upon the alarming is triggered. Besides, some monitoring systems are only equipped with still image encoder to capture discontinuous images for specific purposes. For example, the highway toll stations take pictures for every passing vehicle and store them for possible investigations. Traffic violation monitoring system needs to take about ten license plate pictures of the passing vehicle, which will be used for investigating such traffic violations as passing through red lights, running along the wrong side of the road. The images in one surveillance video can be seen as a specific class of images in that most of the images share similar backgrounds. Therefore, a single image in surveillance video may be better compressed with sparse representation over scene tailored over-complete dictionaries trained by image samples from the surveillance video.

In this paper, we propose a dictionary based surveillance image compression algorithm to reduce the data size in snapshot image transmission. Basically, the conventional transform-domain representation is replaced by a sparse representation over trained dictionaries. We perform adaptive dictionary learning and establish a set of dictionaries which can represent the video surveillance images efficiently. Since the dictionaries are shared by both encoder and decoder simultaneously, only coefficients of sparse representation need to be transmitted, which reduces the size of data significantly. Also, an adaptive image blocking method is introduced to reduce the number of elements used in dictionaries to represent a given block. In practice, an offline learned dictionary for certain typical scenarios is pre-loaded into encoder and decoder. Since video encoder and still image encoder usually co-exist in a system, a more flexible manner is that using the decoded video to adaptively train dictionary. When enough representative images are collected from video decoder and the dictionary training is completed, our algorithm is put into practice, otherwise snapshot images are still compressed with traditional JPEG. Experimental results show that the proposed algorithm enjoys better performance than JPEG and JPEG2000 on the premise of guaranteeing the quality of the reconstructed image.

The remainder of this paper is organized as follows. Section 2 briefly reviews dictionary based sparse representation. Section 3 particularly presents our proposed dictionary based compression method for surveillance images. Experimental results and analyses are provided in Section 4, and we conclude this paper in Section 5.

2. Preliminaries

In dictionary-learning-based sparse representation, a signal $\boldsymbol{y} \in \mathbf{R}^n$ can be represented as a sparse linear combination of elements in an over-complete dictionary $\boldsymbol{D} \in \mathbf{R}^{n \times k}$. The representation of $\boldsymbol{y}$ may either be exact $\boldsymbol{y} = \boldsymbol{Dx}$ or approximate, $\boldsymbol{y} \approx \boldsymbol{Dx}$, satisfying $\|\boldsymbol{y} - \boldsymbol{Dx}\|_p \leqslant \varepsilon$. The vector $\boldsymbol{x} \in \mathbf{R}^k$ contains the representation coefficients of the signal $\boldsymbol{y}$. In approximation methods, typical norms used for measuring the deviation are the l_p-norm for $p = 1, 2$, and ∞. In this paper, we focus on the case of $p = 2$. If the dictionary $\boldsymbol{D}$ is a full-rank matrix and $n < k$, the solution of $\boldsymbol{x}$ is infinite, so some constraints on the solution must be set. We aim to approximate $\boldsymbol{y}$ with the smallest number of dictionary elements. The sparse representation can be found by solving

$$\min_x \|x\|_0 \quad s.t. \quad \|y - Dx\|_2 \leqslant \varepsilon \tag{1}$$

We expect that a dictionary can be trained to better fit the sample data. So the minimization problem in Eq. (1) can be converted to find the best dictionary in the given error value ε for the sparse representation of $\boldsymbol{y}$ as follows:

$$\min_{x,\mathrm{D}} \|x\|_0 \quad s.t. \quad \|y - Dx\|_2 \leqslant \varepsilon \tag{2}$$

The dictionary is trained to provide a better representation of the actual signal when the error value is less than or equal to ε.

3. Proposed method

The proposed algorithm consists of two main processes: An offline dictionary training process and an online image compression process. A block diagram of the compression method is given in Fig. 1.

The trained dictionaries are used to describe the original image in sparse representation module. The results of sparse representation include partition mode, DC value and sparse matrix. They are quantified and encoded, forming the data of the compressed image. For decoding the compressed image, a mirror procedure is manipulated.

3.1. Dictionary training

Dictionary training process is an off-line procedure, preceding any image compression. We use a set of surveillance images from a single surveillance video to learn a dictionary. To agree with the block-wise coding fashion, we need first divide training images into blocks. Since the background variations of the surveillance video are relatively constant, the background parts of a surveillance image can be divided into larger blocks, which thus results in fewer number of dictionary elements in sparse representation. Instead, the foreground parts are described in smaller blocks so that they can be linearly decomposed with higher accuracy. We train three dictionaries in terms of three different sizes of blocks: 16×16, 8×8 and 4×4.

The surveillance images in the training set are divided into different size of blocks, and DC values of each block are calculated and subtracted from all the blocks in the set. We employ K-SVD [9] to train the dictionaries. Prior to the training process, the error value ε in Eq. (2) needs to be determined. The value of ε decides the tolerance of the reconstruction error of sparse approximation, which means that the larger value allows a larger reconstruction error, and vice versa. Therefore, we can figure out a suitable ε to satisfy a desirable quality of reconstructed images. We consider PSNR as a metric to measure the quality of the reconstructed image.

$$PSNR = 10 \times \lg\left(\frac{(2^n - 1)^2}{MSE}\right) \tag{3}$$

Further, MSE (minimum squared error) is calculated by

$$MSE = \frac{\sum_1^{b^2}(I_1 - I_2)^2}{b^2} \tag{4}$$

where $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ denote the pixel values of the original block and the reconstructed block, respectively. b is the side length of each block. For images of 8-bit depth, n is set to 8. We expect that PSNR of the reconstructed block is equal to or more than p. According to the Eqs. (3) and (4), we thus have

$$MSE = \frac{\sum_1^{b^2}(I_1 - I_2)^2}{b^2} \leqslant \frac{(2^n - 1)^2}{10^{\frac{p}{10}}} \tag{5}$$

Since ε satisfies

$$\|y - Dx\|_2 = \sqrt{\sum_1^{b^2}(I_1 - I_2)^2} \leqslant \varepsilon \tag{6}$$

in dictionary training process, the following holds

$$MSE \leqslant \frac{\varepsilon^2}{b^2} \tag{7}$$