Video Indexing, Search, Detection, and Description with Focus on TRECVID

George Awad
National Institute of Standards and Technology
Gaithersburg, Maryland 20899, USA
gawad@nist.gov

Duy-Dinh Le
University of Information Technology,
Vietnam National University HCMC
Ho Chi Minh City, Vietnam
duyld@uit.edu.vn

Chong-Wah Ngo
Department of Computer Science,
City University of Hong Kong
Hong Kong, China
cscwngo@cityu.edu.hk

Vinh-Tiep Nguyen
University of Science, Vietnam
National University HCMC
Ho Chi Minh City, Vietnam
nvtiep@fit.hcmus.edu.vn

Georges Quénot
Univ. Grenoble Alpes
CNRS, Grenoble INP, LIG
F-38000 Grenoble, France
Georges.Quenot@imag.fr

Cees Snoek
University of Amsterdam
Amsterdam, The Netherlands
cgmsnoek@uva.nl

Shin'ichi Satoh
National Institute of Informatics
Japan
satoh@nii.ac.jp

ABSTRACT

There has been a tremendous growth in video data the last decade. People are using mobile phones and tablets to take, share or watch videos more than ever before. Video cameras are around us almost everywhere in the public domain (e.g. stores, streets, public facilities, ...etc). Efficient and effective retrieval methods are critically needed in different applications. The goal of TRECVID is to encourage research in content-based video retrieval by providing large test collections, uniform scoring procedures, and a forum for organizations interested in comparing their results. In this tutorial, we present and discuss some of the most important and fundamental content-based video retrieval problems such as recognizing predefined visual concepts, searching in videos for complex ad-hoc user queries, searching by image/video examples in a video dataset to retrieve specific objects, persons, or locations, detecting events, and finally bridging the gap between vision and language by looking into how can systems automatically describe videos in a natural language. A review of the state of the art, current challenges, and future directions along with pointers to useful resources will be presented by different regular TRECVID participating teams. Each team will present one of the following tasks:

Semantic INdexing (SIN). The TRECVID SIN task [3] ran from 2010 to 2015 and evaluated methods and systems for automatic content-based video indexing. The task was defined as follows: given a test collection, a reference shot segmentation, and concept definitions, return for each target concept a list of at most 2000 shot IDs from the test collection ranked according to their likelihood of containing the target. This tutorial session will give an overview of the SIN task followed by the description of two main approaches, a "classical" one based on engineered features, classification and fusion, and a deep learning-based one [4]. A baseline implementation built by the LIG team and the IRIM group will be introduced and shared.

Zero-example (0Ex) Video Search (AVS). The TRECVID AVS task models the end user search use-case, who is looking for segments of video containing persons, objects, activities, locations, etc. and combinations of the former. Zero-example (0Ex) is basically text-to-video search, where queries are described in text and no visual example is given. Such search paradigm depends heavily on the scale and accuracy of concept classifiers in interpreting the semantic content of videos. The general idea is to annotate and index videos with concepts during offline processing, and then retrieve videos with relevant concepts matching query description [12, 13]. 0Ex video search started since the very beginning of TRECVid in year 2003, growing from around twenty concepts to currently more than ten thousands of classifiers. The queries also evolved from finding a specific thing (e.g., find shots of an airplane taking off) to detecting a complex and generic events (e.g., wedding shower) [18], while dataset size has expanded yearly from less than 200 hours to more than 5,000 hours of videos [17]. This tutorial session will give an overview of the AVS task [1] and 0Ex search paradigm, with topics in development of concept classifiers, indexing and feature pooling, query processing and concept selection, and video recounting. Interesting problems to be discussed include how to determine the number of concepts for query answering, and how to identify query-relevant fragments for feature pooling and video recounting. An overview of the methods used by AVS task participants in 2016 will be presented and a 0Ex baseline system, with a few thousands of concept classifiers (from SIN, ImageNet concept banks) and built on Multimedia Event Detection (MED) and AVS datasets, will be introduced and shared in public domain.

ICMR '17, June 6–9, 2017, Bucharest, Romania
© 2017 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-4701-3/17/06.
DOI: http://dx.doi.org/10.1145/3078971.3079044

Instance Search (INS). The TRECVID INS task [2] aims at exploring technologies that efficiently and effectively search and retrieve specific objects from videos by given visual examples. The task is especially focusing on finding "instances" of object, person, or location, unlike finding objects of specified classes as in the case of the SIN task or ad-hoc video search. This tutorial section will give an overview of the INS task followed by a standard pipeline including short list result generation by bag of visual word technique [20], handling of geometric information and context, efficiency management such as inverted index, and so on [11, 19]. A baseline implementation built by NII team will be introduced and shared.

Multimedia Event Detection (MED). This session will highlight recent research towards detection of events, like 'working on a woodworking project' and 'winning a race without a vehicle', when video examples to learn from are scarce or even completely absent. In the first part of the session we consider the scenario where in the order of ten to hundred examples are available. We provide an overview of supervised classification approaches to event detection, relying on shallow and deep feature encodings, as well as semantic encodings atop of convolutional neural networks predicting concepts and attributes [9]. As events become more and more specific, it is unrealistic to assume that ample examples to learn from will be commonly available [15, 16]. That is why we turn our attention to retrieval approaches in the second part. The key to event recognition when examples are absent is to have a lingual video representation. Once the video is represented in a textual form, standard retrieval metrics can be used. We cover video representation learning algorithms that emphasize on concepts, social tags or semantic embeddings [7, 10, 14]. We will detail how these representations allow for accurate event retrieval and are also able to translate and summarize events in video content, even in the absence of training examples.

Video to Text (VTT). This tutorial session considers the challenge of matching or generating a sentence to a video. The major challenge in video to text matching is that the query and the retrieval set instances belong to different domains, so they are not directly comparable. Videos are represented by audiovisual feature vectors which have a different intrinsic dimensionality, meaning, and distribution than the textual feature vectors used for the sentences. As a solution, many works aim to align the two feature spaces so they become comparable. We will discuss solutions based on low-level, mid-level and high-level alignment for video to text matching [5, 6, 8]. The goal of video to text generation is to automatically assign a caption to a video. We will cover state-of-the-art approaches relying on recurrent neural networks atop a deep convolutional network, and highlight recent innovations inside and outside the network architectures. Examples will be illustrated in the context of the new TRECVID video to text (VTT) pilot task.

CCS CONCEPTS

• **Information systems → Information retrieval**;

KEYWORDS

TRECVID, Semantic Indexing, Multimedia Event Detection, Video Search, Instance Search, Video Description

ACM Reference format:
George Awad, Duy-Dinh Le, Chong-Wah Ngo, Vinh-Tiep Nguyen, Georges Quénot, Cees Snoek, and Shin'ichi Satoh. 2017. Video Indexing, Search, Detection, and Description with Focus on TRECVID. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 2 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079044

REFERENCES

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, and Roeland Ordelman. 2016. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID*, Vol. 2016.

[2] George Awad, Wessel Kraaij, Paul Over, and ShinâĂŹichi Satoh. 2017. Instance search retrospective with focus on TRECVID. *International Journal of Multimedia Information Retrieval* 6, 1 (2017), 1–29.

[3] George Awad, Cees GM Snoek, Alan F Smeaton, and Georges Quénot. 2016. [Invited Paper] TRECVid Semantic Indexing of Video: A 6-Year Retrospective. *ITE Transactions on Media Technology and Applications* 4, 3 (2016), 187–208.

[4] Mateusz Budnik, Efrain-Leonardo Gutierrez-Gomez, Bahjat Safadi, Denis Pellerin, and Georges Quénot. 2016. Learned features versus engineered features for multimedia indexing. *Multimedia Tools and Applications* (2016), 1–18.

[5] Jianfeng Dong, Xirong Li, Weiyu Lan, Yujia Huo, and Cees GM Snoek. 2016. Early Embedding and Late Reranking for Video Captioning. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1082–1086.

[6] Jianfeng Dong, Xirong Li, and Cees GM Snoek. 2016. Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction. In *ArXive*.

[7] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2014. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 17–26.

[8] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2015. Discovering semantic vocabularies for cross-media retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 131–138.

[9] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2017. Video2vec Embeddings Recognize Events when Examples are Scarce. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[10] Amirhossein Habibian and Cees GM Snoek. 2014. Recommendations for recognizing video events by concept vocabularies. *Computer Vision and Image Understanding* 124 (2014), 110–122.

[11] Duy-Dinh Le, S. Phan, V. Nguyen, C. Zhu, D. M. Nguyen, T. D. Ngo, S. Kasamwattanarote, P. Sebastien, M. Tran, D. A. Duong, and Shin'ichi Satoh. 2014. National Institute of Informatics, Japan at TRECVID 2014. In *TRECVID*.

[12] Yi-Jie Lu, Phuong Anh Nguyen, Hao Zhang, and Chong-Wah Ngo. 2017. Concept-Based Interactive Search System. In *International Conference on Multimedia Modeling*. Springer, 463–468.

[13] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. 2016. Event detection with zero example: select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 127–134.

[14] Masoud Mazloom, Efstratios Gavves, and Cees GM Snoek. 2014. Conceptlets: Selective semantics for classifying video events. *IEEE Transactions on Multimedia* 16, 8 (2014), 2214–2228.

[15] Masoud Mazloom, Xirong Li, and Cees GM Snoek. 2016. Tagbook: A semantic video representation without supervision for event detection. *IEEE Transactions on Multimedia* 18, 7 (2016), 1378–1388.

[16] Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. 2016. The imagenet shuffle: Reorganized pre-training for video event detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 175–182.

[17] Xiao-Yong Wei, Yu-Gang Jiang, and Chong-Wah Ngo. 2011. Concept-driven multi-modality fusion for video search. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 1 (2011), 62–73.

[18] Hao Zhang, Yi-Jie Lu, Maaike de Boer, Frank ter Haar, Zhaofan Qiu, Klamer Schutte, Wessel Kraaij, and Chong-Wah Ngo. 2015. VIREO-TNO@ TRECVID 2015: multimedia event detection. In *Proc. of TRECVID*.

[19] Cai-Zhi Zhu, Hervé Jégou, and Shin Ichi Satoh. 2013. Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 1705–1712.

[20] Cai-Zhi Zhu and Shin'ichi Satoh. 2012. Large vocabulary quantization for searching instances from videos. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 52.