VIDEO CLARITY: HIGH SPEED DATA MINING FOR VIDEO

P. D. Fisher[1], A. Alhabib[2], V. Giotsas[1], I. Andreopoulos[2]

[1]BAFTA Research, UK; [2]Department of Electronic and Electrical Engineering, University College London (UCL), UK

ABSTRACT

Video now dominates ICT networks and systems, representing over 64% of global IP traffic[1] and over half of all storage within enterprises and data centers[2]. However, today video cannot be searched in the same way as alphanumeric data – this represents an unyielding 'big data' problem. Current video search relies on resource-intensive human annotations placed in a database, as alphanumeric data. This paper describes a new technology innovated by BAFTA (British Academy of Film and Television Arts) and UCL (University College London), which addresses this issue. The technology extracts a compact video signature representing significant features of the video for search, which can then be used for a plethora of applications such as similarity detection, de-duplication of files, piracy detection, and semantic classification. The video signatures are extremely rich yet highly compact, sized at approximately 5 megabytes per running hour of video. This enables video to be searched at the speed of data, allowing video to become a first-class citizen of ICT networks and systems.

INTRODUCTION

Video feature extraction is not new; there is a long-standing interest among technologists and media industry stakeholders from the 1990s onward in 'automated content analysis'. Early research projects such as "Informedia" (Carnegie Mellon University, 1994-1999)[3] sought to provide meaning from video. The primary focus of the MPEG-7 standard was to provide a "Multimedia Content Description Interface" – a mechanism for storing and communicating features of moving image essence, once known. Although significant interest has been generated and substantial investments made, few products or services capable of commercial adoption have emerged.

The ability to perform automated discovery and search across large bodies of still images and video has become a significant research topic once again, driven by the abundance of media held within Internet-delivered services and dominating data center storage and consumer Internet traffic.

This paper discusses a research initiative currently underway with the aim of both: a) expanding the state of the art in video feature extraction innovation, and b) prioritising requirements unique to the professional media industry. Video Clarity is underway as a collaboration between BAFTA Research, a business unit of the British Academy of Film

and Television Arts (BAFTA), and The Media Institute of University College London (UCL)[4], with generous support from Innovate UK, the UK's Innovation agency.

Video Clarity aims to fully enfranchise video as a searchable data source, allowing video to be searched at the speed of data. Visual similarity detection is the focus of the research, which has many applications including but not limited to:

- De-duplication of files held within data centers and file systems
- Piracy detection, for similar or identical content
- Matching media segments across disparate videos. For example: matching stock footage sources with edited titles, or matching raw camera rushes with editorial output
- Vailidation of unique IDs, anti-tampering assurance

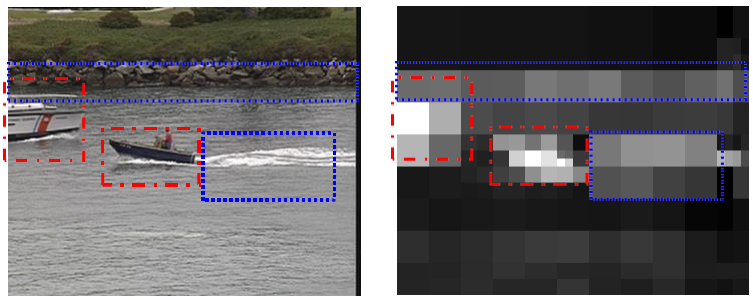
These applications are discussed in more detail, after a brief overview of the technology.

SOURCES OF DATA AND MEANING

In order to provide meaningful search, we must first ask the question "What can a computer learn automatically from video?". There are four generic sources of meaning, discussed briefly here.

Hint extraction

Modern video files contain hidden 'hint' metadata generated for access by networks concerning streaming.[5] This can be used as a source of low-level motion features, for example, the image on the right below shows significant movement detected only by decrypting 'hints' information, without reference to the source imagery:



Early during Video Clarity it was established that hint extraction was over 400x faster than video decoding, even before further optimisation or processing.

Engineering header information

Video file containers contain header information for interchange between systems to provide information on essence contained. Several open source tools[6] are available to retrieve this header information as structured data, providing an overview of contents, for example:

File Format: QuickTime / MOV
Video Format: H.264 / AVC / MPEG-4 AVC / MPEG-4 part 10 | 2.49 Mbps
Audio Format: AAC (Advanced Audio Coding) | 120 Kbps
Duration: 0:07:02.483000
File Size: 132.02 MB
Aspect Ratio: 1920x1080

Embedded semantic metadata

Several existing and emerging metadata standards have chosen to embed metadata within video files, to simplify retrieval or ensure tethering. The SMPTE (http://www.smpte-ra.org/) and MPEG-4 (http://www.mp4ra.org/) registration authorities are among standards organisations acting as registrars these metadata types. The proposed IPTC Video Metadata standard[7] aims to leverage earlier International Press & Telecommunications Council metadata standards in news (NewsML) and photography, each of which won unprecedented adoption. Standards for digital delivery of broadcast production, such as the UK DPP[8], similarly define required container-level semantic metadata.

Video feature extraction

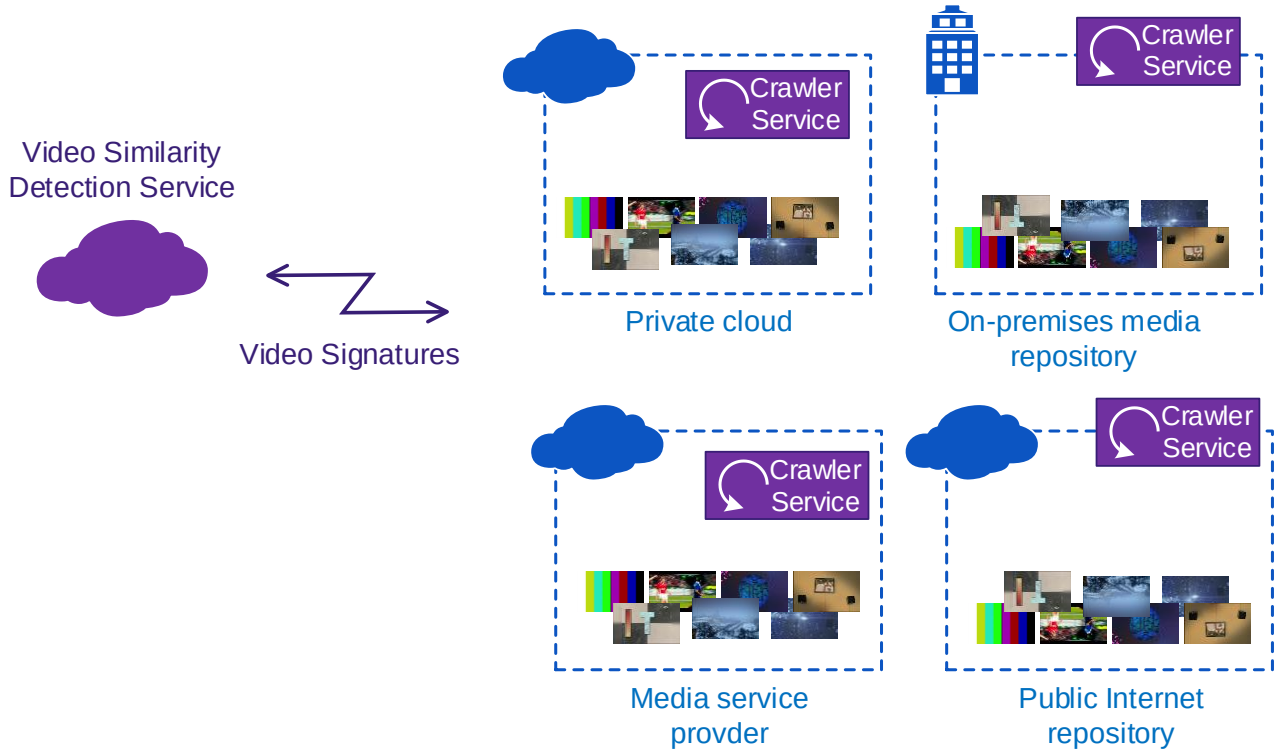
Visual search is an expanding arena for scientific research: conferences and journals concerning media engineering, computer vision and information retrieval abound with innovations in feature extraction, object or face identification, image classification and related topics. An increasingly sophisticated landscape for research is emerging. Leading methods include use of SIFT (scale-invariant feature transform) and VLAD (vector of locally aggregated descriptor)[9] to generate compact video representations for future search. Issues remain concerning the size of these compact representations, impacting search performance, and the robustness and accuracy of results. Current methods can lack robustness, for example "identical" videos may fail to be recognised as such due to changes in aspect ratio, coding artifacts or other forms of interferance.

The Video Clarity initiative has innovated a new method which overcomes several of these difficulties. Termed VLAC (vectors of locally aggregated centers)[10], this is able to identify identical or similar content with degree of precision which substantially exceeds the state of the art. The image above illustrates the local feature centers (left) accurately being compared with video 'in the wild' (right) with different coding and aspect ratio.

VIDEO CLARITY ECOSYSTEM

The video signature generated by Video Clarity technology is highly compact, occupying 5 megabytes of data per hour of moving image content, with the potential for further reduction. Once generated, these video signatures can be held independently of the media. This means that performing visual searches does not depend on the presence of human-viewable media. Hundreds of thousands of hours of video can be searched within seconds, using these compact video signatures.

Given media industry security concerns, Video Clarity architecture has been designed to provide maximum flexibility across organisational boundaries. A 'crawler service', capable of performing local video signature extraction, will create and then communicate only video signatures. The Video Similarity Detection Service can then compare video from multiple disparate domains, without the need for video to be present.



As illustrated above, this means that multiple organisations can perform separate or interoperable video similarity determinations, through establishing a connection:

- A media enterprise, managing a video repository in a private cloud, can run the Crawler Service within their own domain, by choosing a Video Clarity *machine image* for use on a captive server
- A media enterprise with an on-premises media repository, can similarly enable the Crawler Service, either by including an *edge appliance* in their infrastructure or by allowing the remote service periodic *secure access*
- Media service providers and those providing post-production services or distribution routes to market, can similarly allow their holdings to be monitored by using a Video Clarity machine image, edge appliance, or via secure access

- Known media locations accessible via the public Internet can be crawled by the remote service, and may willingly choose to use these services to guard against piracy concerns

Cloud computing is increasingly taking a central role in media management, given the digital nature of media assets and the operational benefits this can provide. An emerging trend confirms a desirable 'dominant design' which takes the computing to the media, rather than the media to the computing. This is particularly relevant where large master and mezzanine -quality files are maintained within the cloud, given respective video sizes:

Reference File Sizes:	Bitrate	1 Hour
Uncompressed SD	188 mbps	79.0GB
Uncompressed HD	1111 mbps	465.7GB
Mezzanine HD (e.g. ProRes/DNxHD)	220 mbps	92.2GB
Mezzanine SD (e.g. IMX50, XDCAM)	50 mbps	21.0GB
Intermediate (e.g. AVC/HEVC)	10 mbps	4.2GB
Intermediate (e.g. AVC/HEVC)	5 mbps	2.1GB

The major cloud infrastructure providers continue to innovate, offering both 'high availability' and tiered storage[11], driving even the largest video assets to the cloud.

In contrast, the storage required for video signatures representing 200,000 hours of video occupies less then one terabyte in a file system.

APPLICATIONS FOR VIDEO SIMILARITY DETECTION

The Video Clarity initiative is primarily concerned with *professional moving images*. These fuel the global $900bn Media and Entertainment industry. We estimate that the total volume of unique hours of professionally-produced or professionally-curated moving images under management today is approximately 200-250 million hours. This includes current releases plus back catalog film and video collections, curated news, commercial footage libraries, and curated archives. This is in contrast to consumer generated video, and uncurated video of all types, from personal videos through to production 'rushes'.

The contrast in scale of these video types is high: at one extreme, 707 feature films were rated for release in the US during 2014 (amounting to approx. 1400 hours of unique material); at the other extreme, on Facebook alone, user-shared video reached 4 billion views per day during Q1 2015, up 400% in 6 months[12], representing a mass of hours of unique video.

Key applications for video similarity detection within professional moving image assets are discussed briefly below.

Piracy Detection

It has become increasingly complex for copyright holders to monitor illicit – either intentionally or inadvertently – use of their assets. Video Clarity research correctly identifies identical or similar content, even under conditions causing other content identification systems to incorrectly 'pass' infringing media.

The video signature extracts a searchable machine-readable summary, which remains robust in the case of, for example:

- Variations in aspect ratio, editorial version, display angles (e.g. in the case of recordings made in cinema or from broadcast display)
- Shot-level infringement detection, rather than depending on full running length
- Quality and encoding differences

De-duplication of stored files

Over time, media may be re-digitised and re-encoded several times. Historic instances and versions may be stored safely, but not deleted when no longer required. Rather than commit substantial human time on file system upkeep, older media is often retained indefinitely. Instead, video similarity assessment can be used to identify matching material and follow desired retention and backup policies. Significant storage expense is incurred managing unnecessary media duplicates today, and optimising this can lead to significant savings.

Verification of unique ID

Universally unique cross-organisation IDs with searchable linked data are desirable. Both ISAN (International Standard Audiovisual Number, an ISO standard), and EIDR (Entertainment ID Registry) have significant databases and committed adopters. Validating that these IDs are correct, by reference to the video signature, will be a powerful application.

Similarity assessment for recommender systems

A major aspect of media content management and distribution systems (such as YouTube, Amazon Instant, others) is search and aggregation of similar content. This increases the efficiency of auto-suggestion mechanisms and can be combined with other tools that make automated recommendations to users. Increased efficiency of such mechanisms leads to increased user engagement and therefore increased revenues. In parallel, business models such as stock footage licensing can make use of visual similarity assessment in order to make recommendations to potential licensors.

Performing 'pre-flight' inspection, identification of best source

Planning for digitization or re-mastering can be accomplished across collection holders. This is an issue among film and television archives, where several archives worldwide may have ostensibly identical copies of the same media, and the best version could be used for digitization and then shared between authorized holders. This is an approach used for library books and journals embraced by libraries worldwide with membership of Stanford's LOCKSS alliance ("Lots Of Copies Keeps Stuff Safe" - http://www.lockss.org/) – resilience without disruption of existing copyright arrangements.

Matching raw materials to edited long form content

Raw materials from cameras are used to produce original content. The production process is very busy and often the rushes are backed up but not adequately labelled, once the production process completes. This makes it difficult to return to these raw materials later to produce new versions or assess the material for stock footage licensing potential. Similarly, historic productions which include archive or footage library material may not be

sufficiently identified and noted, making it difficult for the source archives to track usage of their materials.

FUTURE RESEARCH

Video Clarity has benefited from a clear focus, and is already achieveing results beyond the state of the art. This paper has discussed the generation of compact video signatures based on feature extraction and aggregation. The outcome has been a highly successful method for video similarity detection, capable of early commercial deployment. In parallel, the Video Clarity initiative has allowed a desirable set of low-level features to be implemented with efficiency. These include analyses which may no longer be considered challenging in science, but which contribute to downstream video signature extraction and are highly desireable to users. These include aspects such as shot detection, shot typing (e.g. “show me all the long shots / pans / zooms / establishing shots”), and the identification and extraction of most relevant keyframe, shot by shot. Further high value analyses will be identified over time, in consultation with stakeholders. We anticipate creating a community-viewable ‘hierarchy of needs’ which will help prioritise features of value, and to engage a wide body of technologists in contributing to implementation of these features.

Further work is needed to associate high-level meaning with discovered features. This has two dimensions: the association of subject-index semantic meaning with video (for example, automated keyword generation), and the ability to identify and search discrete objects within the video (e.g. “show me the shots with a car / dog / beach”). Promising work has begun within the research consortium to identify and retrieve objects (using ROI - region of interest) within large video datasets.[13]

Having access to ‘ground truth’ is a significant factor in this research. One interesting source of tagged objects within images is the ImageNet database[14]. This provides links to over 14 million images, each of which has been classified by humans using the WordNet hierarchy of nouns. One example: the word “bird” is currently represented by 850 sub-categories and 812,000 images – valuable ground truth when trying to assess whether an object detected in video is indeed a bird.

The potential for audio spoken word tracks to contribute to semantic meaning is also being investigated, and cooperation with other long-standing audio projects underway.

CONCLUSION

Video represents a ‘big data’ problem today; although occupying over 64% of all Internet traffic and over half of all storage, the ability to perform meaningful search within video is illusive if not impossible. The Video Clarity initiative has innovated a new form of compact video signature extraction, which is achieving results in advance of the state of the art. This paper has discussed the technology along with its context and application. This allows video to become a first-class citizen of ICT networks and storage, and allows for video search at the speed of data.

References

[1] "Cisco Visual Networking Index: Forecast and Methodology, 2014–2019" (27 May 2015); See: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html

[2] "2014 Digital Storage for Media and Entertainment Report", Coughlin Associates, estimates that $6.2bn is spent annually on video storage; Frost & Sullivan estimate the annual value of all storage (2013) to be $10.7bn

[3] "Informedia-I: Integrated Speech, Image and Language Understanding for Creation and Exploration of Digital Video Libraries" (1994-1999), see: http://www.informedia.cs.cmu.edu/dli1/index.html

[4] The Media Institute at UCL supports collaborations between academia and industry, assisting with dissemination and supporting cross-project outputs for re-use. Academic resources are drawn from existing research groups and academic departments, in the case of Video Clarity the Department of Electronic and Electrical Engineering at University College London provided core input (http://www.engineering.ucl.ac.uk/departments/electronic-and-electrical-engineering/)

[5] This builds on previous work conducted by Dr. Yiannis Andreopoulos. See: http://www.ee.ucl.ac.uk/~iandreop/UNV.html and http://www.nsf.gov/awardsearch/showAward?AWD_ID=0541453

[6] One example: https://www.ffmpeg.org/ffprobe.html

[7] "IPTC Video Metadata", presentation here: https://iptc.org/download/events/pmdc2015/51_IPTC-VMD-Draft2-201506.pdf; standard draft for review: www.iptc.org/videometadata-draft2

[8] E.g. Digital Production Partnership in the UK, see: http://www.digitalproductionpartnership.co.uk

[9] For example: "Aggregating local descriptors into a compact image representation", Jegou et al (see: https://lear.inrialpes.fr/pubs/2010/JDSP10/jegou_compactimagerepresentation.pdf)

[10] "VECTORS OF LOCALLY AGGREGATED CENTERS FOR COMPACT VIDEO REPRESENTATION"; Alhabib Abbas, Nikos Deligiannis and Yiannis Andreopoulos; IEEE International Conference on Multimedia and Expo (IEEE ICME 2015)

[11] Amazon Web Services 'Glacier' and Google 'Nearline' are examples of tiered storage offered by cloud infrastructure providers

[12] "Internet Trends 2015", Mary Meeker, KPCB. See: http://www.kpcb.com/internet-trends

[13] "Region-of-Interest Retrieval in Large Image Datasets with Voronoi VLAD", "; Aaron Chadha and Yiannis Andreopoulos; 10th International Conference on Computer Vision Systems (ICVS 2015)

[14] The ImageNet project provides researchers with an extensive set of tagged images for research (http://image-net.org/). See Professor Li Fei Fei's excellent TEDx talk, describing the initiative here: https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=en