Pattern Recognition 47 (2014) 3819–3827

Contents lists available at ScienceDirect

Pattern Recognition

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

Realistic action recognition via sparsely-constructed Gaussian processes

Li Liu [a,b], Ling Shao [a,b,*], Feng Zheng [b], Xuelong Li [c]

[a] College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, PR China
[b] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 3JD, UK
[c] State Key Laboratory of Transient Optics and Photonics, XIOPM, Chinese Academy of Sciences, Xi'an 710119, PR China

ARTICLE INFO

Article history:
Received 6 February 2014
Received in revised form
31 May 2014
Accepted 4 July 2014
Available online 14 July 2014

Keywords:
Action recognition
Gaussian processes
ℓ^1 construction
Local approximation

ABSTRACT

Realistic action recognition has been one of the most challenging research topics in computer vision. The existing methods are commonly based on non-probabilistic classification, predicting category labels but not providing an estimation of uncertainty. In this paper, we propose a probabilistic framework using Gaussian processes (GPs), which can tackle regression problems with explicit uncertain models, for action recognition. A major challenge for GPs when applied to large-scale realistic data is that a large covariance matrix needs to be inverted during inference. Additionally, from the manifold perspective, the intrinsic structure of the data space is only constrained by a local neighborhood and data relationships with far-distance usually can be ignored. Thus, we design our GPs covariance matrix via the proposed ℓ^1 construction and a local approximation (LA) covariance weight updating method, which are demonstrated to be robust to data noise, automatically sparse and adaptive to the neighborhood. Extensive experiments on four realistic datasets, i.e., UCF YouTube, UCF Sports, Hollywood2 and HMDB51, show the competitive results of ℓ^1-GPs compared with state-of-the-art methods on action recognition tasks.

1. Introduction

Human action recognition, nowadays, attracts increasing attention and plays a significant role in various applications, *e.g.*, human–computer interaction, human behavior analysis and video surveillance systems. Traditional action recognition is usually studied with constrained lab settings and a small data set, which assumes that the start and the end of each action are known. It still remains a challenging task for recognizing actions in real-world videos. The wide variety of scene categories, variations in lighting, different view angles and complex backgrounds all lead to obstacles in robust action recognition.

Conventional action recognition approaches are based on hand-crafted features, either global [1] or local [2]. The recent trend is to extract more robust and/or discriminative features through advanced machine learning or profound human knowledge. For example, Le et al. [3] develop an unsupervised deep-learned network to extract the most discriminant features, instead of focusing on using hand-designed local features, such as SIFT or HOG. A similar spatio-temporal feature learning method by convolutional deep net has been proposed by Taylor et al. [4]. Sapienza et al. [5] learn discriminative action subvolumes in a weakly supervised setting. The bag-of-words (BoW) scheme is then followed to represent each action clip. Gilbert et al. [6] use mined hierarchical compound features, which are formed from simple 2D Harris corners. Liu et al. [7] extract action features based on boosted key-frame selection and correlated pyramidal motion feature representations. The AdaBoost learning algorithm is applied to select the most discriminative frames from a large feature pool. In this way, they obtain the top-ranked boosted frames of each video sequence as the key frames which carry the most representative motion information. Moreover, via observing the characteristics of action sequences and studying various existing spatio-temporal descriptors, Wang et al. [8] tailor design the dense trajectory features (DTF), which are formed by the sequence of displacement vectors in trajectory, together with a HOG/HOF descriptor and the motion boundary histogram (MBH) descriptor computed over a local neighborhood along the trajectory. This method is demonstrated to achieve state-of-the-art performance for action recognition.

However, current methods still have the following drawbacks: (1) it is difficult to extract very effective low-level features for a variety of actions; (2) typical methods are not explicitly probabilistic, which makes them inappropriate and unfit for providing an estimation of uncertainty at the inference stage; (3) real-world

* Corresponding author at: College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, PR China. Tel.: +44 1142225841.
E-mail address: ling.shao@ieee.org (L. Shao).

http://dx.doi.org/10.1016/j.patcog.2014.07.006
0031-3203/

action recognition always leads to high computational complexity particularly for some kernel-based machine learning techniques. To overcome these shortcomings, in this work, we introduce a probabilistic framework which effectively combines the Gaussian processes (GPs) regression with sparse ℓ^1 covariance matrix construction for realistic action recognition. Compared with other methods, the GPs-related model is often preferred than SVM in determining the hyper-parameters via evidence maximization and can also provide probabilistic predictions, which is a highly significant property for action recognition in a real and complex background.

GPs [9] have been used for regression tasks in supervised learning systems and applied over a range of applications including data mining, robotics, *etc*. Generally, a Gaussian process is defined as a probability distribution over a function $y(x)$ which is evaluated at an arbitrary set of points with a joint Gaussian distribution. In a wide sense of the world, a stochastic process $y(x)$ is specified by giving the joint probability distribution for any finite set of values in a consistent manner. A key point of Gaussian stochastic processes is that the joint distribution over variables is specified completely by the second-order statistics, namely the mean and the covariance [10]. In most applications, we will not have any prior knowledge about the mean of $y(x)$ and so by symmetry we take it to be *zero*. Therefore, the specification of the Gaussian process is then completed by giving the covariance of $y(x)$ evaluated at any two values of x. The relevant kernel function is given by

$$\mathbb{E}[y(x), y(x')] = k(x, x') \tag{1}$$

In practice, however, GPs have limited applications in large-scale computer vision tasks due to the fact that modeling big training data with stochastic processes is still challenging. This is because inverting a potentially large covariance matrix is computationally expensive during the inference time of GPs. For problems with thousands of observations, a precise inference for conventional GPs becomes intractable and requires approximation. Most previous works based on sparse approximation apply a subset of samples to approximate the posterior distribution for new test samples. These sparse approximation algorithms either rely on the elicitation method to select the subset samples [11] or use obtained pseudo targets during the optimization of log-marginal likelihood of the model [12].

On the other hand, the high complexity problem can be tackled in a different way. Melkumyan and Ramoos [13] proposed a new covariance function which provides intrinsically sparse covariance matrices, instead of applying any sparse approximation. In addition, Ranganathan and Yang [14] also developed a new Gaussian process (GP) regression algorithm, called online sparse matrix Gaussian process (OSMGP) regression, which is exact and allows fast online updates in linear time for covariance functions with local supports. These sparse covariance matrices for GPs are always manually constructed via a k-nearest neighbors constraint, in which the samples x_i and x_j are considered as neighbors if and only if x_i is among the k nearest neighbors of x_j or x_j is among the k nearest neighbors of x_i. k is a positive integer and the similarity between two data samples is commonly measured by the *heat kernel* formulated as follows:

$$K_{ij} = \begin{cases} e^{-\|x_i - x_j\|^2/p} & \text{if } x_i \text{ and } x_j \text{ neighbors,} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where K_{ij} is the covariance matrix element in position (i,j), p is the heat kernel parameter and $\|x_i - x_j\|$ is always measured by the Euclidean distance, which is the pairwise relationship and sensitive to data noise. However, this kind of construction is sensitive to data noise and one noisy feature may dramatically change the data's relationship. Furthermore, when data's distribution is not even, these pairwise-distance based kernels may also involve the far-distance inhomogeneous data together, if the k is large. Besides, the local linear embedding (LLE) proposed by Roweis et al. [15] is also used for sparse matrix construction. The main idea is to reconstruct a sample from its neighboring points and minimize the reconstruction error by the ℓ^2-norm:

$$\min \sum_i \|x_i - \sum_j \alpha_{ij} x_j\|^2, \quad \text{s.t. } \sum_j \alpha_{ij} = 1, \ \forall i \tag{3}$$

where $\alpha_{ij}=0$ if samples x_i and x_j are not neighbors. Nevertheless, this kind of sparse embedding is still suffering the noise on Euclidean distance.

Therefore, in our work we use the ℓ^1-norm constraint to construct the sparse covariance matrix as shown in Fig. 1, which measures the overall data's relationship, instead of employing the pairwise Euclidean distance as in conventional methods. ℓ^1 construction has been utilized for spectral clustering [16], subspace learning, semi-supervised learning [17], *etc*., showing its discriminative advantages for kernel learning approaches: (1) great robustness to data noise, (2) automatic sparsity instead of manual setting, and (3) adaptive neighborhood for each individual data point. We have successfully constructed the ℓ^1 covariance matrix for GPs in our action recognition tasks. The results in Section 3.2 show its significant superiority over baseline and state-of-the-art methods.

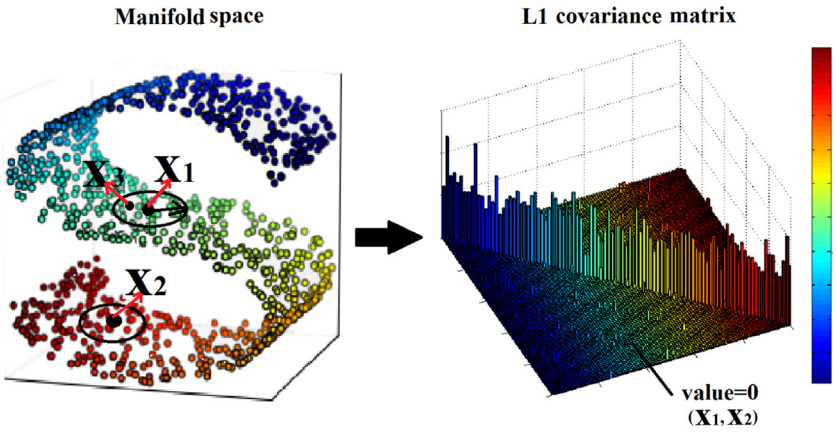

Fig. 1. The left shows data distribution in the original space. From the manifold perspective, the intrinsic structure of data should only be constrained in a local neighborhood (*e.g.*, x_1 and x_3 and the corresponding black circle illustrate the neighborhood of x_1), and the relationship of data between two distant points in the original feature space should be ignored (*e.g.*, x_1 and x_2). The right figure illustrates the proposed sparse ℓ^1 covariance matrix, where the self-similarity has the highest covariance values (*i.e.*, diagonal) and the covariance value between two distant points is zero (*e.g.*, x_1 and x_2).