VideoAnalysis4ALL: An On-line Tool for the Automatic Fragmentation and Concept-based Annotation, and the Interactive Exploration of Videos

Chrysa Collyda
CERTH-ITI
6th Km Charilaou-Thermi Road
Thessaloniki, Greece
ckol@iti.gr

Evlampios Apostolidis
CERTH-ITI
6th Km Charilaou-Thermi Road
Thessaloniki, Greece
apostolid@iti.gr

Alexandros Pournaras
CERTH-ITI
6th Km Charilaou-Thermi Road
Thessaloniki, Greece
apournaras@iti.gr

Foteini Markatopoulou
CERTH-ITI
Thermi, Greece, 57001
Queen Mary University of London
markatopoulou@iti.gr

Vasileios Mezaris
CERTH-ITI
6th Km Charilaou-Thermi Road
Thessaloniki, Greece
bmezaris@iti.gr

Ioannis Patras
Queen Mary University of London
Mile end Campus, E14NS
London, United Kingdom
i.patras@qmul.ac.uk

ABSTRACT

This paper presents the VideoAnalysis4ALL tool that supports the automatic fragmentation and concept-based annotation of videos, and the exploration of the annotated video fragments through an interactive user interface. The developed web application decomposes the video into two different granularities, namely shots and scenes, and annotates each fragment by evaluating the existence of a number (several hundreds) of high-level visual concepts in the keyframes extracted from these fragments. Through the analysis the tool enables the identification and labeling of semantically coherent video fragments, while its user interfaces allow the discovery of these fragments with the help of human-interpretable concepts. The integrated state-of-the-art video analysis technologies perform very well and, by exploiting the processing capabilities of multi-thread / multi-core architectures, reduce the time required for analysis to approximately one third of the video's duration, thus making the analysis three times faster than real-time processing.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; **Video segmentation**; • **Human-centered computing** → **Web-based interaction**;

KEYWORDS

Web-based on-line video analysis, video segmentation, video annotation, video content exploration

ICMR '17, June 06-09, 2017, Bucharest, Romania

ACM ISBN 978-1-4503-4701-3/17/06…$15.00
https://doi.org/http://dx.doi.org/10.1145/3078971.3079015

ACM Reference format:
Chrysa Collyda, Evlampios Apostolidis, Alexandros Pournaras, Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2017. VideoAnalysis4ALL: An On-line Tool for the Automatic Fragmentation and Concept-based Annotation, and the Interactive Exploration of Videos. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 5 pages.
https://doi.org/http://dx.doi.org/10.1145/3078971.3079015

1 INTRODUCTION

Driven by the recent advances in video capturing and sharing technologies, in the last years there is a rapidly growing volume of video content distributed on the web through several channels, such as video-sharing platforms (e.g., YouTube and Vimeo), social networks (e.g., Facebook and Twitter) and on-line archives of content providers (e.g., broadcasters and news organizations). This trend, combined with the users' needs for finding and consuming the most appropriate and desirable content from vast amounts of information, highlights the necessity for making video content searchable and easily accessible, e.g., via some form of links in fragments of it, similar in principle to the hyperlinks between pieces of textual information. To this direction, semantically coherent fragments of a video must be identified and enriched with suitable human-interpretable annotations, that would make these pieces of video content searchable and linkable with related content.

The last years several technologies were introduced for the annotation of videos. Some of them, such as the ELAN [10], ANVIL [14] and EMXARaLDA [18] tools, offer an interactive environment for multi-layered, time-aligned video annotation with transcripts and annotations from various (pre-defined or customizable) categories. Other tools, such as the MobilVox [1], the Sloth [2] and the one in [17] support manual tagging of spatial and temporal fragments of a video with a set of multi-modal annotations (i.e., text, image, audio) in a way similar to the one applied by the YouTube [3] video

[1] https://www.youtube.com/watch?v=WB7M757Pw3I
[2] https://cvhci.anthropomatik.kit.edu/~baeuml/projects/a-universal-labeling-tool-for-computer-vision-sloth/
[3] https://support.google.com/youtube/answer/92710?topic=14354&hl=en

annotation framework. This spatiotemporal video tagging functionality is extended in other approaches, such as the BeaverDam [24] and the VideoJot [6], which enable a region-based, frame-by-frame annotation of videos through the demarcation and labeling of objects that appear within them with the help of bounding boxes or more arbitrary shapes. Furthermore, technologies for commenting and annotating (also streaming) videos through an interactive manual process were developed by the Universities of Harvard [4] and Minnesota [5]. The aforementioned manual solutions for video annotation are labour-intensive and time-demanding, so a set of semi-automatic approaches for reducing the video labeling workload were also introduced. Some of them, such as the Semantic Video Annotation Suite [22], automatically define the shots and the keyframes of the video and assist the manual shot-level annotation of the video by providing a customizable set of MPEG-7 annotations, while others, such as the Vatic [26] framework and the web-based tool of [3], enable the semi-automatic annotation of videos through an interactive user interface that enables the selection, labeling and tracking of specific areas of video frames.

In a slightly different direction, a number of technologies that support the exploration of video collections based on the semantic content of their videos have also been introduced. However, these frameworks require a prior analysis of the entire video collection in order to extract conceptual information about the videos, and they use this information for concept-based video retrieval through a video search engine. For example, the MediaMil system [25] used a lexicon of 100 automatically detected semantic concepts in the videos of the collection and offered a "query-by-concept" mechanism to facilitate users to access news video archives at a semantic level. Another interactive video search engine, VERGE, that extracts and exploits different types of visual information and is capable of retrieving and browsing video collections by integrating multimodal indexing and retrieval modules was presented in [19]. Alternatively, multi-modal approaches that combine different streams of the media content have been also proposed, such as the AXES-LITE video search engine [11], which integrates algorithms for text-based, visual-concept-based and visual-similarity-based retrieval of videos; and, the interactive system of [12], which represents the visual content of a video collection with the help of over 2500 high-quality pre-trained semantic concept detectors and applies text analysis on ASR and OCR data, allowing users to do multi-modal text-to-video and video-to-video search in large video collections. Many more interactive video search engines have been presented, e.g., [23], [13], [15] and [20].

The above overview indicates that most stand-alone existing video labeling tools require the involvement of the user in a labour-intensive and time-demanding video annotation process, while techniques for the automatic analysis and annotation of video have been integrated in prototype video search engines, but these usually do not give to the everyday user the possibility to analyze and annotate his/her own video content. Motivated by the lack of tools that an average user of the Web can employ for performing fine-grained video segmentation and labeling in a fully automatic way, we built an on-line, freely accessible web application that enables users to upload or submit videos of various genres and automatically perform: i) fragmentation of these videos into shots and scenes, ii) semantic annotation of the defined video fragments and iii) interactive exploration of their videos at a fine-grained level with the help of human-understandable visual concepts.

[4] http://annotation.chs.harvard.edu/video.php
[5] https://ant.umn.edu/

2 THE ON-LINE VIDEO ANALYSIS TOOL

The developed on-line tool integrates a set of video analysis technologies (reported in Section 3), and performs temporal fragmentation of a video and semantic annotation of the defined video fragments with the help of a vocabulary of visual concepts. The application allows the user to submit a video for analysis through the user interface depicted in Fig. 1. The submission can be done either via specifying the URL of an on-line available video, or by uploading a local copy of it from the user's machine. A variety of different video formats is supported, including mp4, webm, avi, mov, wmv, ogv, mpg, flv, and mkv. After fetching the video file, the tool decomposes the video into two different granularities, namely shots (i.e., the elementary structural parts of the video) and scenes (i.e., the story-telling parts of the video). Following, a few hundred visual concept detectors are evaluated for each keyframe extracted from the detected shots, and through this process the developed tool defines a shot-level concept-based annotation of the given video file. After submitting a video for analysis, the user can close the user interface and be notified by e-mail when the analysis results are ready; alternatively, he/she can keep the user interface open and monitor the progress of the analysis.

When the analysis is completed the results are presented to the user through the user interface presented in Fig. 2. With the help of this interactive environment the user is able to: i) explore the shot- and scene-level structure of the video, and select video fragments of these two different granularities (Fig. 2(b) shows the window that pops-up after clicking the "See all shots and scenes of the video" link); ii) see the concept-based annotation of each shot of the video (Fig. 2(a) depicts the top-10 concepts for the 49th shot of the video); iii) perform a concept-based search within the collection of detected shots by selecting a concept from the given list of concepts (Fig. 2(c) illustrates the retrieved video shots and the concept-based annotation of a selected one, after searching for the concept "Car"). As shown by the capabilities explained above and presented in Fig. 2, the developed tool automatically defines semantically annotated video fragments that are easily searchable and linkable with related content, with the help of a set of high-level visual concepts. Last but not least, the video files submitted for analysis and the corresponding analysis results are available for inspection via the user interface for approximately 48 to 72 hours after their analysis is completed; after this time period they are automatically deleted from our server.

3 VIDEO ANALYSIS TECHNOLOGIES AND USER INTERFACE

This section gives insights about the video analysis methods integrated in the tool, namely the algorithms for shot segmentation (Section 3.1), scene segmentation (Section 3.2) and concept detection (Section 3.3), and reports on the technologies utilized for building its interactive user interfaces (Section 3.4).

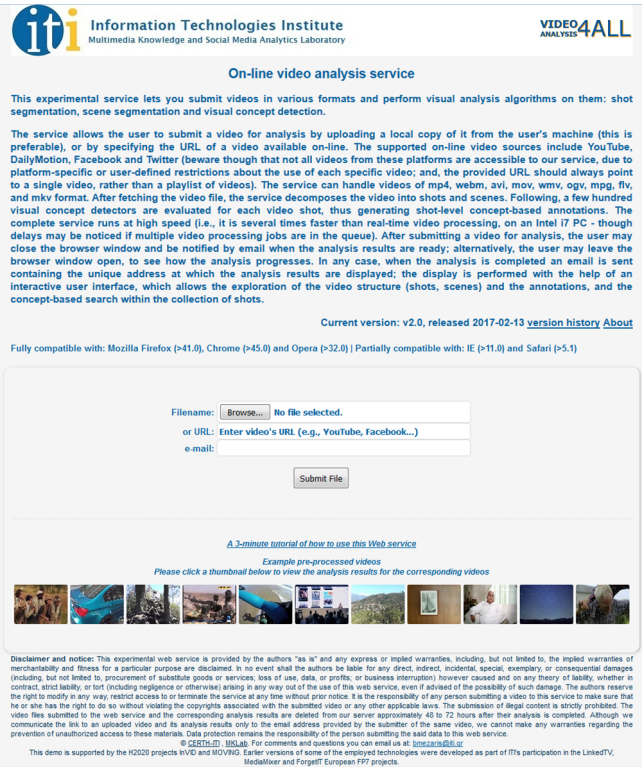


Figure 1: The user interface that allows to submit a video for analysis and to monitor the progress of the process.

3.1 Video shot segmentation

The video is temporally segmented into shots, i.e., sequences of frames captured uninterruptedly by a single camera, based on a variation of the algorithm in [1]. The employed method defines the boundaries of each shot by detecting the abrupt and gradual shot transitions. The latter is performed by evaluating the visual resemblance between consecutive and neighboring frames of the video with the help of local (ORB [5]) and global (HSV histograms) descriptors. Following, the boundaries of each shot of the video are determined through the comparison of the computed similarity scores and patterns against experimentally pre-specified thresholds and models that indicate the existence of abrupt and gradual shot transition. The resulting set of transitions is re-assessed with the help of a flash detector that eliminates falsely identified abrupt transitions due to short-term camera flashes, and a pair of dissolve and wipe detectors (based on [4] and [2] respectively) that remove erroneously detected gradual transitions due to swift camera and/or object movement. The union of the resulting sets of detected abrupt and gradual transitions forms the output of the applied technique. Finally, three representative keyframes are extracted from each shot of the video through a simple frame-sampling strategy that selects three uniformly distributed frames of the shot, and provided for further processing by the scene segmentation and concept detection algorithms of the tool.

3.2 Video scene segmentation

Building upon the outcomes of the shot segmentation analysis, the integrated scene segmentation algorithm from [9] specifies the story-telling parts of the video by grouping shots into sets that correspond to individual scenes of the video, i.e., semantically and temporally coherent segments that cover either a single event or several related events that take place in parallel. This grouping is performed by evaluating the content similarity and the temporal consistency among the shots of the video. Content similarity in the utilized method is expressed by assessing the visual similarity among the keyframes of different shots of the video through the extraction and matching of HSV histograms. Visual similarity and temporal consistency then are taken into account during the shot grouping into scenes that is performed with the help of two extensions of the Scene Transition Graph (STG) algorithm [27]. The first one decreases the computational load of STG-based shot grouping by taking into account shot linking transitivity and by exploiting the fact that scenes are by definition convex sets of shots, while the second extension builds on the former and constructs a probabilistic framework that eliminates the need for manual selection of the STG parameters. Based on these extensions the applied technique can identify the scene-level structure of videos belonging to different genres, and provide results that match well the human expectations.

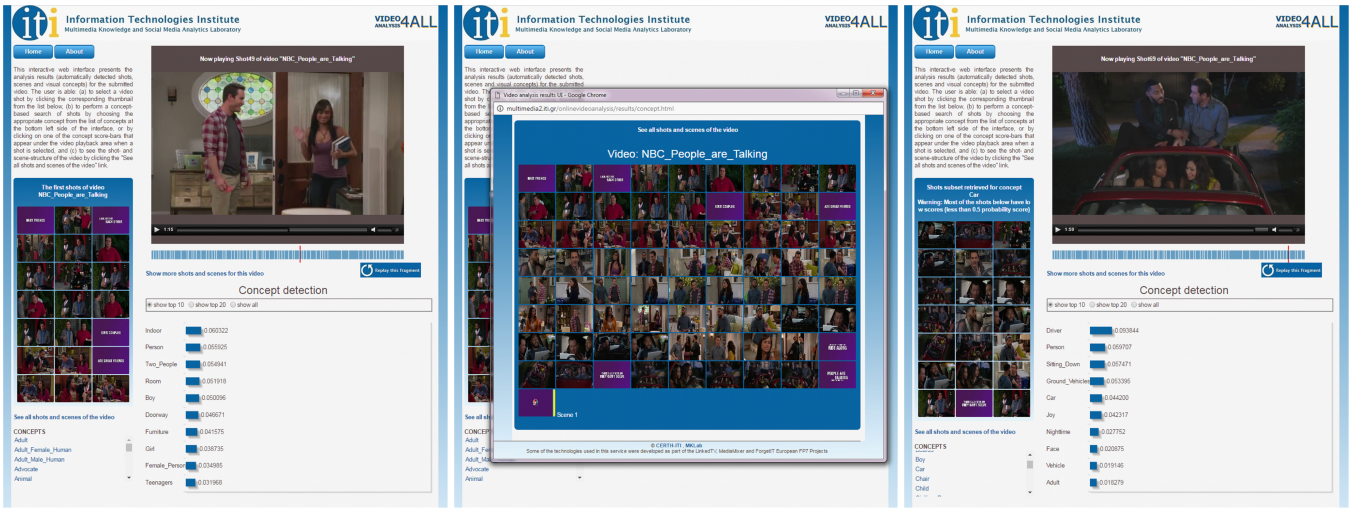
3.3 Video concept detection

The integrated concept detection algorithm annotates each shot of the video by evaluating the existence of a set (several hundreds) of visual concepts in the middle keyframe of the shot. Video concept detection is performed using a modification of the deep multi-task learning algorithm (DMTL_LC) presented in [16]. DMTL_LC combines multi-task learning with deep learning and also constraints the network's concept-related parameters by considering the concept correlations between pairs of concepts. For the developed tool, a pre-trained ImageNet [8] deep network was fine-tuned using the DMTL_LC method on 345 TRECVID SIN concepts [7]. During the analysis, each video keyframe is forward propagated by the fine-tuned network. Then, the output of the network is refined by employing the re-ranking method proposed in [21] and finally, the refined scores are used to annotate the given keyframe and the corresponding shot of the video. Based on these computed shot-level annotations, a scene-level labeling is also automatically defined for each detected scene of the video, by max pooling the scores of the detected concepts in the shots that compose each individual scene.

3.4 UI implementation

The interactive user interface of the tool follows the HTML5 standard and integrates technologies of the JQuery JavaScript library [6] (version 1.9.1). For the presentation of the analysis results the web interface utilizes the JqPlot [7] plotting and charting plugin of the JQuery JavaScript framework, while the selection of the video fragment that is shown in the player is based on the use of media fragment URI references with the new HTML5 video tag. The user interface is fully compatible with Mozilla Firefox (version 41.0 or

[6] https://jquery.com/
[7] http://www.jqplot.com/



(a) The concept-based annotation of a selected shot. (b) The shot- and scene-level structure of the video. (c) The search results for the concept "Car".

Figure 2: The user interface for the presentation and management of the analysis results.

newer), Chrome (version 45.0 or newer) and Opera (version 32.0 or newer), while it is partially compatible with Internet Explorer (version 11.0 or newer) and Safari (version 5.1 or newer). Please note that the developed user interface for the presentation of the analysis results is also fully operable with the latter two browsers, however the playback functionality of the video player might not be optimal in some cases due to slight differences between the way that each browser handles these fragments (i.e., a few frames can be added in the beginning or at the end of a shot in some cases).

4 PERFORMANCE AND TESTING

The back-end services of the demo run on a PC with an Intel i7-4770K at 3.50 GHz, 16GB of RAM and a 384-core NVIDIA GeForce GTX 650 graphics card. By exploiting the multi-thread and multi-core processing capabilities of the available CPU and GPU, the analysis is faster than real-time video processing (where real-time processing would have a processing time equal to the video's duration); though, delays may be noticed if multiple analysis requests are sent to service, since the latter applies a queuing strategy on the incoming analysis requests and the analysis is performed in a one-by-one basis (and not in parallel). In particular the shot and scene segmentation is performed 4 to 6 times faster than real-time processing, depending on the resolution of the given video. The time required for concept detection is related to the number of detected shots (since this type of analysis is performed on a per shot/keyframe basis, as described above). Based on the fact that the service needs approximately 0.15 sec. per keyframe and according to a set of evaluations that included several different types of videos (e.g., news videos, documentaries, sitcoms, talk shows), we can state that the entire video fragmentation and annotation analysis is about three times faster than real-time processing (depending again on the number of the detected shots). We should mention that a bit of extra time is needed for fetching the video file in the service (i.e., for video transfer) and for transcoding it after the analysis, so that it can be displayable by the video player in different browser-player configurations. Our on-line tool for video fragmentation and annotation can be accessed and tested at http://multimedia2.iti.gr/onlinevideoanalysis/service/start.html.

5 CONCLUSIONS

This paper demonstrated the developed on-line tool for automatic video fragmentation and concept-based annotation. Details about the use and functionalities of the tool were given with the help of indicative snapshots of the implemented user interfaces. The integrated methods for video analysis and the employed technologies for building the tool were presented, and information about the performance of the developed technologies was given. The demo will show that our on-line tool is a fully automatic tool for video fragmentation and annotation, and for the creation of semantically annotated video fragments that are searchable using a set of human-interpretable concept labels.

ACKNOWLEDGMENTS

This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-687786 InVID, H2020-693092 MOVING and H2020-732665 EMMA.

REFERENCES

[1] E. Apostolidis and V. Mezaris. 2014. Fast shot segmentation combining global and local visual descriptors. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6583–6587.
[2] K. d. Seo, S. J. Park, and S. h. Jung. 2009. Wipe scene-change detector based on visual rhythm spectrum. *IEEE Transactions on Consumer Electronics* 55, 2 (May 2009), 831–838.
[3] A. Ioannidou, E. Apostolidis, C. Collyda, et al. 2017. A web-based tool for fast instance-level labeling of videos and the creation of spatiotemporal media fragments. *Multimedia Tools and Applications* 76, 2 (2017), 1735–1774.
[4] C.-W. Su, H.-R. Tyan, H.-Y. Mark Liao, et al. 2002. A motion-tolerant dissolve detection algorithm. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, Vol. 2. 225–228 vol.2.

[5] E. Rublee, V. Rabaud, K. Konolige, et al. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*. 2564–2571.

[6] M. Riegler, M. Lux, V. Charvillat, et al. 2014. VideoJot: A Multifunctional Video Annotation Tool. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR '14)*. ACM, New York, NY, USA, Article 534, 534:534–534:537 pages.

[7] O. Paul, A. George, M. Martial, et al. 2015. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *TRECVID 2015 Workshop*. NIST, USA.

[8] O. Russakovsky, J. Deng, H. Su, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.

[9] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, et al. 2011. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 8 (Aug 2011), 1163–1177.

[10] P. Wittenburg, H. Brugman, A. Russel, et al. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.

[11] S. Chen, K. McGuinness, R. Aly, et al. 2012. The AXES-lite video search engine. In *2012 13th International Workshop on Image Analysis for Multimedia Interactive Services*. 1–4. https://doi.org/10.1109/WIAMIS.2012.6226778

[12] S. Xu, H. Li, X. Chang, et al. 2015. Incremental Multimodal Query Construction for Video Search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR '15)*. ACM, New York, NY, USA, 675–678. https://doi.org/10.1145/2671188.2749413

[13] W. Hürst, A. Ip Vai Ching, M. A. Hudelist, et al. 2016. A New Tool for Collaborative Video Search via Content-based Retrieval and Visual Inspection. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, New York, NY, USA, 731–732. https://doi.org/10.1145/2964284.2973824

[14] M. Kipp. 2012. ANVIL: A Universal Video Research Tool. In *Handbook of Corpus Phonology*, J. Durand, U. Gut, and G. Kristofferson (Eds.). Oxford Univ. Press.

[15] Y.-J. Lu, P. A. Nguyen, H. Zhang, and et al. 2017. *Concept-Based Interactive Search System*. Springer International Publishing, Cham, 463–468. https://doi.org/10.1007/978-3-319-51814-5_42

[16] F. Markatopoulou, V. Mezaris, and I. Patras. 2016. Deep Multi-task Learning with Label Correlation Constraint for Video Concept Detection. In *Proceedings of the 2016 ACM on Multimedia Conference (MM '16)*. ACM, NY, 501–505.

[17] M. Martin, J. Charlton, and A. M. Connor. 2016. Mainstreaming video annotation software for critical video analysis. *CoRR* abs/1604.05799 (2016).

[18] C. Meissner and A. Slavcheva. 2013. Review of EXMARaLDA. *Language Documentation and Conservation* 7 (2013), 31–40.

[19] A. Moumtzidou, T. Mironidis, F. Markatopoulou, and et al. 2017. *VERGE in VBS 2017*. Springer International Publishing, Cham, 486–492. https://doi.org/10.1007/978-3-319-51814-5_46

[20] T. D. Ngo, V.-T. Nguyen, V. H. Nguyen, and et al. 2015. *NII-UIT Browser: A Multimodal Video Search System*. Springer International Publishing, Cham, 278–281. https://doi.org/10.1007/978-3-319-14442-9_28

[21] B. Safadi and G. Quénot. 2011. Re-ranking by Local Re-scoring for Video Indexing and Retrieval. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 2081–2084.

[22] P. Schallauer, O. Sandra, and H. Neuschmied. 2008. Efficient semantic video annotation by object and shot re-detection. In *3rd International Conference on Semantic and Digital Media Technologies*.

[23] K. Schoeffmann, M. J. Primus, B. Muenzer, and et al. 2017. *Collaborative Feature Maps for Interactive Video Search*. Springer International Publishing, Cham, 457–462. https://doi.org/10.1007/978-3-319-51814-5_41

[24] A. Shen. 2016. *BeaverDam: Video Annotation Tool for Computer Vision Training Labels*. Master's thesis. EECS Department, University of California, Berkeley.

[25] C. G. M. Snoek, M. Worring, J. van Gemert, et al. 2005. MediaMill: Exploring News Video Archives Based on Learned Semantics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*. ACM, New York, NY, USA, 225–226. https://doi.org/10.1145/1101149.1101188

[26] C. Vondrick, D. Patterson, and D. Ramanan. 2013. Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision* (2013), 184–204.

[27] M. Yeung, B.-L. Yeo, and B. Liu. 1998. Segmentation of Video by Clustering and Graph Analysis. *Comp. Vision and Image Underst.* 71, 1 (1998), 94 – 109.