論文 / 著書情報
Article / Book Information

Title	q-Gaussian Mixture Models for Image and Video Semantic Indexing
Author	Nakamasa Inoue, Koichi Shinoda
Journal/Book name	Journal of Visual Communication and Image Representation, vol. 24, no. 8, pp. 1450-1457
Issue date	2013, 11
DOI	http://dx.doi.org/10.1016/j.jvcir.2013.10.005
Note	This file is author (final) version.

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

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 [12], 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.

Keywords: Semantic indexing, Gaussian mixture models, q-Gaussian mixture models.

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.

Email address: inoue@ks.cs.titech.ac.jp (N. Inoue), shinoda@cs.titech.ac.jp (K. Shinoda). (Koichi Shinoda)

Current approaches to semantic indexing are typically based on bag-of-visual-words (BoW) [2]. In BoW, each low-level feature (e.g. SIFT [9]) 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) [11] which extends BoW to a probabilistic framework is often utilized. For example, the Fisher vector [23] and the GMM supervector [6, 5] represent an image as a concatenation of the GMM parameters and outperform BoW.

Recently, the q-Gaussian distribution, which is derived from Tsallis statistics [12], 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 [12] which has a real-valued parameter q. The q-Gaussian distribution is derived by maximizing the Tsallis q-entropy [12] 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. 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 Sec. 2. The proposed method with the definition of q-GMMs is described in Sec. 3. Experimental results on PASCAL VOC and TRECVID datasets are described in Sec. 4. Conclusion and future work are described in Sec. 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 [9], SURF [10]. k-means clustering is often used for training a visual codebook that consists of several thousands of visual words.

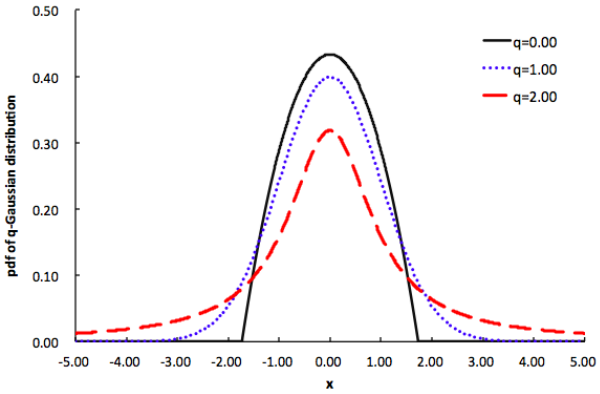

Figure 1. The q-Gaussian distributions. The (normal) Gaussian distribution is obtained when $q = 1$. The tail of a q-Gaussian distribution is longer than that of a Gaussian distribution when $q > 1$.

Soft-assignment approaches are effective for reducing quantization errors in VQ and thus they outperform BoW. Gemert *et al.* [25] 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.* [26, 27] 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.* [11] 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) [29] and super-vector coding [4] use the first order differences between low-level descriptors and visual words in addition to the BoW histogram. Fisher vector [21] represents an image as a concatenation of parameters of a parametric probability model. Perronnin *et al.* [23, 22] 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 [22] 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.* [28] 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, an 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 [30]. A graph-based method based on semi-supervised learning for multiple concept detection in video is presented in [31]. On the other hand, some works focus on reducing the human-labeling cost. For example, an interactive video indexing system using active

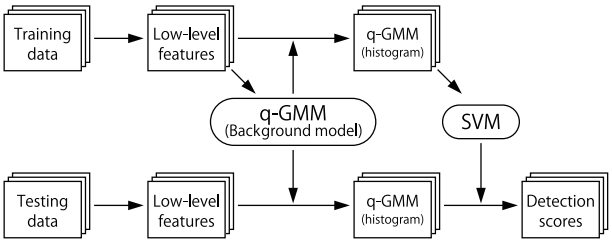

Figure 2. The framework of image and video semantic indexing using q-Gaussian mixture models.

learning is proposed in [32]. Ayache et al. [33] 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 [34, 35].

On the other hand, several previous works focused on applying Tsallis statistics [14, 12, 13] 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 [15, 16, 17, 18]. These works show that long-range correlation between foreground pixels can be modeled by using the Tsallis q-entropy. Fabbrib *et al.* [19] 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

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

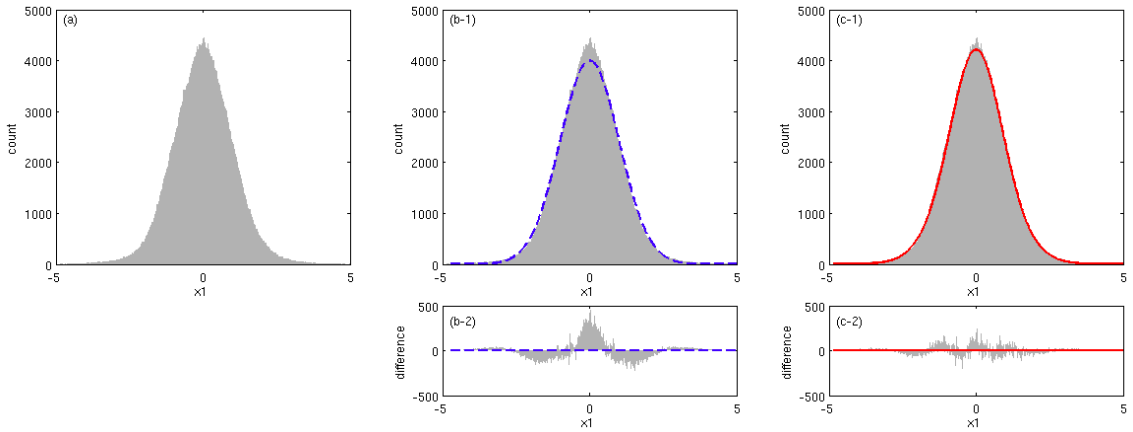

Figure 3. (a): Standardized histogram of the first elements of standardized SIFT descriptors. 1 million low-level descriptors are randomly sampled from training data of PASCAL VOC 2010 dataset. (b-1), (b-2): A fitting result by a Gaussian distribution and its residuals. (c-1), (c-2): A fitting result by a q-Gaussian distribution ($q = 1.12$) and its residuals.

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

$$\mathcal{N}_q(x \mid \mu, \Sigma) = \begin{cases} \dfrac{1}{Z_q}\left(1 - \dfrac{1-q}{3-q}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)^{\dfrac{1}{1-q}}, \\ \qquad \text{if } (x-\mu)^T\Sigma^{-1}(x-\mu) < \dfrac{3-q}{1-q}, \\ 0, \qquad \text{otherwise}, \end{cases} \tag{2}$$

where μ is a mean vector, Σ is a covariance matrix, and Z_q is a normalizing constant to make the integral over Eq. (2) to 1. Note that the q-Gaussian distribution asymptotically approaches to the Gaussian distribution when $q \to 1$.

Fig. 1 shows the shape of q-Gaussian distributions for some q-values. The q-Gaussian distribution has a longer tail than the Gaussian distribution when $q > 1$. The long-tailed distribution obtained for $q > 1$ is expected to be effective for representing complexly correlated data. Fig. 3 shows a histogram of the first values of 1 million low-level descriptors with fitting results. The q-Gaussian distribution is more suitable than the Gaussian distribution to represent distribution of the low-level descriptors.

To further improve the expressiveness of the q-Gaussian distribution, we introduce a mixture model of q-Gaussian distributions, namely a q-Gaussian mixture model (q-GMM), by

$$p_q(x \mid \theta) = \sum_{k=1}^{K} w_k \mathcal{N}_q(x \mid \mu_k, \Sigma_k), \tag{3}$$

where K is the number of mixtures, w_k is a mixture coefficient, and $\theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ is a set of q-GMM parameters.

3.2. Training q-GMM for a Background Model

From a set of low-level features $X = \{x_i\}_{i=1}^N$ in training data, we estimate q-GMM parameters for a background model which is used instead of a codebook for BoW. We propose the expectation maximization (EM) algorithm for q-GMMs as follows.

E-step

Evaluate posterior probabilities c_{ik} as follows:

$$c_{ik} = \frac{\hat{w}_k \mathcal{N}_q(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{k=1}^K \hat{w}_k \mathcal{N}_q(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}. \tag{4}$$

M-step

To derive the parameter-update rules for the M-step, we introduce a Q-function given by

$$Q(\theta) = \log \prod_i p_q(x_i \mid \theta) + \lambda \left(1 - \sum_{k'} w_{k'}\right) \tag{5}$$

where p_q is a pdf of a q-GMM defined by Eq. (3). A lagrangian multiplier λ is introduced to obtain w_k such that

$$\sum_{k'} w_{k'} = 1. \tag{6}$$

The derivations of the Q-function for each parameter are given by

$$\frac{\partial}{\partial w_k} Q(\theta) = \sum_i \frac{\mathcal{N}_q(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} w_{k'} \mathcal{N}_q(x_i \mid \mu_{k'}, \Sigma_{k'})} - \lambda \tag{7}$$

$$= \frac{1}{w_k} \sum_i c_{ik} - \lambda, \tag{8}$$

$$\frac{\partial}{\partial \mu_k} Q(\theta) = \sum_i \frac{a_{ik} w_k \mathcal{N}_q(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} w_{k'} \mathcal{N}_q(x_i \mid \mu_{k'}, \Sigma_{k'})} \Sigma_k^{-1}(x_i - \mu_k) \tag{9}$$

$$= \Sigma_k^{-1} \sum_i a_{ik} c_{ik} (x_i - \mu_k), \tag{10}$$

$$\frac{\partial}{\partial \Sigma_k} Q(\theta) = \frac{1}{2} \sum_i \frac{w_k \mathcal{N}_q(x_i \mid \mu_k, \Sigma_k)}{\sum_{k'} w_{k'} \mathcal{N}_q(x_i \mid \mu_{k'}, \Sigma_{k'})} \Sigma_k^{-2} \left(a_{ik}(x_i - \mu_k)(x_i - \mu_k)^T - \Sigma_k\right) \tag{11}$$

$$= \frac{1}{2} \Sigma_k^{-2} \sum_i c_{ik} \left(a_{ik}(x_i - \mu_k)(x_i - \mu_k)^T - \Sigma_k\right), \tag{12}$$

where

$$a_{ik} = \frac{2}{3 - q - (1 - q)(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)}. \tag{13}$$

The parameter-update rules for the M-step are obtained by setting the derivatives of Q to zero. For mixture coefficients w_k, we obtain

$$\hat{w}_k = \frac{C_k}{\sum_{k=1}^{K} C_k}, \tag{14}$$

from Eq. (8) where $C_k = \sum_i c_{ik}$.

However, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ can not be obtained from Eqs. (10) and (12) analytically since μ_k and Σ_k appear in a_{ik}. Our preliminary experiments show that numerically solving those equations by the steepest descent (SD) method is time-consuming and it's difficult to optimize step size in SD in a high-dimensional space. Hence we assume a_{ik} is a constant A for all i and k, and analytically solve the equations as

$$\hat{\mu}_k = \frac{1}{C_k} \sum_{i=1}^{N} c_{ik} x_i, \tag{15}$$

$$\hat{\Sigma}_k = \frac{A}{C_k} \sum_{i=1}^{N} c_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T. \tag{16}$$

We determine the value of A so that the expectation of $\hat{\Sigma}_k$, $\mathbb{E}[\hat{\Sigma}_k]$, is equal to Σ_k, a covariance matrix of a q-Gaussian distribution $\mathcal{N}_q(\cdot \mid \mu_k, \Sigma_k)$. The expectation of $\hat{\Sigma}_k$ is calculated as

$$\mathbb{E}\left[\hat{\Sigma}_k\right] = \mathbb{E}\left[\frac{A}{C_k} \sum_{i=1}^{N} c_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T\right] \tag{17}$$

$$= \frac{A}{C_k} \sum_{i=1}^{N} c_{ik} \mathbb{V}[x_i] \tag{18}$$

$$= A \frac{3-q}{5-3q} \Sigma_k. \tag{19}$$

where we use the fact

$$\mathbb{V}[x_i] = \frac{3-q}{5-3q} \Sigma_k. \tag{20}$$

Therefore, A is set to

$$A = \left(\frac{3-q}{5-3q}\right)^{-1}. \tag{21}$$

Here, we assume $q < \frac{5}{3}$ since a q-Gaussian distribution has an infinite variance if $q \geq \frac{5}{3}$.

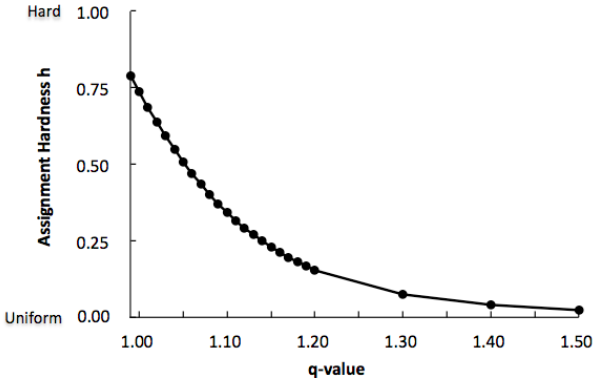

Figure 4. Assignment hardness h with different q-values.

3.3. q-GMM for histogram-based image representation

To represent an image by a feature vector, we create a histogram of low-level features $H(X')$ from a set of low-level features $X' = \{x_i\}_{i=1}^{N'}$ extracted from an image as follows:

$$H(X') = \begin{pmatrix} C_1 \\ C_2 \\ \vdots \\ C_K \end{pmatrix}, \; C_k = \sum_i c_{ik}, \tag{22}$$

where c_{ik} is the posterior probability of x_i being at the k-th q-Gaussian component. It is given by

$$c_{ik} = \frac{\hat{w}_k \mathcal{N}_q(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{k=1}^{K} \hat{w}_k \mathcal{N}_q(x_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}. \tag{23}$$

where $\hat{w}_k, \hat{\mu}_k$, and $\hat{\Sigma}_k$ are q-GMM parameters for the background model. The posterior probabilities c_{ik} can be viewed as weights in soft-assignment of visual words since they satisfy

$$\sum_{k=1}^{K} c_{ik} = 1, \; 0 \le c_{ik} \le 1. \tag{24}$$

Thus, the q-GMM can be regarded as an extension of BoW to a probabilistic framework.

We found that the assignment using the q-GMM comes close to the hard-assignment (i.e., only one of c_{ik} is equal to 1.0 and others are 0.0) as q decreases, and comes close to the uniform-assignment (i.e., all c_{ik} have equivalent values) as q increases. To measure how much the assignment is close to the hard-assignment, we introduce the assignment

Table 1. The targeted semantic concepts for PASCAL VOC 2010 and TRECVID 2010.

PASCAL VOC 2010				
Aeroplane	Bicycle	Bird	Boat	Bottle
Bus	Car	Cat	Chair	Cow
Diningtable	Dog	Horse	Motorbike	Person
Pottedplant	Sheep	Sofa	Train	Tvmonitor
TRECVID 2010				
Airplane Flying	Animal	Asian People	Bicycling	Boat Ship
Bus	Car Racing	Cheering	Cityscape	Classroom
Dancing	Dark-skinned People	Demonstration Or Protest	Doorway	Explosion Fire
Female Human Face Closeup	Flowers	Ground Vehicles	Hand	Mountain
Nighttime	Old People	Running	Singing	Sitting down
Swimming	Telephones	Throwing	Vehicle	Walking

hardness h defined by

$$h = \frac{1}{N'} \sum_{i=1}^{N'} \frac{\max_k c_{ik} - K^{-1}}{1 - K^{-1}}. \tag{25}$$

The assignment hardness h is designed to reach 1.0 for the hard-assignment and 0.0 for the uniform-assignment.

Fig. 4 shows the assignment hardness with different q-values. To improve the final performance of image and video indexing, the assignment should not be too hard nor too uniform. Here, we employ a q-value of 1.05 that has the middle value (0.5) of assignment hardness.

3.4. q-GMM Kernel

Here, we introduce q-GMMs instead of the BoW histograms to represent images and videos. Generally, the number of low-level features extracted from an image is limited and may not be enough to estimate q-GMM parameters robustly. Thus, we use the maximum a posteriori criteria which provides robust parameter estimation. For each image that has low-level features $X' = \{x_i\}_{i=1}^{N'}$, we only update q-GMM mean vectors from the background model as follows:

$$\tilde{\mu}'_k = \frac{\tau\hat{\mu}_k + \sum_{i=1}^{N'} c_{ik} x_i}{\tau + \sum_{i=1}^{N'} c_{ik}}, \tag{26}$$

where N' is the number of the low-level features, $\hat{\mu}_k$ is a q-GMM parameter for the background model, τ is a prefixed hyper-parameter, and c_{ik} is the posterior probability given by Eq. (4).

For a kernel to train support vector machines (SVMs), we introduce the following RBF-based kernel, namely q-GMM kernel,

$$k(X', X'') = \exp\left(-\gamma \sum_{k=1}^{K} \hat{w}_k (\tilde{\mu}'_k - \tilde{\mu}''_k)^T \hat{\Sigma}_k^{-1} (\tilde{\mu}'_k - \tilde{\mu}''_k)\right), \tag{27}$$

where X' is a set of low-level features extracted from an image, $\tilde{\mu}'_k$ is an updated q-GMM mean vector, $\hat{\Sigma}_k, w_k$ are q-

GMM parameters for the background model, and γ is a scaling parameter. The weighted sum of Mahalanobis distance between the k-th q-Gaussian components is utilized in the q-GMM kernel.

To efficiently compute the q-GMM kernel, we define the following super-vector $\phi(X')$ and store it in storage.

$$\phi(X') = \begin{pmatrix} \sqrt{\hat{w}_1}\hat{\Sigma}_1^{-\frac{1}{2}}\tilde{\mu}'_1 \\ \sqrt{\hat{w}_2}\hat{\Sigma}_2^{-\frac{1}{2}}\tilde{\mu}'_2 \\ \vdots \\ \sqrt{\hat{w}_K}\hat{\Sigma}_K^{-\frac{1}{2}}\tilde{\mu}'_K \end{pmatrix}, \tag{28}$$

The dimension of the super-vector is Kd where K is the number of mixture components and d is the dimension of low-level descriptors. The super-vector immediately implies the following simplification of the q-GMM kernel.

$$k(X', X'') = \exp\left(-\gamma \|\phi(X') - \phi(X'')\|_2^2\right). \tag{29}$$

4. Experiments

4.1. Experimental Conditions

In this section, the proposed method is evaluated on two data sets: PASCAL VOC 2010 and TRECVID 2010. The PASCAL Visual Object Classes Challenge (VOC) [8] provides a benchmark for comparison of object classification methods. The PASCAL VOC 2010 classification (validation) challenge data set consists of 4,998 training images and 5,105 testing images of 20 object classes in Table 1. The evaluation measure is Mean average precision (Mean AP), which is the arithmetic mean of APs over all targeted object classes. The TREC Video Retrieval Evaluation (TRECVID) [7] provides a benchmark for comparison of video indexing methods. The TRECVID 2010 Semantic Indexing data set consists of 119,685 training video shots and 146,788 testing video shots. 30 semantic concepts of objects, actions, and scenes in Table 1 and their ground truth labels are provided. Officially provided key-frame images for each video shot is used in our experiments. The evaluation measure is Mean AP among the 30 semantic concepts.

To test our q-GMM on these benchmarks, the low-level image descriptors are densely sampled from 100x100 grid with 3 different scales on an image. The descriptor is a concatenation of 128-dimension SIFT descriptor [9] and 36-dimension hue-histogram descriptor [20]. The dimension of each descriptor is reduced to 32 by applying Principal Component Analysis (PCA) after the concatenation.

The q-GMM for a background model is constructed by applying the proposed EM algorithm to one million randomly sampled descriptors. The number of mixture components K and the hyper-parameter τ in Eq. (26) are set to 512 and 20.0, respectively, in all experiments except the experiment evaluating their influence. Note that the dimension of the q-GMM histogram representation and q-GMM super-vector is 512 and 16,384 (=512×32), respectively.

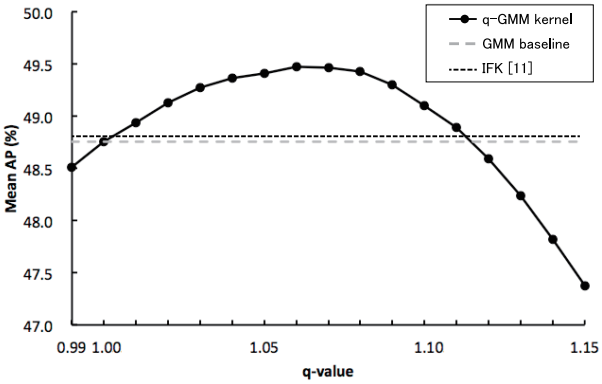

Figure 5. The performance comparison of q-GMM kernels with different q-values on the PASCAL VOC 2010 dataset. The q-GMM kernel outperforms the GMM baseline ($q = 1.00$) and the improved Fisher kernel [22] of GMM means.

4.2. Experimental Results

4.2.1. PASCAL VOC 2010

We first compared our q-GMM histogram representation in Sec.3.3 with the hard-assignment BoW [2]. Mean AP of the BoW was 30.93% and the q-GMM histogram improved it to 32.03%. The q value is set to 1.05 which has the hardness of 0.5 as presented in Sec. 3.3. This shows that the q-GMM is more suitable than the hard-assignment BoW for representing an image by a histogram of low-level descriptors.

Next, we compared our q-GMM kernel in Sec.3.4 with three kernel methods: χ^2-kernel, Fisher kernel (FK) [23], and Improved Fisher kernel (IFK) [22]. For the χ^2-kernel, we computed χ^2-distance between the q-GMM histogram representations. For FK and IFK, we extracted Fisher vectors for GMM means, whose dimension is the same as our super-vector. We applied L2 and power normalization as in [22] for IFK. The parameter of the power normalization was set to 0.4 which performed the best in our experiments. As shown in Table 2, the q-GMM kernel performed the best among these methods and achieved 49.42% in Mean AP. Figure 5 shows the performance of the q-GMM kernel for different q values. As can be seen, the q-GMM outperformed the normal GMM baseline ($q = 1.00$). The best q value and its Mean AP were 1.06 and 49.47%, respectively. This shows the effectiveness of the q-GMM kernel in image classification.

As described in [11], another idea to obtain a discriminative image representation is to train a class-specific model instead of the background model for a visual codebook. However, APs for aeroplane and bicycle were decreased by 1.30% and 2.09%, respectively, when a q-GMM is trained on one million descriptors sampled only from images of the targeted object. There was no significant performance improvement by concatenating both representations: AP for aeroplane was improved by 0.46% but that for bicycle was decreased by 0.31%. In conclusion, it is better to train a visual codebook on images of various object categories. A class-specific model could be useful for other problems such as dog breed classification that focus on a specific category.

Table 2. Performance comparison on PASCAL VOC 2010 dataset. BoW: bag-of-visual-words histogram representation [2] obtained by using vector quantization. Our histogram: q-GMM based histogram representation in Sec. 3.3 for $q = 1.00$ (GMM) and $q = 1.05$. χ^2 kernel: χ^2 kernel on q-GMM histogram representation. FK: Fisher kernel [23] of a GMM. IFK: improved Fisher kernel [22]. Our kernel: q-GMM kernel in Sec.3.4 for $q = 1.00$ (GMM) and $q = 1.05$.

Concept	BoW[2]	Our histogram		χ^2 kernel	FK [23]	IFK[22]	Our kernel	
		GMM	q-GMM				GMM	**q-GMM**
aeroplane	54.38	55.21	55.99	75.10	68.41	81.21	81.72	**82.34**
bicycle	36.53	37.88	38.20	42.52	46.79	52.07	52.62	**53.50**
bird	24.76	25.79	25.77	36.57	34.26	43.82	44.39	**45.07**
boat	37.41	38.29	39.65	50.51	50.37	**58.36**	56.51	57.32
bottle	9.71	10.03	10.28	16.36	18.06	20.51	21.91	**22.62**
bus	56.62	58.56	59.99	66.90	71.80	77.17	76.29	**77.50**
car	35.63	36.51	36.34	45.76	51.08	55.82	56.87	**57.24**
cat	41.75	42.33	42.05	49.62	52.09	56.96	57.18	**57.16**
chair	33.96	34.79	35.04	41.99	43.39	47.10	46.70	**47.76**
cow	7.12	7.69	7.66	12.24	13.48	20.35	21.77	**22.76**
diningtable	14.89	15.36	15.22	24.11	30.76	34.11	34.30	**34.65**
dog	24.36	24.84	24.63	35.30	39.41	44.30	44.61	**44.67**
horse	21.53	22.98	24.06	31.77	35.90	43.71	43.45	**43.81**
motorbike	27.30	28.55	28.19	43.41	49.64	56.27	57.17	**58.83**
person	68.63	69.23	69.29	73.42	73.16	76.69	76.84	**77.50**
pottedplant	8.54	8.38	8.04	10.85	14.31	15.78	16.82	**16.91**
sheep	19.68	20.42	19.37	30.07	32.64	**42.17**	41.63	42.09
sofa	16.17	16.66	17.10	24.09	26.97	33.38	33.47	**33.83**
train	44.59	45.71	46.17	51.33	57.97	62.93	63.17	**64.27**
tvmonitor	35.04	36.20	37.48	45.88	41.96	**48.79**	47.69	48.51
Mean AP	30.93	31.77	32.03	40.39	42.62	48.58	48.76	**49.42**

Table 3. Testing cost and Mean AP for each method. K is the number of mixture components, D is the dimension of low-level descriptor, and N is the averaged number of support vectors of an SVM.

Representation	Dimension	Kernel type	# support vectors N	Testing cost	Mean AP
q-GMM histogram	$K = 512$	linear	1540	$O(K)$	32.03
		χ^2	1700	$O(NK)$	40.39
q-GMM supervector	$DK = 16384$	linear	1251	$O(DK)$	47.09
		q-GMM kernel (RBF)	2261	$O(NDK)$	**49.42**
Fisher vector (mean)	$DK = 16384$	linear	1494	$O(DK)$	46.21
		RBF	2556	$O(NDK)$	48.58

4.2.2. Analysis

Here, we analyze the influence of the number of mixture components and the hyper-parameter τ in Eq.(26) on the performance. Figure 6 compares the q-GMM kernel with the GMM baseline for different numbers of mixture components. It is shown that the q-GMM constantly performed better than the GMM baseline. We observed that the difference between the GMM and the q-GMM is large when the number of mixture components is large. This observation can be explained as follows. For a GMM, as the number of mixture components increases, fewer low-level descriptors are assigned with a significant posterior probability c_{ik} to each Gaussian, i.e., a matrix of c_{ik} becomes sparser and it decreases the performance. On the other hand, for a q-GMM, long-tailed q-Gaussian distributions

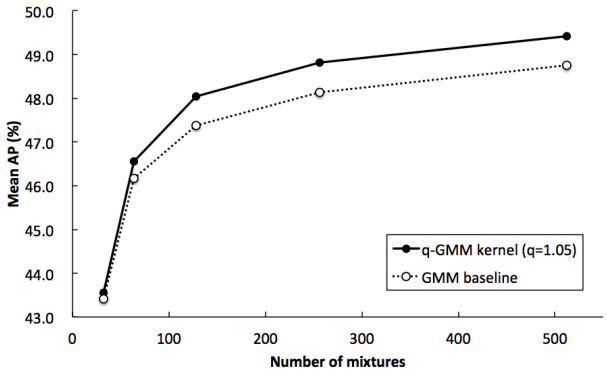

Figure 6. Mean AP on PASCAL VOC 2010 for different numbers of mixture components for q-GMM kernel.

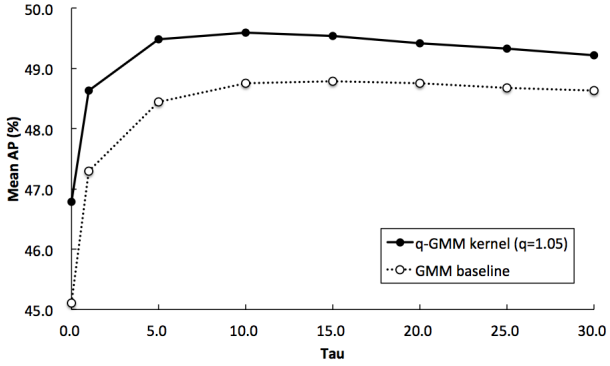

Figure 7. Mean AP on PASCAL VOC 2010 for different hyper-parameter τ in maximum a posteriori adaptation for q-GMM kernel.

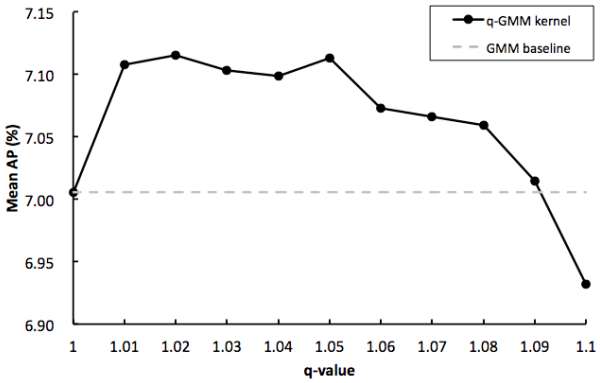

Figure 8. The performance comparison of q-GMM kernels with different q-values on the TRECVID 2010 dataset.

prevent it from becoming sparse. This is the reason why we observed the large difference for a large number of mixture components.

Figure 7 illustrates the influence of the hyper-parameter τ in Eq.(26) on the performance. We observed no significant change in performance when τ was between 10.0 to 25.0. Thus, we conclude that values between 10.0 to 25.0

Table 4. Average precision (AP) by semantic concepts on TRECVID 2010. Results for GMM, q-GMM (q = 1.05), score fusion of GMM and q-GMM (q = 1.05), and feature fusion of 5 types of visual and audio features for q-GMM are reported.

	Single feature			Feature fusion
Concept	GMM	q-GMM	GMM+q-GMM	q-GMM
Airplane Flying	2.75	2.67	**3.01**	15.64
Animal	2.18	**2.53**	2.39	6.44
Asian People	**0.45**	0.18	0.40	3.08
Bicycling	3.10	2.90	**3.31**	5.90
Boat Ship	5.28	**5.51**	5.24	11.01
Bus	**0.80**	0.43	0.57	1.42
Car Racing	**4.30**	3.92	4.03	4.37
Cheering	3.22	**3.56**	3.38	3.81
Cityscape	9.91	10.75	**10.85**	17.43
Classroom	1.24	**1.27**	1.21	0.81
Dancing	3.12	**5.10**	4.53	8.89
Dark-skinned People	12.85	13.21	**13.77**	20.40
Demonstration Or Protest	13.63	13.54	**14.14**	17.86
Doorway	7.84	7.14	**7.95**	12.44
Explosion Fire	**4.61**	4.03	4.28	3.93
Female-Human-Face-Closeup	10.94	10.55	**11.09**	17.79
Flowers	3.57	**3.95**	3.59	3.86
Ground Vehicles	14.59	14.15	**15.46**	20.20
Hand	**4.06**	4.05	3.92	9.30
Mountain	20.14	**20.83**	20.15	20.87
Nighttime	**12.89**	9.99	12.24	15.88
Old People	**2.58**	1.98	2.37	8.20
Running	1.42	1.87	**1.90**	6.88
Singing	6.80	8.11	**8.40**	17.47
Sitting Down	**0.12**	0.08	0.09	0.56
Swimming	32.98	33.07	**33.35**	30.82
Telephones	1.03	**1.39**	**1.39**	1.88
Throwing	3.49	**5.61**	5.45	7.02
Vehicle	**14.61**	14.04	14.28	18.91
Walking	5.66	**6.98**	6.41	13.03
Mean	7.01	7.11	**7.30**	10.87

are reasonable. We also confirmed that the improvement by the q-GMM is robust against the hyper-parameter τ.

4.2.3. TRECVID 2010

Table 4 shows the performance comparison of the q-GMM kernel and the GMM baseline on TRECVID 2010. The q-GMM kernel of q = 1.05 outperformed the GMM and achieved 7.11% in Mean AP. As shown in Figure 8, the q-value of 1.02 performed the best with 7.12% Mean AP on this dataset.

On the other hand, we observed that the performance is improved to 7.49% if we choose the best q-value for each of targeted semantic concept. This shows that supervised q-value optimization has potential for improving the overall performance in future work while our q-value optimization in Sec 3.3, which is based on the assignment hardness, was in an unsupervised way.

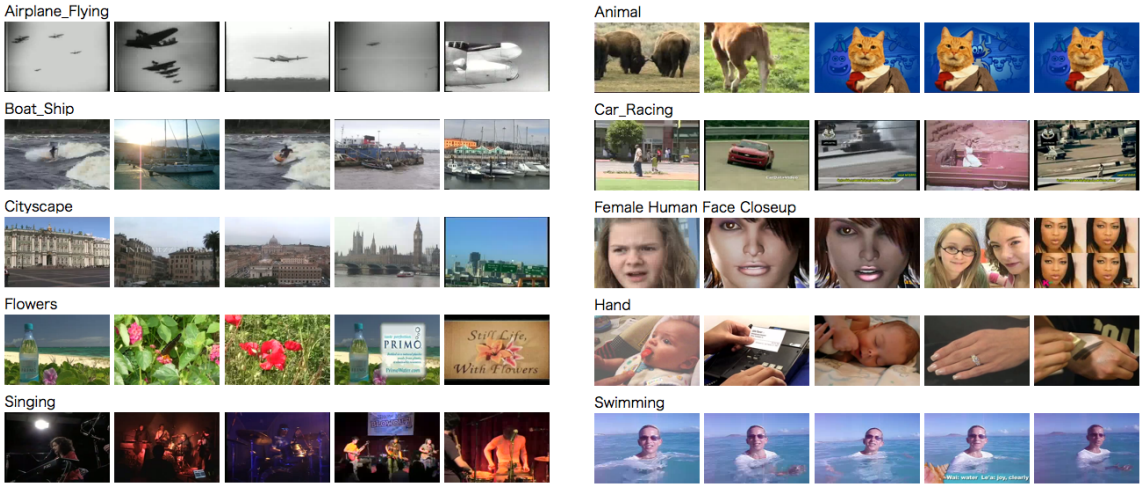

Figure 9. Examples of detected video shots in TRECVID 2010 dataset. Top 5 video shots are shown for ten semantic concepts.

Another idea to improve the overall performance is to use the both of the GMM and the q-GMM. The simplest implementation of this idea is averaging two detection scores obtained from the GMM and the q-GMM. We additionally evaluated it and observed 7.30% Mean AP as shown in Table 4. On the other hand, a disadvantage of averaging scores is that it's time-consuming to compute independent scores for the GMM and the q-GMM. To reduce the computational costs, an efficient method for calculating q-Gaussian probabilities for multiple q-values is needed in future work.

4.2.4. Comparison with other methods

Fig. 10 shows the performance comparison with the other methods in the TRECVID 2010 Semantic Indexing Task [7]. Mean AP of 7.11%, which was obtained by using our q-GMM kernel ($q = 1.05$), ranked 10-th among 87 official runs. Fig. 9 shows some examples of detected video shots. We conclude the q-GMM kernel performed well since the other methods typically used more than 5 types of low-level features while we used only one type of low-level features (SIFT with hue histogram).

Furthermore, we achieved Mean AP of 10.90%, which is better than the best performance on the TRECVID 2010, by combining q-GMM kernels for 4 additional types of low-level features: SIFT with Harris-affine detector, SIFT with Hessian-affine detector, dense HOG, and MFCC audio features.

5. Conclusion

We proposed q-Gaussian mixture models (q-GMMs) and their application to image and video semantic indexing systems. It has been shown in our experiments that the q-GMM kernels outperform both of the BoW method and the normal GMM. The q-GMM kernel achieved 0.494 and 0.109 in Mean Average Precision on the PASCAL VOC 2010 dataset and the TRECVID 2010 Semantic Indexing dataset, respectively. The linear kernel on q-GMM supervectors was shown to be effective in terms of the scalability. Our future work will focus on optimization of q-values for

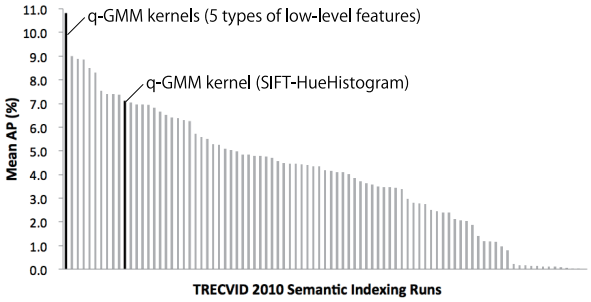

Figure 10. The performance comparison with other methods in TRECVID 2010. We achieved 0.071 in Mean AP by using a q-GMM kernel with SIFT-HueHistogram features and achieved 0.109 with additional 4 types of low-level features.

each semantic concepts. An extension of the Fisher information analysis to Tsallis statistics would be interesting as a promising next step.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 24650079 and 11J04223.

References

[1] N. Inoue, K. Shinoda. q-Gaussian Mixture Models Based on Non-Extensive Statistics for Image And Video Semantic Indexing *In Proc. of ACCV*, 2012.
[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *In Proc. of ECCV SLCV workshop*, pages 59–74, 2004.
[3] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32(9), pp. 1582–1596, 2010.
[4] X. Zhou, et al. Image classification using super-vector coding of local image descriptors. In Proc. of *ECCV*, pp. 141–154, 2010.
[5] N. Inoue, and K. Shinoda. A Fast MAP Adaptation Technique for GMM-supervector-based Video Semantic Indexing Systems. In Proc. of *ACM Multimedia*, pp. 1357–1360, 2011.
[6] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. In *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.
[7] A. F. Smeaton, et al. Evaluation campaigns and trecvid. In Proc. of *ACM Multimedia MIR workshop*, pp. 321–330, 2006.
[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/
[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *In IJCV*, vol. 60 (2), pp. 91–110, 2004.
[10] H. Bay, A. Ess, T. Tuytelaars, and L. Vangool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.
[11] F. Perronnin, C. Dance., G. Csurka, and M. Bressan. Adapted Vocabularies for Generic Visual Categorization. In Proc. of *ECCV*, pp. 464–475, 2006.
[12] C. Tsallis, Possible generalization of boltzmann-gibbs statistics, In *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.
[13] C. Tsallis, R. S. Mendes, A. R. Plastino, The role of constraints within generalized nonextensive statistics, In Physica A, vol. 261, no. 3, pp. 534-554, 1988.
[14] M. Havrda and F. Charvat. Quantification method of classification processes: concept of structural α-entropy, In *Kybernetika*, vol. 3, pp. 30–35, 1967.
[15] M. P. de Albuquerque, I. A. Esquef, and A. R. G. Mello, Image thresholding using Tsallis entropy. In Elsevier Pattern Recognition Letters, vol. 25, pp. 1059–1065, 2004.
[16] P. K. Sahoo, and G. Arora, Image thresholding using two-dimensional Tsallis-Havrda-Charvat entropy, In Elsevier Pattern Recognition Letters, vol. 27, issue. 6, pp. 520–528, 2006.
[17] Q. Lin, and C. Ou, Tsallis entropy and the long-range correlation in image thresholding, In Elsevier Signal Processing, vol. 92, pp. 2931–2939, 2012.
[18] Y. Li, X. Fan, and G. Li. Image segmentation based on Tsallis-entropy and Renyi-entropy and their comparison. In Proc. of *ICII*, pp. 943–948, 2006.
[19] R. Fabbrib, W. N. Goncalvesa, F. J. P. Lopesc, and O. M. Brunoa, Multi-q pattern analysis: A case study in image classification, In Elsevier Physica A: Statistical Mechanics and its Applications, vol. 391, issue. 19, pp. 4487–4496, 2012.
[20] J. van de Weijer and C. Schmid. Coloring local feature extraction. In Proc. of *ECCV*, pp. 334–348, 2006.
[21] T. Jaakkola, and D. Haussler, Exploiting Generative Models in Discriminative Classifiers In Proc. of *NIPS*, pp. 487–493, 1998.
[22] F. Perronnin, S. Jorge, and T. Mensink. Improving the fisher kernel for large-scale image classification. *In Proc. of ECCV*, pages 143–156, 2010.

[23] F. Perronnin, and C. Dance, Fisher kernels on visual vocabularies for image categorization. In Proc. of *CVPR*, pp. 1–8, 2007.
[24] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proc. of *CVPR*, pp. 2169–2178, 2006.
[25] J. C. V. Gemert, J.-m. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. *In Proc. of ECCV*, pages 696–709, 2008.
[26] T. Huang. Linear spatial pyramid matching using sparse coding for image classification. *In Proc. of CVPR*, pages 1794–1801, 2009.
[27] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *In Proc. of CVPR*, pages 3360–3367, 2010.
[28] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. *In Proc. of BMVC*, pages 1–12, 2011.
[29] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. *In Proc. of CVPR*, pages 3304–3311, 2010.
[30] Z.-J. Zha, X.-S. Hua, T. Mei, J Wang, G.-J Qi, and Z. Wang. Joint Multi-Label Multi-Instance Learning for Image Classification. In Proc. of *CVPR*, pp.1–8, 2008.
[31] Z.-J. Zha, T. Mei, J Wang, Z. Wang, and X.-S. Hua. Graph-Based Semi-Supervised Learning with Multiple Labels. In Elsevier JVCI, vol.20, issue.2, pp.97–103, 2009.
[32] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, T.-S. Chua. Interactive Video Indexing With Statistical Active Learning. In IEEE Trans. on Multimedia, vol.14, no.1, pp.17–27, 2012.
[33] S. Ayache and G. Quénot. Video Corpus Annotation Using Active Learning. In Proc. of ECIR, 2008.
[34] R. Datta, D. Joshi, J. Li, J. Wang. Image retrieval: Ideas, Influences, and Trends of the New Age. In ACM Computing Surveys, vol.40, no.2, pp.1–60, 2008.
[35] C. G. M. Snoek and M. Worring. Concept-based Video Retrieval. In Foundations and Trends in Information Retrieval, vol.2, no.4, pp.215–322, 2009.