Neurocomputing 208 (2016) 378–386

Contents lists available at ScienceDirect

Neurocomputing

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

Specific video identification via joint learning of latent semantic concept, scene and temporal structure

Zhicheng Zhao [a,b,*], Yifan Song [a], Fei Su [a]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China
[b] School of Computer Science, Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Received 26 September 2015
Received in revised form
23 May 2016
Accepted 2 June 2016
Available online 11 June 2016

Keywords:
Specific video
CNN
VLAD
LSTM

ABSTRACT

In this paper, based on three typical characteristics of specific videos, i.e., the theme, scene and temporal structure, a novel data-driven identification architecture for the specific video is proposed. To be concrete, at the frame-level, semantic features and scene features from two independent Convolutional Neural Networks (CNNs) are extracted. At the video-level, Vector of Locally Aggregated Descriptors (VLAD) is firstly adopted to encode spatial representation, and then multiple-layer Long Short-Term Memory (LSTM) networks are introduced to represent temporal information. Additionally, a large-scale specific video dataset (SVD) is built for evaluation. The experimental results show that our method obtain impressive 98% mAP. Moreover, in order to validate generalization capability of proposed architecture, extensive experiments on two public datasets, Columbia Consumer Videos (CCV) and Unstructured Social Activity Attribute (USAA), are conducted. Comparison results indicate that our approach outperforms state-of-the-art methods on USAA, and achieves comparable results on CCV.

1. Introduction

The public security has become a main concern of the society, and many nations strengthened the cooperation in security field. However, there was a shortcut to propagate harmful ideas via social networks and video sharing websites, where vast specific videos were easily uploaded and distributed, and finally yielded the negative effects.

Therefore, preventing the spreading of such harmful videos as early as possible is a quite significant task. Some web portals tried to identify specific videos by manual inspection. However, this way is inefficient and unsustainable due to the huge amount videos. Content-based copy detection technique [1] was also adopted to block uploading of suspicious videos, while successful applications were still limited.

As a result, automatic specific video identification should be paid more attention. However, compared with other abundant video understanding works, there is few research to deal with this issue. It currently is an urgent and open question. Different from common videos, specific videos possess distinct characteristics in various aspects. Our method is based on following observations:

1. *Media mode*: Many specific videos were made and edited by stylized media making groups, who have unique editing mode. For instance, obvious logos and slogans regularly appear to remind spectators, and typical temporal structure was edited to differ from common web videos.
2. *Theme*: Although displaying different contents, specific videos could be roughly divided into several themes, and each category covers special semantic concepts and events.
3. *Scene*: The specific video shows discriminative geographical elements. Scenes in such videos are often shot at particular or confidential places: hilly land, ruins and ravages and so on.

Our previous work [2] used Fisher vector [3] to encode dense-SIFT features over video keyframes and then classified videos through detecting pre-defined 16 concepts at the frame-level. Although it achieved promising performance (92.9% of mAP) on a small dataset consisting of 653 specific videos and 4124 normal videos, it is difficult to generalize the method to large-scale datasets because it needs to manually label each keyframe.

In this paper, combining above three clues, we propose a novel framework for specific video identification by joint learning of latent semantic concepts, scenes and temporal structures. First, in order to get spatial representation of videos, based on a user-built specific video dataset (SVD), we train a CNN model to extract semantic features. Meanwhile, a Place-CNN model is introduced to extract scene features. Then, VLAD is applied to encode features, and classifiers are learnt via linear SVM. In temporal

* Corresponding author at: School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China.
E-mail addresses: zhichenz@cs.cmu.edu (Z. Zhao), danielsong1990@gmail.com (Y. Song), sufei@bupt.edu.cn (F. Su).

http://dx.doi.org/10.1016/j.neucom.2016.06.002
0925-2312/

representation, softmax classifiers following multilayer LSTM networks are learnt to realize the classification for long-term temporal features. In fusion scheme, a hierarchical score-based late fusion is adopted to generate the final result. Experiments show that the proposed method is more accurate and faster than that based on hand-crafted features. Furthermore, besides semantic feature, the introductions of scene CNN feature and LSTM temporal feature richen video representation and thus enhance the performance.

The contributions of this paper can be summarized as follows:

1. A data-driven architecture for specific video identification is developed to realize a practical application. Three discriminative clues: latent semantic concept, scene and temporal structure are learnt to aggregate the identification. To the best of our knowledge, this is the first trial to systemically address this problem.
2. A joint spatio-temporal representation is proposed to interpret the event and the activity of the specific video.
3. A large-scale Specific Video Dataset (SVD), which consists of 4429 specific videos and 14,191 normal videos and total length exceeds 2200 h is built for evaluation, and the experimental results demonstrate the effectiveness.
4. Extensive experiments on two cross-domain event datasets (CCV and USAA) are conducted, and promising results validate the generalization capability of the proposed method.

The remaining sections of the paper are organized as follows: in Section 2, we discuss relevant work. Section 3 describes proposed architecture. Section 4 demonstrates the experimental results on three datasets. Finally, we make some conclusions in Section 5.

2. Related work

A great number of literatures have been devoted to the field of video analysis. Most of them were concentrated on video retrieval, semantic concept detection and action recognition [4,5], while the studies on multimedia event and group activity videos were far from enough. Recently, TRECVID multimedia event detection (MED) task [6,7] has attracted growing attentions. Additionally, public video datasets such as Sports-1M [4], UCF-101 [8], CCA [9], USAA [10], MED datasets provided fundamental support for evaluation of event detection. As for specific video identification, however, few work exists and there is no an available dataset either.

Specific video identification could be regarded as a special MED task, which is characterized by representative scenes, actions, objects and interactions. Beyond initial the aggregation of low-level features with statistical models, recent trend of MED turns to explore more discriminative high-level semantic representation [11,12], usually based on the learning of deep features.

2.1. Spatial information representation

Encouraged by the recent results that CNN [13] achieved in image classification [14,15], and detection [16], researchers strived to explore more appropriate CNN descriptors. Successively, deeper and wider ImageNet-like models such as GoogleNet [17] and VGGNet [18] were put forward and boomed the progresses of MED, where fully connected layers (fc) features are usually extracted to describe the image content.

However, fc does not contain explicit spatial information, He et al. [19] proposed a Spatial Pyramid Pooling (SPP) to improve it. Xu et al. [20] extended this method by extracting latent concept descriptors from SPP layer. Compared with descriptors generated from fc, SPP descriptors obtained better results in TRECVID MEDTest [21,22].

In video representation, encoding frame-level features into a global video-level descriptor by using pooling and encoding methods such as Bag-of-Words (BoW), Vector of Locally Aggregated Descriptors (VLAD) [23] and Fisher vector has become common practices.

However, many experiments validated inherent shortages of video classification [4,5,24], if only spatial information was used, thus efforts of exploiting temporal representation were exerted.

2.2. Temporal information representation

A scheme is proposed to segment a video into clips and choose informative ones for pooling [25]. In addition, [5] built an optical flow CNN and achieved promising results in UCF-101 dataset. Karpathy et al. [4] used 3D-convolution over action clips to learn motion features. More recently, the Improved Dense Trajectories (IDTs) [26] were proposed and obtained outstanding results in MED under BoW representation [27].

Other than above methods that can only describe transient temporal information, Ng et al. [24] employed Long Short-Term Memory (LSTM) network [28] to mine long-range temporal relationships for video classification. This method obtained the state-of-the-art results in action datasets UCF-101 and Sports-1M.

In short, in multiple datasets, the deep features showed better interpretation for semantic representation and visualization [29] than traditional hand-crafted features such as HOG [30] and SIFT [31]. However, the generalization capability of them needs to be further validated in cross-domain datasets. Since current deep models mainly depend on elaborate trial and vast computing resource, and theoretical evidences are still insufficient. For specific task and application, a big difficulty is to extract effective spatio-temporal representation based on the deep model. For example, due to extreme complexity, despite of a lot of recent progresses [32–34], MED is still in the infancy.

As a particular MED, specific video identification also faces a slice of challenges such as the extraction of discriminative feature, video representation and classifier designing, etc. To address these problems, we propose a data-drive identification architecture.

3. Data-driven architecture

According to three observations summarized in Section 1, we propose a novel identification architecture which is shown in Fig. 1. At the frame-level, we first introduce two CNN models, i.e., semantic CNN model and scene CNN model, and then extract three kinds of CNN features (SPP, fc6 and fc7) from each model. At the video-level, firstly, we apply VLAD to encode each CNN feature to obtain spatial representation of the video. Secondly, three linear SVM classifiers are learnt to predict scores. Thirdly, in order to extract temporal descriptors for narrative structure of videos, we use semantic CNN's fc6 and fc7 features as the inputs of two LSTM models respectively. Subsequently, the output of each LSTM model is fed to a softmax classifier to classify. Finally, a hierarchical fusion scheme based on mean average precision (mAP) is applied to generate the final result.

In our practice, two CNNs follow the architecture of Krizhevsky et al. [14], and we add a SPP layer after layer 5 (conv5) to enrich spatial information respectively. Moreover, in order to accelerate the extraction of temporal structures, only fc6 and fc7 features of the semantic CNN are used to train LSTM models. The final result is generated by fusing the outputs from spatial and temporal streams.