Neurocomputing 121 (2013) 532–539

Contents lists available at ScienceDirect

Neurocomputing

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

Unsupervised Natural Image Segmentation via Bayesian Ying–Yang Harmony Learning Theory

Shaojun Zhu, Jieyu Zhao*, Lijun Guo, Yuanyuan Zhang

Institute of Computer Science and Technology, Ningbo University, Ningbo 315211, China

ARTICLE INFO

Article history:
Received 17 September 2012
Received in revised form
8 May 2013
Accepted 17 May 2013
Communicated by L. Xu
Available online 12 June 2013

Keywords:
Unsupervised image segmentation
Bayesian Ying–Yang harmony learning
Dirichlet-Normal-Wishart prior
Superpixels
Automatic model selection

ABSTRACT

An unsupervised image segmentation method for natural images is proposed in this paper. We assume that texture features in natural images are distributed as a mixture of Gaussians. In order to cluster the extracted feature vectors, we modify a clustering algorithm based on Bayesian Ying–Yang (BYY) harmony learning theory with Dirichlet-Normal-Wishart prior. This algorithm can determine the number of components automatically during the clustering procedure, as long as we give a large enough initial component number. Our works in this paper have presented a complete pipeline of clustering-based image segmentation including feature extraction, robust feature clustering and methodological effective post processing. The experiments reported in this paper demonstrate that the proposed method is efficient (in terms of visual evaluation and quantitative performance measures) and performs competitively compared to the existing state-of-the-art segmentation methods.

1. Introduction

Image segmentation is a frequent preprocessing step which consists of achieving a compact region-based description of the image scene by aggregating its pixels into spatially coherent regions with similar attributes. This low-level vision task is often the preliminary and also crucial step for many image understanding algorithms and computer vision applications. Due to the grouping nature, it has been popularly addressed as a special type of data clustering problem by various techniques [1,2].

A major challenge in unsupervised image segmentation is to determine the segments numbers automatically. Model selection, i.e., choosing the correct number of components for a given data set, is an important issue, on which numerous approaches have been dedicated [3–6]. The most straightforward strategy of model selection is to fit all possible models to the data, and evaluate all the solutions using a certain criterion. Such criteria, mainly inspired by coding theory, try to avoid data over fitting by penalizing solutions that have high number of components. Examples include Akaike's Information Criterion (AIC) [4], Bayesian Information Criterion (BIC) [5], Minimum Message Length (MML) [6–8], and Minimum Description Length (MDL) [3,9]. However, approaches developed based on these criteria are not computationally efficient as the process of sequentially evaluating the criterion needs large computational cost. Moreover, all of the existing theoretic selection criteria have their limitations. For example, it has been reported that for large values of N (the number of samples), AIC tends to choose models that are too complex. On the other hand, for smaller values of N, BIC tends to choose models that are too simple [10].

Recently, some other gradient based learning algorithms with automated model selection ability have been proposed from the perspective of Bayesian Ying–Yang (BYY) harmony learning principle [11]. The BYY harmony learning principle was proposed in [11] and systematically developed in [12–14]. It acts as a general statistical learning framework not only for unifying several existing typical learning models, but also for tackling the above key challenge with a new learning mechanism that makes automatic model selection implemented simultaneously during parameter learning [15].

It is straightforward to adapt BYY harmony learning into unsupervised image segmentation, simply by assuming that the texture feature vectors extracted from images have some distributions, e.g., Gaussian Mixture Models (GMMs), and clustering them. The original Bayesian–Kullback Ying–Yang Model Selection Criterion has been directly used in image segmentation in [16,17], but the results are poor, those works only use the color feature and did not provide further process for specific problems in image segmentation. Recently, Shi et al. [18] developed BYY algorithms with both Jeffreys prior and Dirichlet-Normal-Wishart prior, their extensive comparisons show that BYY based algorithms

* Corresponding author. Tel.: +86 574 87605303.
E-mail addresses: shj.zhu@gmail.com (S. Zhu), zhao_jieyu@nbu.edu.cn (J. Zhao), guolijun@nbu.edu.cn (L. Guo), zhangyuanyuan@nbu.edu.cn (Y. Zhang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.017

considerably outperforms other Bayesian related approaches, especially in detecting the objects of interest from a confusing background. When using the BYY algorithms on image segmentation directly, high quality segmentation results are obtained in term of Probabilistic Rand Index [19]. However, there exists some limitations when confronting image segmentation problem. The major problem is concerning the speed, image segmentation is a large scale problem, images in Berkeley Segmentation Data Set 500 (BSDS500) [20] are of size 481×321 (or 321×481), that is 154401 pixels. But due to the complex updating rules and the online updating strategy, several inverses of matrices need to be computed for every coming pixel, this greatly slowdown the algorithms. The second problem is that because the features are extracted from cut-off windows around pixels, the segment results cannot fit edges of objects in images very well.

In order to fully investigate the effectiveness of BYY based image segmentation algorithms, we design an novel method for natural images. We develop a clustering algorithm based on BYY harmony learning theory with Dirichlet-Normal-Wishart prior. Our algorithm updates the parameters with batch rules and eliminating some complex terms that need a lot of computing resources, thus greatly accelerate the clustering process. During the clustering procedure, this algorithm keeps the strong capability of determine the number of components automatically, which inherits from BYY systems. Furthermore, we develop post process methods for clustering based image segmentation algorithms. An important improvement comparing with other clustering based methods is that we assign the labels with regard to superpixels [21], but not pixels. We conduct extensive experiments to compare the results with human segmentation using the BSDS500. The segmentation results match very well with those by humans and are competitive with the state-of-the-art segmentation algorithms.

2. Bayesian Ying–Yang learning

Consider a world $\boldsymbol{X}$ with each object in observation represented by a stochastic notation $\boldsymbol{x} \in \boldsymbol{X}$. Corresponding to each $\boldsymbol{x}$, there is an inner representation $\boldsymbol{y} \in \boldsymbol{Y}$ in the representation domain $\boldsymbol{Y}$ of a learning system. We consider the joint distribution of $\boldsymbol{x}, \boldsymbol{y}$, which can be understood from two complement perspectives.

On one hand, we can regard all $\boldsymbol{x}$ are generated from an invisible inner representation $\boldsymbol{x}$ by

$$q(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{y}) q(\boldsymbol{y})\, d\boldsymbol{y} \tag{1}$$

That is, $\boldsymbol{x}$ is generated from an inner distribution $q(\boldsymbol{y})$ in a structure that is designed according to the learning tasks.

On the other hand, we can also regard each $\boldsymbol{x}$ is mapped into an invisible inner representation $\boldsymbol{y}$ by

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} \tag{2}$$

which matches the inner density $q(\boldsymbol{y})$ in a pre-specified structure.

The two perspectives reflect the two types of Bayesian decomposition of the joint density $q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y}) = q(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})$ on $\boldsymbol{X} \times \boldsymbol{Y}$. Without any constraints, the two decompositions should be theoretically identical. However, in practice, the four components $p(\boldsymbol{y}|\boldsymbol{x}), p(\boldsymbol{x}),\ q(\boldsymbol{x}|\boldsymbol{y}), q(\boldsymbol{y})$ are always subject to certain structural constraints. Thus, we have two different but complementary Bayesian representations:

$$p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}), \quad q(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{y}) \tag{3}$$

which, as stated in [12], compliments to the famous Chinese ancient Ying–Yang philosophy with $p(\boldsymbol{x},\boldsymbol{y})$ called Yang machine that consists of the observation space (or called Yang space) by $p(\boldsymbol{x})$ and the forward pathway (or called Yang pathway) by $p(\boldsymbol{y}|\boldsymbol{x})$ and with $q(\boldsymbol{x},\boldsymbol{y})$ called Ying machine that consists of the invisible state space (or Ying space) by $q(\boldsymbol{y})$ and the Ying (or backward) pathway by $q(\boldsymbol{x}|\boldsymbol{y})$. Such a pair of Ying–Yang models is called Bayesian Ying–Yang (BYY) system.

The fundamental learning principle is to make the Ying machine and Yang machine to be best harmony in two senses. One is that the difference between the two Bayesian representations should be minimized, and the other is the resulting BYY system should be least complex. All the unknowns in the system are learned by the harmony principle, which is mathematically implemented by maximizing the following harmony measurement:

$$H(p||q) = \int p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x}) \ln[q(\boldsymbol{x}|\boldsymbol{y}) q(\boldsymbol{y})]\, d\boldsymbol{y}\, d\boldsymbol{x} \tag{4}$$

3. BYY with Dirichlet-Normal-Wishart prior

3.1. Dirichlet-Normal-Wishart prior

One advantage of the Bayesian viewpoint is that the inclusion of prior knowledge arises naturally [22]. We advocate the use of conjugate priors, and, in particular, the Dirichlet distribution for the mixing weights and the Normal-Wishart distributions for the Gaussian models.

The Dirichlet distribution (after Johann Peter Gustav Lejeune Dirichlet), often denoted $\mathrm{Dir}(\boldsymbol{\alpha})$, is a family of continuous multivariate probability distributions parametrized by a vector $\boldsymbol{\alpha}$ of positive reals. It is the multivariate generalization of the beta distribution. Dirichlet distributions are very often used as prior distributions in Bayesian statistics, and in fact the Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.

The Dirichlet distribution has support over the probability simplex, defined by

$$S_k = \left\{ \boldsymbol{\alpha} : 0 \le \alpha_i \le 1,\ \sum_{i=1}^{k} \alpha_i = 1 \right\} \tag{5}$$

The pdf is defined as follows:

$$\mathrm{Dir}(\boldsymbol{\alpha}|\lambda, \xi) = \frac{\Gamma(\xi)}{\prod_{i=1}^{k} \Gamma(\xi\lambda_i)} \prod_{i=1}^{k} \alpha_i^{\xi\lambda_i - 1}, \quad \boldsymbol{\alpha} \in S_k \tag{6}$$

where $\xi > 0$ and $\sum_{i=1}^{k} \lambda_i = 1$.

The Normal-Wishart distribution (or Gaussian–Wishart distribution) is a multivariate four-parameter family of continuous probability distributions. It is the conjugate prior of a multivariate normal distribution with unknown mean and precision matrix. Suppose

$$\boldsymbol{\mu}|\boldsymbol{m}, \beta, \mathbf{T} \sim \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, (\beta\mathbf{T})^{-1}) \tag{7}$$

has a multivariate normal distribution with mean $\boldsymbol{m}$ and covariance matrix $(\beta\mathbf{T})^{-1}$, where

$$\mathbf{T}|\boldsymbol{\Phi}, \nu \sim \mathcal{W}(\mathbf{T}|\boldsymbol{\Phi}, \gamma) \tag{8}$$

has a Wishart distribution. Then $(\boldsymbol{\mu}, \mathbf{T})$ has a normal-Wishart distribution, denoted as

$$(\boldsymbol{\mu}, \mathbf{T}) \sim \mathrm{NW}(\boldsymbol{m}, \beta, \boldsymbol{\Phi}, \gamma) \tag{9}$$

The pdf is defined as follows:

$$\mathrm{NW}(\boldsymbol{\mu}, \mathbf{T}|\boldsymbol{m}, \beta, \boldsymbol{\Phi}, \gamma) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, (\beta\mathbf{T})^{-1}) \mathcal{W}(\mathbf{T}|\boldsymbol{\Phi}, \gamma)$$
$$= \frac{\exp[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{m})^T (\beta\mathbf{T})^{-1} (\boldsymbol{\mu}-\boldsymbol{m})]}{(2\pi)^{k/2} |\beta\mathbf{T}|^{1/2}}$$