Neurocomputing 172 (2016) 64–70

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-level feature representations for video semantic concept detection

Haojie Li [a,*], Lijuan Liu [a], Fuming Sun [b], Yu Bao [a], Chenxin Liu [a]

[a] Dalian University of Technology, China
[b] Liaoning University of Technology, China

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
26 July 2014
Accepted 23 September 2014
Available online 17 June 2015

Keywords:
Video retrieval
Concept detection
Semantic gap
Fisher Vector
LDA

ABSTRACT

Video semantic concept detection is a fundamental problem with many practical applications such as concept-based video retrieval. The major challenge of concept detection lies in the existence of the well-known semantic gap between the low-level visual features and the user's semantic interpretation of visual data. To bridge the semantic gap, in this paper we propose to promote low-level visual features to middle-level representations, expecting that the underlying latent semantic aspects of image data can be discovered, and such aspects can better model the semantic of images. Specifically, we employ latent Dirichlet allocation (LDA) approach to cluster the image data into semantic topics and the distributions of image low-level features on such topics are used as the middle-level feature vectors of images. Meanwhile, a recently developed more efficient probabilistic representation of low-level features, i.e., Fisher Vector is used to complement the LDA representation for video concept detection. The experimental results on the TRECVID 2013 Semantic Indexing dataset have demonstrated the effectiveness of the proposed approach.

1. Introduction

With the amount of multimedia collections increasing at tremendous speed, there is an urgent need to develop automatic video retrieval systems to efficiently consume such massive data. Traditional content-based retrieval methods cannot satisfy people's need and have shown various limitations [1]. In recent years, semantic concept-based video retrieval has attracted attentions of many researchers [2–5]. Here, semantic concept detection, which aims to detect the presence or absence of high-level concepts such as *bus*, *forest*, *sky* in video shots, is the key task. However, though lots of research efforts have been dedicated to the concept detection task, it is still a challenging problem within the multimedia communities [6].

Essentially, concept detection can be regarded as a classification task, in which a binary classifier is usually learned to predict the presence of a certain concept in a video shot or keyframe based on the extracted feature descriptors. Recently, the Bag-of-visual-words (BOV) [7–9], which transforms local image descriptors into fixed dimensional image representations, is the popularly used approach for concept detection. A baseline BOV approach employs k-means method to cluster the local descriptors, e.g., scale-invariant feature transformation (SIFT) [11] from a training dataset into a vocabulary of visual words and, transforms each sample into a histogram of these words by assigning each of its local descriptors to the nearest cluster. There are also many assigning methods for BOV such as the soft-weighting assigning scheme [13] or modeling the vocabulary using GMM [12], etc. However, the BOV method only considers the occurring number of local descriptors assigned to each visual word and lots of information of these descriptors has been lost.

Due to the high degree of the diversity in creator, content, style and production qualities of the video data, which is especially true for social web videos, we need to model the original video frames more comprehensively and naturally. Therefore, we generate the low-level features as dense-SFIT [14] instead of the popularly used sparse SIFT [15] and, in order to avoid the increasing of the dimension of the BOV representation, we propose to adopt Fisher Vector [16,17], which is measured by Fisher kernel, to represent the low-level features. As shown in [18,19] recently, as a generative probabilistic model, Fisher Vector can describe the video frames naturally. By assuming that the samples follow a parametric generative model which is estimated on a training set, Fisher Vector is represented as the gradient of the sample's likelihood with respect to the parameters of the distribution model. Fisher Vector representation has achieved state-of-the-art performance in many tasks [20,21]. In our work, we use Fisher

* Corresponding author.
E-mail addresses: hjli@dlut.edu.cn (H. Li), sunfm@mail.ustc.edu.cn (F. Sun).

http://dx.doi.org/10.1016/j.neucom.2014.09.096
0925-2312/

Vectors to represent the low-level features of video keyframes, expecting that these vectors can also gain effective information for concept detection.

One more important problem with current concept detection or image/video ananotation methods is that they usually used the low-level features to characterize video frames, which tends to lead to the so-called semantic gap, the hard bottleneck in this research [22,23]. Various approaches have been proposed to narrow the semantic gap. Early work on this problem focused concept-specific handcrafted decision rules which mapped restricted sets of low-level visual features to a single high-level concept [24]. [25] proposed a Bayesian probabilistic framework for mapping multimedia representations to high-level semantics, and recently Hinton et al. [26] proposed to use convolutional neural network to learn the features of images automatically based on GPU. Meanwhile, some work [27,28] leveraged the human efforts to boost the machine annotation efficiency while others [29,30] used the associated textural information to enhance the visual retrieval performance.

In this work, we propose to use a middle-level representation, specifically, latent Dirichlet allocation (LDA) [31], to encode low-level image features into higher-level abstractions to bridge the semantic gap in concept detection. In recent research, as a topic model inspired from text mining research, LDA has been widely used in vision processing in various ways [10,33]. LDA is a probabilistic generative model derived from low-level local features, such as SIFT; however, it could describe the co-occurrences between low-level features which may discover some higher-level implicit semantic over the underlying image data. Just like the specificity of an object might rely on the geometrical configuration of a limited number of visual patterns, in our approach we expect that LDA can capture the specificity of a particular concept through modeling the particular co-occurrences of a large number of visual components (e.g., SIFT). We regard the co-occurrences of visual components under a concept as the latent semantic contextual information, which has been learned in some other tasks with various methods [35]. It is noted in our work that, to reduce the computational cost in low-level feature processing, we use only one kind of image feature, i.e., we extract the dense SIFT of video keyframes and then generate two kinds of representations, Fisher Vector and LDA. The contributions of our work are two-fold. First, we propose to use middle-level representation to encode low-level features to narrow the semantic gap for concept detection and, the experimental results on TRECVID 2013 Semantic Indexing dataset have demonstrated its practicality and effectiveness. Second, by fusing the results of Fisher Vector and LDA approaches, we show that descriptors from different probabilistic models, even though from the same underlying low-level features, could capture the different aspects of image data, thus improve the final detection performance.

The rest of the paper is organized as follows. We first introduce the overall framework of our approach in Section 2 and then detail the approach in Section 3. The experiments are conducted in Section 4 and we conclude in Section 5.

2. The proposed framework

In this section, we present the proposed framework of our work. As shown in Fig. 1, we propose to use a topic model based middle-level representation which is inferred from low-level image features, to discover and characterize the higher-level implicit semantic information over the underlying image data. As a complement, we also introduce a recently developed more efficient probabilistic representation of low-level features referred as Fisher Vector to encode the extracted image features. We fuse these two level features generated from image data with different probabilistic models, and expect that they can describe the original frames complementarily and effectively.

As current web videos are mostly with huge variations in contents, we propose to use the densely sampled SIFT descriptors in the low-level feature extraction. While in quantization phrase, we model the original frame with generative probabilistic methods and we expect to generate the representation of frame naturally and effectively. In such phase, we quantize the descriptors from different views, in which the low-level encoding using Fisher Vector [16] estimates a parametric generative model on the training set and characterizes the distribution of the vectors with respect to the model and, the intermediate description vector which is generated over the low-level descriptors using LDA [31] models the associations between varying visual patterns under a given semantic concept. In the last phrase, we propose to train different classifiers for each representation and then fuse the predict results in an unsupervised way [32]. In our experiment, we choose support vector machine (SVM) to train models. For Fisher Vector, we prefer the linear SVM rather than kernel SVM due to its high dimensionality.

3. The approach

3.1. Low-level feature representation: Fisher Vector

The Fisher Vector [16] is a rich image representation, which extends the Bag-of-visual-words (BOV) model by encoding high-order statistics instead of the 0-order, i.e., the occurring frequency of visual words encoded by BOV. To represent samples from dataset, FV uses the gradient space of the parameters from log-likelihood model which describes the generation process of a particular example. The basic idea is to represent a set of data with the gradients of their log likelihood with respect to model parameters.

For a given generative probability model $P(X|\lambda)$, where $X=\left\{x_t|\, x_t\in R^D, t=1\ldots T\right\}$ is a dataset and λ is the vector of model parameters. X can be described as a gradient vector of its log-likelihood function with the given λ:

$$U_X=\nabla_\lambda \log\ P(X|\lambda) \tag{1}$$

U_X is referred as Fisher score which describes X as the steepest ascent direction in the log likelihood function and can be seen as a measurement of the direction to make λ fit better to X. Fisher kernel [16], in which Fisher score plays a important role, is usually used to measure the similarity of making different data values adapt to the existing model with a given set of parameters. The Fisher score maps the example into the feature space and the Fisher kernel can be referred to the latent product in this space, and is defined as

$$K(x_i,x_j)=U_{x_i}^T F_\lambda^{-1} U_{x_j}^T \tag{2}$$

where F_λ is the Fisher information matrix:

$$F_\lambda=E_{x\sim p(x|\lambda)}\left[(\nabla_\lambda \log\ p(x|\lambda))(\nabla_\lambda \log\ p(x|\lambda))^T\right] \tag{3}$$

As F_λ is symmetric and positive define, F_λ^{-1} can be decomposed as

$$F_\lambda^{-1}=L_\lambda^T L_\lambda \tag{4}$$

Then we can rewrite Fisher kernel as

$$K(x_i,x_j)=(L_\lambda U_{x_i})^T\left(L_\lambda U_{x_j}\right) \tag{5}$$

$\Gamma_\lambda=L_\lambda U_{x_i}$ is referred as Fisher Vector. Because of the high cost of computation of inverse, in [21] F_λ is regarded as an identity matrix and the Fisher score is used as an approximation. In fact, if Fisher scores are similar, similar adaptations to the parameters would be required.

Assuming that we have generated the local descriptors $\left\{x_t|\, x_t\in R^D, t=1\ldots T\right\}$ of frame I, we use Gaussian Mixture Model (GMM) [12] to