J. Vis. Commun. Image R. 24 (2013) 1450–1457

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

q-Gaussian mixture models for image and video semantic indexing

Nakamasa Inoue *, Koichi Shinoda

Department of Computer Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan

ARTICLE INFO

Article history:
Received 29 January 2013
Accepted 23 October 2013
Available online 1 November 2013

Keywords:
Semantic indexing
Gaussian mixture models
q-Gaussian mixture models
GMM supervector
Image classification
Bag-of-words
Codebook
Tsallis statistics

ABSTRACT

Gaussian mixture models which extend Bag-of-Visual-Words (BoW) to a probabilistic framework have been proved to be effective for image and video semantic indexing. Recently, the q-Gaussian distribution, derived from Tsallis statistics [11], has been shown to be useful for representing patterns in many *complex* systems in physics. We propose q-Gaussian mixture models (q-GMMs), mixture models of q-Gaussian distributions with a parameter q to control its tail-heaviness, for image and video semantic indexing [1]. The long-tailed distributions obtained for $q > 1$ are expected to effectively represent complexly correlated data, and hence, to improve robustness against outliers. The main improvements over our previous study [1] are q-GMM super-vector representation to efficiently compute the q-GMM kernel, and detailed experimental analysis showing accuracy and testing-cost comparison with recent kernel methods. Our proposed method outperformed BoW and achieved 49.42% and 10.90% in Mean Average Precision on the PASCAL VOC 2010 and the TRECVID 2010 Semantic Indexing, respectively.

1. Introduction

Recent years have seen extensive growth of image and video archives on the Internet. For example, Flickr stores more than 6 billion photos in its database. A specific object or a scene can be easily detected on the Flickr if the photos have detailed meta data such as semantic tags. However, each photo usually only has a few tags since tagging is very time-consuming for users. Further, for video archives, not only tags but also video summarizations in text are needed as meta data in order to search a specific scene precisely.

Semantic indexing is necessary to automatically generate meta data, since semantics are the most important part of the meta data to describe contents of images and videos. Semantic indexing aims to detect objects, scenes, and actions, e.g. "airplane", "cityscape", and "flying". It has been a challenging task due to the semantic gap between low-level features and high-level semantic concepts.

Current approaches to semantic indexing are typically based on bag-of-visual-words (BoW) [2]. In BoW, each low-level feature (e.g. SIFT [8]) extracted from an image is assigned to a visual word, i.e., a code word obtained by vector quantization (VQ). To reduce quantization errors in VQ, a Gaussian mixture model (GMM) [10] which extends BoW to a probabilistic framework is often utilized. For example, the Fisher vector [22] and the GMM supervector [5,4] represent an image as a concatenation of the GMM parameters and outperform BoW.

Recently, the q-Gaussian distribution, which is derived from Tsallis statistics [11], has been shown to be effective for representing patterns in many *complex* systems in physics such as fractals and cosmology. Tsallis statistics is a generalization of the standard Boltzmann–Gibbs (BG) statistics. It introduces Tsallis q-entropy [11] which has a real-valued parameter q. The q-Gaussian distribution is derived by maximizing the Tsallis q-entropy [11] while the Gaussian distribution is derived by maximizing the BG entropy. For the q-Gaussian distribution, q is a parameter that controls its tail-heaviness as shown in Fig. 1. A long-tailed distribution obtained for $q > 1$ is expected to effectively represent complexly correlated data, and hence, to improve the robustness against outliers. Note that since Tsallis q-entropy represents the BG entropy when it takes a value of $q \to 1$, the q-Gaussian distribution represents the Gaussian distribution when $q \to 1$. Many statistical frameworks including BoW can be extended to Tsallis statistics by using the q-Gaussian distribution.

In this paper, we propose a q-Gaussian mixture model (q-GMM) based on the Tsallis statistics and its application to image and video semantic indexing systems [1]. The q-GMM is a mixture model of q-Gaussian distributions and is an extension of the GMM to a mixture of long-tailed distributions. This paper proposes two separate methods based on the q-GMM: a histogram-based representation and a q-GMM kernel. The first method, histogram-based representation, is a direct extension of BoW to the q-GMM. It represents an image as a histogram of low-level features in the same way as BoW but the q-GMM is used as a visual codebook. The second method, q-GMM kernel, is an RBF-kernel in which each image is represented by a q-GMM. The q-GMM super-vector representation is introduced to efficiently compute the q-GMM kernel.

* Corresponding author.
E-mail addresses: inoue@ks.cs.titech.ac.jp (N. Inoue), shinoda@cs.titech.ac.jp (K. Shinoda).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.10.005

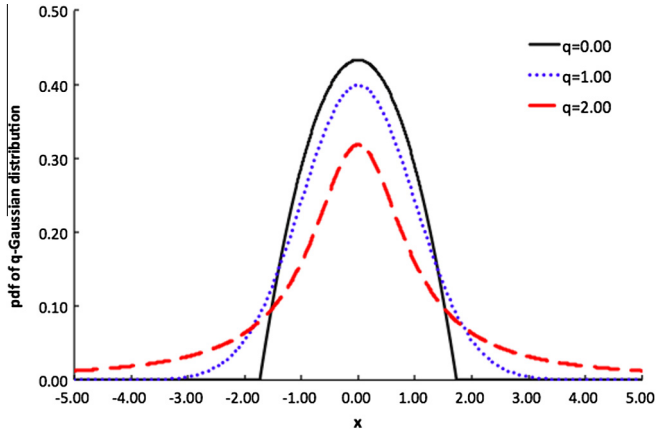

Fig. 1. The q-Gaussian distributions. The (normal) Gaussian distribution is obtained when $q = 1$. The tail of a q-Gaussian distribution is longer than that of a Gaussian distribution when $q > 1$.

The main improvements over our previous study in [1] are this q-GMM super-vector representation and detailed experimental analysis to show accuracy and testing-cost comparison with recent kernel methods, as well as the influence of parameters.

This paper is organized as follows. Related work is described in Section 2. The proposed method with the definition of q-GMMs is described in Section 3. Experimental results on PASCAL VOC and TRECVID datasets are described in Section 4. Conclusion and future work are described in Section 5.

2. Related work

One of the most popular approaches to image and video semantic indexing is the bag-of-visual-words (BoW) approach [2]. In BoW, an image is represented as a histogram of visual words obtained by applying vector quantization (VQ) to each low-level descriptors, e.g. SIFT [8], SURF [9]. k-means clustering is often used for training a visual codebook that consists of several thousands of visual words.

Soft-assignment approaches are effective for reducing quantization errors in VQ and thus they outperform BoW. Gemert et al. [23] proposed a kernel codebook in which each low-level descriptor is assigned to all visual words in a soft manner with weighting. Hang et al. [24,25] used sparse coding which assigns a low-level descriptor to several tens of visual words by solving a constrained least square fitting problem. Perronnin et al. [10] used a Gaussian mixture model (GMM) for a codebook which holds mean and variance information for each visual word. The GMM is a straight-forward extension of BoW to a probabilistic framework.

Recently, high-dimensional image representations have been proven to be effective in image classification. Vector of locally aggregated descriptor (VLAD) [27] and super-vector coding [3] use the first order differences between low-level descriptors and visual words in addition to the BoW histogram. Fisher vector [20] represents an image as a concatenation of parameters of a parametric probability model. Perronnin et al. [22,21] achieved the best performance in the PASCAL VOC image classification challenge by using a GMM as the parametric probability model for the Fisher vector. They reported that an normalization technique [21] is needed since the Fisher vectors become sparser as the number of Gaussians increases. They proposed the L2 + power normalization to make the Fisher vectors dense. The normalized fisher vector is called "improved Fisher kernel (IFK)". Chatfield et al. [26] reported that IFK is the best of these recent image representations.

Another direction to improve image classification performance is to develop discriminative learning methods. With above image representations, a one-versus-all support vector machine (SVM) is most widely used since it is a powerful classifier with theoretical foundations. Kernel tricks such as an RBF-kernel and χ^2-kernel significantly improve classification performance of the SVM. Some recent works focus on multi-label learning that aims to train detectors of multiple objects at the same time. For example, Zha et al. proposed a multi-label and multi-instance learning method using hidden conditional random fields for image classification in [28]. A graph-based method based on semi-supervised learning for multiple concept detection in video is presented in [29]. On the other hand, some works focus on reducing the human-labeling cost. For example, an interactive video indexing system using active learning is proposed in [30]. Ayache et al. [31] proves a web-based active learning system for annotating video corpus in TRECVID. More discussion of image retrieval and video retrieval can be found in survey in [32,33].

On the other hand, several previous works focused on applying Tsallis statistics [13,11,12] to image processing. Tsallis q-entropy, which is a generalization of the Boltzmann–Gibbs (BG) entropy, is used for image thresholding for foreground extraction in [14–17]. These works show that long-range correlation between foreground pixels can be modeled by using the Tsallis q-entropy. Fabbrib et al. [18] applied Tsallis q-entropy to image texture classification. They reported that texture classification accuracy is improved by using Tsallis q-entropies for multiple q-values as a feature vector. To the best of our knowledge, we are the first to apply Tsallis statistics to the bag-of-visual-words framework.

3. q-Gaussian mixture models

The procedure of the proposed image and video semantic indexing based on q-Gaussian mixture models (q-GMMs) is shown in Fig. 2. First, low-level features (e.g. SIFT features) are extracted from image/video data. Second, a q-GMM for a background model is estimated from low-level features in training data. Finally, each image is represented by a histogram of low-level features in which the background model is used as a visual codebook.

3.1. q-Gaussian mixture models

The q-Gaussian distribution, which has a parameter q to control its tail-heaviness, is derived by maximizing Tsallis q-entropy S_q given by

$$S_q = -\frac{1}{1-q}\left(1 - \int p(x)^q dx\right). \quad (1)$$

The Boltzmann–Gibbs (BG) entropy is obtained from Tsallis q-entropy S_q for $q \to 1$. The probability density function of the q-Gaussian distribution $\mathcal{N}_q$ is given by

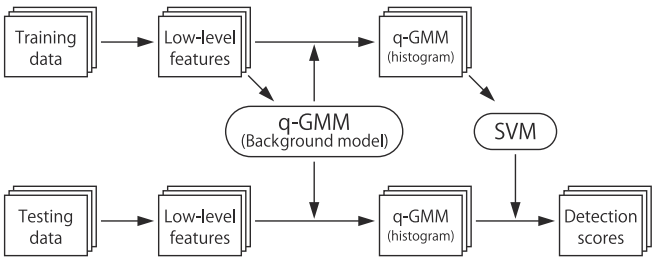

Fig. 2. The framework of image and video semantic indexing using q-Gaussian mixture models.