IJCSI International Journal of Computer Science Issues, Volume 15, Issue 5, September 2018
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.5281/zenodo.1467648

Non-local Image Denoising by Using Bayesian Low-rank Tensor Factorization on High-order Patches

Lihua Gui[*], Xuyang Zhao[*♯], Qibin Zhao[♭♯] and Jianting Cao[*]

[*] **Graduate School of Engineering, Saitama Institute of Technology**
Fukaya, Japan
[♯] **Tensor Learning Unit, RIKEN AIP**
Tokyo, Japan
[♭] **School of Automation, Guangdong University of Technology**
Guangzhou, China

Abstract

Removing the noise from an image is vitally important in many real-world computer vision applications. One of the most effective method is block matching collaborative filtering, which employs low-rank approximation to the group of similar patches gathered by searching from the noisy image. However, the main drawback of this method is that the standard deviation of noises within the image is assumed to be known in advance, which is impossible for many real applications. In this paper, we propose a non-local filtering method by using the low-rank tensor decomposition method. For tensor decomposition, we choose CP model as the underlying low-rank approximation. Since we assume the noise variance is unknown and need to be learned from data itself, we employ the Bayesian CP factorization that can learn CP-rank as well as noise variance solely from the observed noisy tensor data, The experimental results on image and MRI denoising demonstrate the superiorities of our method in terms of flexibility and performance, as compared to other tensor-based denoising methods.

***K*eywords:** *Tensor factorization, CP factorization, Image denoising.*

1. Introduction

Image denoising is an important task in image processing field, many techniques try to solve this problem. Recently, non-local filtering techniques have attracted a lot of interest and demonstrated the superiority in terms of performance [1–5]. The key technique in image denoising is to infer the optimal bases from a group of similar patches. More specifically, for any reference patch, the bases can be learned from a set of patches selected within a specific distant range, which are similar to that patch. These image denoising methods have been also extended to video denoising, which enables the patches from adjacent video frames to be considered in gathering the similar patches. However, the existing methods are all based on 2D patches, thus is not effective for multiway data that is naturally represented as a tensor.

Multidimensional data is natural represent by tensor, compared to matrix tensor factorization can capture the multilinear latent factors effectively and take the structure information into account explicitly. The theory and algorithms of tensor factorization have been widely studied during the past decade and were successfully applied to many real-world applications, such as image completion [6–14], signal processing [15–21], brain machine interface (BMI) [22–24], image classification [25, 26], face recognition [27], machine learning [28], etc.

The higher order singular value decomposition (HOSVD) is an extension of the matrix SVD technique to the multiway tensor [29]. Recently, the HOSVD has been successfully applied to image and video denoising [4] [30] [31] as a multilinear transform basis. However, the limitations of HOSVD-based denoising are that the noise standard deviation must be known in advance, which results in difficulties in practical applications.

To solve this problem, we leverage the Bayesian approach to learn the noise variance from original data without using priori knowledge. In contrast to HOSVD model which is a Tucker tensor decomposition model, we apply the CP decomposition which has a more compact representation ability than Tucker decomposition. Since the computation of CP Rank of a tensor is proven to be a NP hard problem, we specify the sparsity priors over the latent components, which can thus obtain the minimum number of components via Bayesian inference. Similarly, we also place a non-informative hyper-prior

Qibin Zhao and Jianting Cao are the corresponding authors.

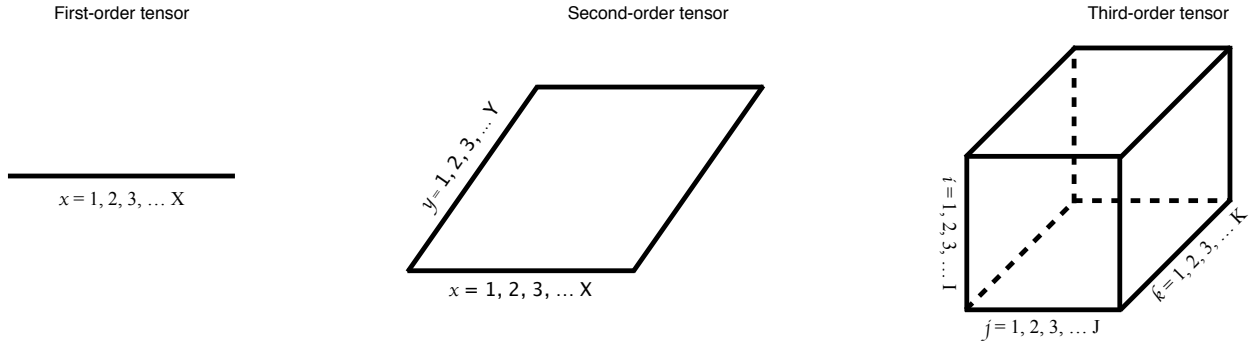

Fig. 1: First, second, thrid-order tensor

over the noise precision parameter which leads to the possibility of inferring it from data.

The rest of the paper is organized as follows. We first briefly present some basic tensor operations and algebra in Section II. Section III describes Bayesian CP factorization (BCPF) based denoising method. The experimental results are shown in Section IV. We summarize our method in Section V.

2. Tensor decompositions

Tensor is a multidimensional array which is a generalization of vectors and matrices to higher dimensions. First-order tensor is a vector, second-order tensor is a matrix, and third and higher order tensor are called a tensor. The tensors of first, second, third-order are shown in Fig. 1.

Tensor decompositions originated from Hitchcock [32] [33]. Under the work of Tucker [34] [35] [36], Carroll and Chang [37], Harshman [38], Appellof and Davidson [39], the tensor theory and tensor decompositions (factorizations) algorithms have been successfully applied to various fields, the examples include signal processing, computer vision and etc.

2.1 Notation

Tensor is a multidimensional array, the order of a tensor is the number of dimensions [40]. Tensor of order one (vector) is denoted by boldface lowercase letters, e.g., $\mathbf{a}$, the i-th element of a one-order tensor is denoted by a_i. Tensor of order two (matrix) is denoted by boldface capital letters, e.g., $\mathbf{A}$, the (i, j) element of a two-order tensor is denoted by a_{ij}. Tensor of order three or higher (higher-order tensor) is denoted by boldface Euler script letters, e.g., $\boldsymbol{\mathcal{X}}$, the (i, j, k) element of a three-order tensor is denoted by x_{ijk}. Indices typically range from 1 to their capital version, e.g., $i = 1, ..., I$.

2.2 Tensor Algebra

The $Frobeniusnorm$ of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times .... \times I_N}$, is the square root of the sum of the square of all elements (1)

$$\|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1 i_2 \cdots i_N}^2}. \quad (1)$$

The $inner\ product$ of two same sized tensors $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined by

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} ... \sum_{i_N=1}^{I_N} x_{i_1 i_2 \cdots i_N} y_{i_1 i_2 ... i_N}. \quad (2)$$

It follows immediately that $\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle = \|\boldsymbol{\mathcal{X}}\|_F^2$.

The $Hadamard\ product$ is an elementwise product between two tensors that must be same sizes. Given $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{I \times J}$, the Hadamard product is denoted by $\mathbf{A} \circledast \mathbf{B} \in \mathbb{R}^{I \times J}$, which is computed by

$$\mathbf{A} \circledast \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1J}b_{1J} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2J}b_{2J} \\ a_{31}b_{31} & a_{32}b_{32} & \cdots & a_{3J}b_{13} \\ \cdots & \cdots & & \cdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \cdots & a_{IJ}b_{IJ} \end{bmatrix}. \quad (3)$$

The Hadamard product of $N \geq 3$ items is defined as

$$\circledast_{n=1}^{N} \mathbf{A}^{(n)} = \mathbf{A}^{(1)} \circledast \mathbf{A}^{(2)} \circledast \cdots \circledast \mathbf{A}^{(N)}. \quad (4)$$

The $Kronecker\ product$ of matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$ becomes a matrix of size $IK \times JL$, denoted by $\mathbf{A} \otimes \mathbf{B}$ and computed by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1J}B \\ a_{21}B & a_{22}B & \cdots & a_{2J}B \\ a_{31}B & a_{32}B & \cdots & a_{3J}B \\ \cdots & \cdots & & \cdots \\ a_{I1}B & a_{I2}B & \cdots & a_{IJ}B \end{bmatrix}. \tag{5}$$

The $Khatri-Rao\ product$ of matrices $\mathbf{A} \in \mathbb{R}^{I\times K}$ and $\mathbf{B} \in \mathbb{R}^{J\times K}$ is a matrix of size $IJ \times K$, denoted by $\mathbf{A} \odot \mathbf{B}$. In particular, the Khatri-Rao product of $N \geq 3$ matrices in a reverse order is defined by

$$\bigodot_{n=1}^{N} \mathbf{A}^{(n)} = \mathbf{A}^{(N)} \odot \mathbf{A}^{(N-1)} \odot \cdots \odot \mathbf{A}^{(1)}. \tag{6}$$

The Khatri-Rao product of a group of matrices, except the nth matrix is denoted by $\mathbf{A}^{(\backslash n)}$ and computed by

$$\bigodot_{k=1,k\neq n}^{N} \mathbf{A}^{(k)} = \mathbf{A}^{(N)} \odot \cdots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \cdots \odot \mathbf{A}^{(1)}. \tag{7}$$

2.3 CP decompositions

CANDECOMP/PARAFAC (CP) decomposition method is proposed by Carroll and Chang [37] and PARAFAC (parallel factors) proposed by Harshman [38]. Usually, we refer to the CANDECOMP/PARAFAC decomposition as CP [41]. CP decomposition is to represent a tensor as a sum of rank-one tensors. For instance, given a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I\times J\times K}$, we wish to represent it by

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]. \tag{8}$$

The element-wise form of (8) is written as

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}, \quad \forall\, i = 1,...,I, \forall\, j = 1,...,J, \forall\, k = 1,...,K. \tag{9}$$

where $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$ and $\mathbf{c}_r \in \mathbb{R}^K$, $\forall r$ = 1,...,R. The rank of a tensor $\boldsymbol{\mathcal{X}}$, denoted R = rank($\boldsymbol{\mathcal{X}}$), is define as the smallest number of rank-one tensors that can exactly represent $\boldsymbol{\mathcal{X}}$. The scheme of CP decompositions is illustrated in Fig. 2.

2.4 Tucker decompositions

The Tucker decomposition was proposed in 1963 [42], and refined in subsequent articles by Levin [39] and Tucker [35,36]. Tucker decomposition can be considered as an extension of PCA (Principal Components Analysis) to a high order tensor, which decomposes a tensor into a core tensor multiplied (or transformed) by several matrices along each mode. For instance, given a three-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I\times J\times K}$, Tucker decomposition is written as

$$\begin{aligned} \boldsymbol{\mathcal{X}} &= \boldsymbol{\mathcal{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \\ &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr} \circ \mathbf{a_p} \circ \mathbf{b_q} \circ \mathbf{c_r} \\ &= [\![\boldsymbol{\mathcal{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]. \end{aligned} \tag{10}$$

The element in tensor can thus be computed and represented by

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr} a_{ip} b_{jq} c_{kr}, \quad \forall\, i = 1,...,I, \forall\, j = 1,...,J, \forall\, k = 1,...,K. \tag{11}$$

Here, $\mathbf{A} \in \mathbb{R}^{I\times P}$, $\mathbf{B} \in \mathbb{R}^{J\times Q}$ and $\mathbf{C} \in \mathbb{R}^{K\times R}$ are the factor matrices (which are usually orthogonal) and can be considered as the principal components in each mode. Tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{P\times Q\times\ R}$ is called the $core\ tensor$ and its entries show the level of interaction between the different components. The last equality in (10) using the shorthand $[\![\boldsymbol{\mathcal{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ was introduced in [40]. The scheme of Tucker decompositions is illustrated in Fig. 3.

3. Bayesian CP factorization based tensor denoising

3.1 Non-local Tensor Denoising

We consider a given tensor $\boldsymbol{\mathcal{T}}$ corrupted by Gaussion noise $\boldsymbol{\mathcal{N}}(0, \sigma)$, our objective is to recover the underlying clean tensor $\boldsymbol{\mathcal{V}}$. The main procedure includes three steps that are

- At each tensor element and for a fixed sub-tensor size, a group of similar sub-tensors is selected and constructed to be a higher order tensor.
- The proposed Bayesian CP factorization is employed for each stack to obtain an estimate of a denoised stack.
- The sub-tensors are reassembled in original location to obtain a denoised tensor.

Given a reference sub-tensor $\boldsymbol{\mathcal{P}}$ from the noisy tensor $\boldsymbol{\mathcal{T}}$, we choose other sub-tensors in the tensor $\boldsymbol{\mathcal{T}}$ that are similar to $\boldsymbol{\mathcal{P}}$.

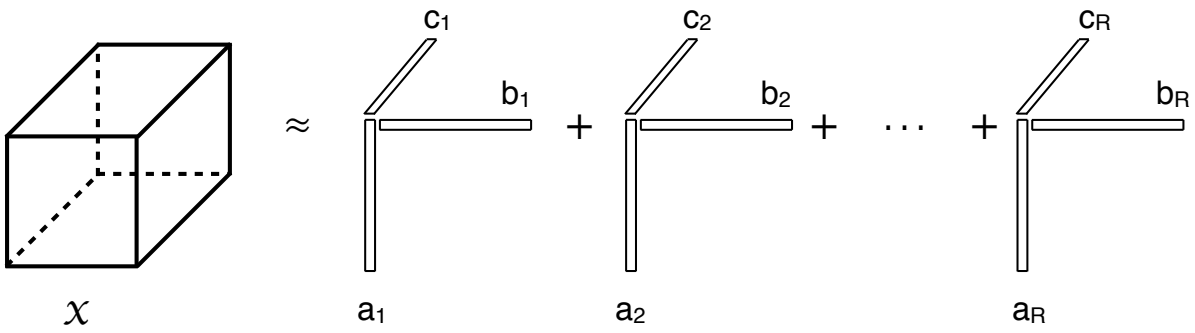

Fig. 2: CP decomposition of a third-order tensor

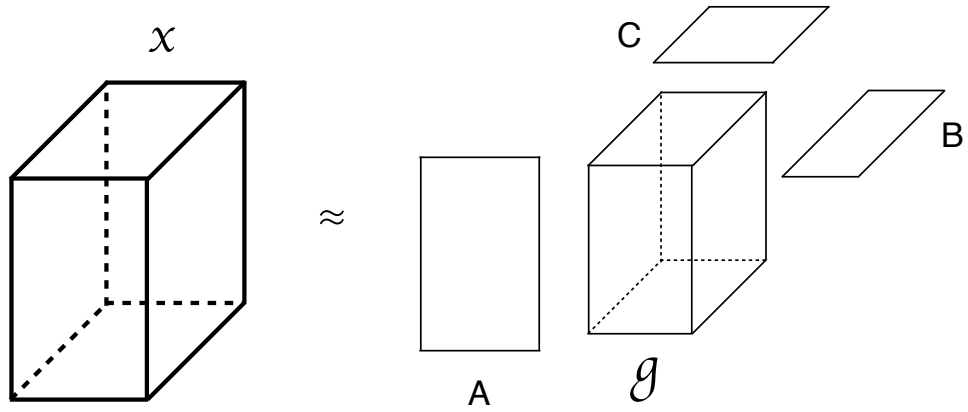

Fig. 3: Tucker decomposition of a third-order tensor

The similarity can be simply measured by Euclidean distance. There are two choices for selecting similar sub-tensors. One is to use a distance threshold $\tau_d = 3\sigma^2 s$, where σ^2 denotes the noise variance and s denotes the size of each sub-tensor. The other one is to use a fixed number of sub-tensors ordered by the distance with the reference sub-tensor. Assume that there are K such sub-tensors (including $\boldsymbol{\mathcal{P}}$) which are labeled as $\{\boldsymbol{\mathcal{P}}_i\}$ where $1 \leq i \leq K$. These sub-tensors were assumed to be noise corrupted versions of $\boldsymbol{\mathcal{P}}$. If a set of sub-tensors are similar to each another, denoising can be performed by leveraging this fact and filter them jointly. Based on this, we group together similar sub-tensors and organize them as a higher order tensor $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{\mathcal{P}}_i | i = 1, \ldots, K\}$.

Now we consider how the filtering of $\boldsymbol{\mathcal{Y}}$ can be performed. The concept of jointly filtering multiple patches has been implemented in the BM3D algorithm but with fixed bases. However, we extend this concept to learn the spatially adaptive bases. By assuming that the group of similar sub-tensors were generated from a same clean sub-tensor, we can easily identify the low-rank properties of $\boldsymbol{\mathcal{Y}}$. Therefore, the low-rank tensor factorization can be employed to learn the bases independently for each group of similar sub-tensors. One straightforward way is to apply HOSVD to solve this problem. However, the truncated HOSVD method requires that the parameter for thresholding the transform coefficients must be known in advance, which results in difficulties in the practical applications. Hence, in this paper, we propose a Bayesian tensor factorization based on CP model and the low-rank assumption. In addition, we assume that noise variance is unknown and must be learned from noisy data automatically. After leaning the latent multiliear factor matrices from Bayesian CP factorization, we can reconstruct the group of similar sub-tensors as the denoised results for $\boldsymbol{\mathcal{Y}}$. Then, all the sub-tensors in $\boldsymbol{\mathcal{Y}}$ are jointly denoised. This procedure will be repeated for each reference sub-tensor $\boldsymbol{\mathcal{P}}_n$ in a sliding window fashion and the denoised sub-tensors are averaged to obtain the denoised result for tensor $\boldsymbol{\mathcal{T}}$.

3.2 Bayesian Low-Rank Tensor Factorization

We introduce the Bayesian CP factorization for jointly filtering of multiple sub-tensors in $\boldsymbol{\mathcal{Y}}$. Without loss of generality, let $\boldsymbol{\mathcal{Y}}$ be an Nth-order tensor of size $I_1 \times I_2 \times \cdots \times I_N$. We assume $\boldsymbol{\mathcal{Y}}$ is a noisy observation of true tensor $\boldsymbol{\mathcal{X}}$, that is, $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\varepsilon}$, where the noise term is assumed to be an i.i.d. Gaussian distribution, i.e., $\boldsymbol{\varepsilon} \sim \prod_{i_1,\ldots,i_N} \mathcal{N}(0, \tau^{-1})$, and the latent tensor $\boldsymbol{\mathcal{X}}$ is generated by a CP model, defined by

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(N)} = [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!]. \tag{12}$$

where $\circ$ denotes the outer product of vectors and $[\![\cdots]\!]$ is a shorthand notation, also termed as the Kruskal operator. CP factorization can be interpreted as a sum of R rank-one tensors, and the smallest integer R is defined as *CP rank*. $\{\mathbf{A}^{(n)}\}_{n=1}^{N}$ denote a group of factor matrices. For clarity, we denote mode-n factor matrix $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$ by row-wise or column-wise vectors (13)

$$\begin{aligned} \mathbf{A}^{(n)} &= \left[\mathbf{a}_1^{(n)}, \ldots, \mathbf{a}_{i_n}^{(n)}, \ldots, \mathbf{a}_{I_n}^{(n)}\right]^T \\ &= \left[\mathbf{a}_{\cdot 1}^{(n)}, \ldots, \mathbf{a}_{\cdot r}^{(n)}, \ldots, \mathbf{a}_{\cdot R}^{(n)}\right]. \end{aligned} \tag{13}$$

The likelihood of CP model can be factorized over tensor elements, which is given by (14)

$$\begin{aligned} \mathcal{Y}_{i_1 i_2 \ldots i_N} \;\Big|\; \{\mathbf{A}^{(n)}\}, \tau \sim \\ \mathcal{N}\left(\left\langle \mathbf{a}_{i_1}^{(1)}, \mathbf{a}_{i_2}^{(2)}, \cdots, \mathbf{a}_{i_N}^{(N)} \right\rangle, \; \tau^{-1}\right). \end{aligned} \tag{14}$$

where τ is the noise precision, and $\left\langle \mathbf{a}_{i_1}^{(1)}, \mathbf{a}_{i_2}^{(2)}, \cdots, \mathbf{a}_{i_N}^{(N)} \right\rangle$ is a generalized inner-product among N vectors. The observation model in (14) shows that $\mathcal{Y}_{i_1 \cdots i_N}$ is represented by a group of R-dimensional latent vectors $\left\{\mathbf{a}_{i_n}^{(n)} | n = 1, \ldots, N\right\}$, which results in that the multilinear interactions can be considered. As compared to matrix factorization, tensor factorization allows us to model the multilinear structure by the inner product of $N \geq 3$ vectors.

The number of latent components, i.e., $Rank_{CP}(\boldsymbol{\mathcal{X}}) = R$, is a tuning parameter whose selection is very difficult in practical applications. To avoid manually adjusting this parameter, we aim to develop an automatic model selection, which can find the rank of the latent tensor $\boldsymbol{\mathcal{X}}$ solely from the observed data $\boldsymbol{\mathcal{Y}}$. Taking into account the low-rank property, the number of latent components is desired to be minimal. Therefore, we employ specific sparsity-inducing priors over latent components and control the variance of each component by individual hyperparameters. Through Bayesian inference, the variance of unnecessary components can be reduced to zero. This strategy is related to automatic relevance determination (ARD) or sparse Bayesian learning. The difference lie in that our method employs a group of sparsity-inducing priors over each mode-n factors and the hyperparameters are common among these priors. Hence, the low-rank constraint can be imposed jointly to the factor matrices.

For each mode-n factor matrix, we specify a prior distribution that is governed by hyperparameters $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_R]$, among which λ_r corresponds to rth component. The prior distribution over latent factors is thus given by

$$\begin{aligned} \mathbf{a}_{i_n}^{(n)} \;|\; \boldsymbol{\lambda} \sim \mathcal{N}\left(\mathbf{a}_{i_n}^{(n)} \;|\; \mathbf{0}, \; \boldsymbol{\Lambda}^{-1}\right), \\ \forall n \in [1, N], \; \forall i_n \in [1, I_n]. \end{aligned} \tag{15}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ is a diagonal matrix that is also called the precision matrix. This precision matrix is jointly shared by all latent factor matrices. Since the precision parameters $\boldsymbol{\lambda}$ is unknown, and need to be learned automatically, we employ the hyperprior over $\boldsymbol{\lambda}$, given by

$$\lambda_r \sim \mathrm{Ga}(\lambda_r | c_0^r, d_0^r), \quad \forall r \in [1, R]. \tag{16}$$

where the Gamma distribution is given by $\mathrm{Ga}(x|a,b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}$. The number of components (i.e., R) is usually initialized to be a maximum possible value. By employing a Bayesian inference framework, the effective number of components can be inferred automatically solely from observed data. Because the hyperparameters of sparsity priors over all factor matrices are common, the same number of components can be obtained for each factor matrix, resulting in that the minimum number of rank-one terms can be learned. Hence, the CP rank of the tensor can be effectively inferred while performing low-rank tensor factorization.

Since the noise variance is assumed to be unknown, we can also specify a hyperprior over the noise parameter τ, which is given by (17)

$$\tau \sim \mathrm{Ga}(\tau | a_0, b_0). \tag{17}$$

To simplify the notations, we collect and denote all unknown variables by $\Theta = \{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}, \boldsymbol{\lambda}, \tau\}$. Finally, the joint distribution of Bayesian low-rank tensor factorization model can be written as (18)

$$\begin{aligned} p(\boldsymbol{\mathcal{Y}}, \Theta) = &\rho\left(\boldsymbol{\mathcal{Y}} \;\Big|\; \{\mathbf{A}^{(n)}\}_{n=1}^{N}, \tau\right) \\ &\prod_{n=1}^{N} \rho\left(\mathbf{A}^{(n)} \;\Big|\; \boldsymbol{\lambda} p(\boldsymbol{\lambda}) p(\tau)\right). \end{aligned} \tag{18}$$

Generally, maximum a posteriori (MAP) estimation of Θ can be obtained by optimizing. In contrast to the MAP estimation, we aim to develop a Bayesian inference method to infer the full posterior distribution of unknown variables in Θ, which is computed by (19)

$$p(\Theta | \boldsymbol{\mathcal{Y}}) = \frac{p(\Theta, \boldsymbol{\mathcal{Y}})}{\int p(\Theta, \boldsymbol{\mathcal{Y}}) \, d\Theta}. \tag{19}$$

3.3 Bayesian Model Inference

Since the exact Bayesian inference in is obviously analytically intractable, we must resort to the approximate inference framework. In this section, we employ the variational Bayesian (VB) inference strategy to perform model inference for tensor factorization model.

We assume that $q(\Theta)$ is an approximation of the true posterior distribution $p(\Theta|\boldsymbol{\mathcal{Y}})$, which is optimized by KL divergence between them, which can be shown to be

$$\begin{aligned}\mathrm{KL}\Big(q(\Theta)\,\Big|\Big|\,p(\Theta\mid\boldsymbol{\mathcal{Y}})\Big) &= \int q(\Theta)\ln\left\{\frac{q(\Theta)}{p(\Theta|\boldsymbol{\mathcal{Y}})}\right\}d\Theta \\ &= \ln p(\boldsymbol{\mathcal{Y}}) - \int q(\Theta)\ln\left\{\frac{p(\boldsymbol{\mathcal{Y}},\Theta)}{q(\Theta)}\right\}d\Theta.\end{aligned} \tag{20}$$

where $\ln p(\boldsymbol{\mathcal{Y}})$ denotes the marginal likelihood, and $\boldsymbol{\mathcal{L}}(q) = \int q(\Theta)\ln\left\{\frac{p(\boldsymbol{\mathcal{Y}},\Theta)}{q(\Theta)}\right\}d\Theta$ can be defined as its *lower bound*. Therefore, instead of minimizing the KL divergence directly, we can maximize the lower bound alternatively due to the fact that the model evidence is a constant and not related to any unknown variables.

By employing the mean-field approximation, we assume that the variational distribution can be factorized as (21)

$$q(\Theta) = q_\lambda(\boldsymbol{\lambda})q_\tau(\tau)\prod_{n=1}^{N} q_n\left(\mathbf{A}^{(n)}\right). \tag{21}$$

Therefore, it can been shown that the posterior distribution of factor matrices is also a Gaussian distribution and the distributions corresponding to each row are independent, which is written as

$$q_n(\mathbf{A}^{(n)}) = \prod_{i_n=1}^{I_n}\boldsymbol{\mathcal{N}}\left(\mathbf{a}_{i_n}^{(n)}\;\middle|\;\tilde{\mathbf{a}}_{i_n}^{(n)}\mathbf{V}^{(n)},\ \forall n\in[1,N]\right). \tag{22}$$

where the variational parameters are computed by

$$\begin{aligned}\tilde{\mathbf{A}}^{(n)} &= \mathbb{E}_q[\tau]\boldsymbol{\mathcal{Y}}_{(n)}\mathbb{E}_q\left[\mathbf{A}^{(\backslash n)}\right]\mathbf{V}^{(n)}, \\ \mathbf{V}^{(n)} &= \left(\mathbb{E}_q[\tau]\mathbb{E}_q\left[\mathbf{A}^{(\backslash n)T}\mathbf{A}^{(\backslash n)}\right] + \mathbb{E}_q[\boldsymbol{\Lambda}]\right)^{-1},\end{aligned} \tag{23}$$

where $\boldsymbol{\mathcal{Y}}_{(n)}$ denotes the mode-n matricization of $\boldsymbol{\mathcal{Y}}$ and

$$\mathbf{A}^{(\backslash n)} = \bigodot_{k\neq n}\mathbf{A}^{(k)}, \tag{24}$$

where the size of $\bigodot_{k\neq n}\mathbf{A}^{(k)}$ is $\prod_{k\neq n}I_k \times R$. Thus, $\mathbb{E}_q[\mathbf{A}^{(\backslash n)T}\mathbf{A}^{(\backslash n)}]$ denotes the expectation of covariance matrix, while the covariance matrix corresponds to the Khatri-Rao product of all factor matrices except the nth-mode.

Therefore, the parameters of posterior distribution over factor matrices can be approximated by, which can be also used to compute the posterior moments, such as $\forall n, \forall i_n$, $\mathbb{E}_q\left[\mathbf{A}^{(n)}\right]$, and $\mathbb{E}_q\left[\mathbf{A}^{(n)}\mathbf{A}^{(n)T}\right], \mathbb{E}_q\left[\mathbf{A}^{(n)T}\mathbf{A}^{(n)}\right]$.

For the inference of $\boldsymbol{\lambda}$, we can derive that the posterior distribution over $\lambda_r, \forall r\in[1,R]$ can be obtained by

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \prod_{r=1}^{R}\mathrm{Ga}(\lambda_r|c_M^r, d_M^r), \tag{25}$$

where the variational parameters are computed by

$$\begin{aligned}c_M^r &= c_0^r + \frac{1}{2}\sum_{n=1}^{N} I_n, \\ d_M^r &= d_0^r + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_q\left[\mathbf{a}_{\cdot r}^{(n)T}\mathbf{a}_{\cdot r}^{(n)}\right].\end{aligned} \tag{26}$$

The expectation term in above equations denotes the norm of the rth component from mode-n matrix, which can be easily computed by

$$\mathbb{E}_q\left[\mathbf{a}_{\cdot r}^{(n)T}\mathbf{a}_{\cdot r}^{(n)}\right] = \tilde{\mathbf{a}}_{\cdot r}^{(n)T}\tilde{\mathbf{a}}_{\cdot r}^{(n)} + I_n\left(\mathbf{V}^{(n)}\right)_{rr}. \tag{27}$$

For inference of hyperparameter τ, it can be derived that the variational posterior is a Gamma distribution, given by

$$q_\tau(\tau) = \mathrm{Ga}(\tau|a_M, b_M), \tag{28}$$

the variational parameters of the posterior distribution are computed by

$$\begin{aligned}a_M &= a_0 + \frac{1}{2}\prod_n I_n, \\ b_M &= b_0 + \frac{1}{2}\mathbb{E}_q\left[\Big|\Big|\,\boldsymbol{\mathcal{Y}} - [\![\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}]\!]\,\Big|\Big|_F^2\right]\end{aligned} \tag{29}$$

3.4 Initialization of model parameters

In this probabilistic tensor decomposition model, it is important to initialize the hyperparameters. Specifically, $\mathbf{c}_0, \mathbf{d}_0$, a_0, b_0 are set to 10^{-6} yielding a noninformative prior. The mode-n factor matrices $\{\mathbf{A}^{(n)}\}_{n=1}^N$ can be either randomly drawn from $\mathcal{N}(\mathbf{0},\mathbf{I})$ or initialized by SVD method, i.e., $\mathbf{A}^{(n)} = \mathbf{U}^{(n)}\boldsymbol{\Sigma}^{(n)^{\frac{1}{2}}}$, where $\mathbf{U}^{(n)}$ is the singular vectors and $\boldsymbol{\Sigma}^{(n)}$ is the singular values matrix.

4. Experiments Results

4.1 Image Denoising

We use color image (Lena, Peppers, Barbara) denoising to evaluate our method BCPF. For the noise model $\mathcal{N}(0, \sigma)$, we select $\sigma \in \{0.4, 0.8, 1.2\}$. In the experiment, we use images of size $256 \times 256 \times 3$ consisting of R, G, B channels. The method performance is evaluated by PSNR which is defined by $10 \log_{10}(Max_I^2/MSE)$ (Max_I is maximum possible pixel value of the image, and MSE denotes the mean squared errors).

The result are shown in Table 1 and the noisy and denoised images are shown in Fig. 4. The size of sub-tensors is selected to be $4 \times 4 \times 3$ and the maximum number of similar sub-tensors is set to 30. We observe that our method can obtain high quality of denoised images when noise level is low, especially, it can obtain a relatively good quality even when the noise level is extremely high.

Table 1: The denoising performances evaluated by PSNR for Lena, Peppers, Barbara images under three different noise levels.

Lena		Noise standard deviation		
	Methods	0.4	0.8	1.2
	BCPF	32.90	32.39	31.72
	HOSVD	30.70	30.54	30.46
Peppers		Noise standard deviation		
	Methods	0.4	0.8	1.2
	BCPF	31.37	30.89	30.77
	HOSVD	30.73	30.49	29.84
Barbara		Noise standard deviation		
	Methods	0.4	0.8	1.2
	BCPF	30.32	30.89	30.78
	HOSVD	30.55	30.54	30.36

4.2 MRI Denoising

Magnetic resonance imaging (MRI) is a medical imaging which is widely employed in the clinical diagnosis [43]. Because of the movements of the subject or electronic interference, MRI data always contain noise, the denoising of MRI data is thus important for the diagnosis quality. In this experiment, we use the public MRI data (http://brainweb.bic.mni.mcgill.ca/brainweb/), the size of MRI data is $181 \times 217 \times 165$, and we use the sub-tensor size in $4 \times 4 \times 4$ and the maximum number of similar sub-tensors is set to 30. The result are shown in Table 2 and the noisy and denoised images are shown in Fig. 5.

Table 2: The denoising performances evaluated by PSNR for MRI denoising under three different noise levels (0.05, 0.1, 0.15).

	Noise standard deviation		
Methods	0.05	0.10	0.15
BCPF	35.97	33.81	33.00
HOSVD	36.64	33.71	32.86

5. Conclusion

In this paper, we propose a Bayesian tensor factorization based denoising framework and apply it to image and MRI denoising tasks. In contrast to most existing denoising methods, we use sub-tensors instead of 2D patches. Moreover, the transform bases of a group of sub-tensors can be learned by the probabilistic CP factorization with a low-rank assumption. As compared to other methods, our method enables us to infer automatically the noise variance, which indicates that our method is more practical. Experimental results show that our method can outperform HOSVD based denoising method.

Acknowledgment

This work was supported by JSPS KAKENHI (Grant No. 17K00326 and 18K04178) and NSFC China (Grant No. 61773129).

References

[1] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[2] Z. Wang and D. Zhang, "Progressive switching median filter for the removal of impulse noise from highly corrupted images," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 46, no. 1, pp. 78–80, 1999.

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[4] A. Rajwade, A. Rangarajan, and A. Banerjee, "Using the higher order singular value decomposition for

Fig. 4: Visualization of Image data. From top to bottom rows, the original, noisy and denoised images are shown under the condition of $\sigma = 0.4$.

video denoising," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2011, pp. 344–354.

[5] K. Kundu, "Image denoising using patch based processing with fuzzy triangular membership function," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 3, p. 44, 2015.

[6] L. Gui, Q. Zhao, and J. Cao, "Brain image completion by bayesian tensor decomposition," in *Digital Signal Processing (DSP), 2017 22nd International Conference on*. IEEE, 2017, pp. 1–4.

[7] L. Yuan, Q. Zhao, and J. Cao, "High-order tensor completion for data recovery via sparse tensor-train optimization," *arXiv preprint arXiv:1711.02271*, 2017.

[8] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[9] L. Yuan, Q. Zhao, and J. Cao, "Completion of high order tensor data with missing entries via tensor-train decomposition," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 222–229.

[10] L. Yuan, J. Cao, Q. Wu, and Q. Zhao, "Higher-dimension tensor completion via low-rank tensor ring decomposition," *arXiv preprint arXiv:1807.01589*, 2018.

[11] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.

[12] M. Filipović and A. Jukić, "Tucker factorization with missing data with application to low-n-rank tensor completion," *Multidimensional systems and signal processing*, vol. 26, no. 3, pp. 677–692, 2015.

[13] L. Yuan, C. Li, D. Mandic, J. Cao, and Q. Zhao, "Rank minimization on tensor ring: A new paradigm in scalable tensor decomposition and completion," *arXiv preprint arXiv:1805.08468*, 2018.

[14] L. Yuan, Q. Zhao, L. Gui, and J. Cao, "High-dimension tensor completion via gradient-based optimization under tensor-train format," *arXiv preprint arXiv:1804.01983*, 2018.

[15] L. De Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1096–1105, 2008.

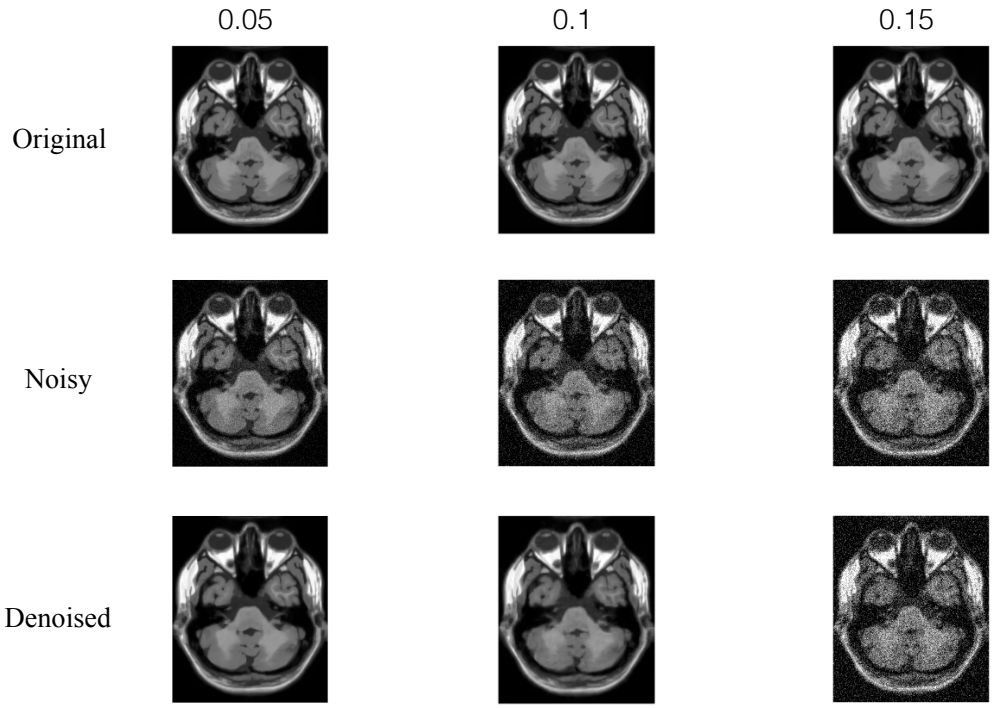

Fig. 5: Visualization of MRI denoising results. From top to bottom rows, the original, noisy and denoised slice of MRI data are shown under the condition of $\sigma = \{0.05, 0.1, 0.15\}$.

[16] L. Gui, Q. Zhao, and J. Cao, "Tensor denoising using bayesian cp factorization," in *Information Science and Technology (ICIST), 2016 Sixth International Conference on*. IEEE, 2016, pp. 49–54.

[17] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.

[18] L. Gui, G. Cui, Q. Zhao, D. Wang, A. Cichocki, and J. Cao, "Video denoising using low rank tensor decomposition," in *Ninth International Conference on Machine Vision (ICMV 2016)*, vol. 10341. International Society for Optics and Photonics, 2017, p. 103410V.

[19] D. Muti and S. Bourennane, "Multidimensional filtering based on a tensor approach," *Signal Processing*, vol. 85, no. 12, pp. 2338–2353, 2005.

[20] L. De Lathauwer and B. De Moor, "From matrix to tensor: Multilinear algebra and signal processing," in *Institute of Mathematics and Its Applications Conference Series*, vol. 67. Citeseer, 1998, pp. 1–16.

[21] Y.-l. Zhang and G. Li, "An efficient algorithm to estimate mixture matrix in blind source separation using tensor decomposition," *International Journal of Computer Science Issues (IJCSI)*, vol. 11, no. 1, p. 19, 2014.

[22] Y. Liu, M. Li, H. Zhang, H. Wang, J. Li, J. Jia, Y. Wu, and L. Zhang, "A tensor-based scheme for stroke patients? motor imagery EEG analysis in BCI-FES rehabilitation training," *Journal of neuroscience methods*, vol. 222, pp. 238–249, 2014.

[23] J. Mocks, "Topographic components model for event-related potentials and some biophysical considerations," *IEEE transactions on biomedical engineering*, vol. 35, no. 6, pp. 482–484, 1988.

[24] Y. Zhang, Q. Zhao, G. Zhou, J. Jin, X. Wang, and A. Cichocki, "Removal of eeg artifacts for bci applications using fully bayesian tensor completion," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 819–823.

[25] A. Shashua and A. Levin, "Linear image coding for regression and classification using the tensor-rank principle," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[26] H. Hota, S. Shukla, and G. K. Kiran, "Review of intelligent techniques applied for classification and prepro-

cessingof medical imagedata," *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 1, p. 267, 2013.

[27] X. Geng, K. SmithMiles, Z. Zhou, and L. Wang, "Face image modeling by multilinear subspace analysis with missing values," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 3, pp. 881–892, 2011.

[28] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S. Amari, "Bayesian robust tensor factorization for incomplete multiway data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 736–748, 2016.

[29] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[30] A. Rajwade, A. Rangarajan, and A. Banerjee, "Image denoising using the higher order singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 849–862, 2013.

[31] X. Zhang, Z. Xu, N. Jia, W. Yang, Q. Feng, W. Chen, and Y. Feng, "Denoising of 3d magnetic resonance images by using higher-order singular value decomposition," *Medical image analysis*, vol. 19, no. 1, pp. 75–86, 2015.

[32] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[33] ——, "Multiple invariants and generalized rank of a p-way matrix or tensor," *Journal of Mathematics and Physics*, vol. 7, no. 1-4, pp. 39–79, 1928.

[34] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *Problems in measuring change*, vol. 15, pp. 122–137, 1963.

[35] ——, "The extension of factor analysis to three-dimensional matrices," *Contributions to mathematical psychology*, vol. 110119, 1964.

[36] ——, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[37] J. D. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[38] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis," *Foundations of the PARAFAC procedure*, 1970.

[39] C. J. Appellof and E. R. Davidson, "Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents," *Analytical Chemistry*, vol. 53, no. 13, pp. 2053–2056, 1981.

[40] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[41] H. A. Kiers, "Towards a standardized notation and terminology in multiway analysis," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 14, no. 3, pp. 105–122, 2000.

[42] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," in *Problems in measuring change.*, C. W. Harris, Ed. Madison WI: University of Wisconsin Press, 1963, pp. 122–137.

[43] R. Srivaramangai, P. Hiremath, and A. S. Patil, "Preprocessing mri images of colorectal cancer," *International Journal of Computer Science Issues (IJCSI)*, vol. 14, no. 1, p. 48, 2017.