Categorizing Big Video Data on the Web: Challenges and Opportunities

Yu-Gang Jiang
School of Computer Science
Fudan University
Shanghai, China
http://www.yugangjiang.info

***Abstract*—Video categorization is a very important problem with many applications like content search and organization, smart content-aware advertising, open-source intelligence analysis, etc. This paper discusses selected representative research progresses in categorizing big video data, with a focus on the user-generated videos on the Internet. We identify two major challenges in this vibrant field and envision promising directions that deserve in-depth future investigations. The discussions in this paper are brief but hopefully useful for quickly understanding the current progress and knowing where we should go in the next couple of years.**

I. Introduction

Capturing, sharing and viewing videos are a part of our everyday life. Techniques for classifying videos according to their high-level semantics are urgently needed in many applications. One important use-case is Web video search and organization. In addition, knowing a video's semantics will help select and place more suitable (and more profitable) advertisements on online video sharing sites.

In this short paper, we briefly discuss related techniques on video categorization. In particular, we focus more on user-generated videos (UGVs), which are normally short and contain a dominate category like "birthday party" and "kids playing blocks". This is different from professional videos like movies or news that contain mixed semantics and require temporal segmentation before content categorization. UGVs are important because of their dominate role in the Internet video-sharing activities. Compared with the professionally generated videos, most UGVs do not come with high-quality textual descriptions, which makes the task of automatic categorization especially important and desired in practical applications.

II. A Short Review

First, we briefly discuss the representative and recent techniques and benchmark datasets in the area of Web video categorization. For a more comprehensive review, please refer to [1].

A. Features

Successful video categorization systems rely heavily on prominent features, which are usually expected to be robust to withstand intra-class variations and discriminative to correctly recognize different categories. Videos are known to be naturally multimodal, consisting of both visual and acoustic channels. The visual channel not only depicts the appearance information of objects but also captures their movements over time, while the acoustic channel usually contains background sounds and conversations. The two channels are known to be highly complementary and should be jointly used in video categorization. Various features have been developed in recent years, covering static appearance features, motion features and acoustic features. In the following, we discuss several selected representative ones.

1) Static Appearance Features: Static appearance features can be extracted from video frames separately. Though without exploring temporal information, they are widely adopted for video analysis due to low computational cost and decent performance in many applications. Numerous existing image features can be computed, among which the well-known SIFT [2] has been the most popular one. Several other features are also widely used. For instance, color-SIFT, a variant of SIFT, has been frequently adopted to model color information. In addition, the Histogram of Orientated Gradients (HOG) [3] is also popular.

Recently, the off-the-shelf Convolutional Neural Network (CNN) based representations have also been utilized as static appearance features for video categorization. Jain *et al.* extracted frame-level features from a CNN model pre-trained on the ImageNet [4] data for action recognition and reported promising performance [5]. The appealing results clearly demonstrated that CNN features are powerful and should be considered for video categorization [6].

2) Motion Features: Different from the static appearance features, motion features incorporate the temporal information to model movements and the temporal evolution of video contents, which are of great value for understanding human actions and complex events.

A straightforward and natural way to obtain the motion features is to extend frame-based image features to the 3D spatial-temporal space. For example, Laptev *et al.* [7] extended the Harris corner patch detector to locate Space-Time Interest Points (STIP), which are space-time volumes where pixel values dramatically vary in both space and time. Instead of detecting interest points in the 3D space, Wang *et al.* [8] proposed the dense trajectory features by tracking densely sampled patches at different scales using optical flow fields to obtain trajectories, upon which four local descriptors are

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.17

computed to encode motion and appearance information. This feature has been popular and dominated all the popular benchmarks with outstanding performance. More recently, Simonyan *et al.* proposed to model the temporal information in videos with a CNN by substituting video frames with stacked optical flow images as inputs [9]. This approach has demonstrated competitive results compared with the dense trajectories.

3) Acoustic Features: The acoustic features can provide valuable and highly complementary information to the visual counterpart. Mel-frequency cepstral coefficients (MFCC), which represent the short-term power spectrum of an audio signal, have demonstrated top-notch performance in many applications like speech recognition. In the context of video classification, Jiang *et al.* encoded MFCC descriptors with the bag-of-words (BoW) representation as a complementation of visual features and achieved appealing performance on event detection [10]. Similar representations have been popularly used in recent works on video categorization.

B. Classifiers

Given the feature representations, video categorization becomes a typical classification problem, which can be achieved by various classifiers. Particularly, Support Vector Machines (SVMs) have been the most popular classifier due to simplicity and good generalizability. Linear SVMs can be efficiently trained but may suffer from poor performance when the data is not linearly separable. For non-linear kernels, the χ^2 kernel [11] and the intersection kernel [12] are fairly popular.

In contrast to the SVMs, there is a recent trend of applying deep learning to perform video classification. Wu *et al.* designed a multimodal deep neural network to explore inter-feature and inter-class relationships in videos [13]. Karparthy *et al.* proposed a multi-resolution CNN for end-to-end action recognition by stacking frames over time [14]. Simonyan *et al.* utilized a two-stream framework, in which two CNNs are trained on frames and optical flow images receptively to model spatial and temporal information.

C. Benchmark Datasets

1) Kodak Consumer Video Dataset: The Kodak consumer videos were recorded by around 100 customers of the Eastman Kodak Company [15]. The dataset consists of 1,358 video clips labeled with 25 concepts (including activities, scenes and single objects) as a part of the Kodak concept ontology [15].

2) MCG-WEBV: MCG-WEBV is a large set of YouTube videos collected by the Chinese Academy of Sciences [16]. There are 234,414 web videos with annotations on several topic-level events like"a conflict at Gaza", which are too complicated to be recognized relying merely on content analysis. This dataset is mostly adopted for video topic detection and ranking, by taking advantage of textual descriptions around videos.

3) Columbia Consumer Videos (CCV): The CCV dataset was constructed in 2011, aiming to stimulate the research on Internet consumer video analysis [17]. It contains 9,317 user generated videos from YouTube, which are annotated into 20 classes, including objects (e.g., "cat" and "dog"), scenes (e.g., "beach" and "playground"), sports events (e.g., "basketball" and "soccer") and social activities (e.g., "birthday" and "graduation").

4) TRECVID MED Dataset: Driven by the practical needs of analyzing high-level events in videos, the annual NIST TRECVID activity created a Multimedia Event Detection (MED) task since 2010. Each year a new or an extended dataset is constructed for worldwide system comparisons. In 2014, the MED development dataset contains 20 events, such as "birthday part","bike trick", etc. According to the NIST train/test split, there are around 5K videos for training and 23K videos for testing, totaling more than 1,200 hours.

5) UCF-101 & THUMOS-2014: The UCF-101 [18] dataset is a widely adopted benchmark for action recognition in videos, which consists of 13,320 video clips (27 hours in total). There are 101 annotated classes that can be divided into five types: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments and Sports. More recently, the THUMOS-2014 Action Recognition Challenge [19] extended upon the UCF-101 dataset, by adopting videos from the UCF-101 dataset for training. Additional Web videos were collected, including 2,500 background videos, 1,000 validation videos and 1,574 test videos.

6) Sports-1M Dataset: The Sports-1M [14] dataset was collected by Google researchers with a focus on sports videos, consisting of 1 million YouTube videos and 487 classes, such as "bowling", "cycling", "rafting", etc. The dataset is not manually labeled. The annotations were automatically produced by analyzing textual descriptions from Web users.

7) Fudan-Columbia Video Dataset (FCVID): Our recently released FCVID dataset[1] contains 91,223 Web Videos annotated into 239 classes, covering a wide range of topics organized in a hierarchy of 11 high-level groups, including "Art","Beauty & Fashion", "Cooking & Health", "DIY", "Education & Tech", "Everyday Life", "Leisure & Tricks", "Music", "Nature", "Sports" and "Travel". Different from Sports-1M, FCVID is manually annotated with reliable labels. Multiple annotators were involved to minimize subjectivity.

We further summarize and compare these datasets in Table I.

TABLE I
SEVERAL POPULAR BENCHMARK DATASETS FOR WEB/CONSUMER VIDEO CATEGORIZATION, SORTED BY THE YEAR OF CONSTRUCTION.

Dataset	# Videos	# Classes	Year	Manually Labeled
Kodak	1,358	25	2007	✓
MCG-WEBV	234,414	15	2009	✓
CCV	9,317	20	2011	✓
UCF-101	13,320	101	2012	✓
THUMOS-2014	18,394	101	2014	✓
MED-2014	≈28,000	20	2014	✓
Sports-1M	1M	487	2014	✗
FCVID	91,223	239	2015	✓

[1] http://bigvid.fudan.edu.cn/FCVID/

III. Challenge and Opportunity

Although significant progresses have been made in the past decade, the current video categorization techniques are far from satisfactory. The major challenge remains to be the semantic gap between the computable low-level features and the high-level semantic categories, which cannot be well bridged by the state-of-the-art approaches. On the aforementioned benchmarks CCV, UCF-101, and THUMOS-2014, the best reported results are 69.3% [13], 88.0% [9] and 70.8% [5], respectively. Results on Web-scale data are expected to be much worse because of the many more noises and confusions. This level of performance is obviously not sufficient under many practical scenarios.

With the growing popularity of the deep learning approaches, we believe that a big leap on the categorization performance may be achieved by ***designing new deep learning paradigms that are suitable for video analysis***. So far, deep learning has demonstrated very impressive results on many tasks including image annotation [20], [21], speech recognition [22] and text analysis [23]. However, for videos, we have not seen very strong performance reported using this approach.

The main reason is that videos have very unique spatial-temporal characteristics, which cannot be fully captured by the popular CNN architecture. The CNN can be deployed on sampled video frames but the important temporal information cannot be modeled. Currently, the best reported result on the UCF-101 benchmark is from the two-stream CNN approach mentioned earlier [9], which is the fusion of two traditional CNNs running on spatial and temporal streams separately. This approach only produced similar performance to the traditional hand-crafted dense trajectory features. In addition, existing works also showed that directly extending CNN to the 3D spatial-temporal domain does not work well too [24], [14]. Therefore, we envision that a new network architecture is needed for video analysis.

While designing a new deep neural network architecture is difficult but possible, training a new network is not an easy task. One critical problem is that we do not have sufficient training data in the video domain. As discussed earlier, currently the most comprehensive Web video categorization benchmark with manual annotations is probably the new FCVID dataset we recently collected. Compared with the image domain where the largest ImageNet dataset has over 14 million annotated images [4], the 91K FCVID videos are still too few. Therefore, ***creating a large and well-designed video database is a challenge but also an opportunity***, because—once we have sufficient training data—the performance of deep learning based video categorization may be greatly improved to a surprising level.

In summary, there are two directions that urgently demand in-depth future investigations, large-scale benchmark construction and deep learning approaches specially for video data analysis. Note that the former does not just require the tedious annotation efforts. It also demand a really smart design to ensure a good coverage, detectability by computer algorithms in the next several years, and suitability to practical application needs.

Finally, we would like to underline that the above discussions only reflect a partial view of this broad and complex topic. Beyond that, there could be many more interesting and important research problems in this vibrant field.

References

[1] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 2, pp. 73–101, 2013.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[5] M. Jain, J. van Gemert, and C. G. M. Snoek, "University of amsterdam at thumos challenge 2014," in *ECCV THUMOS Challenge 2014*, 2014.

[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *CoRR*, 2014.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[8] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[10] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," *NIST TRECVid Workshop*, 2010.

[11] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *ACM CIVR*, 2007.

[12] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *CVPR*, 2008.

[13] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *ACM Multimedia*, 2014.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[15] A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: concept definition and annotation," in *ACM MIR Workshop*, 2007.

[16] J. Cao, Y.-D. Zhang, Y.-C. Song, Z.-N. Chen, X. Zhang, and J.-T. Li, "MCG-WEBV: A benchmark dataset for web video analysis," *Technical Report, CAS Institute of Computing Technology*, 2009.

[17] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *ACM ICMR*, 2011.

[18] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 2012.

[19] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *CoRR*, 2014.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE TASLP*, 2012.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *NIPS*, 2013.

[24] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," in *ICML*, 2010.