Neurocomputing 189 (2016) 53–65

Contents lists available at ScienceDirect

Neurocomputing

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

Event Bank based multimedia representation via latent group logistic regression minimization

Changyu Liu [a,d,e], Dapeng Li [b,*], Bin Lu [c], Juntao Xiong [a]

[a] College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China
[b] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[c] School of Computer Science, Wuyi University, Jiangmen 529020, China
[d] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[e] Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA

ARTICLE INFO

Article history:
Received 16 September 2014
Received in revised form
16 November 2015
Accepted 2 December 2015
Communicated by: Yang Tang
Available online 15 December 2015

Keywords:
Event Bank
Multimedia event detection (MED)
Latent group logistic regression (latent GLR)
Coordinate descent and gradient descent
$l_{2,1}$-norm regularized logistic loss function

ABSTRACT

In order to perform multimedia event detection (MED) tasks in uncontrolled videos, a very large number of labeled videos are required for training the event classifier, which would become quite challenging especially when there are lots of events. Because an event involves usually several spatial temporal objects, one intuitive solution is to model those objects from a large number of labeled images which can be obtained very easily from standard image datasets, such as the ImageNet challenge dataset, and to model their spatial temporal relationships from a relatively small number of labeled videos which can be also obtained very easily from standard video datasets, such as the TRECVID MED 2012 dataset. In this paper, we propose accordingly a latent group logistic regression (latent GLR) mixture model for those objects and an event bank descriptor for their spatial temporal relationships. Furthermore, we develop an efficient iterative training algorithm to learn model parameters of the individual latent GLR mixture model, which combines the coordinate descent approach and the gradient descent approach to minimize the $\ell_{2,1}$-norm or group regularized logistic loss function. We also conduct extensive experiments to evaluate the object detection performance by using the latent GLR mixture model on the ImageNet challenge dataset and the event detection performance by using the event bank descriptor on the TRECVID MED 2012 dataset. The results demonstrate the effectiveness of both proposed approaches.

1. Introduction

The research on multimedia analysis has been focusing more and more on video analysis than image analysis, due to rich semantic information contained in video clips. There are many research work performed recently in the video event detection domain, which can be roughly categorized into simple event detection and MED [1]. However, most of them are conducted to detect simple events. Duan et al. [2] proposed an interactive approach that discovers local attributes that are both discriminative and semantically meaningful from image datasets annotated only with fine-grained category labels and object bounding boxes. Wang et al. [3] designed a new motion feature based on motion relativity and visual relatedness for detecting simple events such as dancing. Vitaladevuni et al. [4] proposed a Comparison Hadamard random projection for improving the efficiency of locality sensitive hashing within Orthogonal Matching Pursuit. Xu et al. [5] studied detecting sport events like goal by using web-casting text and broadcast video. These researches have contributed to the video event detection in different ways. However, they are not suitable for complex event detection tasks that are performed on uncontrolled videos.

In order to encourage research on detecting more generic and complicated events, the NIST launched the MED task in the TRECVID competition since 2010. The MED task has attracted rapidly many well-known teams from academia and industry, such as the 2010 Columbia-UCF team [6], the 2015 CMU Informedia team [7], and the 2013 AT&T Research team [8]. According to the review paper [9], extracted features, recognition models [10] and evaluation strategies are three key components in high-level event recognition tasks. Furthermore, compared to the other two components, extracted features, which carry robust appearance and motion information, play a more critical role in the video analysis. Although many well-known teams have put a great deal of effort into improving the performance in the past five years, the research on MED is still in its infancy, mainly due to following issues:

[*] Corresponding author.
E-mail addresses: yezhich@gmail.com (C. Liu), dapengli@njupt.edu.cn (D. Li), lbscut@gmail.com (B. Lu), xiongjt2340@163.com (J. Xiong).

http://dx.doi.org/10.1016/j.neucom.2015.12.002
0925-2312/© 2015 Elsevier B.V. All rights reserved.

(1) Compare to simple events or actions, complex events are accompanied by more complex spatial temporal relationships, which cannot be easily captured by traditional approaches. (2) State-of-the-art event recognition methods usually requiring extensive computation, which cannot be deployed on large-scale datasets [11]. (3) In order to detect multiple events in uncontrolled videos, a substantial number of labeled videos are required for training event models. Although multimedia collections available to people, such as video clips provided by online portals like YouTube, are expanding with the rapid development of science and technology, gathering substantial labeled videos for MED is a quite challenging task especially when the number of events is very large. The major contributions of the paper lie in proposing a new representation approach for solving above first issue and third issue in MED from the first key component aspect.

It is very easy to obtain a large number of labeled images. For example, there are 1.2 million labeled images available in the ImageNet challenge dataset [12]. Whether could we adopt those labeled image as supplementaries of labeled videos in MED? Because an event involves usually several spatial temporal objects, one affirmative yet intuitive answer to the question is to detect those objects by models trained from a very large number of labeled images which can be obtained very easily from standard image datasets, such as the ImageNet challenge dataset, and to model their spatial temporal relationships from a relatively small number of labeled videos which can be also obtained very easily from standard video datasets, such as the TRECVID MED 2012 dataset [13]. Accordingly, we propose a latent group logistic regression (latent GLR) mixture model for detecting objects and an event bank descriptor for encoding their spatial temporal relationships in this paper. We also develop an efficient iterative training algorithm to learn model parameters of the individual latent GLR mixture model by using both the coordinate descent approach and the gradient descent approach.

Fig. 1 shows the overall feature extraction framework by using the Event Bank descriptor.

based on the latent GLR detector model.

As we can see from Fig. 1, there are four major steps to extract spatial temporal features from given input video clips by using the Event Bank descriptor, which are: (1) Train and test beforehand N (e.g., $N=1000$) latent GLR object detector models from standard image datasets (e.g., ImageNet challenge 2012 dataset), as shown in Section 5.3. (2) Preprocess the input video clips, i.e., resize video frames and select key frames, as shown in Section 4.5. (3) Extract features from preprocessed video clips by using the Event Bank descriptor. Specifically, get firstly 3D interest points by spatial temporal point detectors, such as Harris3D detector [14], Cuboid detector [15], and Hessian3D detector [16]. In the paper, we adopt the Harris3D interest point detector, as shown in Section 4.1. Then, get 2D feature pyramids by spatial descriptors, such as SURF descriptor [17], and HOG descriptor. In the paper, we adopt the latter, as shown in Section 4.2. After that, get the latent GLR score pyramid, as shown in Section 4.3. At last, generate the Event Bank feature vector, as shown in Section 4.4. (4) Perform k-means clustering on extracted Event Bank features of training video clips to generate a visual dictionary, then encode globally extracted Event Bank features of input video clips with a spatial bag of interest points tiling (SBIT) approach based on the dictionary, and output the encoded features as spatial temporal representations of events, as shown in Section 4.5.

Inspired by the object bank [18,19] which is a global scene descriptor based on latent SVM detectors and the action bank [20] which is a global action descriptor based on energy detectors, the event bank which is a local event descriptor based on latent GLR detectors adopts 1000 pre-trained latent GLR mixture models to represent multimedia events. Compared to traditional object detectors, latent GLR detectors show following benefits: (1) Subtle structural variations, e.g., differences between a running person and a walking person, are allowed for detection by using latent variable based optimization. (2) Outliers in training examples are removed by using $\ell_{2,1}$ regularized logistic regression function. (3) Feature vectors of event bank are jointly sparse by using $\ell_{2,1}$ regularization, which can't be achieved by using ℓ_1 regularization or ℓ_2 regularization. For training and testing event classifier, a sparse solution would be efficient.

While both the object bank [18,19] and the action bank [20] offer rich and high-level representations, a severe issue when we want to extend these two approaches to concatenate a large number of detectors for MED lies in the curse of dimensionality. Note that, for each object detector, the object bank [18,19] generates a 252-

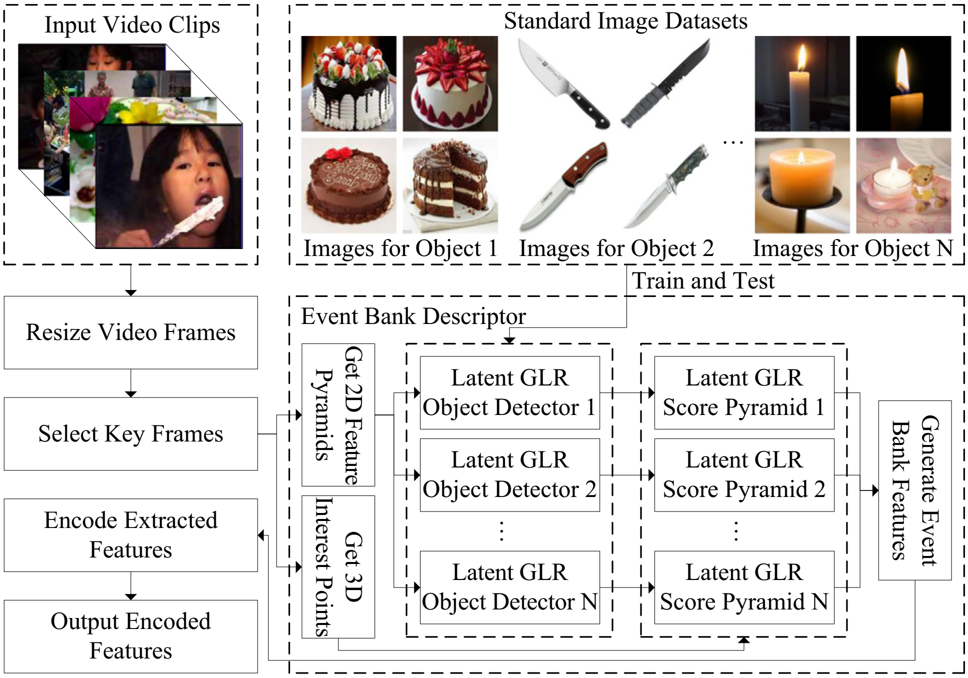

Fig. 1. The overall feature extraction framework by using the Event Bank descriptor based on the latent GLR detector model.