MOVING IMAGE DEMOGRAPHICS: DATA MINING TO SECURE VALUE

P. D. Fisher

The Media Institute, University College London, UK

ABSTRACT

Existing moving image collections face an onslaught of challenges to remain relevant and licensable: readiness for new colour standards and display sizes, rights-readiness and versioning in a cloud-connected world, and the ability to be found through discovery and search. Deep Learning technology creates a wealth of new potential to drive out data to address these key issues concerning professional moving image collections. While powerful, learning systems rely on a vast and disparate set of exemplar media and ground truth data to produce strong results. Broadcasters, producers, news organisations and archives are ideally placed to provide media for secure research, and to benefit from research results. There are opportunities to generate orthogonal resulting data, such as cross-collection media demographics and shared linked data pools. This paper discusses the issues in the context of recent TMI/UCL Deep Learning projects and a demographics study, 'Archive Watch', conducted by The Media Institute and FOCAL International.

INTRODUCTION

As an industry, we already welcome a handful of resources for information-sharing, for example: the IMDB website and ID registries EIDR and ISAN provide bare-bones public data on released titles; news organisations operate business-to-business (B2B) services for data and media interchange. JSON-LD and numerous standards provide an opportunity for cross-industry discovery and data sharing. Beyond media, industries such as air travel, fast-moving consumer goods and others have a history of cooperation to exchange B2B data within their industries, for mutual benefit. This paper suggests there may be novel shared benefit from sharing content and data within media industries, for research purposes. The resulting 'data mining' enables new scientific advances and speed to market in metadata generation, and the potential for industry-wide insights and business intelligence. Overall, this secures and enhances the value of media assets for the future.

This paper builds on research projects in the fields of automated video analysis and deep learning, in particular past project Video Clarity and current project DELVE-VIDEO, and a study concerning media demographics and the opportunity for cross-collection analysis and insights, ArchiveWatch. Each of these projects was conducted with the generous support of InnovateUK, the UK's Innovation Agency.[1]

In this paper, the scale and nature of today's media assets is discussed, followed by the requirements for automated metadata extraction. The potential for deep learning methods is outlined, and an industry-specific example is used to illustrate the main premise: that a modicum of shared data can benefit the media industry and speed research overall.

MEDIA DEMOGRAPHICS

There are over 200 million hours of ***unique*** professionally-produced and curated moving images maintained in repositories and vaults around the world today. This is in addition to the multitude of versions and instances of these originals, plus an avalanche of un-curated recordings (e.g. production rushes, unedited news and sport). Content spans the life cycle from production to release to archive, and the operational workflow at each stage in which editorial and curatorial decisions are made. Industry revenues peak at the point of release, and archive stages bear the primary burden of onward media management.

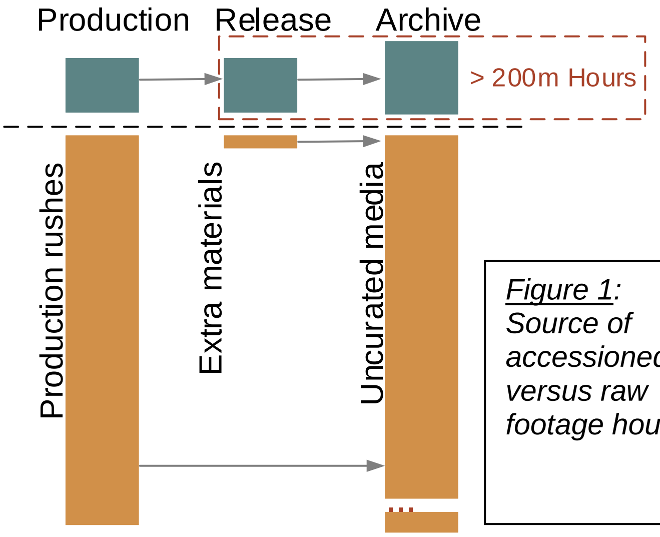

Figure 1: Source of accessioned versus raw footage hours)

New distribution methods are bringing an abundance of new opportunities for content producers and owners. Networked OTT ("Over the Top") delivery provides an opportunity to provide robust audience measurement and tracking, providing insights into content popularity and new reach into niche markets. New platforms and services have benefited from these feedback loops, enabling investment growth in production. For example, Netflix currently produce over 1,000 hours of original content annually, more than CBS or HBO.[2]

Video dominates the Internet: IP traffic worldwide will be 82% comprised of video by 2020, rising from 70% today (Cisco VNI). Real-time entertainment makes up 78% of North American fixed broadband traffic (Netflix 37%, YouTube 18%, others 23%; Sandvine). For the first time, in 2016, Internet advertising exceeded television advertising spend worldwide (KP Internet Trends 2017). It has been estimated that nearly 20% of digital advertising spend now relates to video, and will be the fastest-growing category through 2020 (eMarketer, 2016).

Professional moving image communities face unique challenges maintaining and re-investing in their collections, beyond first release. Revenues often depend on the extent and quality of digitisation, which cannot be assumed; for example, only 16% of European film collections exist in digital form.[2]

Searchability depends on availability of and access to metadata. Substantial expert time has been invested in defining standards for descriptive (semantic), rights and engineering metadata across multiple communities of interest (e.g. MPEG, SMPTE, EBU, IPTC, others) over many years. However, the ability to perform cross-collection search on these ‘normalised’ data structures has not yet been accomplished ‘in the wild’, for example by crawling available published data.

AUTOMATED METADATA EXTRACTION: A HIERARCHY OF NEEDS

Interest in automated metadata extraction is palpable and rising, clearly pointing to its value. The advent of Deep Learning has made this economic and feasible. Earlier algorithmic approaches may not have achieved the performance needed for commercial success, however the longstanding industry interest these has had an important benefit: a long and clear history of expressed requirements.

Automated metadata extraction can be expressed in terms of a hierarchy of needs: semantic, rights and engineering.[3] One aim of the DELVE-VIDEO project is to codify this hierarchy: asking “what are the most important features of moving image content that machine learning systems should address?”. This prioritises and builds potential for both open and added-value toolsets for all content owners.

Each metadata category – semantic, rights, engineering – can be seen as starting with a primary item [data item zero] and building data from basic characteristics through to complex, highly nuanced analyses. For example, were books being discussed, this might comprise the initial set: (semantic) ‘title’, (rights) unique ISBN number, and (technical) number of pages. Some information is intrinsic, and therefore possible for computer extraction, some is extrinsic, and so needs human entry as a starting point. The following diagram shows illustrative metadata for video, with each set of data growing in complexity:

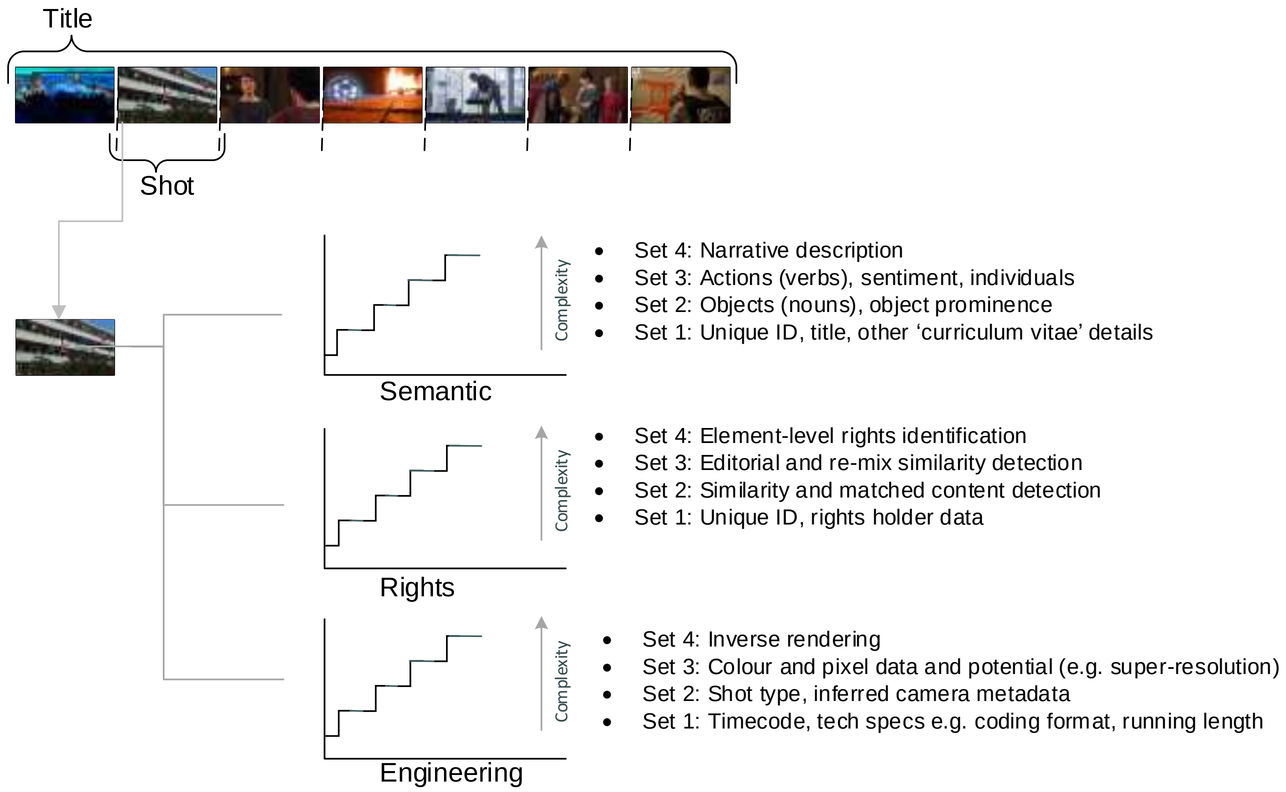
Title
Shot
Complexity
- Set 4: Narrative description
- Set 3: Actions (verbs), sentiment, individuals
- Set 2: Objects (nouns), object prominence
- Set 1: Unique ID, title, other ‘curriculum vitae’ details
Semantic
- Set 4: Element-level rights identification
- Set 3: Editorial and re-mix similarity detection
- Set 2: Similarity and matched content detection
- Set 1: Unique ID, rights holder data
Rights
- Set 4: Inverse rendering
- Set 3: Colour and pixel data and potential (e.g. super-resolution)
- Set 2: Shot type, inferred camera metadata
- Set 1: Timecode, tech specs e.g. coding format, running length
Engineering

Figure 2: Illustrative video metadata by category

Many video metadata analyses sit at the junction between intrinsic and extrinsic, and so require ‘ground truth’ materials in order to provide training. Sets of annotated media are invaluable. For example, the ImageNet[5] database includes 1503 images tagged and validated by humans as ‘Bengal tigers’, along with a multitude of other ‘nouns’ for image recognition, providing valuable ground truth for machine learning researchers.

Figure 3: Exemplar images of “Bengal tigers” in ImageNet

In the inaugural article of Apple’s new machine learning journal, it is proposed that photo-realistic synthetic imagery be generated to emulate real-world ground truth sources, in view of the lack of “large, diverse and accurately annotated” datasets.[6]

IMPACT OF DEEP LEARNING

The terms ‘artificial intelligence’ and ‘machine learning’ appear with increasing frequency within the media technology community. For example, at a recent IPTC (International Press and Telecommunications Council) Annual Conference, attended by the world’s leading news organisations, half a day was dedicated to “Auto-tagging and Visual Search”. Much of this growth is due to a S-curve change in which deep learning is significantly exceeding the results of earlier algorithmic methods. In another example, the annual SMPTE Conference 2017 opens with an all-day symposium on artificial intelligence.

In business, an S-curve illustrates the transition from one technology basis to another as cumulative R&D investment results in technology performance exceeding that of previous innovations in a shorter time. For example, CRT display technology was displaced by LCD and later LED and OLED technology. Similarly, perpendicular magnetic recording (PMR) displaced longitudinal magnetic recording (LMR) in hard disks.

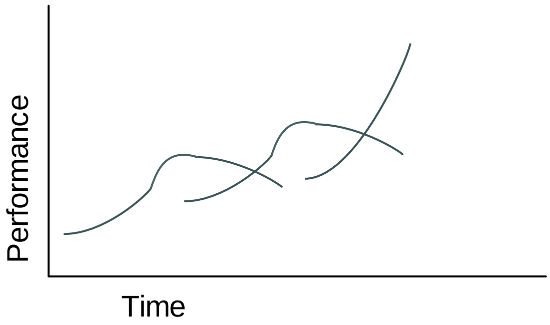

Figure 4: illustrative S-curves

Initially, an emerging S-curve performs less well than established technology but as R&D effort increases the prior performance is exceeded. Deep Learning represents the beginning of an exceptional new S-curve, poised for phenomenal performance gains as it matures. Substantially less (although highly skilled) R&D effort is needed to achieve results in excess of previous methods, making advanced analyses commercially affordable and performant, often for the first time.

The emergence of dominant design paradigms in deep learning, using CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) architectures, has resulted in rapid technology maturity.[7] This has given rise to a succession of tools (e.g. TensorFlow, Caffe, MX-Net), allowing scientific developers to create a model and then simplify ongoing training and refinement for deep learning insights.

High level APIs for developers not creating in-depth algorithmic models themselves have also flourished, in particular among cloud service vendors. For example: AWS Rekognition, Google Cloud AI, Microsoft Azure Cognitive Services, and IBM Watson offer variously a set of speech recognition or video and image analysis tools. Overall, this creates a flexible set of adoption points for developers. The growing body of cloud API developers also provides a ready route to market for video analysis innovators.

Deep learning differs from earlier purely algorithmic methods in that the results are derived from the both the design of the algorithmic model and the 'black box' of automated learning. Systems are 'smart' in similar ways to the human brain: the rationale for knowledge and insights is too nuanced to be altogether delineated.

Deep learning is increasingly used in order to extract compact video signatures, which can then be used for high-speed search and discovery. While this initially relied on purely algorithmic methods, deep learning is raising accuracy to exceed human levels of similarity matching.

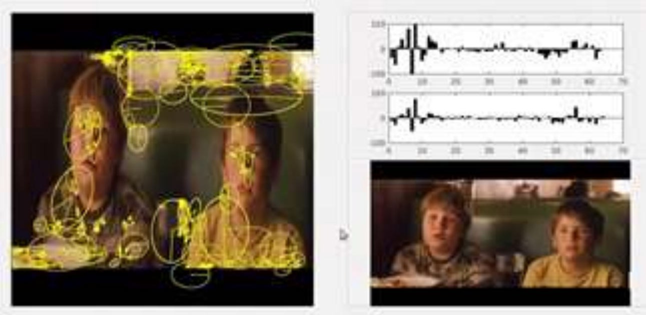

Figure 5: Example of video signature matching for divergent aspect ratios, from Video Clarity project

The generation and use of compact video signatures to provide a basis for search and discovery is gaining ground, and standardisation activity is highly active within ISO/MPEG. Termed "Compact Descriptors for Video Analysis"[8], the specification and reference software commit to "being robust to partial occlusions as well as changes in viewpoint, camera parameters, and lighting conditions". For example, video filmed at a cinema with chair backs occluding the view, should be correctly recognised. These factors are extremely important, and were a research accomplishment of Video Clarity, however not yet present in today's services and commercial toolsets. As with other ISO/MPEG standards, there is significant scope for commercial products to add value, however the standard aims to provide a basis for heterogeneous organisations and architectures to interact.

A compact video signature of less than one megabyte per running hour is able to replace video of 00s of megabytes or several gigabytes per hour for the purposes of search. There are speed and security benefits: the original content can't be 'seen' or reconstituted, therefore search in third party environments is not prone to piracy. Search across a vast body of content holdings can be accomplished rapidly. In TMI/UCL experiments, searches for identical or similar content across 100,000 hours of content have been performed in under 10 seconds, using a 64-core ecosystem.

As with periodic re-encoding, as technology improves over time, compact signatures will need to be re-extracted at intervals. This may provide the impetus to generate and share resolvable search data, for discovery. Further, enabling reference media or compact signatures to be shared for research purposes creates abundance of ground truth. To illustrate the potential for this, the context of a specific industry segment is outlined below.

EXAMPLE: FOOTAGE ARCHIVE INDUSTRY

Worldwide, as previously noted, it is estimated that over 200m unique hours of professionally -produced and -curated moving image content is managed in physical and digital vaults today. This footage spans both 'active' collections in current release and the footage archive industry, which includes commercial stock footage libraries, cultural organisations, broadcast and film archives. While valuable, as evidenced by continued investments in preservation and asset management by 000s of organisations, much of this content suffers from poor commercialization and access. Industry members typically seek to license footage for re-use, either as a fully commercial endeavor or to subsidize digitisation and archive management costs. The market for footage licensing for repurposing under-performs significantly, with less than 1% of all available content under management licensed annually.[9] Thus, this small proportion of revenue-generating content must support the remaining 99% in asset management.

With demand for content rising and burgeoning new routes to market, growth opportunities abound. Current market dampeners such as inadequate digitisation and lack of metadata can potentially be reversed by allowing content to be used as 'ground truth' for research. This generates increased revenues to content owners and underpins valuable new research potential.

The ARCHIVE WATCH project investigated the feasibility of creating a 'living' data source indexing and monitoring these 200m hours of moving image content over time. The project hypothesis was that market growth depends on solving three critical barriers to content marketability: lack of rights-readiness, lack of discoverability, and inability to assess qualitative fitness for purpose. These relate to the three broad categories of metadata outlined in this paper: semantic, rights and engineering.

The project proposed an ongoing "tools for data" ecosystem, in which content owners would be able to benefit from automated metadata extraction in exchange for providing or validating results as ground truth, to further fuel research.

The potential for the cross-collection search this enables, not only serves market demand for friction-free discovery, but provides a basis for sector-level business intelligence. New information syntheses may provide answers to questions such as "how many orphan works are there?", "how much wildlife footage was shot at night?" or "rank available cityscape media by location".

CONCLUSIONS

The media industry has reached a genuine 'inflection point', as IP-based entertainment ecosystems and deep learning -fueled research reach maturity together. Creating an environment for 'on demand' searchability in the wild is not trivial, but the potential for this outcome is beginning to emerge. As has been seen, learning systems rely on a vast set of well-annotated media for ground truth, but once this is available the ability to generate metadata automatically is increasingly achievable. Broadcasters, producers, news organisations and archives are ideally placed to provide media for secure research, and to be the first to benefit from research results.

REFERENCES

[1] Project “Video Clarity: High-speed meaning extraction in large video datasets” (2014-6) was conducted by partners TMI (The Media Institute)/UCL (University College London) and BAFTA (the British Academy of Film and Television Arts); Project “DELVE-VIDEO: Deep Learning -based Bitstream Analysis for Value Discovery in Video” (2017-8) is underway between partners Dithen, Soundmouse and TMI; Project “ArchiveWatch” (2016-7) was conducted by FOCAL International and TMI.

[2] Reported following a Netflix live streaming announcement, e.g.: https://venturebeat.com/2017/02/08/netflix-announces-1000-hours-of-new-original-content-for-2017/

[3] See: http://www.obs.coe.int/en/-/only-16-of-europe-s-film-heritage-collections-has-been-digitised; also European Audiovisual Observatory report: “The Exploitation of Film Heritage Works in the Digital Era”

[4] For example, see the IPTC Video Metadata Hub: https://iptc.org/standards/video-metadata-hub/

[5] The ImageNet dataset includes 14m annotated images (see: http://image-net.org/); see Professor Fei Fei Li’s TED talk providing an overview here: https://youtu.be/40riCqvRoMs

[6] See: “Improving he Realism of Synthetic Images” https://machinelearning.apple.com/2017/07/07/GAN.html

[7] Convolutional Neural Networks (CNNs) have shown state-of-the-art performance in image and video retrieval problems [Hinton *et al.,* NIPS 2012], e.g., allowing for face recognition accuracy that exceeds that of human experts [BBC News, Intelligent Machines, Sept. 2015]. Recurrent Neural Networks (RNNs) that have been the driving force behind audio, text and natural language processing, and led to the first reliable methods for these problems [LeCun *et al.*, arXiv 1312.1847, 2015, LeCun *et al*., Nature, 2015].

[8] See: http://mpeg.chiariglione.org/standards/exploration/compact-descriptors-video-analysis and “Compact Descriptors for Video Analysis: the Emerging MPEG Standard” *Ling-Yu Duan, Vijay Chandrasekhar, Shiqi Wang, Yihang Lou, Jie Lin, Yan Bai, Tiejun Huang, Alex Chichung Kot, and Wen Gao* https://arxiv.org/pdf/1704.08141.pdf

[9] Source: Screen Digest “Global Trade in Audio-visual Archives” (2010), confirmed by ArchiveWatch project (2017)