V3C1 Dataset: An Evaluation of Content Characteristics

Fabian Berns
University of Münster
fabian.berns@uni-muenster.de

Luca Rossetto
University of Basel
luca.rossetto@unibas.ch

Klaus Schoeffmann
University of Klagenfurt
klaus.schoeffmann@aau.at

Christian Beecks
University of Münster
christian.beecks@uni-muenster.de

George Awad
NIST; Dakota Consulting, Inc
george.awad@nist.gov

ABSTRACT

In this work we analyze content statistics of the V3C1 dataset, which is the first partition of the *Vimeo Creative Commons Collection* (V3C). The dataset has been designed to represent true web videos in the wild, with good visual quality and diverse content characteristics, and will serve as evaluation basis for the Video Browser Showdown 2019-2021 and TREC Video Retrieval (TRECVID) Ad-Hoc Video Search tasks 2019-2021. The dataset comes with a shot segmentation (around 1 million shots) for which we analyze content specifics and statistics. Our research shows that the content of V3C1 is very diverse, has no predominant characteristics and provides a low self-similarity. Thus it is very well suited for video retrieval evaluations as well as for participants of TRECVID AVS or the VBS.

KEYWORDS

V3C, video collection, video analytics, content statistics, TRECVID

ACM Reference Format:
Fabian Berns, Luca Rossetto, Klaus Schoeffmann, Christian Beecks, and George Awad. 2019. V3C1 Dataset: An Evaluation of Content Characteristics. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3323873.3325051

1 INTRODUCTION

The *Vimeo Creative Commons Collection* (V3C) [24] is a large-scale video dataset that has been collected from high-quality web videos with a time span over several years in order to represent true videos in the wild. It consists of 28 450 videos with a duration of 3 801 h in total. The first part of this dataset (V3C1) has been used by the Video Browser Showdown (VBS) 2019 [17] already and will be used for the Ad-Hoc Video Search (AVS) task at TRECVID 2019 as well [5]. For both campaigns V3C1 will serve as a basis over three years (VBS 2019-2021 and TRECVID 2019-2021) before it is planned to be extended with further parts of the V3C dataset.

V3C1 contains 1 000 h of video content and approximately one million shots that were created by the authors of the dataset using the open-source multimedia retrieval engine Cineast [22]. In this paper we perform a thorough analysis of content characteristics of the V3C1 dataset in order to provide a basis for future users of the dataset (e.g., participants or organizers of evaluation campaigns or studies). Such an investigation is important as it reveals content statistics and characteristics that enable better assessment of relevance and impact in terms of evaluation results. For example, [23] has shown that many public video datasets have quite specific content characteristics in several aspects (e.g., duration, resolution, genre, content classes, upload date etc.), which do significantly differ from videos in the wild and therefore limit the generalizability of evaluation results to the real world. Also, a comprehensive analysis of content characteristics is crucial for assessing the appropriateness of a dataset for a given task and comparing it to other datasets. It provides the basic source of information when defining queries (e.g., for TRECVID or VBS) and performing evaluations. For example, only when we know that a concept is not predominant in the dataset, it makes sense to create an AVS query related to this concept. Similarly, only if we know the number of shots showing three faces, for example, we can assess the recall of retrieval tools or classifiers related to faces. Therefore, in this paper we highlight such content semantics and provide related statistics and assess the content diversity of the V3C1 dataset.

2 RELATED WORK

Benchmarking video datasets has become prominent in recent years in order to provide a solid ground truth for further research based on those datasets [13, 21]. While V3C1 will be used during TRECVID 2019, *Internet Archive videos with Creative Commons* (IACC.1-3) datasets [19] were used the years before. IACC.1 contains only videos with a constant bit rate of 512 kb/s and captures around 64 000 video files gathered between 1996 and 2009 via the platform Internet Archive[1]. Those videos have an average duration of 12.720 min and a maximum duration of about 500 min. There are means of categorizing those videos available for IACC.1, but as it is pointed out in [19] there is no overall ontology and for a lot of videos no or incomplete categorization available. This is also true for IACC.2 and IACC.3.

The *Yahoo Flickr Creative Commons 100 Million* (YFCC100M) [29] is a dataset sourced from media sharing platform Flickr[2] between 2004 and 2014 and encompasses a collection of 800 000 videos, besides a much broader collection of images. While only 52 % of those videos have user generated annotations, it is stated in [29] that they added information about visually detected concepts to the videos metadata. Due to former Flickr upload restrictions most of the videos in the YFCC100M dataset have a duration of 90 s or

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.
ICMR '19, June 10–13, 2019, Ottawa, ON, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325051

[1] URL: https://archive.org/

[2] URL: https://www.flickr.com/

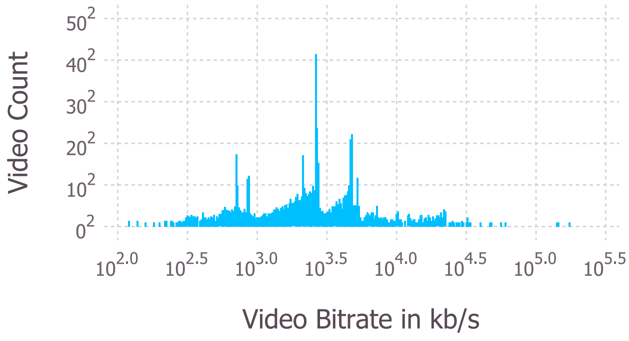

Figure 1: Distribution of the video bit-rate

less. The dataset encompasses a wide range of different resolutions and aspect ratios [6].

YouTube-8M is a dataset of around 8 million videos gathered via the video platform Youtube[3] [1]. The average duration of those videos is 3.75 min in the range of 2 to 8.33 min [26] and for every video metadata according to a predefined vocabulary is given. As other sources describe these metadata as noisy [8] its quality may be assessed as controversial. The actual dataset does not contain the videos themselves, but URLs (Uniform Resource Locator) to retrieve those videos via Youtube [15]. While those exemplary named datasets have their strengths and apply to certain preferences, they also have some shortcomings, that may be overcome by using the given V3C1 dataset.

3 VIDEO META CHARACTERISTICS

In this section we want to address the videos' meta features, i.e., properties, which are not related to their visual content. While other video datasets such as IACC.1 represents only videos with a bitrate of 512 kb/s, Figure 1 shows a broader distribution of the bitrate among V3C1's videos. Although it shows a broad distribution, there are some visible preferences for certain bitrates. Peaks in the distribution of the video bitrate are especially apparent for roughly 700, 850, 2 100, 2 600, 4 700 and 5 200 kb/s.

Each given video is separated into shots, i.e., visually similar segments that are represented by certain key frames. The number of shots per video depends on its duration and visual content [24]. Figure 2 illustrates this distribution of shots per video. On average a video consists of 114.84 shots per video, in the range of 5 to 5 011 shots per video.

Figure 3 shows the distribution of the duration of these segmented video shots and illustrates a strong tendency towards shorter shots. Despite some peaks in the overall share of shots at around 25 and 30 s, most shots are shorter than 10 s.

The given video dataset comprises videos uploaded to Vimeo between 2006 and 2018 and thus provides more current data than all the datasets mentioned in Section 2. While most of the data was gathered between 2011 and 2018 (despite a drop between 2015 and 2016), it also contains videos uploaded prior to that period, albeit amounting to a rather low share (see Figure 4)

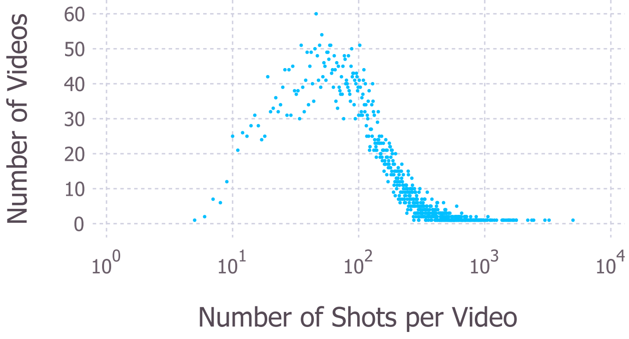

Figure 2: Distribution of shots per video

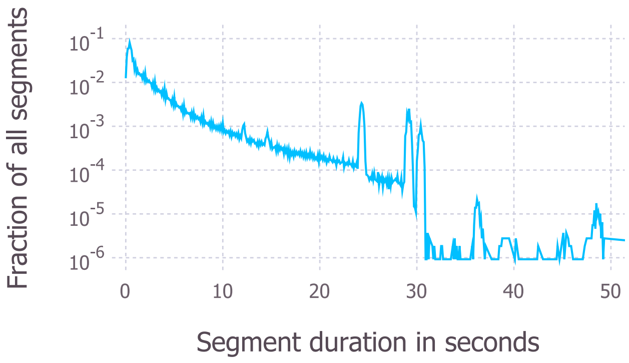

Figure 3: Distribution of segment durations

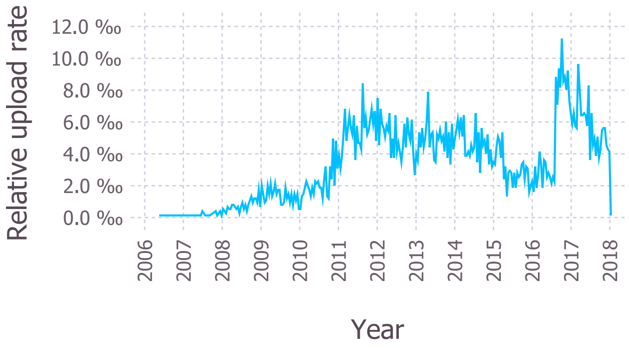

Figure 4: Distribution of video upload date

Figure 5 shows the distribution of the resolution of all the videos contained in the V3C1 dataset. While *1280x720* and *1920x1080* are obviously the dominant resolutions in the dataset, there is a high variance in terms of low resolutions. In particular there are some visible peaks in the figure in the standard VGA (Video Graphics Array) resolution range [12].

It may be concluded that the given dataset provides a collection of well-structured videos (segmented into shots), which is offering a wide range of video bit rates, gathered over a large time range

[3] URL: https://www.youtube.com/

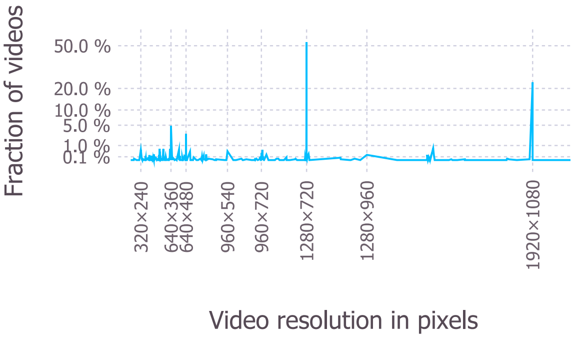

Figure 5: Distribution of video resolution

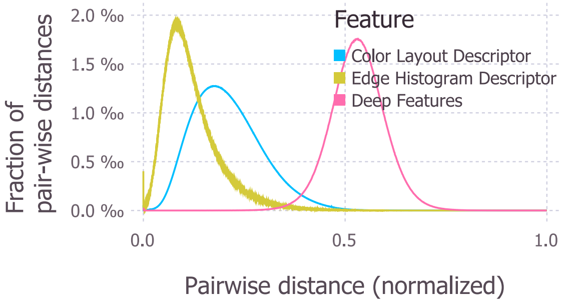

Figure 6: Pairwise distance between video shots

(compared to other video datasets) and comprising videos for a variety of resolutions [4].

4 VIDEO CONTENT CHARACTERISTICS

While the previous section is mainly concerned with the meta characteristics of the videos within the given V3C1 collection, this section focuses on content-based properties that were gathered via a thorough image analysis and inter-image comparison.

To reveal further insights of the V3C1 dataset, we investigate the visual and semantic content of key frames, that are provided with the dataset. As a first step we computed pairwise distances between all key frames by using three visual content descriptors: *Color Layout Descriptor* (CLD) [10, 11], *Edge Histogram Descriptor* (EHD) [18, 20] and *Deep Features* (i.e., weights of the last-fully connected layer, also known as *neural codes*) generated by classification of the key frames with GoogLeNet/InceptionNet [25, 27]. While CLD especially focuses on representing the prevailing colors of an image as a feature vector [10], EHD creates such a vector based upon the image's geometry and its main structure measured according to predominant edges [18]. Figure 6 shows that there is a higher similarity between video shots according to their texture (EHD) than according to their color layout (CLD). Manhattan / L_1 distance was used as distance metric for comparing the given features. For the similarity measured according to EHD as well as CLD and the deep features there is a discrete peak at distance zero. This anomaly occurs due to some key frames being identical, e.g., solid black shots, etc.

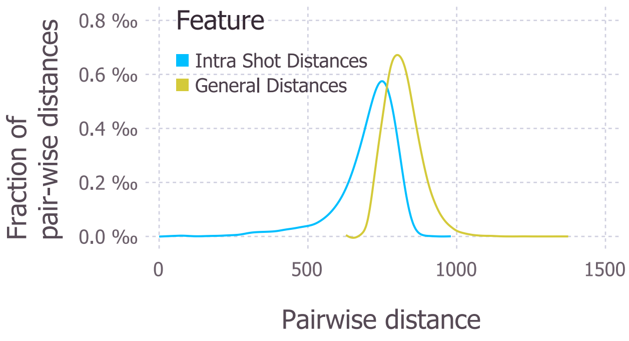

Figure 7: Distribution of self-similarity among shots per video

Table 1: Dominant Colors of Shot Key frames

Dominant Color	Number of Key frames
Blue	32 058
Cyan	8 109
Green	10 912
Magenta	1 899
Orange	33 188
Red	30 569
Violet	108
Yellow	1 424

Figure 7 illustrates the self-similarity among shots of a video measured according to the deep features of its key frames (cf. [25]). This *Intra Shot Distance* is shown along with the so called *General Distance*, which is the one displayed as "deep feature" in Figure 6, for comparison. Figure 7 shows that among shots of the same video the similarity is higher than the similarity among all shots of the whole video collection. This leads to the interpretation that, although the given deep features are not based on a dedicated visual feature representation like EHD or CLD, those features can be used as a mean for visual content analysis to a certain degree. Researchers using the V3C1 video collection have to keep in mind that this only holds true to a certain degree, as there is still some overlapping between the given curves of Figure 7. Table 1 outlines which colors are predominant for which amount of key frames. It shows that dominantly blue, red and orange key frames are prevailing within the video collection, while other colors are subordinate.

We analyzed these key frames using *NASnet*, a state-of-the-art convolutional neural network developed at Google Inc [30]. The ImageNet image collection provides a set of classes [7] that may be used to categorize images according to their visual content (cf. [14]). We applied this mean of categorization utilizing NASnet with the 1 000 ImageNet-classes of the ILSVRC challenge in the CNN's softmax layer to each key frame and attached that meta data to the according shot.

Figure 8 shows the distribution of the detected classes per video. It is obvious that the vast majority of videos have more than 100 concepts, which underlines the high diversity of the content. Figure 9 illustrates the top 20 detected classes and shows which classes

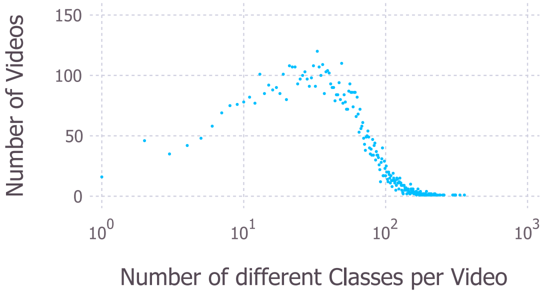

Figure 8: Distribution of detected classes per Video

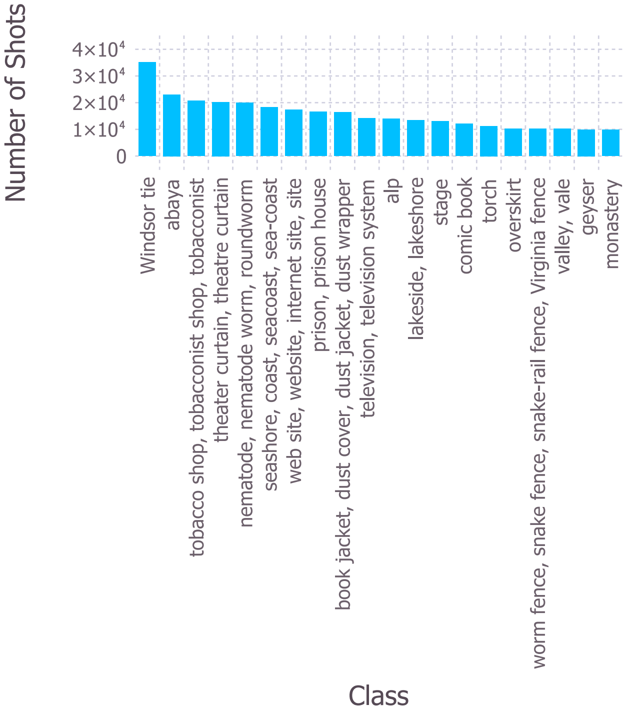

Figure 9: Top 20 detected Classes

are the most prominent ones among the given dataset. It shows that different video characteristics are spread equally within the collection, as even the most frequently detected classes only account for about 35k video shots - roughly 0.32 % of the whole dataset.

To assess the confidence level for the detection of the top 20 detected classes per key frame, we provide the respective confidence range in Figure 10. As often posed in literature too high confidence rates are signs for an overfitting of the given neural network [16, 28]. Thus those displayed confidence rates are satisfying.

Beyond these classes we also analyzed the semantics of each shot's key frame. Table 2 notably identifies, whether human faces or text was detected on each key frame and to which degree. This detection differentiates between some and a lot text, as well as different amounts of faces. It is evident that most key frames do not contain faces (roughly 75 %) while those that contain faces have a strong tendency to just include one. A fraction of about 11 % of all key frames contains text at all, whereas a heavy bias towards key frames with just a little text exists. This semantic analysis of the key frames was done via Apple's Core Image technology [2] both for identifying faces [9] and texts [3].

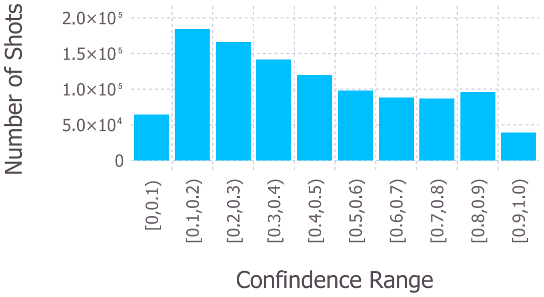

Figure 10: Confidence of best detected class per shot

Table 2: Semantics of Keyframes of Shots

Semantics	Number of Key frames
No face	815 435
One face	204 049
Two faces	43 291
Three faces	11 618
Four faces	4 334
Some text	103 934
A lot of text	18 862

5 CONCLUSIONS AND FUTURE WORK

In the broad field of data analytics one of the key components of research is having the right and appropriate datasets in place. Using standard and open datasets enables researchers to reproduce analytical experiments based on these datasets and thus validate the respective research. In this paper we have analyzed content characteristics of the V3C1 dataset [24] to provide ground truth for further related research. Most importantly, our analysis has shown that the content of V3C1 is very diverse in several aspects (upload time, visual concepts, resolutions, colors, etc.), it has no predominant characteristics and provides a low self-similarity (i.e., few near duplicates). Such properties make the dataset very well suited for video retrieval evaluations.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2018. YouTube-8M: A Large-Scale Video Classification Benchmark. (2018). http://arxiv.org/pdf/1609.08675v1

[2] Apple Inc. 2016. About Core Image. (2016). https://developer.apple.com/library/archive/documentation/GraphicsImaging/Conceptual/CoreImaging/ci_intro/ci_intro.html

[3] Apple Inc. 2019. CITextFeature: Core Image. (2019). https://developer.apple.com/documentation/coreimage/citextfeature

[4] Zlatka Avramova, Danny de Vleeschauwer, Pedro Debevere, Sabine Wittevrongel, Peter Lambert, Rik van de Walle, and Herwig Bruneel. 2011. On the performance of scalable video coding for VBR TV channels transport in multiple resolutions and qualities. *Multimedia Tools and Applications* 53, 3 (2011), 487–517. DOI: http://dx.doi.org/10.1007/s11042-010-0506-2

[5] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges QuÃĺnot, Joao Magalhaes, David Semedo, and Saverio Blasi. 2018. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. In *Proceedings of TRECVID 2018*. NIST, USA.

[6] Jun-Ho Choi and Jong-Seok Lee. 2016. Analysis of Spatial, Temporal, and Content Characteristics of Videos in the YFCC100M Dataset. In *Proceedings of the 2016 ACM Workshop on Multimedia COMMONS*, Bart Thomee (Ed.). ACM, New York, NY, 27–34. DOI: http://dx.doi.org/10.1145/2983554.2983559

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[8] Basura Fernando and Stephen Gould. 2017. Discriminatively Learned Hierarchical Rank Pooling Networks. *International Journal of Computer Vision* 124, 3 (2017), 335–355. DOI: http://dx.doi.org/10.1007/s11263-017-1030-x

[9] Nick Haber, Catalin Voss, Azar Fazel, Terry Winograd, and Dennis P. Wall. 2016. A practical approach to real-time neutral feature subtraction for facial expression recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE Winter Conference on Applications of Computer Vision (Ed.). IEEE, [Piscataway, NJ], 1–9. DOI: http://dx.doi.org/10.1109/WACV.2016.7477675

[10] Hamid A. Jalab. 2011. Image retrieval system based on color layout descriptor and Gabor filters. In *ICOS 2011*. IEEE, [Piscataway, NJ], 32–36. DOI: http://dx.doi.org/10.1109/ICOS.2011.6079266

[11] E. Kasutani and A. Yamada. 2001. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *2001 international conference on image processing*. IEEE, 674–677. DOI: http://dx.doi.org/10.1109/ICIP.2001.959135

[12] Asmar A. Khan and Shahid Masud. 2009. Memory Efficient VLSI Architecture for QCIF to VGA Resolution Conversion. In *Advances in image and video technology*, Toshikazu Wada, Fay Huang, and Stephen Lin (Eds.). Lecture notes in computer science, 0302-9743, Vol. 5414. Springer, Berlin, 829–838. DOI: http://dx.doi.org/10.1007/978-3-540-92957-4_72

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73. DOI: http://dx.doi.org/10.1007/s11263-016-0981-7

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[15] Joonseok Lee, Apostol (Paul) Natsev, Walter Reade, Rahul Sukthankar, and George Toderici. 2018. The 2nd YouTube-8M Large-Scale Video Understanding Challenge. (2018). https://static.googleusercontent.com/media/research.google.com/de//youtube8m/workshop2018/c_01.pdf

[16] Pengchao Li, Liangrui Peng, and Juan Wen. 2016. Rejecting Character Recognition Errors Using CNN Based Confidence Estimation. *Chinese Journal of Electronics* 25, 3 (2016), 520–526. DOI: http://dx.doi.org/10.1049/cje.2016.05.018

[17] Jakub Lokoc, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, and George Awad. 2018. On influential trends in interactive video retrieval: Video Browser Showdown 2015-2017. *IEEE Transactions on Multimedia* (2018).

[18] Atif Nazir, Rehan Ashraf, Talha Hamdani, and Nouman Ali. 2018. Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor. In *2018 International Conference on Computing 2018*. 1–6. DOI: http://dx.doi.org/10.1109/ICOMET.2018.8346343

[19] Paul Over, George Awad, Alan F. Smeaton, Colum Foley, and James Lanagan. 2009. Creating a web-scale video collection for research. In *Proceedings of the 1st workshop on Web-scale multimedia corpus*, Benoit Huet (Ed.). ACM, New York, NY, 25. DOI: http://dx.doi.org/10.1145/1631135.1631141

[20] Dong Kwon Park, Yoon Seok Jeon, and Chee Sun Won. 2000. Efficient use of local edge histogram descriptor. In *Proceedings ACM Multimedia 2000 workshops*, Shahram Ghandeharizadeh, Shih-Fu Chang, Stephen Fischer, Joseph Konstan, and Klara Nahrstedt (Eds.). Association for Computing Machinery, New York NY, 51–54. DOI: http://dx.doi.org/10.1145/357744.357758

[21] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. 2016. Learning Features by Watching Objects Move. (2016). http://arxiv.org/pdf/1612.06370v2

[22] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2014. Cineast: a multi-feature sketch-based video retrieval engine. In *Multimedia (ISM), 2014 IEEE International Symposium on*. IEEE, 18–23.

[23] Luca Rossetto and Heiko Schuldt. 2017. Web video in numbers-an analysis of web-video metadata. *arXiv preprint arXiv:1707.01340* (2017).

[24] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C – A Research Video Collection. (2019), 349–360.

[25] Guo Sheng, Huang Weilin, Wang Limin, and Qiao Yu. 2017. Locally Supervised Deep Hybrid Model for Scene Recognition. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 26, 2 (2017), 808–820. DOI: http://dx.doi.org/10.1109/TIP.2016.2629443

[26] Tej Singh and Dinesh Kumar Vishwakarma. 2018. Video benchmarks of human action datasets: a review. *Artificial Intelligence Review* 43, 3 (2018), 1. DOI: http://dx.doi.org/10.1007/s10462-018-9651-1

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (2016). http://arxiv.org/pdf/1602.07261v2

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). http://arxiv.org/abs/1512.00567

[29] Bart Thomee, Benjamin Elizalde, David A. Shamma, Karl Ni, Gerald Friedland, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M - The New Data in Multimedia Research. *Commun. ACM* 59, 2 (2016), 64–73. DOI: http://dx.doi.org/10.1145/2812802

[30] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc Le V. 2017. Learning Transferable Architectures for Scalable Image Recognition. (2017). http://arxiv.org/pdf/1707.07012v4