Neurocomputing 173 (2016) 298–305

Contents lists available at ScienceDirect

Neurocomputing

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

Hierarchical Representation Learning based spatio-temporal data redundancy reduction

Min Wang [a,*], Shuyuan Yang [b], Bin Wu [c]

[a] National Key Lab of Radar Signal Processing, China
[b] Key Lab of Intelligent Perception and Image Understanding of Ministry of Education, China
[c] Information Technology Department, School of Electrical Engineering, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 1 July 2014
Received in revised form
2 February 2015
Accepted 7 February 2015
Available online 16 September 2015

Keywords:
Visual big data
Spatio-temporal data
Hierarchical Representation Learning
Incremental learning
Parallelized realization

ABSTRACT

Spatio-temporal data is characteristic of large volume and high redundancy, which will require large amounts of space for storage and computing power for analysis. In this paper, inspired by the sparse, multi-scale and hierarchical characteristics of visual perception in the Human Vision System (HVS), we advance a new Hierarchical Representation Learning (HRL) based spatio-temporal data redundancy reduction approach. In our method, the most informative and representative data can be identified in a cascade manner via a hierarchical and sparse self-representation model. The parallelized realization of the proposed scheme is discussed. The proposed method is investigated on some large volume spatio-temporal data, and the experimental results prove its efficiency and superiority to some state-of-the-art results.

1. Introduction

The past decade has witnessed an explosive growth of images and videos data in Internet, computer vision, remote sensing, machine learning and other fields [1–3]. For example, Facebook, as an online social networking service with ten years history, has more than 140 billion images. Moreover, the uploaded images increase 300 million every day. Video often brought larger volume of data for the continuous collection of spatio-temporal streams. For example, YouTube is the biggest video-sharing website in the world that has 120 billion videos. Because the registered users can upload an unlimited number of videos, YouTube has an increase of 72 h of videos every minute. This data deluge has produced a new concept of “Big Data”, which usually include data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time [4–8].

Among all the Big Data come from different fields, visual related data especially spatio-temporal visual data, accounts for a large and important part. These data are so large and complex that it becomes difficult to process using on-hand database management tools (for storage) or traditional data processing (transmission, explanation and so on) applications. Consequently inferring laws from these spatio-temporal visual data sets to reveal relationships, dependencies and perform predictions of outcomes and behaviors is a very difficult task [9–15]. Spatio-temporal visual data are characteristic of large-scale, multi-source, high-dimensionality and low-density value. Consequently, getting the concise, valuable information from a sea of data is necessary. In order to save the computation and storage burden and overcome the curse of dimensionality in processing high-dimensional visual data, lots of dimensionality reduction theories and algorithms have been developed, such as Linear Discriminant Analysis (LDA) and its variants [16–18,35,36], Negative Matrix Factorization (NMF) [19] and so on [20,21,37,38]. They aim at finding and exploiting low-dimensional structures in high-dimensional data, and play an increasingly important role in the analytics of visual big data. However, most of dimensionality reduction approaches focus on exploring the redundancy of spatio-temporal visual data in the spatio domain, which is the foundation of many video compression algorithms [22]. However, in many applications of spatio-temporal videos such as security monitoring, change detection and so on, there are high redundancies in videos in the temporal domain. Identifying the most salient and task-relevant “needles” in the spatio-temporal visual data sea and reducing the redundancy can well reveal relationships and dependencies of data. Consequently, “Big Data” will be reduced to “Small Data” and do not lose the original information [23].

Making a close inspection of the recent process made in data mining we will find that more and more advancements can be attributed to introducing bio-inspired algorithms, which is a flourishing area of computing. It is well known that the Human Visual System (HVS) is able to locate objects of interest in complex scenes [24]. In psychology and computational neuroscience, the visual salience is proved to be a key component of various visual mechanisms. Therefore the saliency attention can be used as a prior for solving

[*] Corresponding author.
E-mail address: syyang2009@gmail.com (S. Yang).

http://dx.doi.org/10.1016/j.neucom.2015.02.099
0925-2312/© 2015 Elsevier B.V. All rights reserved.

more difficult video signal processing tasks [25–28]. In this paper, inspired by the sparse, multi-scale and hierarchical characteristics of visual perception in the HVS, we advance a new Hierarchical Representation Learning (HRL) based spatio-temporal data redundancy reduction approach. Representative data are gradually selected via hierarchical self-representation and sparse coding. A Multiple Measurement Vector (MMV) algorithm is advanced and the parallelized realization is also discussed. The proposed method is investigated on some videos, and the experimental results prove its efficiency and superiority to some state-of-the-art results.

The contribution of this paper is threefold. First, by simulating the hierarchical, sparse and multi-scale rules of visual organization, we cast a sparse, representative, and meaningful assumption on the large-volume videos for modeling big data. Second, a Hierarchical Representation Learning (HRL) model is proposed for saliency description and a sparse optimization algorithm is employed to detect this saliency. Third, the parallelized realization is presented. Finally we investigate the proposed method on some datasets to prove its efficiency.

2. Hierarchical Representation Learning based redundancy reduction

As mentioned in Section 1, visual coding has the sparse, hierarchical and multi-scale characteristics [29,30]. Inspired by them, we establish a sparse and Hierarchical Representation Learning model for spatio-temporal data redundancy reduction.

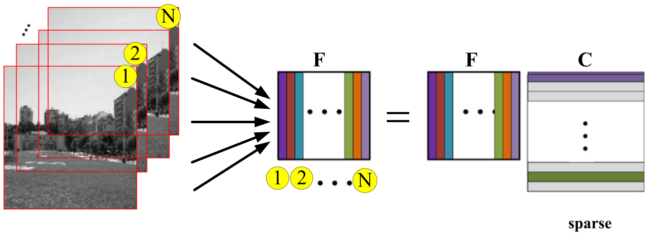

Fig. 1. Self-representation learning for data reduction.

2.1. Hierarchical Representation Learning (HRL) model

In our method, we denote the spatio-temporal stream as $\mathbf{F}=\left[\mathbf{I}^1,\mathbf{I}^2\ldots,\mathbf{I}^N\right]\in\Re^{m\times N}$ $(m=p\times q)$, where N is the number of frames in the stream and $\mathbf{I}^i$ represent the ith vectorized frame whose size is $p\times q$. The salient data extraction is equal to selecting a few "important" frames from a set of frames with many redundant and useless information that are unrelated with visual tasks. In order to locate meaningful and representative salient frames in videos, we consider the following optimization problem to develop a self-representative model for **F**

$$\begin{cases} \min\limits_{\mathbf{C}} \|\mathbf{C}\|_{1,q} \\ s.t.\quad \mathbf{F}=\mathbf{FC};\quad 1^T\mathbf{C}=1^T \quad (q\ge 1) \end{cases} \tag{1}$$

where $\mathbf{C}\in\Re^{N\times N}$ is a coefficient matrix, and each element $\mathbf{C}_{i,j}$ represents the importance or contribution of the jth frame in representing the ith frame. The coefficients obtained can not only indicate the locations of the representatives but also provide information about the relative importance of each representative frame. The constraint $||\mathbf{C}||_{row,0}\le Q$ confines the number of non-zero rows to at most k. That is, we wish to find at most S representatives to best "describe" the whole image. Another constraint $1^T\mathbf{C}=1^T$ is used to ensure the selection of representatives to be invariant with respect to a global transition of the dataset. Fig. 1 shows the self-representation based dimensionality reduction scheme. In order to solve the optimization, Eq. (1) can be rewritten as

$$\min_{\mathbf{C}} \|\mathbf{F}-\mathbf{FC}\|_2^2+\lambda\|1^T\mathbf{C}-1^T\|_2^2 \quad \text{s.t.} \quad \|\mathbf{C}\|_{row,0}\le Q \tag{2}$$

This is a convex optimization problem, and a Simultaneous Orthogonal Matching Pursuit (SOMP) algorithm can be used for solving the problem in Eq. (2) [31]. As soon as Q non-zero rows are

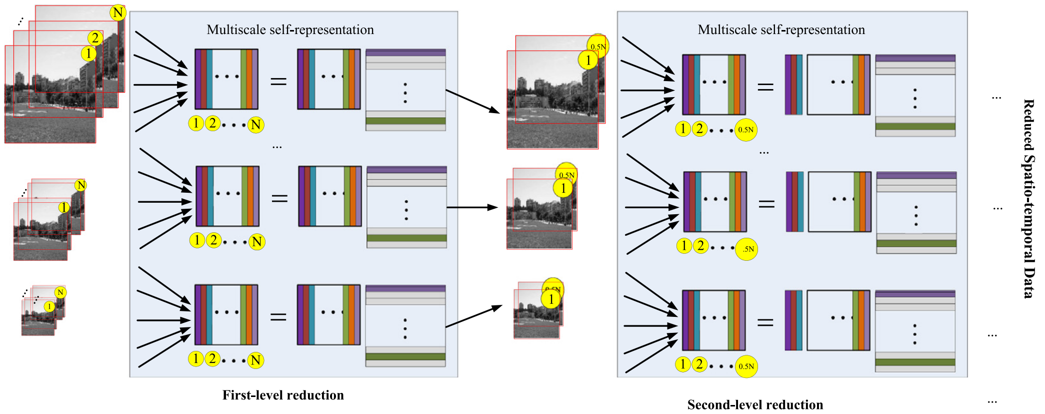
Multiscale self-representation
Multiscale self-representation
N
2
1
1 2 . . . N
0.5N
1
1 2 . . . 0.5N
First-level reduction
Second-level reduction
Reduced Spatio-temporal Data

Fig. 2. Multiscale and hierarchical representation for redundancy reduction.

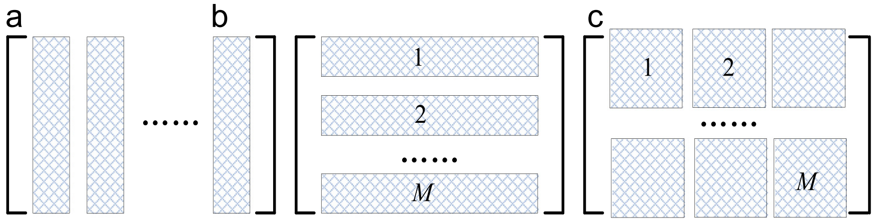

Fig. 3. Division schemes of large volume spatio-temporal data. (a) Division along spatio domain, (b) division along temporal domain, and (c) division in spatio-temporal domain.